



HAL
open science

Anatomo-functional dynamics of the hippocampal subfields across the lifespan: contributions of explainable artificial intelligence

Clément Poiret

► **To cite this version:**

Clément Poiret. Anatomo-functional dynamics of the hippocampal subfields across the lifespan: contributions of explainable artificial intelligence. Human health and pathology. Université Paris Cité, 2023. English. NNT: 2023UNIP7209 . tel-04688243

HAL Id: tel-04688243

<https://theses.hal.science/tel-04688243v1>

Submitted on 4 Sep 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Université Paris Cité
Ecole Doctorale 158
InDev, NeuroDiderot, Inserm U1141
UNIACT, NeuroSpin
CEA Paris-Saclay, Frederic Joliot Institute

Anatomo-Functional Dynamics of the Hippocampal Subfields Across the Lifespan: Contributions of Explainable AI

Thèse de doctorat en Neurosciences

par Clément POIRET
Dirigée par Dr. Marion NOULHIANE
Soutenance prévue publiquement le
20 Octobre 2023

Devant un jury composé de :

Dr. Philippe CIUCIU	CEA Saclay, Inria Saclay	Rapporteur
Pr. Sophie DUPONT	Université Paris Sorbonne	Rapporteuse
Pr. Emrah DUZEL	University College London	Examineur
Dr. Maria DEPREZ	King's College London	Examinatrice
Pr. Martine GAVARET	Université Paris Cité	Examinatrice
Dr. Marion NOULHIANE	Université Paris Cité	Directrice

Abstract

The hippocampus is critical for episodic memory, as well as cognitive functions such as spatial navigation and learning. It contains anatomically and functionally distinct subfields, the dentate gyrus, the cornu ammonis, and the subiculum. In vivo study of the hippocampal subfields with magnetic resonance imaging (MRI) can illuminate the emergence of episodic memory in children and its disease-related alterations. However, their segmentation is challenging due to image quality or contrast, and because of a lack of unified manual segmentation guidelines. We developed the Hippocampal Segmentation Factory (HSF), an automated segmentation tool using deep learning, enabling efficient segmentation of the hippocampal subfields. HSF was trained on diverse populations spanning ages 4-100 years, including healthy subjects, temporal lobe epilepsy, or Alzheimer's disease. This enables robust segmentation for various applications. Using HSF, we delineated hippocampal subfield trajectories throughout life using datasets from the HCP initiative. We also described the subfields functional connectivity evolution from ages 4-25 years using resting-state fMRI. To handle result uncertainties, we applied post hoc explainable AI and studied the Rashōmon effect, — how different AI models produce different explanations for the same prediction. We analyzed this effect on generic datasets using SHAP. Overall, this thesis facilitates the analysis of the development, disease, and function of the hippocampal subfield by providing a fast, robust segmentation tool and insights into AI uncertainties. The developed methods enable a deeper study of the hippocampal subfields. Beyond neuroscience, this thesis provides machine learning professionals with generalizable guidelines for uncertainty analysis in explainable AI.

Keywords: Semantic Segmentation, Lifespan development, Structural MRI, Functional connectivity, Explainable AI

Résumé

L'hippocampe joue un rôle essentiel dans la mémoire épisodique, la navigation spatiale et l'apprentissage. Il est composé de sous-champs anatomiquement et fonctionnellement distincts, le gyrus denté, la corne d'Ammon, et le subiculum. L'étude *in vivo* des sous-champs hippocampiques par imagerie par résonance magnétique (IRM) peut permettre de décrire l'émergence de la mémoire épisodique ainsi que les altérations de celle-ci liées à diverses pathologies. Cependant, leur segmentation est difficile à cause de la qualité et du contraste des images, et du manque de consensus quant à la segmentation manuelle. Nous avons donc développé Hippocampal Segmentation Factory (HSF), un outil de segmentation automatisé utilisant l'apprentissage profond, permettant une segmentation efficace des sous-champs hippocampiques. HSF a été entraîné sur diverses populations âgées de 4 à 100 ans, incluant des sujets sains, des épilepsies du lobe temporal ou encore des patients atteints de la maladie d'Alzheimer. Ceci permet une segmentation robuste pour diverses applications. Grâce à HSF, nous avons délimité les trajectoires des sous-champs hippocampiques tout au long de la vie en utilisant les données issues du projet HCP. Nous avons également décrit l'évolution de la connectivité fonctionnelle des sous-champs entre 4 et 25 ans en utilisant l'IRM fonctionnelle de repos. Pour gérer les incertitudes de modélisation, nous avons étudié l'effet Rashomon, — comment différents modèles d'IA produisent des explications différentes pour la même prédiction. Nous avons analysé cet effet sur des bases de données génériques à l'aide de SHAP. Cette thèse facilite l'analyse du développement, des maladies et du fonctionnement des sous-champs hippocampiques en fournissant un outil de segmentation rapide et robuste, permettant une étude plus approfondie des sous-champs hippocampiques. Au-delà des neurosciences, cette thèse fournit des lignes directrices généralisables pour l'analyse d'incertitude dans l'IA explicable.

Mots-clés: Segmentation Sémantique, Trajectoires développementales, IRM Structurel, Intelligence Artificielle Explicable

Résumé Substantiel

L'hippocampe, situé dans le lobe temporal médian, joue un rôle essentiel dans la mémoire épisodique, la navigation spatiale et l'apprentissage. Il est composé de sous-champs anatomiquement et fonctionnellement distincts : le gyrus denté, la corne d'Ammon, et le subiculum. L'étude *in vivo* de ces sous-champs par IRM peut permettre de mieux comprendre le développement de la mémoire épisodique ainsi que les altérations hippocampiques liées à diverses pathologies. Cependant, leur segmentation sur IRM est difficile à cause de la qualité et du contraste des images. De plus, il n'existe pas de consensus clair sur les protocoles de segmentation manuelle en IRM.

L'objectif de cette thèse était de caractériser la maturation structurelle et fonctionnelle des sous-champs hippocampiques au cours de la vie, en mettant l'accent sur l'enfance. Pour cela, de nouvelles méthodologies en intelligence artificielle appliquées à la neuroimagerie ont été développées. Premièrement, un outil de segmentation automatisé par deep learning nommé Hippocampal Segmentation Factory (HSF) a été introduit. Deuxièmement, HSF a permis de révéler les trajectoires développementales des volumes des sous-champs dans de larges échantillons. Troisièmement, la connectivité fonctionnelle des sous-champs a été analysée par IRM fonctionnelle de repos, révélant des changements prolongés pendant l'enfance, en correspondance avec les trajectoires anatomiques et la spécialisation des réseaux fonctionnels. Globalement, cette thèse a permis de mieux comprendre le développement prolongé des circuits hippocampiques sous-tendant la mémoire en proposant des outils méthodologiques innovants. Cette thèse s'appuie sur l'utilisation de larges bases de données d'IRM structurelles et fonctionnelles du cerveau humain à différents âges (de 4 à 100 ans), dans divers états de santé (par exemple, sain, avec épilepsie du lobe temporal, ou encore la maladie d'Alzheimer). La plupart de ses bases de données disposants de segmentation manuelles des sous-champs hippocampiques, des algorithmes supervisés d'apprentissage profond ont pu être développés.

Partie 1. Anatomie

Première contribution expérimentale. Tout d'abord, les réseaux de capsules 3D avec portes attentionnelles ont été explorés pour leur robustesse intrinsèque grâce à l'encodage de paramètres d'équivariance (Poiret et al., (2023). *Attention-Gated 3D CapsNet for Robust Hippocampal Segmentation*. Journal of Medical Imaging. Submitted). Les capsules représentent les entités visuelles via des vecteurs plutôt que des scalaires, permettant de reconnaître des objets malgré des transformations. Cette approche prometteuse pour segmenter l'hippocampe malgré les variations anatomiques n'a cependant pas été concluante à cause des limitations en calcul et mémoire des opérations sur les vecteurs.

Seconde contribution expérimentale. Cela a conduit au développement de HSF, plus efficace grâce à des réseaux convolutionnels (Poiret, et al., (2023). *A fast and robust hippocampal subfields segmentation: HSF revealing lifespan volumetric dynamics*. Frontiers in Neuroinformatics.). HSF intègre les avancées en vision par ordinateur comme l'attention visuelle, qui permet au réseau de se concentrer sur les régions pertinentes. Il a été entraîné sur des segmentations manuelles hétérogènes et amélioré par des annotations humaines pour capturer des variations liées à des cas atypiques, comme par exemple des scléroses hippocampiques sévères. HSF surpasse les outils existants

en termes de précision et rapidité d'inférence. Sa conception modulaire permet l'intégration de nouveaux modèles, assurant son évolution. HSF est open-source pour favoriser son adoption par la communauté scientifique.

Enfin, l'application à grande échelle se montre d'une importance significative pour relier l'anatomie des sous-champs à la cognition. Le développement d'outils transparents et efficaces est nécessaire pour concrétiser les bénéfices de l'IA en médecine, tout en gérant les risques éthiques. HSF vise à catalyser les études hippocampiques tout en assurant un développement responsable. Pour tester nos hypothèses à grande échelle, nous nous sommes concentrés sur les bases de données d'IRM structurales et fonctionnelles provenant principalement de l'initiative Human Connectome Project (HCP). HCP fournit des jeux de données d'IRM à grande échelle sur des échantillons sains à différents âges :

- HCP Développement : IRM structurales et fonctionnelles au repos acquises chez 588 sujets âgés de 5 à 21 ans. L'échantillonnage fin de cette période développementale est crucial pour cartographier la maturation des circuits cérébraux.
- HCP1200 : IRM structurales et fonctionnelles chez 1201 jeunes adultes âgés de 22 à 35 ans. Cette tranche d'âge permet d'étudier le cerveau à maturité.
- HCP Aging : IRM structurales et fonctionnelles chez 1246 sujets âgés de 36 à 100 ans. L'exploration de cette large gamme d'âge est essentielle pour quantifier le vieillissement cérébral.

L'utilisation de ces jeux de données à grande échelle et standardisés était indispensable pour caractériser finement l'évolution typique de l'anatomie et de la fonction hippocampiques au cours de la vie. HSF a été appliqué pour segmenter automatiquement les sous-champs hippocampiques dans les IRM structurales des sujets HCP. Grâce à l'inférence rapide de HSF, il a été possible de traiter efficacement les milliers de sujets HCP. Pour chaque IRM et chaque sous-champ (gyrus denté, CA1, CA2, CA3 et subiculum), HSF fournit un masque de segmentation au niveau du voxel. Le volume de chaque sous-champ est ensuite calculé, permettant d'obtenir à grande échelle des mesures volumétriques précises pour les sous-champs hippocampiques à travers plus de 3750 sujets HCP couvrant la quasi-totalité de la vie.

Les volumes des sous-champs hippocampiques ont ensuite été modélisés en fonction de l'âge à l'aide d'une analyse en "natural cubic splines" (NCS). Cette technique de régression non paramétrique permet de capturer des relations complexes non linéaires entre deux variables, sans avoir à spécifier une forme fonctionnelle a priori. Les NCS ont révélé que chaque sous-champ suit une trajectoire volumétrique distincte au cours de la vie. Par exemple, le gyrus denté présente une croissance prolongée jusqu'à l'adolescence alors que le subiculum atteint un volume mature dès l'âge de 5 ans. Ces courbes développementales non linéaires mettent en évidence l'importance de considérer l'hétérogénéité des sous-champs plutôt que l'hippocampe comme une entité homogène. Elles fournissent des indices quant aux mécanismes sous-jacents tels que la neurogenèse et la myélinisation. En revanche, l'anatomie n'est qu'une partie du problème, et le versant fonctionnel de l'hippocampe devait aussi être pris en compte.

Partie 2. Connectivité Fonctionnelle

Première contribution expérimentale L'application de méthodes d'IA explicable à l'IRM fonctionnelle au repos pour en extraire des biomarqueurs développementaux fiables pose plusieurs défis:

- Le bruit inhérent à l'IRMf peut induire des corrélations fallacieuses entre régions.
- La nature multidimensionnelle des données de connectivité rend l'analyse ambiguë, donnant lieu à des effets de type "Rashōmon" où différents modèles produisent des explications divergentes.
- Le manque de vérité terrain empêche la validation directe des réseaux identifiés.
- La reproductibilité des méthodes sur des jeux de données indépendants n'est pas toujours établie.
- Pour surmonter ces obstacles, il est essentiel de quantifier la convergence entre explications de modèles, et de répliquer les résultats après validation hors-échantillon.

Avant d'appliquer l'approche au cas de l'IRMf hippocampique, la méthodologie d'IA explicable proposée dans ce travail a été validée sur des problèmes de classification génériques provenant de divers domaines (Poiret et al., (2023). *Can we Agree? On the Rashomon Effect and the reliability of Post-Hoc Explainable AI*. arXiv:2308.07247). L'analyse de la similarité entre explications pour des modèles entraînés sur des tailles d'échantillon croissantes a permis de quantifier la quantité de données nécessaire pour obtenir des explications stables. De plus, l'agrégation d'explications de plusieurs modèles et la mesure de convergence vers un consensus ont démontré la capacité à extraire des caractéristiques prédictives fiables. Ces expériences sur des cas simplifiés ont servi de preuve de concept avant d'aborder la complexité des données d'IRMf, renforçant la confiance dans la méthode.

Seconde contribution expérimentale L'ensemble de ses études ont permis d'améliorer les descriptions anatomo-fonctionnelles des sous-champs hippocampiques tout au long de la vie (Poiret, and Noulhiane (2023). *Charting Hippocampal Development with Robust Explainable AI and Resting-State fMRI*. In prep.). L'application de HSF à 3750 IRM de 4 à 100+ ans a révélé des trajectoires volumétriques distinctes pour chaque sous-champ. Le gyrus denté, essentiel pour la séparation de pattern, montre une croissance prolongée jusqu'à l'adolescence. À l'inverse, le subiculum présente une maturation précoce puis une atrophie avec l'âge avancé. Ces différences reflètent probablement une maturation cytoarchitectonique et myéloarchitectonique asynchrone. Le développement prolongé de certains sous-champs supporte l'amélioration des capacités de mémoire pendant l'enfance. La vulnérabilité du subiculum chez les personnes âgées pourrait servir de biomarqueur précoce. Globalement, modéliser l'hétérogénéité des sous-champs éclaire les mécanismes développementaux typiques et atypiques sous-tendant la mémoire.

L'analyse en IRM fonctionnelle de repos a révélé une diminution des interactions entre les sous-champs hippocampiques et les régions corticales sensorielles et attentionnelles pendant l'enfance. La convergence des explications des modèles a identifié 10 connexions clés avec le cortex temporal médian, frontal et pariétal. Le gyrus denté montre les changements les plus importants, cohérents avec sa maturation structurelle prolongée. La diminution de connectivité reflète probablement

l'optimisation des circuits hippocampiques pour la mémoire épisodique et la navigation spatiale. Il existe une forte correspondance entre maturations structurelles et fonctionnelles spécifiques aux sous-champs.

Cette thèse présente des contributions majeures sur le plan méthodologique et sur le plan neuroscientifique :

- D'un point de vue théorique, nous avons proposé un nouvel outil de segmentation hippocampique robuste aux différentes modalités d'acquisition et aux populations atypiques, et avons proposé une méthodologie pour améliorer la qualité des interprétations basées sur de l'intelligence artificielle explicable,
- D'un point de vue pratique, nous avons utilisé ledit outil pour décrire précisément les trajectoires volumétriques des sous-champs hippocampiques tout au long de la vie, et l'évolution des connexions fonctionnelles des sous-champs au cours de l'enfance et de l'adolescence.

Cette thèse a apporté des contributions méthodologiques majeures, avec le développement de HSF pour la segmentation automatisée des sous-champs hippocampiques, et l'introduction d'un cadre d'IA explicable multimodèle pour gérer l'effet Rashōmon. HSF intègre des avancées en vision par ordinateur comme l'attention visuelle, et a été validé contre les logiciels alternatifs présents dans la littérature. La méthode pour quantifier la convergence d'explications rivales permet d'extraire des connaissances fiables de données complexes et multidimensionnelles comme l'IRMf.

D'un point de vue neuroscientifique, cette thèse a permis de révéler les trajectoires volumétriques de chaque sous-champ hippocampique de 4 à 100 ans. Leur maturation prolongée et asynchrone reflète l'évolution des capacités mnésiques pendant l'enfance. De plus, l'analyse de la connectivité fonctionnelle a identifié une diminution des interactions des sous-champs avec les régions corticales sensorielles et attentionnelles, traduisant une spécialisation des circuits neuronaux. HSF ouvre de nouvelles opportunités pour les études hippocampiques à grande échelle et dans diverses populations. Sa robustesse facilite son application chez des patients présentant des variations anatomiques atypiques. HSF pourrait permettre le suivi longitudinal d'interventions comme la chirurgie épileptique. Couplé à d'autres modalités d'IRM, HSF servira de socle pour relier structure, fonction et cognition.

Cependant, plusieurs limitations doivent être soulignées. Des données longitudinales seront nécessaires pour confirmer les trajectoires individuelles. La diversité des échantillons doit être augmentée. D'autres mesures quantitatives multimodales apporteront une perspective plus complète. Une validation plus poussée des biomarqueurs de connectivité est requise. Globalement, intégrer davantage de données comportementales et génétiques reste essentiel. De plus, le déploiement responsable d'outils d'IA comme HSF en pratique clinique nécessite de garantir leur sûreté, leur robustesse à des conditions variées, et leur adoption par les praticiens. Des problèmes éthiques clés comprennent le respect de la vie privée des patients, la transparence des algorithmes, et l'accès équitable aux innovations. Une approche pluridisciplinaire centrée sur l'humain sera cruciale pour une intégration bénéfique de ces technologies prometteuses.

En résumé, cette thèse a introduit des outils novateurs et des approches de modélisation permettant de suivre la maturation coordonnée de l'anatomie et de la fonction des sous-champs de l'hippocampe tout au long de la vie. Les contributions expérimentales établissent une base pour étudier les contributions des sous-champs hippocampiques dans la santé et la maladie à une échelle sans précédent.

Acknowledgement

First and foremost, I would like to express my profound gratitude to my advisor, Marion Noulhiane, for her unwavering support, mentorship, and guidance throughout the course of this scientific journey. Your insights and words of encouragement have often inspired me and were of great comfort during challenging times.

I would like to extend my appreciation to the members of my doctoral thesis jury who participated in the evaluation and positively impacted the scientific quality of my work, the rapporteurs Philippe Ciuciu and Sophie Dupont, and the examiners, Martine Gavaret, Emrah Duzel, and Maria Deprez.

I would like to thank Antoine Grigis and Edouard Duchesnay who helped me improve my scientific rigor, providing insightful knowledge and expertise when I needed it. I also thank Michel Bottlaender and Matthieu Faillot for their participation in the construction of HSF, and Antoine Bouyeure, Sandesh Patil, Julia Micaux, Jérémie Allinger, and Cécile Boniteau for their active engagement in our “HippoMnesis” team, from manual segmentation of the hippocampus, to methodological implementations, and technical discussions. Thanks to Frédéric Lemaître for guiding me on this path and for his advices.

I am also deeply thankful to my members of the InDev team, notably Yann Leprince for his advice and our geeky discussions while walking toward the R2. Thanks to all PIs of the team, David Germanaud, Lucie Hertz-Pannier, Catherine Chiron, and Jessica Dubois. Your critical thinking and dedication to high standards helped shape my thoughts and the present work.

I also would like to thank Julien Lefèvre and Karim Jerbi, members of my thesis monitoring committee, who took the time to listen to my ideas, hear my needs and interrogations, and made me benefit from their expertise.

I am grateful to the Deep Learning group of NeuroSpin and the students who made it happen. It has been and will remain a place of quality to learn AI-related methods and techniques. My sincere thanks to the students of the InDev team who contributed to the always pleasant atmosphere, to our discussions, and to your insights.

I am grateful to the Fondation de France for providing the financial support and resources that were vital for this research, to the IDRIS and Genci for the access to the Jean Zay supercomputer, and to NeuroSpin for its computational support via the Kraken cluster, and the newborn “InDev” workstation. And finally, I would like to thank all the awesome contributors and principal investigators of the open databases we used, and the open source softwares and libraries that were used in this work.

Contents

1	Introduction	15
1.1	General introduction	15
1.1.1	Anatomy of the Hippocampal Subfields	16
1.1.2	Connectivity of the Hippocampal Subfields	19
1.2	Segmenting the Hippocampal Subfields in MRI	21
1.3	Functional Connectivity of the Hippocampal Subfields in rs-fMRI	23
1.4	Aims and Hypotheses	26
1.4.1	Structure of the Dissertation	26
1.4.2	General Objective	26
1.4.3	Aims and Hypotheses	26
1.4.4	Experimental Contributions	28
2	Material and Methods	31
2.1	Databases	31
2.2	Preprocessing	33
2.2.1	Structural MRI	33
2.2.2	Functional MRI	33
2.2.3	Manual Segmentation	34
2.2.4	Divergent Classes Handling	35
2.2.5	Parcellation Atlas	35
2.3	Computational Resources	35
3	Experimental Contributions	37
3.1	Segmenting the Hippocampal Subfields: Anatomical Trajectories	37
3.1.1	Attention-Gated 3D CapsNet for Robust Hippocampal Segmentation	40
3.1.2	A Fast and Robust Hippocampal Subfields Segmentation: HSF Revealing Lifespan Volumetric Dynamics	64
3.1.3	Additional Discussion	78
3.2	Exploring the Functional Maturation of the Hippocampal Subfields	80
3.2.1	Can we Agree? On the Rashōmon Effect and the Reliability of Post-Hoc Explainable AI	82
3.2.2	Charting Hippocampal Development with Robust Explainable AI and Resting-State fMRI	97
4	Discussion	109
	Bibliography	123
A	Appendix 1: Hippocampal Subfield Volumes and Memory Discrimination in the Developing Brain	139

List of Figures

1.1	Anteroposterior Segmentation and Hippocampal Subfields	17
1.2	Coronal slices of the hippocampus	17
1.3	The hippocampal circuitry involved in memory encoding and retrieval	19
1.4	Structural and functional maturation of hippocampal subfields using neuroimaging and artificial intelligence: manuscript overview	27
2.1	Example of a formerly excluded subject	33
3.1	Achitecture of the dissertation: Anatomy	38
3.2	Random segmentation examples of the three datasets	60
3.3	Our proposed Attention Gates	60
3.4	Attention-Gated SegCaps for volumetric segmentation (3D-AGSCaps)	61
3.5	Example of attention map given an input x and its reconstruction y	61
3.6	Comparison between automatic segmentations against manual labeling	62
3.7	Mean segmentation quality with an increasing degree of random rotations	63
3.8	Technical description of ROILoc	68
3.9	Complete overview of HSF	69
3.10	Segmentation example from a random subject	71
3.11	Cumming estimation plots comparing HSF	71
3.12	Lifespan dynamics of hippocampal subfields	73
3.13	Normalized anteroposterior composition of subfields	74
3.14	Qualitative evaluation of HSF on Out-Of-Distribution samples	79
3.15	Achitecture of the dissertation: Functional Connectivity	81
3.16	Usual Perception of the Prediction-Explanation trade-off	84
3.17	A simplified example of a Rashomon set in a two-dimensional hypothesis space	85
3.18	Proposed framework for XAI through model validation and consensus finding	86
3.19	Impact of the sample size on model performance	90
3.20	Impact of the sample size on top_j similarity between model explanations	91
3.21	Impact of the convergence of models towards the best consensus available	92
3.22	Overview of the pipeline for modeling age from hippocampal subfields connectivity	101
3.23	Illustrative example of a superposition between the Glasser Atlas and HSF's ROIs	101
3.24	Learning curves showing model performance as a function of sample size	103
3.25	Relationship between model error and explanation agreement	103
3.26	Analysis of model explanation agreement with an increasing number of top features	104
3.27	Cortical regions with predictive functional connectivity to hippocampal subfields	104
4.1	Structural and functional maturation of hippocampal subfields using neuroimaging and artificial intelligence: manuscript overview	110

Forgetting is not merely a simple vis inertiae, it's rather an active inhibitory faculty, a positive faculty in all strength of the term. (...) Temporarily closing the doors and windows of consciousness; to set ourselves apart from the noise; to create a bit of silence, of tabula rasa within our consciousness to make room for the new, to enable ruling, anticipating, deciding in advance, this is the utility of active forgetfulness, a kind of usher, guardian of psychic order, of tranquility, of etiquette: one can immediately see why without forgetfulness there could be neither happiness, nor serenity, nor hope, nor pride, nor *present*.

Nietzsche, *The Genealogy of Morals*

1 Introduction

1.1 General introduction

The hippocampus, a multifaceted structure that resides within the medial temporal lobe, is involved in a wide range of cognitive functions, including learning and memory (Eichenbaum & Cohen, 2004; Jarrard, 1993; Scoville & Milner, 1957; Whitlock, 2006), spatial navigation (Burgess et al., 2002; Eichenbaum et al., 1999; Jarrard, 1995) and even emotional response (Douglas, 1967; Richter-Levin & Akirav, 2000). Learning and memory are two critical, interconnected mental processes. The first is the process of acquiring new information or knowledge, while the latter is the process of retaining, storing, and recalling that information over time, which means that the hippocampus is essential for forming, organizing, and storing these memories. It helps to consolidate new information. More specifically, it is the unique place in the human brain where adult neurogenesis occurs, particularly within the dentate gyrus — one of its subfields — contributing significantly to memory formation and learning processes (Aimone et al., 2006).

Memory itself is not a unary construct, as it can be divided into multiple subcategories, such as episodic and semantic memories (Tulving, 1972). Semantic memory refers to the general knowledge of the world that we have accumulated throughout our lives. It includes common facts, concepts, and ideas. Episodic memory, however, is the memory of specific events — episodes linked to emotions and spatio-temporal details — that have occurred in our lives. Episodic memory is underpinned by the mnemonic functions handled by the hippocampus, namely pattern separation and pattern completion (Rolls, 2016). Pattern separation is the process by which one distinguishes between different inputs or experiences, even if they are very similar. On the other hand, pattern completion is the process that allows one to retrieve and reconstruct a complete memory from partial or degraded inputs. Both pattern separation and pattern completion are essential to our ability to learn from our experiences and recall them accurately. They represent the brain's flexibility in handling information, being able to keep things separate when needed but also to connect the dots when information is incomplete.

However, episodic memory represents a perpetually dynamic and evolving process, unfurling across one's life. Healthy aging process has a profound impact on the hippocampus and thus on memory and cognitive function. Before age two, the period known as infantile amnesia dominates, meaning memory encoding and retention is almost impossible. As the child matures, from ages two to six, they traverse the phase of childhood amnesia, marked by the forgetting of the majority, albeit not the entirety, of episodic memories Bouyeure and Noulhiane, 2020. The strong neurogenesis that occurs in the dentate gyrus of children has been proposed to be the main reason for their increased forgetting and poorer memory retention, compared to adults (Frankland et al., 2013). Episodic memory reaches its peak during early adulthood and starts to decline after the end of the first adult decade (21 to 31 years) (Cansino, 2009). On the other hand, the hippocampus is involved in several mental health conditions, including depression and anxiety (Sheline et al., 1999), Alzheimer's disease (Rao et al., 2022), drug-resistant epilepsy (Cendes, 2005), posttraumatic stress disorder (Shin, 2006), and

schizophrenia (Van Erp et al., 2016), where hippocampal atrophy presents as one of the earliest macroscopic changes. Moreover, the functions operated by the hippocampus are influenced by extrinsic variables. Elements such as physical exercise or an intellectually challenging environment with many affordances, can foster adult neurogenesis, synaptic plasticity, and cognitive faculties associated with learning and memory. This highlights the importance of lifestyle determinants in maintaining the robustness of hippocampal health (Erickson et al., 2011; Gregorians & Spiers, 2022; Kubie et al., 2020).

With its implication to core behavioral functions, its impact on child development, and its implication in many health conditions, the study of the hippocampus and its subregions using *in vivo* MRI is of great importance. This non-invasive neuroimaging approach helps visualize the complex structure of the hippocampus, providing unprecedented insight into its functional organization and role in cognitive processes. Being able to quantify and monitor changes in hippocampal volume, morphology, and connectivity contributes to our understanding of how this important brain structure evolves throughout life and responds to different environmental changes. In addition, MRI can help us understand the pathophysiology of many neurological and psychiatric disorders in which hippocampal abnormalities are often detected. Examining the hippocampus with *in vivo* MRI provides a deeper understanding of the hippocampal contribution to learning and memory. However, analyzing the hippocampal subfields in MRI presents manifold challenges to which the present work proposes solutions. We dedicated the present thesis not only to the developmental trajectories of the hippocampal subfields but also to the development of novel and innovative methodologies.

1.1.1 Anatomy of the Hippocampal Subfields

The hippocampus, anatomically and functionally interconnected with the entorhinal, perirhinal, and parahippocampal cortices, constitutes the medial temporal lobe. It can be conceptually divided in two primary ways (Figure 1.1): an anteroposterior division, following a longitudinal specialization (Vos de Wael et al., 2018) that separates the hippocampus into head, body and tail sections; or a subdivision into distinct subfields, namely the Dentate Gyrus (DG), the Cornu Ammonis (CA1-4), and the Subiculum (Sub), each of these subfields possessing a unique myelo- and cyto-architectony (Duvernoy et al., 2013). The motivation behind the study of the hippocampus in subfields rather than in anteroposterior delineations, comes from the distinct anatomo-functional properties each subfields possesses, promoting a more nuanced understanding of the role of the hippocampus in cognitive processes and neurological disorders (e.g. Pluta et al., 2012 and La Joie et al., 2013).

To further elaborate on the distinctive morphology of its subfields, the hippocampus exhibits bilaminar characteristics, with CA curled over the DG (Figure 1.1). Despite the considerable discrepancies in the literature on the delineation of the hippocampal subfields (e.g. de Flores et al., 2015), we can divide CA into four segments. CA1 is the continuation of the subiculum, followed by CA2, CA3, and finally CA4, which is completely enveloped within the concavity of the DG (Duvernoy et al., 2013). As current *in vivo* resolution and signal-to-noise ratio in MRI do not allow for a clear distinction between CA4 and DG, it should be mentioned that a common choice in the literature is to merge CA4 and DG since Blackstad, 1956 grouped them into the “area dentata”. Observing the hippocampus from anterior to posterior (Figure 1.2), we see both the subiculum and CA1 emerging in the head of the hippocampus, followed by the dentate gyrus. Just before the body, CA2 and CA3 appear (Berron et al., 2017). Although the tail is especially difficult to

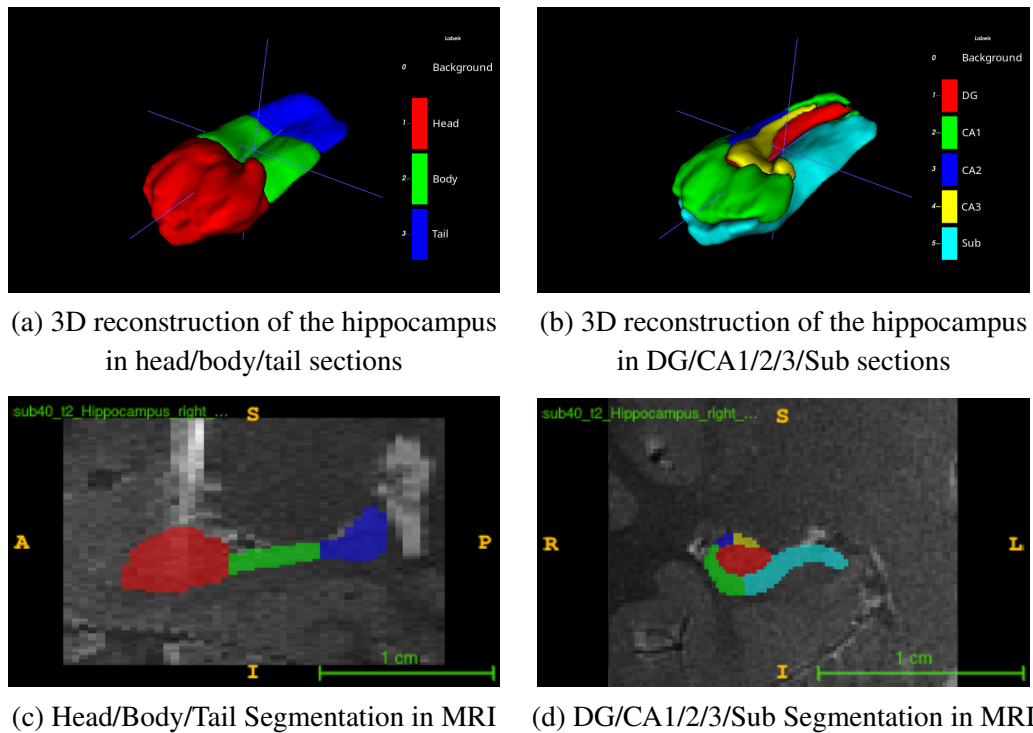


Figure 1.1: Anteroposterior Segmentation and Hippocampal Subfields. A comparison of the two most common ways to delineate the hippocampus, either in head, body, and tail regions, or in subfields with the Dentate Gyrus (DG), the Cornu Ammonis (CA1/2/3), and the Subiculum (Sub). Image from the HIPlay7 dataset.

accurately segment in MRI due to its complex anatomy, the body usually ends with thinning of CA2-3, progressive disappearance of the DG at the beginning of the tail, leaving only the subiculum and CA1 left at the posterior end of the hippocampus (Dalton et al., 2017).

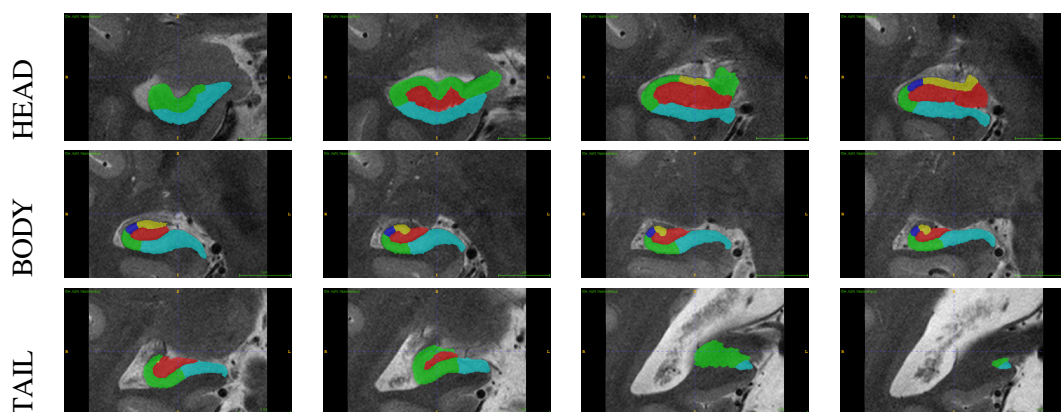


Figure 1.2: Coronal slices of the hippocampus from anterior (top left image), to posterior (bottom right image). First, second, and third rows are respectively the hippocampal head, body, and tail. The dentate gyrus is in red, CA1, CA2, and CA3 are respectively in green, blue, and yellow, and the Subiculum is in cyan. Image from the HIPlay7 dataset.

Adding to overall complexity, the hippocampus generally increases in volume during childhood up to early adulthood (Uematsu et al., 2012). More specifically, Ziegler et al., 2012 observed an increase in gray matter volume during adulthood in the hippocampus up to 41 years, with a maximum of 62 years for DG and CA, followed by fast atrophy. This is in agreement with the described quadratic relationship between the total volume of the hippocampus and age followed by an inflection point at 63 years of age, a strong negative correlation between volume and age (X. Yang et al., 2013). Furthermore, studies in non-human primate animals (e.g. Lavenex and Banta Lavenex, 2013) have shown that subfields such as DG, CA2-3 and subiculum grow asynchronously until adulthood, suggesting differential maturation. However, this question has only recently been addressed in human children and adolescents with inconsistent results. According to Ellis et al., 2021, DG shows a very rapid growth in infants, doubling in size, associated with an increase in the volumes of CA1 and CA3 during development (8-14 years of age). This contrasts with a stable or slight linear decrease in subicular volumes (Lee et al., 2014; Ziegler et al., 2012). In normal aging, the data suggest a volumetric decrease in all subfields that predominate in DG (de Flores et al., 2015; Foster et al., 2019). Subicular atrophy is also a hallmark of aging, and even more so in Alzheimer's disease (Chételat et al., 2008; La Joie et al., 2010; Lace et al., 2009), partly due to a decrease in the number of neurons (Simis et al., 1997). To address the lack of studies that examine the development of the hippocampal subfields in relation to memory discrimination abilities in childhood, we conducted an initial investigation of this question (Appendix A, Bouyeure et al., 2021). If previous work suggested that memory discrimination continues to improve during childhood, the timeline is unclear. Long maturation of the hippocampal subfields is believed to support gains in memory processes, but direct evidence linking subfields development and memory discrimination is scarce, especially research separating dentate gyrus and CA3 contributions. We examined memory discrimination and manually segmented volumes of the hippocampal subfields in 26 children aged 5-12 years. We found that memory discrimination performance improved linearly with age during this developmental period. Examination of the volumes of the subfields showed different developmental pathways, with CA1 and subiculum volumes increasing linearly with age, while CA2-3 and dentate gyrus volumes were not correlated with age. Regression analyzes revealed that larger volumes of CA2-3 and subiculum were associated with better memory discrimination, even when controlling for age. Furthermore, the relationship between subiculum volume and memory discrimination was moderated by age. In the discussion, we suggest that our results confirm a role for CA2-3 in pattern separation and highlight a potential contribution of the subiculum during development. In general, our findings indicate that prolonged maturation of certain subfields of the hippocampal is related to improvements in memory discrimination abilities in childhood.

But the anatomy of the subfields is only a part of their complexity. The DG and subiculum are the gates of the hippocampus to other parts of the brain. In addition to their intrinsic computational functions, they transmit information through the monosynaptic and trisynaptic pathways from the entorhinal cortex to the other subfields (Figure 1.3). The essential functions that the hippocampus performs rely on specific connectivity, not only intra-hippocampus but also with other regions of the brain responsible for the gathering of information that will be processed and eventually stored as memories.

retrieval, while independently, and thanks to its recurrent collaterals, CA3 enables the retrieval of comprehensive memory representations even in the presence of partial cues, — a process known as pattern completion (Ngo et al., 2021). This permits the recollection of memories from signals that are similar, albeit not identical. In contrast, the CA1 region amalgamates and compares inputs from both CA3 and the entorhinal cortex (Yassa & Stark, 2011), thereby being responsible for the detection of matches or mismatches. This function facilitates the comparison of input patterns with stored memory traces, enabling judgments on novelty or familiarity. It is also responsible for consolidation and late memory retrieval (Mueller et al., 2011). The Subiculum, however, receives most of its input from CA1. However, it has functional properties that seem to be independent of the rest of the hippocampus, allowing it to separate multiple types of information and, hence, support memory and spatial functions (Aggleton & Christiansen, 2015).

Early and late retrieval refer to different time periods after learning during which a memory can be recalled. Early retrieval is more dependent on hippocampal mechanisms, as memories have not yet been fully consolidated, which means the hippocampus is needed to reconstruct the memory trace. Late retrieval occurs after consolidation, when the cortical regions can directly retrieve memory, meaning that the hippocampus plays less of a role, as it is only needed when contextual details are recalled. This consolidation process pertains to the gradual reorganization and strengthening of post-learning memory traces, necessitating communication between the hippocampus and the neocortex to solidify representations for long-term storage (Lavenex & Amaral, 2000; Winocur et al., 2010). This means that different anatomical systems allow multiple functions such as episodic memory, affective and social learning, sensory processing, and integration (Aggleton, 2012).

The hippocampus interacts bidirectionally with medial temporal cortical regions such as the entorhinal, perirhinal, and parahippocampal cortices, forming distinct processing streams for object and spatial information. As stated above, the entorhinal cortex is an interface between the hippocampus and the neocortex, two agents communicating for different purposes. This communication is essential to theories — such as the Standard Consolidation Theory (SCT) — defining memories as being first encoded by the hippocampus, then laid down as a hippocampal-neocortical ensemble with the sparsely encoded hippocampal neurons referencing and activating the neocortical neurons to recreate the content of an experience (Sekeres et al., 2018; Teyler & DiScenna, 1986; Teyler & Rudy, 2007). Therefore, the hippocampus interacts with perceptual and spatial systems in the posterior neocortex to represent fine-grained details of memories, but also interacts with conceptual systems like the medial prefrontal cortex to represent coarser, more global aspects of memories (Robin & Moscovitch, 2017; Sekeres et al., 2018). In addition, the hippocampus is also connected to subcortical structures such as the amygdala, nucleus accumbens, thalamus, and mammillary bodies, allowing emotional and motivational modulation of memory (Koelsch, 2014; Pessoa, 2017).

Although volumetric changes in the hippocampal subfields associated with healthy aging have been characterized, little is known about how functional connectivity of the subfields may be affected. Recent work by Dalton et al., 2019 examined the functional connectivity of the hippocampal subfields in young and old adults. They found a reduced connectivity between CA1 and the subiculum in older adults compared to the younger group. More research is needed to clarify how subfields functional connectivity is continually altered by development and aging processes. Investigating the functional connectivity of the subfields longitudinally throughout the lifespan could elucidate the timeline of these connectivity changes and, in fine, shed light on the neurobiological mechanisms of age-related memory changes.

1.2 Segmenting the Hippocampal Subfields in MRI

Semantic Segmentation is a concept in the field of computer vision which refers to the process of dividing an image into segments or “parts” to simplify image analysis. It is said “semantic”, because the result of such an algorithm is an image that has the same shape and structure as the original, but instead represents groups of pixels that belong to a semantically meaningful class. In a nutshell, semantic segmentation is like teaching a computer to see and understand an image just like a human would, by dividing it into meaningful parts and understanding what each part represents. In the context of magnetic resonance imaging (MRI), these categories could be various types of tissues, as in our case, the hippocampal subfields. This delineation process enables the study of structural patterns, which may in fine lead to a better comprehension, diagnosis, and prognosis of the aforementioned diseases, as segmentation allows one to easily derive the geometry, shape, and size of a given region of interest.

The segmentation process within MRI is a challenging task due to a multitude of factors. First, managing interindividual variability is crucial, complicating the development of a universally applicable algorithm for semantic segmentation. Additionally, the quality of the image itself poses concerns. As reported by L. E. M. Wisse et al., 2021, the segmentation of the hippocampal subfields may be reliable only for submillimetric scans. However, low resolution is not the only culprit, as artifacts originating from various sources such as motion, patient’s condition, or even hardware-related noises, could potentially disrupt the segmentation procedure. Finally, the inherent complexity of the task cannot be overlooked. The differentiation between various tissue types poses a significant challenge due to an insufficient contrast, a source of uncertainty that only a histological approach — impossible to carry out in-vivo — could resolve. More specifically, structures such as CA1 and Subiculum, despite their distinct histological characteristics, appear indistinguishable on MRI due to identical contrast properties (Canada et al., 2023; Yushkevich, Amaral, et al., 2015).

The intricacy of the task has allowed the creation of numerous manual segmentation protocols. However, these methodologies often exhibit a marked degree of incompatibility with one another (L. E. Wisse et al., 2017; Yushkevich, Amaral, et al., 2015). The vast majority of these protocols incorporate the use of geometrical heuristics, primarily in an endeavor to map histological features onto MRI. This is evidenced by Berron et al., 2017; Dalton et al., 2017, among others. Notwithstanding, manual segmentation, although widely regarded as the gold standard, is an intricate, labor-intensive, and subjective task. It is susceptible to errors, thereby compromising its reproducibility. This issue is particularly pronounced at the head and tail of the hippocampus, where tissue ambiguities are the most important. The border between the subiculum and CA1, for example, highlights such discrepancies, being the most variable part of the segmentation (Yushkevich, Amaral, et al., 2015). Added to this, an additional source of high variability is the delineation of the constitutive Cornu Ammonis substructures. Some methodologies merge CA2, CA3, and CA4 La Joie et al., 2010, others combine CA3 and CA4 with DG (Mueller et al., 2007), some only CA2 and CA3 (Van Leemput et al., 2009), or simply merge CA4 and DG (Berron et al., 2017; Dalton et al., 2017; L. Wisse et al., 2012). There are even instances where the head and tail are excluded from segmentation due to the heightened complexity in these regions (e.g. Berron et al., 2017; L. Wisse et al., 2012). This scarcity of standardization further obfuscates the process and impairs the comparability of the results between studies.

Although there is a wealth of manual segmentation guidelines, a number of tools are also available for the automated segmentation of hippocampal subfields. As explicated by Chiappiniello et al., 2021 and Fragueiro et al., 2023, the selection of an automated segmentation pipeline can influence both the volumes and the test-retest reproducibility of the volumes of the human brain hippocampal subfields, invariably leading to divergent interpretations. The predominant tool for handling this task is a powerful and multi-purpose tool named FreeSurfer (Iglesias et al., 2015). However, due to its useful automated features exceeding the scope of this thesis, it generates a multitude of outputs, thereby necessitating an extended computation time, rendering it inappropriate for scientific studies with a singular focus on a specific substructure of the human brain. Its inference time, approximately ten hours per subject, is comparatively slower than manual segmentation of the hippocampal subfields. Despite previous studies affirming that FreeSurfer's segmentation quality is satisfactory for studying hippocampal subfields (Schmidt et al., 2018), others have demonstrated that FreeSurfer's segmentation quality is inferior compared to more contemporary tools (de Flores et al., 2015; DeKraker et al., 2022), with segmentations that conflict with established anatomical boundaries, resulting in significantly different volumetry, particularly in the head and tail of the hippocampus (Wisse et al., 2014). In addition to FreeSurfer, these novel tools include ASHS (Yushkevich, Pluta, et al., 2015), HIPS (Romero et al., 2017), and more recently HippUnfold (DeKraker et al., 2021). They offer superior segmentations, more closely aligned with manual segmentation, but they also present their own set of challenges. For example, if left unsupervised, ASHS can misclassify or exclude some or all regions in final slices, or, in certain instances, add slices, relative to the range defined and segmented by the manual segmentation protocol (Bender et al., 2018).

Moreover, none of them implement cutting-edge, end-to-end deep learning, which has been established to be more resilient and generalizable to new observations, particularly in complex and nonlinear tasks (O'Mahony et al., 2020). Recent studies have underscored the potential benefits of end-to-end deep learning for hippocampal segmentation (Chen & Liu, 2021; Q. Qiu et al., 2019; Z. Yang et al., 2020; Zhu et al., 2019), promising fast inference time (less than a minute per subject versus several hours for FreeSurfer), increased accuracy, and robustness to anatomical variations. Regrettably, most deep learning solutions are presently offered merely as a proof-of-concept, with either no public implementation, no pre-trained models, or are trained on limited and specific datasets, thereby restricting their generalizability. The literature lacks an end-to-end deep learning segmentation model trained on a heterogeneous database to guarantee segmentation quality across contrast, magnetic field intensity, age range, and health condition.

Despite the aforementioned challenges, the development of a novel automated segmentation pipeline represents a promising opportunity. Such a tool could enhance the reproducibility of the research and ameliorate the onerous and time-consuming nature of manual segmentation protocols, allowing researchers to dedicate more time to hypothesis testing. Moreover, the techniques utilized for semantic segmentation are widely applicable across medical imaging modalities, and hence research in this domain can catalyze progress in ancillary fields. Most critically, automated semantic segmentation of the hippocampus could allow earlier detection and diagnosis of neurological diseases, leading to improved patient outcomes. By providing granular and personalized understanding of subtle anatomical variations, it can pave the way for more targeted and individualized treatment regimens. In summary, the confluence of computer vision and medical imaging via semantic segmentation constitutes an auspicious frontier that can significantly

propel hippocampal research and clinical practice. The development of a robust end-to-end deep learning solution for hippocampal subfields segmentation would constitute a momentous milestone in the realization of these possibilities.

With the emergence of sophisticated segmentation techniques, current research has improved its ability to handle the intricacies of the hippocampal subfields. Reliable segmentation allows for granular morphological analyses as well as probing the functional roles of the hippocampal subfields. In particular, by enabling the isolation of subfields from MRI data, segmentation methods permit the analysis of functional connections and networks anchored in specific subfields of the hippocampus. Functional connectivity assessed using resting-state fMRI offers a window into the intrinsic functional architecture of the hippocampal circuits. Consequently, there has been a surge of enthusiasm in the exploration of their functional characteristics. This interest is fueled by the prospect that subfield-specific functional connectivity disruptions may provide sensitive biomarkers for neurological diseases, which constitutes a pivotal research avenue.

1.3 Functional Connectivity of the Hippocampal Subfields in rs-fMRI

As formulated by Bijsterbosch, 2017, when a particular region of the brain is activated, it consumes oxygen. Utilizing this fundamental principle, functional Magnetic Resonance Imaging (fMRI) gauges cerebral activity by discerning alterations in blood oxygenation, providing us with a dynamic representation of brain activity. The study of functional connectivity is indispensable, as it provides a comprehensive understanding of how different areas of the brain interact and communicate, thereby revealing the intrinsic organization of the brain. When an individual is in a resting state, that is, not engaged in any task-specific endeavors, their brain undertakes a substantial amount of its background work, such as the consolidation of memories and the processing of experiences. The relevance of resting-state fMRI (rs-fMRI) is underscored by its ability to provide valuable insights into the brain's internal functioning during states of passive cognition and its potential to highlight abnormalities associated with various neurological and psychiatric disorders. By examining the statistical relationship between the time series derived from fMRI during a resting state, we can explore the complex network of interactions that underlie the brain's inherent functionality. One such potential statistical relationship may be a measure of correlation, where a high temporal correlation between two regions suggests that they are working together or are constituents of the same network. These types of study have been shown to be significant, providing information on the functional organization of the hippocampus in activities such as spatial navigation (Rodriguez, 2009), episodic memory formation (Greicius et al., 2003), and stress response regulation (Sandner et al., 2021).

rs-fMRI has become an increasingly used technique over the past decade to study intrinsic brain organization and connectivity patterns. In contrast to task-based fMRI, resting state requires subjects to simply lie in the scanner without performing an active paradigm, making data acquisition relatively straightforward. This presents several advantages: specialized equipment is not needed for stimulus presentation, scan sessions are easier for investigators to set up, and strict expertise in experimental design is not as critical. In particular, the resting state is more accessible across diverse subject populations compared to task fMRI. Patient groups who may struggle to perform cognitively demanding paradigms can more readily undergo resting-state scans. The lack of complex

task demands also makes resting state feasible to apply across the lifespan, from early infancy through old age. These practical benefits suggest a strong potential for the resting state to be used in clinical settings as an objective biomarker for mental disorders. Even when behavioral output appears to be unaffected, underlying disruptions in intrinsic connectivity can serve as an early indicator of pathology or risk. Longitudinal changes in resting networks could chart disease progression and act as a marker of treatment response. Therefore, the ease of data acquisition combined with the possibility to detect subtle brain abnormalities position the resting state fMRI as a promising tool for diagnosis, monitoring of the course of the disease, and evaluating interventions. Although resting-state networks show robust and consistent patterns across subjects, an important question is how stable these networks are within the same individual over time. Test-retest reliability studies have found moderate to high intrasubject reproducibility of resting-state networks (Braun et al., 2012). However, if resting-state patterns are influenced by the underlying neuroanatomical structure, they are also affected by other local neuronal dynamics (Deco et al., 2011), — some variability is observed, underscoring the potential influence of current mental state on intrinsic connectivity. Functional connectivity derived from analysis of resting state could be uncorrelated with the underlying anatomy (Honey et al., 2009; Honey et al., 2010). Several limitations must be considered when interpreting resting-state fMRI. Confounding factors like head motion and physiological noise can induce spurious correlations between regions. BOLD fMRI provides only an indirect measure of neural activity, limiting inferences about the specific neurophysiological processes underlying connectivity. There is also a lack of ground truth for validation, as the “true” networks remain unknown. Open sharing of high-quality datasets, such as those provided by the Human Connectome Project (HCP), is invaluable to enable rigorous testing of the reliability and reproducibility of connectivity methods. Overall, while rsfMRI holds great promise, care must be taken to validate findings in data sets, replicate results, control for confounding information, and integrate information from complementary modalities.

A variety of analytical techniques have been developed to examine rs-fMRI data. Broadly, these can be categorized into voxel-based and node-based (or network-based) approaches. Voxel-based methods, such as independent component analysis (ICA), aim to identify spatially distributed networks by finding voxels with similar temporal dynamics. Maps of large-scale networks can be derived such as default mode, dorsal attention, salience, and executive control networks. In node-based analysis, the brain is first segmented into distinct nodes or regions of interest. Timeseries are extracted for each node, connectivity is calculated between all node pairs to estimate network edges, and these are assembled into a matrix describing the full network topology. Edges can simply reflect connectivity strength, often based on Pearson’s correlation between node time-series given their ease of computation and interpretability. However, other measures, such as partial correlation and Granger causality can estimate directional or causal relationships between nodes (Bijsterbosch, 2017; Smith, Vidaurre, et al., 2013; van den Heuvel & Hulshoff Pol, 2010; J. Yang et al., 2020).

The data-driven and multivariate nature of artificial intelligence (AI) and machine learning makes these approaches well suited to analyze complex resting-state fMRI connectivity patterns. Machine learning algorithms can identify distributed connectivity features that differentiate patients from healthy controls even when there are no visible behavioral symptoms, as exemplified by Gallo et al., 2023; Tejwani et al., 2017; Yan et al., 2020. Additionally, sophisticated methodologies such as deep learning possess the ability to deduce biomarkers by isolating pivotal patterns within the data. By interpreting these acquired features, we can accumulate knowledge and foster a more profound understanding of the underlying processes. Beyond binary classification, machine learning can be applied in a more nuanced manner to predict specific symptom domains or capture network-based

associations with continuous clinical variables. Together, these emerging AI applications underscore the wealth of information contained within functional connectivity data from the resting state that can be harnessed to explore brain-disorder relationships and support precision diagnosis. However, care must be taken to avoid overfitting and ensure generalizability of predictive models across diverse cohorts of patients.

Recent advances in high-resolution structural MRI theoretically enable unprecedented segmentation of the boundaries of the hippocampal subfield, providing new opportunities to explore the functional connectivity of these cytoarchitecturally distinct subregions during resting state. However, accurately delineating such small and heterogeneous structures presents great challenges. The simple use of an atlas — as traditionally done with bigger regions — proves inadequate, particularly given the inherently coarse spatial resolution of fMRI. Nevertheless, characterizing intrinsic networks associated with hippocampal subfields may illuminate their differential involvement in memory encoding and retrieval, spatial cognition, and psychiatric physiopathology. Realizing this potential requires both (i) a sophisticated subfields segmentation tool that reliably maps the structure of the hippocampus to avoid mislocalization of fMRI time series and (ii) analytical methods that properly model functional correlations while accounting for uncertainty and minimizing spurious associations. The complexity of this multivariate endeavor makes it susceptible to reliability issues. Thus, elucidating the intricate functional connectivity fingerprints of hippocampal subfields — distinctive patterns that reliably distinguish individuals based on their age — requires integrating precise structural maps with sophisticated computational modeling. Furthermore, well-curated datasets and rigorous independent validation will be imperative for the determination of biologically valid biomarkers from these specialized resting-state networks. Overall, improved subfields segmentation techniques are essential to unlock the potential of rs-fMRI to illuminate the hippocampal microcircuitry. While some studies have investigated functional connectivity between the subfields of the hippocampus, their scope has been limited. Dalton et al., 2019 and Stark et al., 2021 examined subfields connectivity differences along the longitudinal axis and with neighboring extra-hippocampal cortices. Task-based fMRI work has provided support for subfields roles in episodic memory (Maass et al., 2014; Yassa et al., 2010). In resting-state fMRI, several studies have segmented the hippocampus into head, body, and tail rather than subfields (Damoiseaux et al., 2016; Das et al., 2013). Hao et al., 2020 assessed subfields connectivity but used a coarse resolution of 3x3x3 mm. Other subfields resting-state studies have relied on group comparisons between young and old rather than examining continuous age-related changes (De Flores et al., 2017; T. Qiu et al., 2022; Stark et al., 2021). The work of Sanders et al., 2023 comes closest to the current goal of relating whole hippocampal age-related connectivity to memory, but did not segment the hippocampal subfields. Therefore, to the best of our knowledge, no study has yet characterized how the functional connectivity of the hippocampal subfields changes continuously across the lifespan in relation to episodic memory abilities.

1.4 Aims and Hypotheses

1.4.1 Structure of the Dissertation

The remaining chapters are organized as follows. Chapter 2 (Material and methods) covers the datasets, manual segmentation methods, and computational resources leveraged in this work. Chapter 3 presents the two main contributions of the present dissertation (as illustrated in Figure 1.4):

1. An anatomical contribution involving the development of methodologies to segment the hippocampal subfields and their application to study the volumetry of the hippocampus across the lifespan,
2. A functional contribution applying these segmentation tools to characterize the development of intrinsic functional connectivity networks anchored in hippocampal subfields, measured with resting-state fMRI during childhood.

Finally, Chapter 4 concludes with a general discussion of key findings, limitations of the current work, and promising directions for future research into hippocampal subfields development.

1.4.2 General Objective

The goal of this dissertation is to characterize the structural and functional maturation of the hippocampal subfields across the lifespan, with a particular emphasis on changes during childhood development. The hippocampus is a critical structure for learning, memory, and broader aspects of cognition. Importantly, the hippocampal subfields follow distinct developmental trajectories that contribute to the maturation of memory and cognitive functions. This work aims to delineate anatomical and functional changes in hippocampal subfields from early childhood through older adulthood using advanced neuroimaging paired with artificial intelligence techniques. The use of large public datasets acquired as part of the Human Connectome Project provides the opportunity to study these changes at a scale not previously possible.

1.4.3 Aims and Hypotheses

The first objective concerns the anatomy of the hippocampal subfields. Our goal is to construct and validate an end-to-end deep neural network architecture optimized for the semantic segmentation of hippocampal subfields from structural T1- and T2-weighted MRI scans. To this end, we first explore the methodological space to find a data-efficient and compute-efficient deep learning method, notably through Capsule Networks and Convolutional Neural Networks. These models will be trained on a large aggregated dataset of manual subfields segmentations encompassing diverse ages, conditions, field strengths, and scanning parameters to improve generalizability. They will incorporate recent techniques such as attention mechanisms and human feedbacks to capture multi-scale contextual information and regularize training. It is hypothesized that our models will significantly improve on the accuracy of current leading automated segmentation tools, and approach the reliability of manual protocols, providing a robust option for widespread adoption. Then, we want to capitalize on the model's efficient inference to delineate subfields

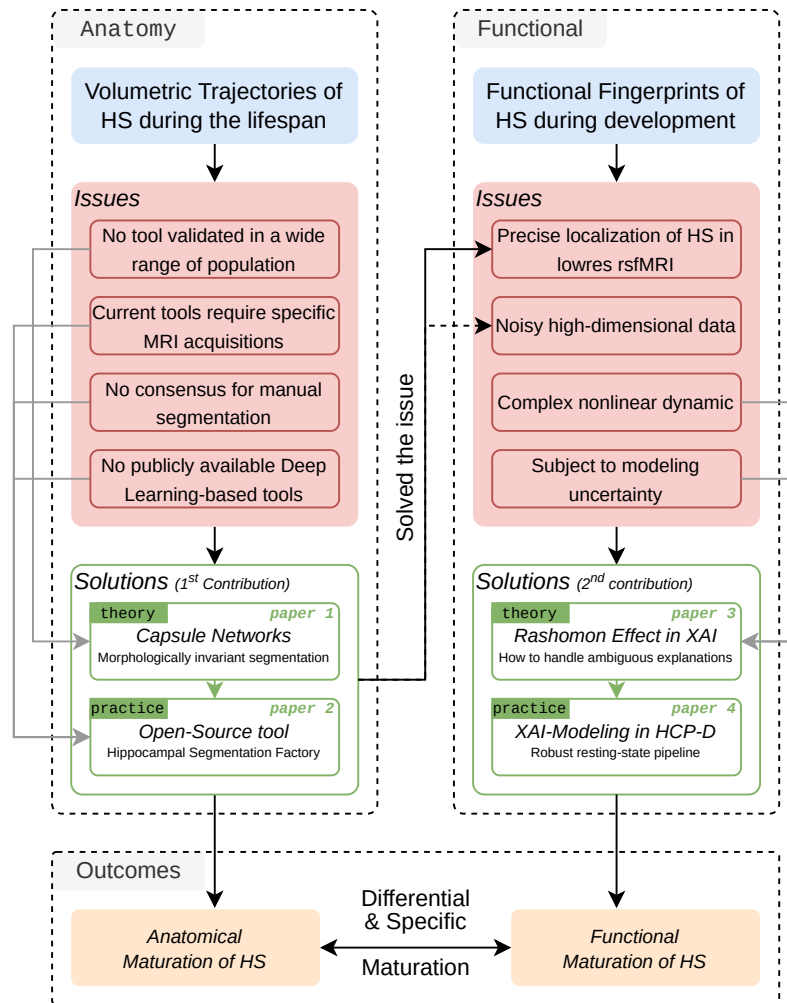


Figure 1.4: Architecture of the Ph.D. dissertation aimed at characterizing anatomical and functional changes in hippocampal subfields across the lifespan. The work is organized into four main parts with its accompanying paper: (1) a methodological exploration of innovative and disruptive segmentation methods; (2) the development of an automated deep learning tool for hippocampal subfields segmentation from structural MRI; (3) the application of this model to explore volumetric developmental trajectories in a large lifespan dataset; (4) the investigation of the maturation of functional connectivity of the subfields using resting-state fMRI and explainable AI modeling. Together, these novel tools and analytical approaches provide insights into the prolonged development of hippocampal circuits supporting gains in memory and cognition.

volumes in 3750+ subjects, aged 4 to 100 years, from the Human Connectome Project lifespan datasets. This unprecedented sample size for the segmentation of the hippocampal subfields will characterize fine-grained, nonlinear volumetric trajectories across the lifespan for each subfield. It is hypothesized that development will be prolonged for dentate gyrus and CA2/3, supporting gains in pattern separation, while CA1 and subiculum will show earlier maturation. In older age, subiculum atrophy is expected to emerge as an early biomarker, aligning with its particular vulnerability.

Trajectory modeling will quantify the acceleration, peaks, and deceleration of growth and decay of the subfields. Sex differences will also be investigated. Together, observed developmental patterns will link subfields' maturation timelines to known improvements and decline in learning and memory over the lifespan.

The second objective concerns the functional connectivity of the hippocampal subfields.

The second aim of this dissertation is to shed light on the maturation of the functional connectivity of the hippocampal subfields using resting-state fMRI, particularly during childhood and adolescent development. The hypothesis is that the intrinsic connectivity fingerprints of the subfields, measured by correlations in spontaneous BOLD signal fluctuations, will show differential developmental patterns related to their distinct computational roles. It is expected that changes in subfields connectivity will correspond to enhancements observed in episodic memory behaviors over childhood. To analyze the complex multivariate relationships between subfields connectivity, and age, machine learning methods will be applied. However, a key challenge is the Rashōmon effect in model interpretations: different AI models can produce divergent explanations for the same prediction. To account for uncertainties, we will develop a methodology to find consensus between competing explanations and validate the reliability of connectivity-age associations. This represents a novel approach to handle model ambiguity in the context of explainable AI. Overall, this study of the developmental dynamics of the functional connectivity of the hippocampal subfields using resting-state fMRI and advanced computational modeling will provide new insights into the maturation of medial temporal lobe memory circuits.

1.4.4 Experimental Contributions

This dissertation provides several key experimental contributions to address our anatomofunctional goals (Figure 1.4), each of which led to a research paper:

- Segmenting the Hippocampal Subfields: Anatomical Trajectories:
 1. Section 3.1.1, *Attention-Gated 3D CapsNet for Robust Hippocampal Segmentation* — The aggregation of a large and heterogeneous training dataset, encompassing manually segmented subfields labels from diverse ages, health conditions, field strengths, and segmentation protocols. Together, these allowed training of a robust model for widespread application. It validated a novel end-to-end deep learning tool that leverages recent advances in computer vision to enable automated and efficient segmentation of hippocampal subfields anatomy from structural MRI data,
 2. Section 3.1.2, *A Fast and Robust Hippocampal Subfields Segmentation: HSF Revealing Lifespan Volumetric Dynamics* — The use of this newly introduced tool to conduct an analysis of the volumetric trajectories of the hippocampal subfields across the lifespan in over 3750 subjects from the HCP, providing new insights into anatomical maturation and decay.

- Exploring the Functional Maturation of the Hippocampal Subfields:
 3. Section 3.2.1, *Can we Agree? On the Rashōmon Effect and the Reliability of Post-Hoc Explainable AI* — The development of a Machine Learning framework to enforce robustness to modeling variability caused by an effect called the “Rashōmon Effect” where ambiguity cause multiple models to base their predictions on a varying, inconsistent, subset of features,
 4. Section 3.2.2, *Charting Hippocampal Development with Robust Explainable AI and Resting-State fMRI* — The exploration of changes in intrinsic functional connectivity of hippocampal subfields, analyzed through our introduced explainable AI framework to account for uncertainties inherent in resting-state fMRI modeling.

In general, the development of this automated and open source segmentation pipeline — named HSF and available at <https://hsf.rtfid.io/> — provides the foundation to relate anatomical changes in hippocampal subfields volumes to the maturation of their functional connectivity patterns. As subfields networks measured with resting-state fMRI show prolonged development over childhood in parallel with gains in memory performance, the volumetric growth trajectories can help explain the neurobiological mechanisms underlying these functional enhancements. Together, multimodal characterization of structural and functional hippocampal circuits illuminates their coordinated maturation supporting memory formation and cognitive development.

2 Material and Methods

The experimental contributions presented in this dissertation relied on a combination of MRI datasets, manual segmentation protocols, computational resources, and customized image processing pipelines. Structural and functional MRI data from multiple public and private sources were aggregated to compile a diverse dataset of human brain images across the lifespan, acquired on scanners ranging from 1.5T to 7T field strengths. Manual segmentation of hippocampal subfields was performed primarily following the protocol by Berron and colleagues (2017), with additional datasets following comparable guidelines. Manual labels were reconciled to match the Berron protocol. The experiments leveraged high-performance computing clusters that provide hundreds of GPUs for parallel processing. Custom pipelines tailored preprocessing, analysis, and evaluation procedures to the specific objectives of each study. This methodological foundation enabled the development of advanced tools for the analysis of the hippocampal subfields through a synergistic application of human expertise and artificial intelligence.

2.1 Databases

To produce the aforementioned experimental contributions (Figure 1.4), the present work could not have been made possible without the use of both public and private datasets. Addressing the methodological intricacies associated with the segmentation of hippocampal subfields, data acquisition, and segmentation pose significant challenges. However, as Gashler et al., 2008 puts it, a “small heterogeneous is better than large and homogeneous”. Therefore, this work relies on the aggregation of both internal and public datasets, combining a diverse population, multiple scanners, health conditions, image contrasts, magnetic field intensity, resolution, and segmentation guidelines. The rationale behind our decision to train on as diverse data as feasible was the assertion that interpolation is less complex than extrapolation, — training a model on excessively homogenized data will precipitate its failure to generalize on out-of-distribution samples. The datasets used in this work are described in Table 2.1. The IXI dataset is accessible under the Creative Commons CC BY-SA 3.0 license at <https://brain-development.org/ixi-dataset/>. The HIPlay7 dataset is an internally curated dataset (funding registration ANR-16-NEUC-0001-01). The HCP datasets were partially provided by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) financed by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research and by the McDonnell Center for Systems Neuroscience at Washington University.

fMRI data came from the HCP Development dataset. The rs-fMRI data were acquired using a multiband echo-planar imaging sequence with the following parameters: repetition time (TR) = 720 ms, echo time (TE) = 33 ms, flip angle = 52 degrees, and isotropic spatial resolution of 2 mm. Whole brain coverage was achieved using 72 slices with a field of view (FOV) of 208 x 180 mm.

Dataset	Ages	Contrasts	Fields	Resolutions (mm)	Condition	1	2	3
IXI ^{1,2}	20-86	T1 & T2	1.5 & 3T	0.94*0.94*1.2				×
Winterburn et al., 2013	29-57	T1 & T2	3T	T1: 0.6*0.6*0.6 T2: 0.3*0.3*0.3				×
Kulaga-Yoskovitz et al., 2015	21-55	T1 & T2	3T	T1: 0.6*0.6*0.6 T2: 0.4*0.4*2.0		×		×
Yushkevich, Pluta, et al., 2015		T1 & T2	3T	T1: 1.0*1.0*1.0 T2: 0.4*0.4*2.0	MCI			×
Hindy et al., 2016	18-30	T2	3T	0.44*0.44*1.5				×
Simpson et al., 2019		T1	3T	1.0*1.0*1.0		×		
Bouyeure et al., 2021 ¹	4-12	T1 & T2	3T	T1: 0.9*0.9*0.9 T2: 0.45*0.45*1.0		×		×
Yushkevich et al., 2010	38-82	T1 & T2	4T	T1: 1.0*1.0*1.0 T2: 0.4*0.5*2.0	MCI/AD			×
HIPlay7 ¹	12-21	T1 & T2	7T	T1: 0.75*0.75*0.75 T2: 0.125*0.125*1.2	TLE			×
L. Wisse et al., 2016	50-68	T2	7T	T1: 1.0*1.0*1.0 T2: 0.7*0.7*0.7				×
Berron et al., 2017	19-32	T1 & T2	7T	0.44*0.44*1.1				×
Haeger et al., 2020 ^{1,2}	50-70	T2	7T	0.3*0.3*1.2				×
Shaw et al., 2020 ^{1,2}	23-29	T1 & T2	7T	T1: 0.5*0.5*0.5 T2: 0.2*0.2*0.8				×
Lagarde et al., 2021	50-84	T2	7T	0.3*0.3*1.2	SC/MCI/AD			×
HCP-Development ^{1,2}	4-21	T1 & T2	3T & 7T	0.7*0.7*0.7				×
HCP-1200 ^{1,2}	21-35	T1 & T2	3T & 7T	0.7*0.7*0.7				×
HCP-Aging ^{1,2}	35-100	T1 & T2	3T & 7T	0.7*0.7*0.7				×

Table 2.1: Specifications of the aggregated datasets. ¹ indicates manual segmentation following the guidelines provided by Berron et al., 2017. ² indicates that only a small portion of the dataset has been used because manually segmenting the entire dataset would have been intractable. “1” stands for deep learning experiments, “2” HSF training process, and “3” the functional connectivity analysis.

Phase encoding was done in the left-right direction, with balanced acquisitions of left-right and right-left encoded volumes in an interleaved fashion. The total duration of each rs-fMRI scan was 14.4 minutes. More details can be found in Smith, Beckmann, et al., 2013.

2.2 Preprocessing

2.2.1 Structural MRI

As shown in De Raad et al., 2021, the preprocessing of MRI images must be tailored to each application, therefore no general guideline can be given. As we aspired for our tools to be robust on a broad set of images, and given that most public datasets are already preprocessed, we opted for minimal preprocessing encompassing Z-Normalization and padding to ensure that the final image size is a multiple of 8, to benefit from hardware acceleration. For our internal datasets (namely HIPlay7 and datasets from Bouyeure et al., 2021, Haeger et al., 2020, and Lagarde et al., 2021), all images were manually checked for quality issues. Notably, some subjects originally excluded due to anatomical anomalies or artifacts have been reincluded to train our models (e.g., Figure 2.1). Regarding images from the HCP datasets, we used their preprocessed images.

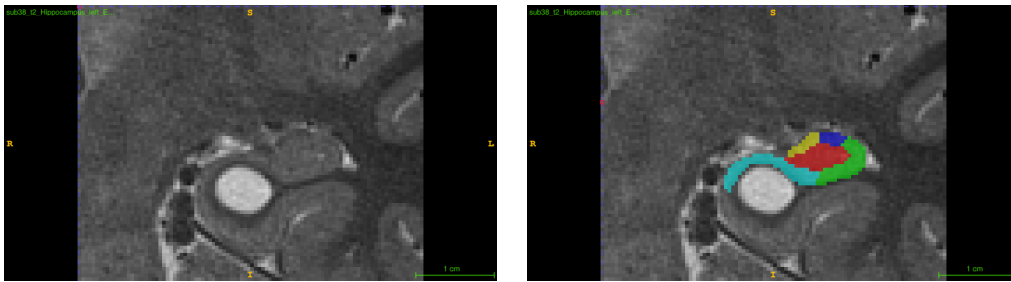


Figure 2.1: Example of a formerly excluded subject. Due to a cyst under the hippocampus that affected the morphology of the subiculum, this subject was originally excluded from the MemoDev dataset (Bouyeure et al., 2021). In our case, this atypical morphology has been included to bring diversity to the training set.

2.2.2 Functional MRI

For preprocessing of the rs-fMRI data, several steps were taken (Glasser et al., 2013): 1) Correction for geometric distortions caused by gradient nonlinearity; 2) Realignment of volumes to compensate for subject motion using rigid body transformations; 3) Correction of susceptibility-induced distortions using field maps and FSL's TOPUP tool; 4) Registration of functional data to structural T1-weighted images using boundary-based registration; 5) Resampling to standard grayordinate space by projecting cortical gray matter voxels onto surface vertices and subcortical voxels into volumetric parcels; 6) Surface smoothing using a novel geodesic Gaussian algorithm with 2 mm FWHM kernel; 7) No additional temporal filtering beyond high-pass filtering above 0.25 Hz was applied as part of minimal preprocessing. The outputs were CIFTI-format dense timeseries resampled in a common grayordinate space to enable group analyses.

2.2.3 Manual Segmentation

The hippocampal subfields were manually segmented following the detailed protocol developed by Berron et al., 2017. This protocol was chosen over other existing protocols such as L. Wisse et al., 2012 because it incorporates recent neuroanatomical findings to allow for more distinct segmentation rules and improved delineation of smaller structures in both the hippocampal head and body.

Specifically, the protocol leverages new data on subdivisions and anatomical variability in the hippocampal head and body from ex vivo MRI and extensive histological studies by Ding and colleagues (Ding et al., 2016; Ding & Van Hoesen, 2015). These seminal studies revealed novel insights into the sequential order of appearance and precise location of subregions in the hippocampal head based on digitations and folding patterns. Notably, Ding and Van Hoesen, 2015 also presented data with sections cut perpendicular to the long axis of the hippocampus, matching the orientation of typical in vivo scans. This aids translation of ex vivo anatomical boundaries to in vivo MRI segmentation, overcoming a significant barrier faced by earlier protocols relying on histological data sectioned differently.

In the Berron et al. protocol, the hippocampal head, body, and tail are first identified based on internal landmarks like the uncal apex and disappearance of the inferior and superior colliculi. The subiculum is the first subfield segmented in the head, starting 1-2 mm posterior to the hippocampal head apex where a hypointense line divides it from CA1. As the uncal sulcus progressively opens, rules incorporate new borders between subiculum, CA1, and dentate gyrus based on the sulcus and number of hippocampal digitations, following neuroanatomical findings from Ding and Van Hoesen, 2015.

Four slices anterior to the end of the head, CA3 and CA2 segmentation begins. Although CA2 emerges prior to CA3, both are segmented simultaneously in the protocol to boost reliability. The border between CA1 and CA2 is defined geometrically orthogonal to the structure at one voxel medial to the DG edge. At the uncal separation, the CA2-CA3 border is constructed halfway between the most medial CA point and lateral DG extent, allowing it to approximate anatomical variability.

Moving posteriorly into the hippocampal body, CA3 and dentate gyrus are separated along the endfolial pathway visible with 7T MRI, allowing more accurate delineation than previous protocols. For 3T scans where the pathway is not discernible, alternative geometric rules are thoughtfully provided. Subfields are merged as “hippocampal tail”, ending conservatively when the colliculi fully disappear to ensure consistent endpoints.

Comprehensive slice-by-slice examples are depicted for common sulcal variants to facilitate protocol learning. Supplementary guidelines ease application in anatomically complex cases, like decision trees for perirhinal cortex segmentation. Additionally, detailed instructions alongside thorough reliability testing make this high-resolution 7T protocol valuable for precise hippocampal subfields delineation based on new neuroanatomical insights.

2.2.4 Divergent Classes Handling

As the datasets enumerated in Table 2.1 may have manual segmentation based on different protocols, we preprocessed those labels to achieve convergence with the Berron’s protocol. To homogenize the labels, three common divergences warrant noting: 1) the segmentation of CA4, 2) the segmentation of the stratum radiatum/stratum lacunosum-moleculare (SLRM), and 3) the division of the subiculum into pre and parasubiculum. For the first case, we merged CA4 with the DG. Regarding the SLRM, the label is removed and each empty voxel is reassigned to the nearest neighboring class. Finally, for the subiculum, we merged both the pre and parasubiculum into a single class. Cysts are removed when feasible. Note that some protocols fuse classes such as CA2 and CA3, classes separated by Berron’s protocol. Additionally, some protocols do not segment the head or tail of the hippocampus, instead assigning them specific classes. These two special cases are presented in chapter 3.1 as they are handled in the HSF training process.

2.2.5 Parcellation Atlas

The functional connectivity analysis of the hippocampal subfields leveraged the surface-based multi-modal parcellation atlas developed by Glasser et al., 2016. Briefly, this 180-area per hemisphere atlas was generated using multi-modal MRI data from the Human Connectome Project, including cortical thickness and myelin maps for architecture, task fMRI data for function, and resting-state fMRI for connectivity. Gradients between areas were calculated across modalities to delineate borders. Initial borders were drawn by neuroanatomists then optimized algorithmically. Resting-state networks were used to associate each cortical area with auditory, somatosensory, visual, task positive, or task negative systems. The surface-based atlas enables more accurate localization than volume-based approaches and avoids mixing signals across tissue types or convoluted cortical folds.

A key advantage of this atlas is the use of areal feature-based alignment to precisely match cortical areas across subjects during registration. The improved intersubject alignment enabled generation of a robust group parcellation. Once generated from the group, the parcellation can be mapped to individual subjects using a cortical areal classifier driven by multi-modal fingerprints of each area. The classifier had high sensitivity and specificity for detecting individual areas. This allowed the atlas to be applied to our dataset for network-based analyses of the hippocampal subfields connectivity anchored in this precise surface-based coordinate system using the Connectome Workbench.

2.3 Computational Resources

Our experimental contributions have been supported by several institutions. Aside from the computational resources provided by the InDev team (NeuroSpin, France) used to run small algorithms and statistical analyzes, small deep learning experiments were conducted on the Kraken Cluster (8 RTX8000 GPUs) of the NeuroSpin facility. Larger computations, such as the complete HSF training process, were executed on the Jean-Zay HPE SGI 8600 cluster of the IDRIS supercomputer Center (GENCI-CNRS, Orsay, France). The latter provides hundreds of GPU nodes comprising NVidia V100 and A100 GPUs. More specifically, we employed the GPU nodes with 2 AMD Milan EPYC 7543 processors (32 cores at 2.80 GHz) — namely 64 cores per node — 512 GB

of memory per node, and 8 Nvidia A100 SXM4 80 GB GPUs leveraging a Distributed Data Parallel strategy. Distributed Data Parallel is a technique used in computing where data is partitioned and shared across multiple nodes, and computations on these data chunks are conducted simultaneously. This approach is particularly beneficial for handling large datasets and complex computations, as it can significantly accelerate processing time.

3 Experimental Contributions

Understanding the form and function of the hippocampus is critical given its vital role in learning, memory, and emotional regulation. As developed in the Section 1.1.1, the hippocampus is a remarkably heterogeneous structure, composed of anatomically and functionally distinct subfields with differential vulnerabilities across neurodevelopment, aging, and disease (Small et al., 2011). Unfortunately, in-vivo segmentation of hippocampal subfields in magnetic resonance imaging (MRI) has remained a major challenge, particularly in pediatric and clinical populations, due to insufficient image resolution, partial voluming effects, and variability in hippocampal anatomy (Yushkevich, Pluta, et al., 2015). To overcome these limitations, we developed and validated a new automated segmentation tool to delineate hippocampal subfields throughout the lifespan in MRI (Section 3.1.2). This innovative tool leverages nearly all public datasets of hippocampal segmentations, enabling robust subfields segmentations in conventional resolution MRI across different ages and health conditions.

Equipped with this cutting-edge subfields segmentation method, we conducted the first volumetric study of hippocampal subfields from early childhood to late adulthood (Poiret, Bouyeure, et al., 2023). Quantifying the developmental trajectories of each subfield revealed divergent patterns. This mirrors known differences in their structural maturation timecourses. Additionally, modeling subfield-specific atrophy patterns in aging provided novel insights into vulnerability differences. These anatomical findings demonstrate the importance of examining hippocampal substructures rather than treating the hippocampus as a homogeneous entity.

Critically, we further utilized the subfields segmentations to characterize the functional maturation of hippocampal subfields using resting-state functional MRI (Section 3.2). This revealed that the DG, the most age-corrected subfield structurally, also displayed the most changing functional connectivity patterns in childhood. In contrast, other subfields showed protracted functional development extending into adolescence, paralleling their ongoing structural maturation. The tight correspondence between subfield-specific anatomical and functional development emphasizes the importance of modeling hippocampal heterogeneity. Moving forward, our validated subfields segmentation approach can be widely applied to study the abnormalities of the subfields across diverse disorders impacting the hippocampus.

3.1 Segmenting the Hippocampal Subfields: Anatomical Trajectories

Accurate segmentation of the subfields of the hippocampus is essential to study their distinct roles in learning, memory, and disease. However, manual segmentation is tedious, prone to rater inconsistencies, time-consuming, and therefore does not scale beyond small studies. Popular automated segmentation tools like FreeSurfer (Iglesias et al., 2015), ASHS (Yushkevich, Pluta,

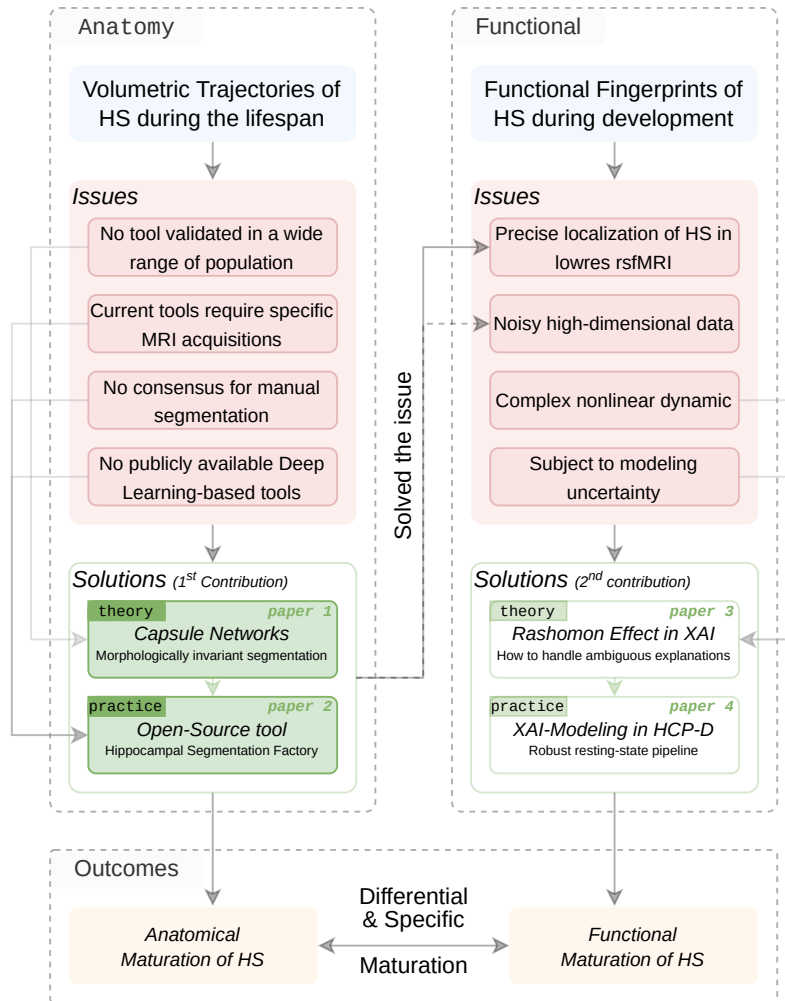


Figure 3.1: Achitecture of the dissertation: Anatomy. The aim of this chapter is to characterize the anatomical changes in hippocampal subfields across the lifespan. This chapter possess two research papers, (1) a methodological exploration of innovative and disruptive segmentation methods; (2) the development of an automated deep learning tool for hippocampal subfields segmentation from structural MRI.

et al., 2015), HIPS (Romero et al., 2017), and HippUnfold (DeKraker et al., 2022) have been developed using traditional computer vision techniques. However, these have significant limitations, such as long inference time, geometric rules that make them brittle to anatomical variations, or a lack of evolvability, — the capacity to evolve easily over time to better suit the needs of most studies. In contrast, deep learning has become the dominant approach for biomedical image segmentation, with techniques like U-Nets achieving expert-level performance (Ronneberger et al., 2015). But deep learning has not been widely adopted for segmenting hippocampal subfields, primarily due to the massive memory and compute requirements of 3D MRI data, and the need for large labeled datasets, even if proof-of-concepts gradually emerge in the literature (e.g., Chen and Liu, 2021; Q. Qiu et al., 2019; Z. Yang et al., 2020; Zhu et al., 2019). If traditional tools like FreeSurfer or HippUnfold have implemented deep learning steps, no tool is currently available in an end-to-end

fashion — i.e., a computational pipeline where every module is differentiable — an approach proven successful in other domains (e.g. Chen et al., 2023; Dong et al., 2022; Tang et al., 2022). However, developing scalable deep learning tools that can efficiently process large 3D MRIs on affordable hardware is vital to enable large-scale studies relating the anatomy of the subfields to behavior and disease. Scalability is the capacity of systems to accommodate increased workloads while maintaining performance, reliability, and cost-efficiency concurrently. When applied to deep learning, scalability encapsulates the ability to manage extensive deep learning models and an important amount of data proficiently within a distributed infrastructure. With exponential growth of biomedical datasets, the needs for scalability in computer vision systems increase to be practically useful for research and clinical care.

One potential way to develop on the need of scalable algorithms in computer vision, is to build on the literature trying to make such algorithms robust to natural variations occurring in images. In this context, Capsule networks represent a departure from conventional convolutional neural networks (CNNs) in their ability to encode higher-level semantic features rather than just spatial hierarchies of features. Whereas CNNs use scalar outputs to represent the presence of low-level features like edges and textures, capsule networks utilize vector outputs to encapsulate more complex concepts such as part-whole relationships, pose, deformation, and other attributes. We first tried to develop hippocampal segmentation with the key motivation of equipping networks with intrinsic invariances that come from modeling spatial relationships between parts and wholes. By encoding pose, deformation, and other instantiation parameters in output vectors, capsules can recognize entities even under novel conditions not seen during training. Dynamic routing algorithms allow lower-level capsules to agree on instantiation parameters for higher-level capsules. This parsing process based on part-whole agreement is more powerful than CNN pooling operations (Kwabena Patrick et al., 2019; LaLonde et al., 2020; Sabour et al., 2017). Overall, capsule networks aim to overcome limitations in CNNs' generalization and representation power by modeling semantics, spatial hierarchies, and transformations in richer vector representations. Although promising results have been shown on datasets such as MNIST (Lecun et al., Nov./1998), more research is still needed to match CNN performance on complex real-world data. However, capsules represent an exciting direction towards less data-hungry and more human-aligned feature learning.

Thus, the goal of this chapter is to develop and validate a scalable deep learning approach to segment the hippocampal subfields. We experimented with state-of-the-art deep learning methods (first green section of Figure 3.1) to develop a new tool optimized for computational efficiency (second green section of Figure 3.1). We also make this tool open-source to provide the community a first step towards scalable subfields segmentation. Leveraging this tool, we segmented three large lifespan datasets with 3750 subjects ages 4 to 100+ years. Analysis of these segmentations revealed distinct volumetric trajectories for each subfield throughout aging. These trajectories provided valuable information on the development, maturation, and decay of medial temporal circuits underlying memory throughout life. Overall, this work establishes that efficient deep learning tools can scale subfields segmentation to large datasets, enabling a more powerful study of how the anatomy of the subfields relates to cognition.

3.1.1 Attention-Gated 3D CapsNet for Robust Hippocampal Segmentation

Clement Poiret, Antoine Bouyeure, Sandesh Patil, Cécile Boniteau, Edouard Duchesnay, Antoine Grigis, Frederic Lemaitre, and Marion Noulhiane (2023). Attention-Gated 3D CapsNet for Robust Hippocampal Segmentation. *Journal of Medical Imaging*. Submitted.

We first explored Capsule Networks (CapsNets) (Sabour et al., 2017), which can recognize objects despite transformations such as rotations, scaling, or deformation. This is promising for segmenting hippocampal subfields, as they are robust to anatomical variations (LaLonde et al., 2020), such as incomplete hippocampal inversion. Capsules work by representing visual entities such as hippocampal subfields as vectors instead of scalars. The vector’s length represents the presence of an entity, while its orientation encodes equivariant properties like pose, deformation, etc. For example, a capsule encoding a subiculum would produce the same vector outputs despite changes in subicular pose or deformation, with only the orientation changing. This statistical equivariance to rotations, deformations, etc. gives CapsNets inherent robustness, unlike CNNs which only exhibit translation equivariance. However, the routing of information between capsules is computationally intensive. The iterative routing-by-agreement algorithm weights the capsules by agreement, allowing the selective passage of information to higher-level capsules.

Abstract

Goal: The hippocampus is a key structure involved in memory and is involved in various neuropathologies. Accurate segmentation of the hippocampal subfields in MRI is challenging due to small sizes and anatomical variability. We aimed to develop a robust deep learning model for hippocampal subfields segmentation, particularly in atypical cases like incomplete hippocampal inversion (IHI).

Material and Methods: We implemented the first public 3D Capsule Network with attention gates (3D-AGSCaps) and compared it to 3D convolutional neural networks (CNNs) to segment the hippocampus (both in anterior/posterior and in subfields) in three datasets with manual labels. We evaluated segmentation performance under increasing random rotational perturbations to simulate IHI.

Results: On typical MRI, 3D-AGSCaps performed comparably to CNN in subfields segmentation. However, with increasing rotational perturbations, 3D-AGSCaps showed significantly higher Dice similarity, lower Hausdorff distance, and higher volumetric similarity compared to CNNs.

Conclusions: 3D-AGSCaps exhibits greater robustness than CNNs when segmenting hippocampi under atypical conditions like IHI, likely due to built-in equivariance. Our model could enable accurate segmentation in diverse datasets, including those with developmental variations.

Attention-Gated 3D CapsNet for Robust Hippocampal Segmentation

Clement Poiret^{a,b}, Antoine Bouyeure^{a,b}, Sandesh Patil^{a,b}, Cécile Boniteau^{a,b}, Edouard Duchesnay^a, Antoine Grigis^a, Frederic Lemaitre^{c,d}, and Marion Noulhiane^{a,b*}.

5

^aUNIACT, NeuroSpin, Institut Joliot, CEA Paris-Saclay, France.

^bInDEV team, U1141 NeuroDiderot, Inserm, Université Paris Cité, France.

^cCETAPS EA 3832, Université de Rouen, France.

^dCRILOBE, UAR 3278, CNRS-EPHE-UPVD, Moorea, Polynésie Française.

10

*Corresponding author: *e-mail*: marion.noulhiane@cea.fr

Abstract

Purpose: The hippocampus is organized in subfields (HSF) involved in learning and memory processes, and widely implicated in pathologies at different ages of life, from neonatal hypoxia, to temporal lobe epilepsy or Alzheimer's disease. Getting a highly accurate and robust delineation of sub-millimetric regions like HSF to investigate anatomo-functional hypothesis is a challenge. One of the main difficulties encountered by those methodologies is related to the small size and anatomical variability of HSF, resulting in the scarcity of manual data labeling. Recently introduced, Capsule Networks solve analogous problems in medical imaging, providing Deep Learning architectures with rotational equivariance. Nonetheless, Capsule Networks are still 2D, and unassessed for the segmentation of HSF.

Approach: We released the first public 3D Capsule Network (3D-AGSCaps, <https://github.com/clementpoiret/3D-AGSCaps>) and compared it to equivalent architectures using classical convolutions on the automatic segmentation of HSF on small and atypical datasets (incomplete hippocampal inversion, IHI). We tested 3D-AGSCaps on three datasets with manually labeled hippocampi.

Results: Our main results were: (1) 3D-AGSCaps challenged analogous CNNs; (2) 3D-AGSCaps has been more robust to random rotational perturbations. This may greatly facilitate the study of atypical subjects, including healthy and pathological cases like those presenting an IHI.

Conclusion: We expect our newly introduced 3D-AGSCaps to allow a more accurate and fully automated segmentation on atypical populations, small datasets, as well as on and large cohorts where manual segmentations are nearly intractable.

Keywords: Hippocampal Subfields, Convolutional Neural Networks, Deep Learning, Equivariance, MRI

40

1. Introduction

Located in the medial temporal lobe, the hippocampus is crucial for learning and memory processes¹. The hippocampus is also a key player in diverse neuropathologies with high prevalence in the population at distinct ages of life-span, such as neonatal hypoxia, age-dependant cognitive decline, Alzheimer’s disease, or medial-temporal lobe epilepsy². The hippocampus includes distinct cyto- and myelo-architectonic hippocampal subfields (HSF): the Dentate Gyrus (DG), four parts of the Cornu Ammonis (CA4 to CA1), and the Subiculum (Sub). Recent research focused on the distinct roles of HSF in memory functions, as well as in the progressive spatial evolution of neurological diseases.

50 Segmenting the Hippocampus, Stakes, and Methods

Such research involve an accurate delineation (or segmentation) of the HSF, which consists of assigning a class to every voxel of a given image. In the context of MRI segmentation, bilateral regions of interest (ROI) like the HSF are assigned the same labels, to divide an image into a set of semantically meaningful, homogeneous, and non-overlapping regions of similar attributes such as intensity, depth, color, or texture³. This delineation process enables the study of structural patterns which may in fine lead to a better comprehension, diagnosis and prognosis of such diseases, as segmentation allows one to easily derive the geometry, shape and size of a given ROI. To date, the segmentation of the hippocampus can capture anatomical variability such as the Incomplete Hippocampal Inversion (IHI)⁴, a developmental abnormality occurring in consequent subsets of the healthy or pathological population, where the hippocampal body and the collateral sulcus can be rotated up to 90°⁵. However, this methodology remains so time-consuming that it cannot be considered as routine clinical practice. While distinct techniques of various complexity have been developed to segment HSF on MRIs⁶⁻⁸, the field suffers from labeled data scarcity as manual segmentation is a time-consuming and error-prone process partially caused by inconsistent guidelines. Because manual segmentation is the only way to gather labeled datasets to train neural networks, the aforementioned difficulties greatly limit the size of available datasets, thus reducing the probability to learn in specific cases where IHI are found, like in Temporal Lobe Epilepsy.

Nowadays, segmentation tasks are now handled close to exclusively through specific and supervised Convolutional Neural Networks (CNN), an architecture called UNet⁹, leveraging the properties of an auto-encoder architecture to quickly achieve a segmentation with an expert-level accuracy, and small sample size. Nevertheless, those models suffer from several pitfalls. It has been showed that the performances of such computer-vision models are prone to image corruptions like noise or rotations¹⁰⁻¹². Albeit recent works validated Deep Learning as a great candidate for automated segmentation of HSF^{13,14}, IHI, which are unassessed in recent automated segmentation methods, may cause troubles to most conventional CNNs. With a transformation g , an image x , and a model f , equivariance is defined as $g(f(x))=f(g(x))$. Similarly, invariance is achieved if and only if $f(g(x))=f(x)$. CNNs are efficient in modeling structural patterns in a given image, especially thanks to their built-in translation equivariance: a translation of a specific pattern in the input image, shifts the output of the convolutional layer. As previously found, transformations such as rotations can impair CNNs performances¹⁵. On one hand, a standard approach to approximate rotational equivariance would be to use data augmentation by providing multiple rotated versions of the training set. However, it involves learning redundant parameters corresponding to similar patterns at varying angles¹⁶. In addition, data augmentation may increase overfitting risks^{17,18}, meaning the improvement on standard CNNs would only be marginal on small training sets and may sometimes lead to a drop in accuracy on unperturbed images¹⁰. Partly to

solve this issue, recent works introduced Capsule Networks (CapsNets), a novel kind of Neural Network designed to benefit from natural or augmented variability more efficiently^{19,20}.

Capsule Networks

90 CapsNets are replacing standard convolutions with capsules. A capsule aims to replace scalar activation values with vectors (of which the number of dimensions constituting its space is sometimes referred to as the number of atoms). The L2-Norm of a given vector is equivalent to the activation of a standard convolution, but now a network can encode information into the orientation of the vector. This intriguing characteristic promotes the emergence of a key property: the theoretical ability to learn
95 equivariances to features (sometimes called "instantiation parameters") such as rotations, local deformations²⁰, or even to more subtle features like sphericity, lobulation or textures²¹. Intuitively, whereas convolutions learn multiple kernels to detect different versions of the same object (e.g. a rotated hippocampus), capsules embed those different versions under the same weights through vectors' orientations, leading to less redundancy in the network, and greater expressive power for the same
100 number of parameters. Moreover, the usual feedforward pass is altered by using a Routing-by-Agreement between capsule layers²⁰. This Routing-by-Agreement recurrently weights the feedforward pass by selectively passing information from capsules in the layer l to the layer $l+1$. Each capsule in l will vote for the potential output of capsules in $l+1$. Then, activations between all capsules in l and a specific capsule in $l+1$ are weighted by their L2-Norm to the centroid of the predictions: similar
105 predictions are likely to be sent to a single parent capsule. In 2D, CapsNets have showed improved robustness against physical alterations such as rotations²², placing themselves as possible candidates to explore for MRI processing. To date, CapsNets are yet to be publicly implemented and benchmarked on a 3D segmentation task. For example, if a capsule represents a high-level object like a hippocampus, for every patch of a given image the vector's norm represents the probability of presence of the object.
110 Then, its direction encodes relevant instantiation parameters. On the other hand, their activation (L2-norm) stays invariant. This behavior leads to an increased expression power and a higher sample efficiency²³.

CapsNets have already been implemented in the biomedical field with promising results, where the
115 authors were able to overcome the shortcomings of CNNs on a brain tumor and lung nodule type classification tasks^{24,25}. By processing MRIs in a slice-by-slice manner, they achieved a classification accuracy of 78% against 61.97% for a CNNs of comparable architecture. To date, CapsNets for image segmentation are poorly investigated. The authors of the SegCaps model were the first ones to successfully perform segmentation with capsule layers²¹. They made this possible by building a deeper
120 model than the original implementation using locally-constrained routing and transformation matrix sharing to reduce the number of parameters and memory consumption. To build a UNet-shaped model, they also introduced transposed capsules. Following this segmentation paradigm, recent works handled coronary artery segmentation from intravascular optical coherence tomography in a slice-by-slice manner²⁶. While they did not achieved SotA accuracy on their dataset, they managed to get honorable
125 segmentations with a model of nearly 5M parameters, while SotA models were between 30M and 40M parameters. This may suggest that capsules can effectively benefit from learned equivariances in segmentation tasks. Nevertheless, both implementations act in a 2D space, on binary segmentation tasks. CapsNets able to perform 3D segmentation tasks are yet to be implemented. If 3D segmentation is not yet handled by CapsNets, several works experimented with 3D capsules in other tasks. Newer
130 developments used 3D CapsNet to perform object recognition with a shallow architecture²⁷, where they found that 3D CapsNets were more data efficient than analogous CNN architectures. Additional works

found a consistent improvement of 3D capsules over SotA models for 3D point set classification, especially for noisy observations^{28,29}. Finally, 3D CapsNets were applied with success in the biomedical field for lung nodule malignancy prediction with a highly competitive accuracy²⁴.

135

Thus, 3D capsules come out as relevant candidates for tasks where the number of observations is limited, variable or noisy, and where models are operating in resource-constrained environments. Notwithstanding the fact that 3D CapsNets are an active research area, no implementation is currently publicly accessible.

140

Attention-Gated Networks

The idea behind Attention Gates (AG) is to allow a CNN to implicitly learn how to suppress or highlight specific regions in an input image, with minimal computational overhead³⁰. Initially developed as an extension to the standard U-Net model, it generates through additive attention a soft-attention grid, composed of gating coefficients $\alpha_i \in [0,1]$. Finally, those gating coefficients are multiplied by the input feature map. The original study reported significant improvements related to their additive attention gates on their segmentation task³⁰. Later, similar improvements were obtained for 3D coronary computed tomography (CT) angiography segmentation³¹, or liver CT image segmentation³². While the attention mechanism has been implemented in the Routing-by-Agreement algorithm³³⁻³⁶, Attention-Gated CapsNets are yet to be assessed. The original Routing-by-Agreement algorithm aims at weighting information sent from a layer l to a layer $l+1$. We think that it could potentially work synergistically with AG by modulating on-the-fly the activation of a capsule layer.

145

150

The aim was to extend CapsNets for segmentation tasks in 3-dimensional spaces applied to MRI segmentation of the hippocampus to investigate the robustness of our new model, namely 3D Attention-Gated SegCaps (3D-AGCaps) on developmental particularities, such as the IHI. In this aim, our approach was as follows: (1) we validated 3D-AGCaps on hippocampal segmentation against the equivalent architectures using classical convolutions in-place of capsule layers; (2) we investigated the robustness of 3D-AGCaps to various random rotational perturbations of the MRI acquisitions, simulating IHI.

155

160

We hypothesized that,

1. 3D-AGCaps will challenge Convolutional architectures on hippocampal segmentation on typical MRI;
2. 3D-AGCaps will exhibit robustness to random rotational perturbations (replicating atypical conditions, like IHI) thanks to their implicit ability to learn equivariances over various instantiation parameters.

165

Methods

Datasets Description

170

We used two public and one in-house datasets with manually labeled hippocampi by expert raters (Table 1). The three datasets were manually labeled by experts, from which the first one is an anteroposterior hippocampal segmentation of 263 hippocampi, while the other two are 50 hippocampi segmented in subfields. The Kulaga-Yoskovitz dataset has been segmented from head to tail according

175 to an in-house segmentation protocol. MemoDev has hippocampal bodies manually segmented (AB, SP, MN, CP) following³⁷. Examples of both types of hippocampal segmentation are showed in figure 1.

Data acquisition for our in-house dataset, MemoDev, was performed under the regulations of an appropriate Ethical Committee board (CPP 2011-A00058-33).

180

Name	N	Acquisition Parameters	Segmentation	Ref.
<i>Simpson</i>	263	3T; 3D T1-weighted MPRAGE sequence; TI/TR/TE, 860/8.0/3.7 ms; 170 sagittal slices; voxel size, 1.0 mm ³ ;	- Anterior - Posterior	38
<i>Kulaga-Yoskovitz</i>	50	3T; 3D T1-weighted MPRAGE sequence; TI/TR/TE, 1500/3000/4.32 ms; 176 sagittal slices; voxel size, 1.0 mm ³ ;	- DG - CA - Sub	39
<i>MemoDev</i>	50	3T; Coro-T2-weighted TSE sequence; TR/TE, 3970/89 ms; 46 coronal slices; voxel size, 0.4*0.4*1.2mm;	- DG - CA1 - CA2/3 - Sub	40

Table 1: Description of the datasets used. DG: Dentate Gyrus, CA: Cornu Ammoni, and Sub: Subiculum.

*** Insert figure 1 about here ***

From 2D to 3D Capsule Networks in MRI Segmentation of the Hippocampus

185 CapsNets are compute-intensives, both in terms of computational complexity and memory requirements²¹. If they are solving issues inherent to CNNs, this is the major drawback for the adoption of capsules. Therefore, we used the public 2D SegCaps implementation of LaLonde et al., (2020) to migrate the architecture from 2D to 3D in PyTorch. Their implementation offers important addons to reduce the number of parameters of CapsNets, like (de)convolutional capsules, and a locally-constrained routing.

190 In addition to the reimplementaion by adding a spatial dimension, we introduced a novel activation function to handle multiclass classification tasks. The Squash function²⁰ has been originally introduced to rescale the L2-Norm of the Capsules to $[0; 1]$ without changing their directions, such as:

$$v_j = \frac{\|s_j\|^2}{1 + \|s_j\|^2} \frac{s_j}{\|s_j\|}$$

195

To segment multiple classes, we want the L2-Norm of each capsule c to represent a probability distribution over the brain regions. Therefore we introduced the Softmax-Squash (or SMSquash) function, that we used for our last capsule layer, defined as:

$$v_j = \frac{e^{\|s_j\|}}{\sum e^{\|s_j\|}} \frac{s_j}{\|s_j\|}$$

200

However, migrating SegCaps from 2D to 3D worsened the computational burden of CapsNets. This led us to introduce Attention-Gated Capsules to route the information with greater precision while reducing the number of Routing-by-Agreement iterations to improve the efficiency of our model.

205

3D Attention-Gated SegCaps (3D-AGSCaps)

To complement the Routing-by-Agreement algorithm, we introduced a variation of the AG (figure 2) coming from Oktay et al., (2018), which helps the network to focus on the target brain structures. Our AG, implemented at the concatenation of the volumes of the downsampling and the upsampling path, aims to modulate the L2-norm of the capsules.

210

*** Insert figure 2 about here ***

215

The gating signal g from the layer $l - 1$ is upsampled using Transposed Capsules²¹. Then, g and x of the corresponding layer l of the downsampling path are combined to form an attention grid of which the size matches the number of atoms. Information coming from the downsampling path is then multiplied to the attention grid to modulate capsules' L2-norms. Convolutions are followed by SwitchNorm layers⁴¹.

220

Our final model, 3D-AGSCaps is depicted in figure 3. It retakes a UNet-like architecture where we used our AG in-place of the concatenation of both downsampling and upsampling paths, and assessed its efficacy through an ablation study. The resulting implementation in PyTorch is publicly available under an MIT license at <https://github.com/clementpoiret/3d-agscaps>.

225

To perform ablation studies, we tested multiple variations of our proposed model to analyze the impact of our AG.

230 Implementation Details and Evaluation Metrics

The models have been implemented in Python, using PyTorch. The training is done with PyTorch Lightning. Data augmentation is handled with TorchIO. We trained our models with Automatic Mixed Precision (16-bit), and validated them using a 10-Fold Cross-Validation. We kept an hold-out test set

235 for further analysis to keep samples free from potential unwanted tuning. Complete implementation for training and validation details are listed in the table 2.

*** Insert figure 3 about here ***

	<i>Simpson (N = 263)</i>	<i>Kulaga-Yoskovitz (N = 50)</i>	<i>MemoDev (N = 50)</i>
Training	N = 225	N = 36	N = 36
	64 Epochs, Batch Size 8		
	AdamW, Learning Rate 1e-3		
	Cosine Annealing Scheduler (no restart, no warm-up)		
	Stochastic Weight Averaging		
Validation	N = 25	N = 4	N = 4
	10-Fold Cross-Validation		
Test	N = 13	N = 10	N = 10
	Hold-out test set		
Preprocessing	1. Crop/Pad around hippocampus		
	2. Z-Normalization		
	1. Left/Right Flips (p=0.5)		
	2a. Affine Transformations (p=0.8)*		
	2b. Elastic Deformations (p=0.2)*		
	3. Gaussian Noise (p=0.5)		
Augmentation	4. Random Contrast (p=0.5)		

Table 2: Implementation and validation details for each dataset. Given a fixed (hold-out) test-set, training and validation sets are defined using a standard 10-Fold Cross-Validation on the remaining samples. (*) denote an “either/or” scheme, i.e.: affine and elastic transformations can’t be applied at the same time, but one of them is always applied.

240 Following our Cross-Validation protocol to validate our models, we assessed the effect of rotational data augmentation by training 10 times all the models with a different maximum amount of random rotations of the training set. Finally, to assess the behavior of the models in the presence of atypical hippocampi, we randomly rotated our test-sets (table 2) with an increasing amount of maximum amplitude (from 0° to 180°). As this process involves random perturbations of the test sets, we repeated
 245 this process 10 times to better estimate the variability induced by our artificial alterations of the MRIs. Segmentations are assessed with the Dice Coefficient (DC), the Volumetric Similarity (VS), and the Hausdorff Distance (HD) computed with PyMia. Given a manual segmentation y_m and a predicted segmentation y_p :

- the DC is an overlap metric ranging from 0 (no overlap), to 1 (full overlap) defined as

$$250 \quad DC = \frac{2|y_m \cap y_p|}{|y_m| + |y_p|},$$

- the HD is a metric of surfacic distance ranging from 0 to +inf. With the directed hausdorff distance between two point sets X and Y such as $hd(X, Y) = \max_{x \in X} \min_{y \in Y} \|x - y\|_2$, the HD is defined as $HD(y_m, y_p) = \max(hd(y_m, y_p), hd(y_p, y_m))$
- the VS is a comparison between volumes of two segmentations ranging from 0 (complete dissimilarity between volumes) to 1 (exact match between volumes). With the volume of a region S , it is defined as $VS = 2 \frac{|S_m \cap S_p|}{|S_m + S_p|} \cdot 100\%$.

255

260

We computed each metric on a per-channel basis to assess the quality of each class, then averaged across classes to get a general score. In order to evaluate our hypotheses, we performed a two-way ANOVA with model types and rotation angles of the test images as independent variables. p -values are corrected using a Benjamini-Hochberg False Discovery Rate. While CNNs were trained using a Focal Tversky loss L_s^{42} . Given an input x , our segmentation loss L_s is defined with TP and TN the true positives and negatives, FN and FP the false positives and negatives, and $\alpha=0.3$, $\beta=0.7$, $\gamma=3/4$ such as:

265

$$L_s = \left(1 - \frac{TP}{TP + \beta FN + \alpha FP} \right)^\gamma$$

270

Values of the hyper-parameters α and β are following the recommendations of their original paper: weighing more the FN enhanced convergence by shifting the focus on minimizing the FN. According to the authors, it helped balance precision-recall scores and gave better DC on a similar architecture of ours. 3D-AGSCaps uses a combination of L_s and the mean squared error of the reconstruction of the hippocampus. Thus, our loss L for our 3D-AGSCaps is defined as:

$$L = L_s + \frac{1}{n} \sum (x_i - \hat{x}_i)^2$$

275

Comparison with Analogous Convolutional Models

We benchmarked 3D-AGSCaps against the best-known models used in the hippocampal segmentation literature in a 3D approach:

280

- UNet (16.3M)⁴³, the baseline of most segmentation models, consisting of an auto-encoder architecture with skip connection between layers of the same depth,
- Residual UNet (35.0M)^{44,45}, grouping every couple of convolutions with the aim to stabilize the training of deeper networks,
- and their counterparts DUNet (16.7M) and Residual DUNet (35.5M)¹⁴, replacing the second to last skip connection with a Dilated Dense Network of convolutions to improve the information flow between the encoder and the decoder.

285

Experimental Results

Ablation studies

290 Results of ablation studies (table 3) across the three datasets revealed that the best overlap (DC) between HSF was obtained by models with the AG (e.g. 0.872 ± 0.028 vs 0.834 ± 0.058 for Kulaga-Yoskovitz). The ten-fold Cross-Validation results are reported in the table 3. For the quality of the reconstruction (table 4), we showed a significant impact of the reconstruction ($p=0.034$, $T=5.262$, $BF_{10}=3.196$, cohen’s $d=0.032$). An example of the reconstruction can be seen in the figure 4. However, we found no significant differences regarding the value of α ($p>0.05$, table 4).

295

Dataset	Model	DC	HD	VS	MSE
Kulaga-Yoskovitz	Baseline (2.3M)	0.834 ± 0.058	22.173 ± 17.975	0.921 ± 0.071	1.006 ± 0.004
	AG (2.4M)	0.872 ± 0.028	15.350 ± 15.516	0.962 ± 0.034	1.005 ± 0.003
MemoDev	Baseline (2.3M)	0.659 ± 0.163	17.182 ± 26.403	0.835 ± 0.165	0.896 ± 0.061
	AG (2.4M)	0.654 ± 0.175	10.502 ± 13.288	0.828 ± 0.181	0.897 ± 0.059
Simpson	Baseline (2.3M)	0.877 ± 0.037	3.160 ± 2.234	0.937 ± 0.037	0.238 ± 0.031
	AG (2.4M)	0.880 ± 0.037	2.875 ± 1.874	0.941 ± 0.036	0.236 ± 0.028

Table 3: Ablation study of our Attention Gates (AG). Baseline models are Capsule Networks without AG. DC: Dice Coefficient, HD: Hausdorff Distance, VS: Volumetric Similarity, and MSE: Mean Squared Error assessing the reconstruction head. Results are presented in mean \pm standard deviation.

*** Insert figure 4 about here ***

Dataset	α	DC	HD	VS
Kulaga-Yoskovitz	0.0	0.869 ± 0.029	11.784 ± 11.831	0.960 ± 0.031
	0.1	0.870 ± 0.029	10.143 ± 9.661	0.959 ± 0.031
	1.0	0.869 ± 0.030	11.167 ± 10.622	0.960 ± 0.031
	10.0	0.870 ± 0.029	9.843 ± 10.476	0.960 ± 0.030
MemoDev	0.0	0.664 ± 0.160	17.011 ± 22.268	0.865 ± 0.141
	0.1	0.664 ± 0.161	12.970 ± 17.867	0.866 ± 0.139
	1.0	0.662 ± 0.165	13.775 ± 20.456	0.869 ± 0.144
	10.0	0.667 ± 0.161	14.848 ± 19.280	0.859 ± 0.149

	0.0	0.878 \pm 0.037	3.977 \pm 4.088	0.944 \pm 0.036
	0.1	0.879 \pm 0.037	3.993 \pm 4.339	0.944 \pm 0.036
	1.0	0.879 \pm 0.036	4.237 \pm 4.909	0.944 \pm 0.036
Simpson	10.0	0.878 \pm 0.036	3.794 \pm 3.592	0.943 \pm 0.036

Table 4: Impact of the reconstruction module. Evolution of segmentation with different weights α of the loss across all three datasets, with DC the Dice Coefficient, HD the Hausdorff Distance, and VS the Volumetric Similarity. Results are presented in mean \pm standard deviation.

300 3D-AGSCaps: Comparison with Analogous Convolutional Models on Typical MRIs

We started by comparing 3D-AGSCaps against different data-augmentation strategies. Best results on test-sets are obtained with little (15°) to no rotational augmentation: 3D-AGSCaps and both dilated models were showing better segmentation quality without training-time rotational augmentation, while simpler models (UNet and Residual UNet) slightly benefited from 15° maximum augmentation. On typical MRIs, we noted an overall superiority of residual models (namely 3D-AGSCaps, Residual UNet, and Residual DUNet) compared to the single ones (table 5). However, among residual models, we failed to show a significant difference ($p>0.05$) on DC, but 3D-AGSCaps showed a higher HD and VS (table 5, with a qualitative comparison figure 5). We additionally monitored the computational resources required during the training phases (table 6).

a. (DC)	AGSCaps	UNet	Residual UNet	DUNet	Residual DUNet
0°	0.839 \pm0.098	0.664 \pm 0.289	0.815 \pm 0.138	0.712 \pm0.280	0.849 \pm0.096
15°	0.811 \pm 0.120	0.676 \pm0.284	0.840 \pm0.105	0.593 \pm 0.345	0.819 \pm 0.126
45°	0.669 \pm 0.218	0.415 \pm 0.317	0.681 \pm 0.269	0.461 \pm 0.281	0.711 \pm 0.234
90°	0.460 \pm 0.342	0.331 \pm 0.328	0.482 \pm 0.369	0.283 \pm 0.315	0.481 \pm 0.361
180°	0.453 \pm 0.342	0.274 \pm 0.372	0.368 \pm 0.347	0.252 \pm 0.343	0.306 \pm 0.357

b. (HD)	AGSCaps	UNet	Residual UNet	DUNet	Residual DUNet
0°	11.376 \pm9.045	26.316 \pm 19.955	5.699 \pm 2.694	19.107 \pm16.640	4.934 \pm2.113
15°	18.502 \pm 11.598	14.390 \pm16.215	5.029 \pm2.575	21.393 \pm 15.304	5.991 \pm 2.637
45°	29.278 \pm 14.429	28.512 \pm 15.197	16.382 \pm 10.921	27.881 \pm 15.240	13.209 \pm 8.453
90°	25.868 \pm 12.671	29.960 \pm 14.665	24.034 \pm 14.328	30.375 \pm 13.002	26.238 \pm 13.827
180°	26.912 \pm 13.513	30.750 \pm 11.788	27.891 \pm 13.320	29.926 \pm 12.436	31.018 \pm 12.752

c. (VS)	AGSCaps	UNet	Residual UNet	DUNet	Residual DUNet
0°	0.953 ±0.043	0.739 ±0.302	0.928 ±0.113	0.815 ±0.281	0.961 ±0.040
15°	0.971 ±0.030	0.754 ±0.308	0.963 ±0.044	0.683 ±0.376	0.929 ±0.088
45°	0.888 ±0.112	0.543 ±0.350	0.834 ±0.244	0.600 ±0.318	0.843 ±0.177
90°	0.660 ±0.282	0.501 ±0.352	0.618 ±0.356	0.374 ±0.324	0.587 ±0.388
180°	0.772 ±0.227	0.445 ±0.372	0.525 ±0.367	0.382 ±0.328	0.447 ±0.360

Table 5: Effect of rotations as data augmentation during training time on segmentation quality on a test-set. Each model is trained multiple times with a varying amount of random rotations as part of the data augmentation pipeline, and then evaluated on (a.) the Dice Coefficient, (b.) the Hausdorff Distance, and the (c.) Volumetric Similarity on an unseen test-set. Results are presented as mean ±std. Bold results highlight the maximum amplitude of random rotations leading to the best performances.

	AGSCaps	UNet	Residual UNet	DUNet	Residual DUNet
# Parameters	2.285M	16.318M	35.069M	16.723M	35.475M
# FLOPS	0.493T	0.148T	0.180T	0.215T	0.246T
Epoch Duration (seconds)	59.063	17.813	21.563	19.688	23.438
Training Time (minutes)	63	19	23	21	25

Table 6: Computational comparisons. Timings were monitored on a Nvidia RTX8000, and are given only for information purposes.

315

3D-AGSCaps: Robustness to Random Rotational Perturbations (IHI)

After assessing segmentation quality on typical MRIs, we evaluated generalization on randomly rotated MRIs of our test-set. An example of an MRI comprised in our test-set is depicted in the figure 5, with and without deformation. Across our three datasets, an ANOVA showed no evidence of significant differences between all segmentation models for observations with little (15°) to no rotation ($p > 0.05$). However, for rotations greater or equal to 45°, segmentation models start to differentiate (figure 6). 3D-AGSCaps showed a higher DC ($p = 0.004$, $BF_{10} = 12.588$, cohen's $d = 0.179$) than its CNN counterparts, a lower HD ($p = 0.001$, $BF_{10} = 1.638$, cohen's $d = -0.120$), and a higher volumetric similarity ($p < 0.001$, $BF_{10} = 1e+15$, cohen's $d = 0.356$).

325

*** Insert figure 5 about here ***

*** Insert figure 6 about here ***

330 Discussion

The aim was to validate a public implementation of 3D-AGSCaps which offers a more accurate and fully automated segmentation on atypical populations and small datasets. We showed that 1/ 3D-AGSCaps challenged analogous convolutional architectures on hippocampal segmentation on typical MRI that is especially relevant in clinical population, and 2/ that 3D-AGSCaps exhibited robustness to random rotational perturbations (replicating atypical conditions, like IHI) thanks to their implicit ability to learn equivariances over various instantiation parameters. On one hand, because 3D-AGSCaps has been on-par with all other convolutional networks, we confirmed its ability to perform hippocampal segmentation on T1w and T2w MRIs. On the other hand, we showed that with an increasing quantity of random rotational perturbations, 3D-AGSCaps provided better segmentations than CNNs. Therefore, 3D-AGSCaps exhibits interesting properties in clinical settings: a better robustness to atypical images even when trained on small cohorts with only few patients.

3D-AGSCaps: Implementation and Ablation

We implemented a 3D SegCaps on a segmentation task and showed that 3D-AGSCaps is capable of hippocampal segmentation, with up to 15 times fewer parameters (35.5M parameters for a Residual DUNet, against 2.3M for 3D-AGSCaps, table 6). During our experiments, we found that a single iteration of the Routing-by-Agreement algorithm leads to the best results. This is a known effect, as previous works reported that usual Routing-by-Agreement algorithms may not behave as expected, unaffected classification results, and often producing worse results than baseline algorithms⁴⁶.

Most CapsNets use an atypical regularization and explanation technique. In addition to outputting the results of our task of interest, they output a reconstruction of the original input, optimized through a specific term in the loss function (figure 4). The relative weight of this loss term in the final loss functions remains unclear in the literature. Interestingly, we found that the reconstruction did not yield any significant enhancement of the segmentation, even if it reduced outliers produced in the MemoDev dataset (table 4). However, if regularization is the main goal of the reconstruction module, other more efficient techniques should provide the same benefits without forcing the use of additional layers. If its goal is to achieve some sort of explainability of capsules' atoms, tuning the coefficient defined in the loss lead to no significant improvement in reconstruction quality. Qualitatively, reconstructions were of a relatively poor quality (e.g. figure 4), but the object of interest is recognizable enough for the sake of explanations. This is certainly caused by the MSE term of the loss function that assumes pixel independence without accounting for spatial relationships. MSE has been shown to produce low-quality reconstructions compared to more recent and specific loss functions like the SSIM or LWSSIM⁴⁷.

To deal with the problem of the exponentially increasing number of parameters when switching from a 2D to a 3D space, we introduced AG (figure 2) with the aim of improving the efficiency of information routing. Our results showed an enhancement in segmentation quality, with a better overlap (mean DC increased by .012), fewer outlier voxels (mean HD reduced by 4.713), and a better VS (increased by .012).

3D-AGSCaps: Comparison with Analogous Convolutional Models both on Typical and Atypical MRIs

375 As a traditional approach to achieve rotational equivariance would be to use data augmentation, we assessed the effect of data augmentation on both our network (encoding equivariances of visual patterns through the orientation of the capsules) and SotA networks with training-time rotational augmentation (learning equivariance by learning the same pattern multiple times for different angles). Interestingly, we found that rotational augmentation mostly deteriorated segmentation quality (table 5). At first sight, this fact may seem counter-intuitive as data augmentation should improve the
380 generalization of Deep Learning models. However, this is coherent with part of the literature stating that data augmentation may increase overfitting risks^{17,18}, leading to a marginal improvement on small training sets⁴⁸ or even lead to a drop in accuracy on unperturbed images¹⁰. This highlights pieces of evidence that efficient and robust segmentations on small training sets will benefit from networks showing built-in capacities to handle equivariances.

385 Given the best amount of training-time data augmentation for each model, 3D-AGSCaps did not show a significant improvement for typical MRIs on DC (table 5) compared with architecturally equivalent models¹⁴, but showed a higher HD and VS. Overall, all models handled our segmentation task equally well with most DC superior to 0.8, but it is worth noting that we did not gather any clear evidence for
390 statistically significant differences between models introduced by Zhu et al., (2019) and classic Residual UNet models. Simpler and non-residual models (UNet and DUnet) were consistently left behind.

3D-AGSCaps: Robustness to Random Rotational Perturbations (IHI)

395 Finally, we assessed our 3D-AGSCaps against SotA models regarding behaviors facing plausible alterations of the hippocampus such as the IHI. Therefore, we monitored segmentation quality with an increasing maximum angle of random rotations up to 90°, following a realistic range (figure 6) on test observations. Alongside the previously discussed lack of differences for typical hippocampi (i.e.: no rotation) heterogeneity was highlighted by increasing the amount of rotation. For angles as small as
400 45°, 3D-AGSCaps stood out, giving a better DC with a higher VS, followed by the two models introduced by Zhu et al., (2019) and then the UNet and Residual UNet baselines. Those evidences support our hypothesis stating that CapsNets can segment the hippocampus with more robustness towards alterations such as rotations, which is beneficial when working with clinical settings affected by data-scarcity issues.

405 It should be noted, however, that this improved robustness comes at the cost of computational requirements. This cost is mainly driven by the storage of activation values as vectors instead of scalars. As of now, CapsNets suffers from scalability issues, consuming up to 10 times the amount of GPU memory compared to CNNs with analogous architectures. This computational overhead also
410 increased training time, from 19 minutes with a batch-size of 8 for a UNet, to 1 hour for 3D-AGSCaps on an Nvidia RTX 8000 (table 6). In order to process MRIs with a CapsNet such as 3D-AGSCaps, the use of a specific preprocessing of the input has to be performed, such as automatic detection of an enclosing box of both hippocampi to crop the MRI and reduce its memory footprint. This is the reason why we also published a third-party tool called ROILoc⁴⁹ (available at
415 <https://github.com/clementpoiret/ROILoc>), as a modest solution to this limit.

Therefore, the use of CapsNets for MRI processing have to be justified by an underlying hypothesis such as the presence of IHI or hippocampal sclerosis in a pathological population. While this type of architecture seems promising, we believe it would be important to further investigate the computational efficiency of CapsNets to find ways to address these limitations. For example, it could be interesting to explore the GLOM architecture⁵⁰, although still prototypical, but introduced specifically to solve some of the difficulties posed by the capsule design. Alternatively, as one of the main issues to solve this complex problem lies in the scarcity of labeled datasets, other tracks might be interesting to explore. In this way, self-supervised pretraining could help with the relative uselessness of rotational data augmentation, and semi-supervised training such as the recently introduced Annotation-efficient Deep Learning (AIDE) framework⁵¹ seems to provide a simple way to handle segmentation tasks with scarce and noisy labeling.

To date, the segmentation of the hippocampus can capture anatomical variability such as the Incomplete Hippocampal Inversion (IHI)⁴, a developmental abnormality occurring in consequent subsets of the healthy or pathological population such as in temporal lobe epilepsy or hippocampal sclerosis. The IHI is gradual, locally impacting shapes of the hippocampus. Because Capsules in our architecture have a kernel size of 3, they can encode instantiation parameters as finely as a local cube of 3^3 voxels. Therefore by construction, 3D-AGSCaps can handle the naturally occurring variations of the IHI going up to 90° rotations, by modelling the verticality and roundness of the hippocampal body and the collateral sulcus (local and global rotational statistical equivariance of capsules), or the medial positioning of the hippocampus (translation equivariance of convolutions)⁵.

Although CapsNets are now known to be more robust to rotational perturbations of brain tissues thanks to this present work, they might also exhibit other clinical robustnesses. While it has been shown in previous works that capsules can learn a statistical equivariance on tissular properties, they might be relevant to the study of other hippocampal conditions where different difficulties are encountered, such as hippocampal sclerosis, temporal lobe epilepsy, or the Alzheimer's disease.

Conclusion

With respect to not-so-rare atypical variations of the hippocampus, we assessed the usefulness of Capsule Networks in hippocampal segmentation both in an anteroposterior and in subfields manner. With our newly introduced architecture, 3D-AGSCaps, we validated the first public implementation of 3D Capsules. On one hand, we confirmed the ability to perform hippocampal segmentation on T1w and T2w MRIs with 3D-AGSCaps, even if we found no evidence of superior segmentation quality for typical hippocampi (i.e.: without rotation). On the other hand, we demonstrated that with an increasing quantity of random rotational perturbations, 3D-AGSCaps provided better segmentations than analogous CNNs thanks to their implicit ability to learn equivariances over various instantiation parameters. Unfortunately, we also found capsules to be unequivocally demanding for GPU memory, which is the main drawback of this methodology. This concern raises the need for further investigations to bring back scalability into this promising methodology offering enhanced robustness, especially given that the hippocampus is a small brain region demanding for higher resolution MRIs.

Acknowledgements

460 The authors declare not to have any conflicts of interest. This work has been partly funded by the Fondation de France. Clément Poiret has a PhD funding from Paris Cité University.

References

1. Squire, L. R. *et al.* Role of the hippocampus in remembering the past and imagining the future. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 19044–19048 (2010).
2. Toda, T., Parylak, S. L., Linker, S. B. & Gage, F. H. The role of adult hippocampal neurogenesis in brain health and disease. *Mol Psychiatry* **24**, 67–87 (2019).
3. Despotović, I., Goossens, B. & Philips, W. MRI Segmentation of the Human Brain: Challenges, Methods, and Applications. *Computational and Mathematical Methods in Medicine* **2015**, 1–23 (2015).
4. Raininko, R. & Bajic, D. “Hippocampal Malrotation”: No Real Malrotation and Not Rare. *AJNR Am J Neuroradiol* **31**, E39–E39 (2010).
5. Cury, C. *et al.* Incomplete Hippocampal Inversion: A Comprehensive MRI Study of Over 2000 Subjects. *Front. Neuroanat.* **9**, (2015).
6. Iglesias, J. E. *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* **115**, 117–137 (2015).
7. Romero, J. E., Coupé, P. & Manjón, J. V. HIPS: A new hippocampus subfield segmentation method. *NeuroImage* **163**, 286–295 (2017).
8. Yushkevich, P. A. *et al.* Automated volumetry and regional thickness analysis of hippocampal subfields and medial temporal cortical structures in mild cognitive impairment: Automatic Morphometry of MTL Subfields in MCI. *Hum. Brain Mapp.* **36**, 258–287 (2015).
9. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv:1505.04597 [cs]* (2015).

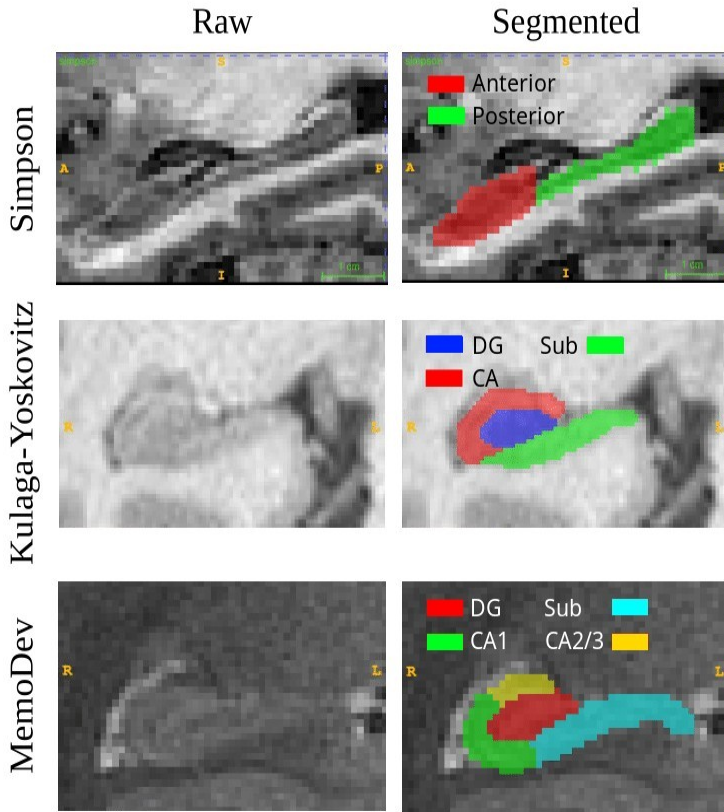
10. Engstrom, L., Tran, B., Tsipras, D., Schmidt, L. & Madry, A. Exploring the Landscape of Spatial Robustness. *arXiv:1712.02779 [cs, stat]* (2019).
11. Kanbak, C., Moosavi-Dezfooli, S.-M. & Frossard, P. Geometric robustness of deep networks: analysis and improvement. *arXiv:1711.09115 [cs]* (2017).
12. Xiao, C. *et al.* Spatially Transformed Adversarial Examples. *arXiv:1801.02612 [cs, stat]* (2018).
13. Qiu, Q., Gong, G., Wang, L., Duan, J. & Yin, Y. Feasibility of Automatic Segmentation of Hippocampus Based on Deep Learning in Hippocampus-Sparing Radiotherapy. *International Journal of Radiation Oncology*Biography*Physics* **105**, E137–E138 (2019).
14. Zhu, H. *et al.* Dilated Dense U-Net for Infant Hippocampus Subfield Segmentation. *Front. Neuroinform.* **13**, 30 (2019).
15. Punjabi, A., Schmid, J. & Katsaggelos, A. K. Examining the Benefits of Capsule Neural Networks. *arXiv:2001.10964 [cs, stat]* (2020).
16. Marcos, D., Volpi, M. & Tuia, D. Learning rotation invariant convolutional filters for texture classification. *2016 23rd International Conference on Pattern Recognition (ICPR) 2012–2017* (2016) doi:10.1109/ICPR.2016.7899932.
17. O’Gara, S. & McGuinness, K. Comparing Data Augmentation Strategies for Deep Image Classification. *IMVIP 2019: Irish Machine Vision & Image Processing* 9 (2019) doi:10.21427/148b-ar75.
18. Shorten, C. & Khoshgoftaar, T. M. A survey on Image Data Augmentation for Deep Learning. *J Big Data* **6**, 60 (2019).
19. Hinton, G. E., Krizhevsky, A. & Wang, S. D. Transforming Auto-Encoders. in *Artificial Neural Networks and Machine Learning – ICANN 2011* (eds. Honkela, T., Duch, W., Girolami, M. & Kaski, S.) vol. 6791 44–51 (Springer Berlin Heidelberg, 2011).
20. Sabour, S., Frosst, N. & Hinton, G. E. Dynamic Routing Between Capsules. *arXiv:1710.09829 [cs]* (2017).
21. LaLonde, R., Xu, Z., Irmakci, I., Jain, S. & Bagci, U. Capsules for biomedical image segmentation. *Medical Image Analysis* **68**, 101889 (2020).

22. Li, D., Zhao, X., Yuan, G., Liu, Y. & Liu, G. Robustness comparison between the capsule network and the convolutional network for facial expression recognition. *Appl Intell* **51**, 2269–2278 (2021).
23. Kwabena Patrick, M., Felix Adekoya, A., Abra Mighty, A. & Edward, B. Y. Capsule Networks – A survey. *Journal of King Saud University - Computer and Information Sciences* S1319157819309322 (2019) doi:10.1016/j.jksuci.2019.09.014.
24. Afshar, P. *et al.* 3D-MCN: A 3D Multi-scale Capsule Network for Lung Nodule Malignancy Prediction. *Sci Rep* **10**, 7948 (2020).
25. Afshar, P., Mohammadi, A. & Plataniotis, K. N. Brain Tumor Type Classification via Capsule Networks. *arXiv:1802.10200 [cs]* (2018).
26. Balaji, A., Kelsey, L., Majeed, K., Schultz, C. & Doyle, B. Coronary Artery Segmentation from Intravascular Optical Coherence Tomography Using Deep Capsules. *arXiv:2003.06080 [cs, eess, stat]* (2021).
27. Ahmad, A., Kakillioglu, B. & Velipasalar, S. 3D Capsule Networks for Object Classification from 3D Model Data. in *2018 52nd Asilomar Conference on Signals, Systems, and Computers* 2225–2229 (IEEE, 2018). doi:10.1109/ACSSC.2018.8645256.
28. Cheraghian, A. & Petersson, L. 3DCapsule: Extending the Capsule Architecture to Classify 3D Point Clouds. in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)* 1194–1202 (IEEE, 2019). doi:10.1109/WACV.2019.00132.
29. Kakillioglu, B., Ren, A., Wang, Y. & Velipasalar, S. 3D Capsule Networks for Object Classification With Weight Pruning. *IEEE Access* **8**, 27393–27405 (2020).
30. Oktay, O. *et al.* Attention U-Net: Learning Where to Look for the Pancreas. *arXiv:1804.03999 [cs]* (2018).
31. Shen, Y. *et al.* Coronary Arteries Segmentation Based on 3D FCN With Attention Gate and Level Set Function. *IEEE Access* **7**, 42826–42835 (2019).

32. Li, C. *et al.* Application of U-Shaped Convolutional Neural Network Based on Attention Mechanism in Liver CT Image Segmentation. in *Medical Imaging and Computer-Aided Diagnosis* (eds. Su, R. & Liu, H.) vol. 633 198–206 (Springer Singapore, 2020).
33. Choi, J., Seo, H., Im, S. & Kang, M. Attention routing between capsules. *arXiv:1907.01750 [cs]* (2019).
34. Huang, W. & Zhou, F. DA-CapsNet: dual attention mechanism capsule network. *Sci Rep* **10**, 11383 (2020).
35. Mazzia, V., Salvetti, F. & Chiaberge, M. Efficient-CapsNet: Capsule Network with Self-Attention Routing. *arXiv:2101.12491 [cs]* (2021).
36. Tsai, Y.-H. H., Srivastava, N., Goh, H. & Salakhutdinov, R. Capsules with Inverted Dot-Product Attention Routing. *arXiv:2002.04764 [cs, stat]* (2020).
37. Dalton, M. A., Zeidman, P., Barry, D. N., Williams, E. & Maguire, E. A. Segmenting subregions of the human hippocampus on structural magnetic resonance image scans: An illustrated tutorial. *Brain and Neuroscience Advances* **1**, 239821281770144 (2017).
38. Simpson, A. L. *et al.* A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv:1902.09063 [cs, eess]* (2019).
39. Kulaga-Yoskovitz, J. *et al.* Multi-contrast submillimetric 3 Tesla hippocampal subfield segmentation protocol and dataset. *Sci Data* **2**, 150059 (2015).
40. Bouyeure, A. *et al.* Hippocampal subfield volumes and memory discrimination in the developing brain. *Hippocampus* hipo.23385 (2021) doi:10.1002/hipo.23385.
41. Luo, P., Ren, J., Peng, Z., Zhang, R. & Li, J. Differentiable Learning-to-Normalize via Switchable Normalization. *arXiv:1806.10779 [cs]* (2019).
42. Abraham, N. & Khan, N. M. A Novel Focal Tversky loss function with improved Attention U-Net for lesion segmentation. *arXiv:1810.07842 [cs]* (2018).
43. Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T. & Ronneberger, O. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. *arXiv:1606.06650 [cs]* (2016).

44. Bhalerao, M. & Thakur, S. Brain Tumor Segmentation Based on 3D Residual U-Net. in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries* (eds. Crimi, A. & Bakas, S.) vol. 11993 218–225 (Springer International Publishing, 2020).
45. Rassadin, A. G. Deep Residual 3D U-Net for Joint Segmentation and Texture Classification of Nodules in Lung. *arXiv:2006.14215 [cs, eess]* **12132**, 419–427 (2020).
46. Paik, I., Kwak, T. & Kim, I. Capsule Networks Need an Improved Routing Algorithm. *arXiv:1907.13327 [cs]* (2019).
47. Lu, Y. The Level Weighted Structural Similarity Loss: A Step Away from the MSE. Preprint at <http://arxiv.org/abs/1904.13362> (2019).
48. Candemir, S., Nguyen, X. V., Folio, L. R. & Prevedello, L. M. Training Strategies for Radiology Deep Learning Models in Data-limited Scenarios. *Radiology: Artificial Intelligence* **3**, e210014 (2021).
49. Poiret, C. clementpoiret/ROIloc: A Registration-based Approach to ROI Location in T1w and T2w MRIs. (2021) doi:10.5281/ZENODO.5506958.
50. Hinton, G. How to represent part-whole hierarchies in a neural network. *arXiv:2102.12627 [cs]* (2021).
51. Wang, S. *et al.* Annotation-efficient deep learning for automatic medical image segmentation. *Nat Commun* **12**, 5915 (2021).

Figures



465

Figure 1: Random segmentation examples of the three datasets: Simpson (axial slice), Kulaga-Yoskovitz (coronal slice), and MemoDev (coronal slice). Letters indicate spatial directions: left (L), right (R), anterior (A), posterior (P), superior (S), and inferior (I).

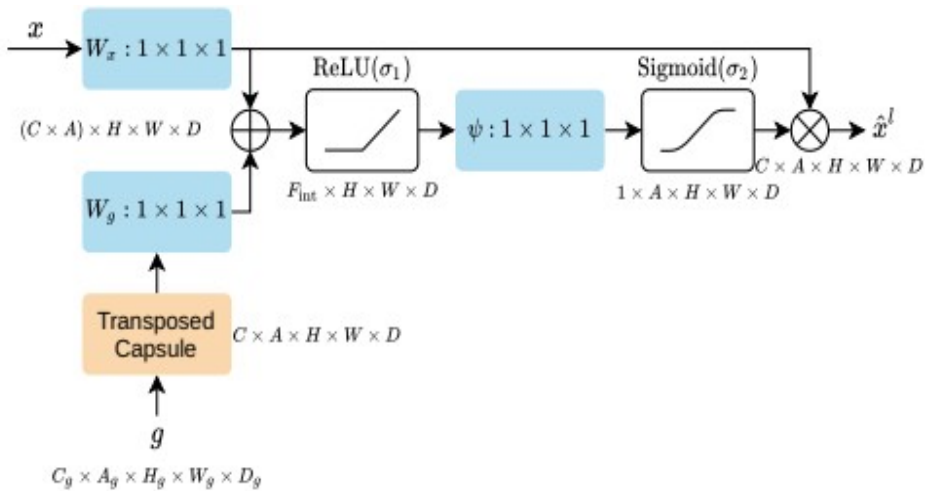


Figure 2: Our proposed Attention Gates. The input x comes from the downsampling path. The gating signal g comes from the layer $l-1$ in the upsampling path. g passes through a Transposed Capsule to match x 's size. Blue rectangles represent 3D Convolutions and SwitchNorm3D.

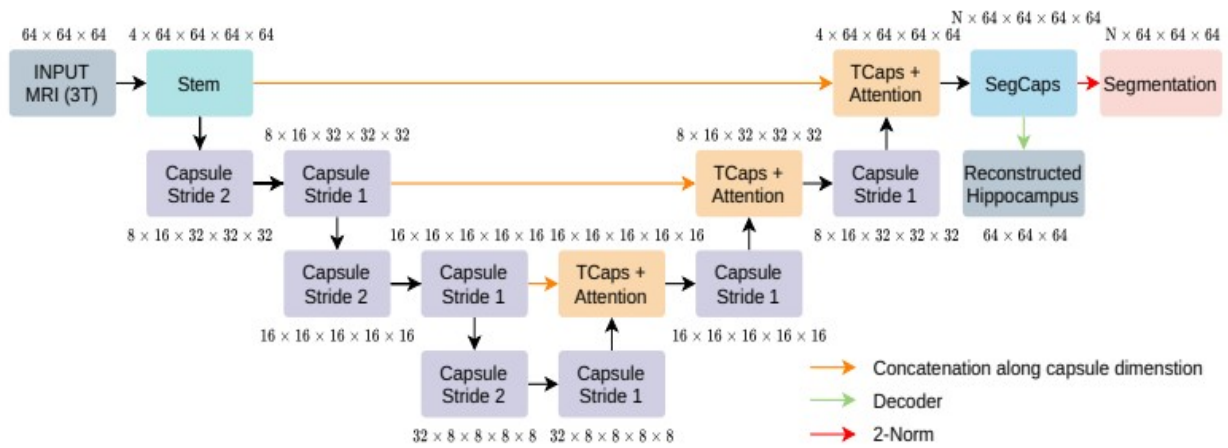


Figure 3: Attention-Gated SegCaps for volumetric segmentation (3D-AGSCaps). Attention gates are implemented after our Transposed Capsules (orange rectangles) to ensure both inputs are of the same size. Our network takes an MRI as input (of size 64^3 in this example), and outputs a reconstruction of the original Hippocampus (without the background class) alongside the segmentation.

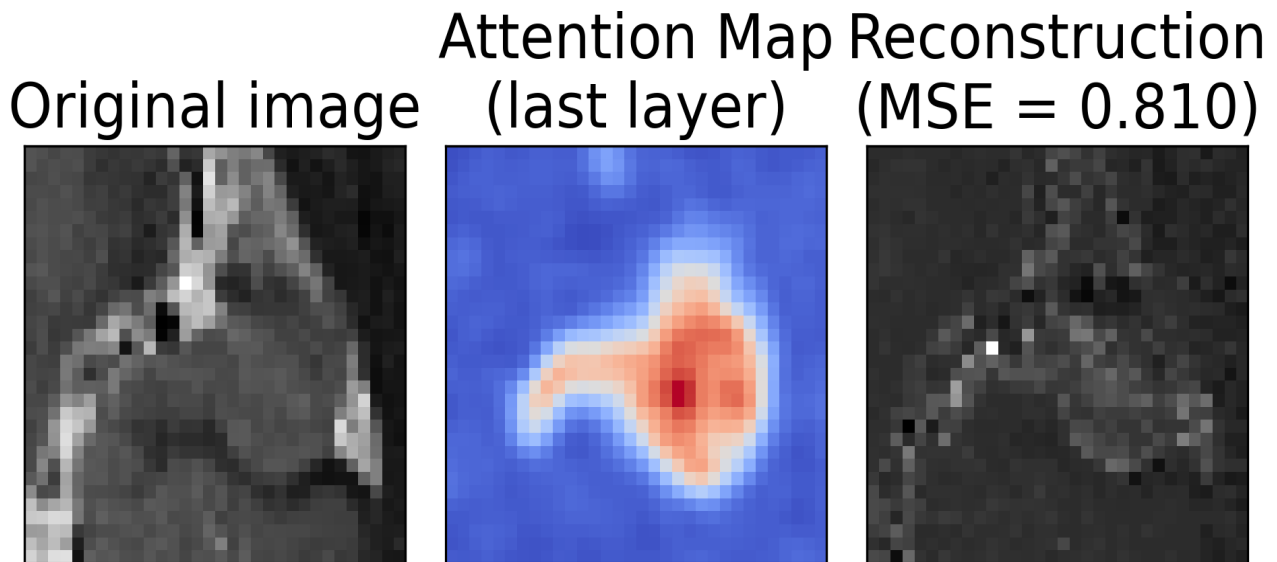


Figure 4: Example of attention map given an input x and its reconstruction \hat{x} (Kulaga-Yoskovitz dataset). The attention map comes from the very last layer just before entering the reconstruction decoder and the last two Capsule Layers.

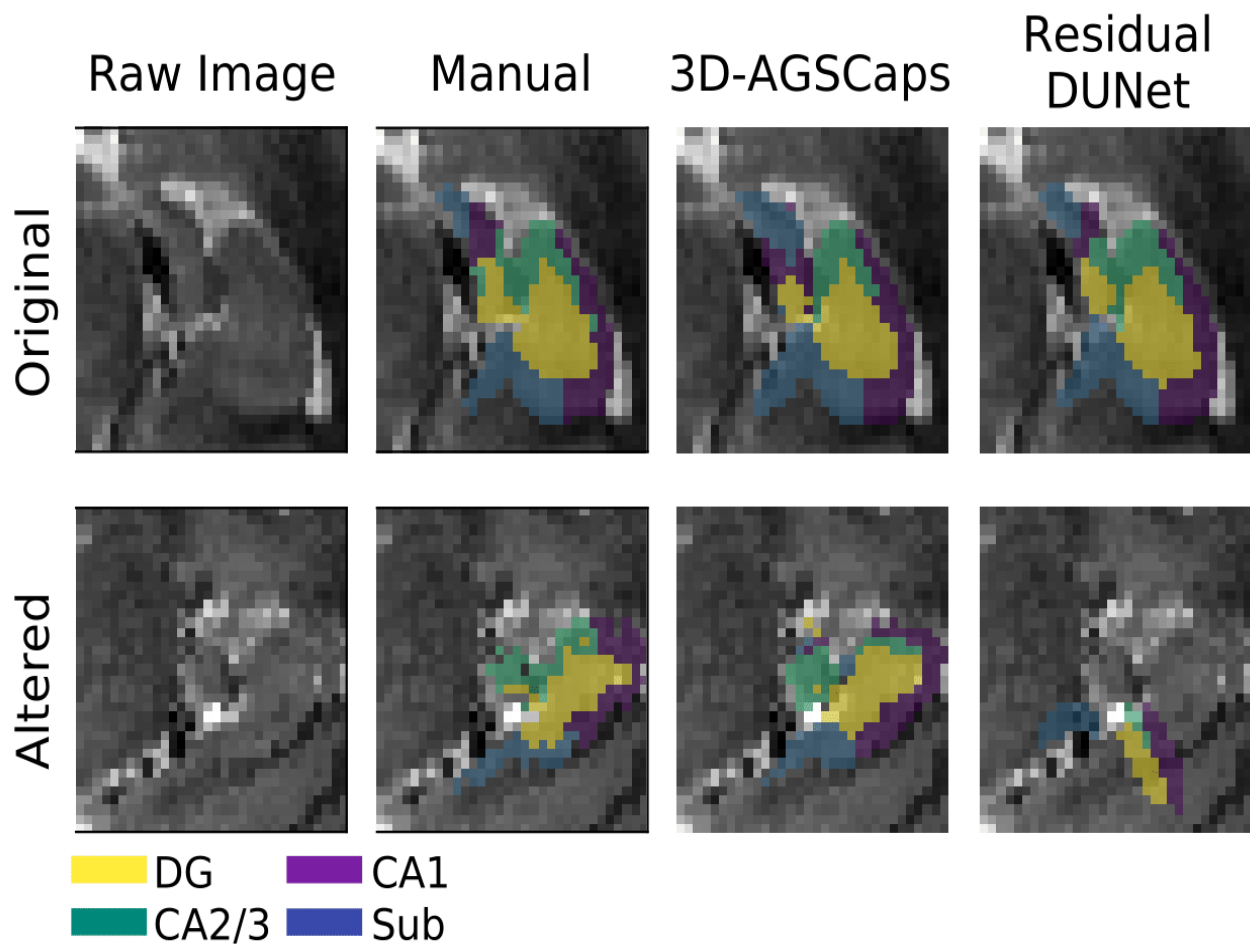


Figure 5: Comparison between automatic segmentations against manual labeling. A single unseen hippocampi is segmented when left unaltered (original), and when altered with a random rotation (45° of maximum amplitude in a random axis). We present a manual segmentation, a segmentation from our model 3D-AGSCaps, and a segmentation produced by a Residual DUNet from Zhu et al., (2019).

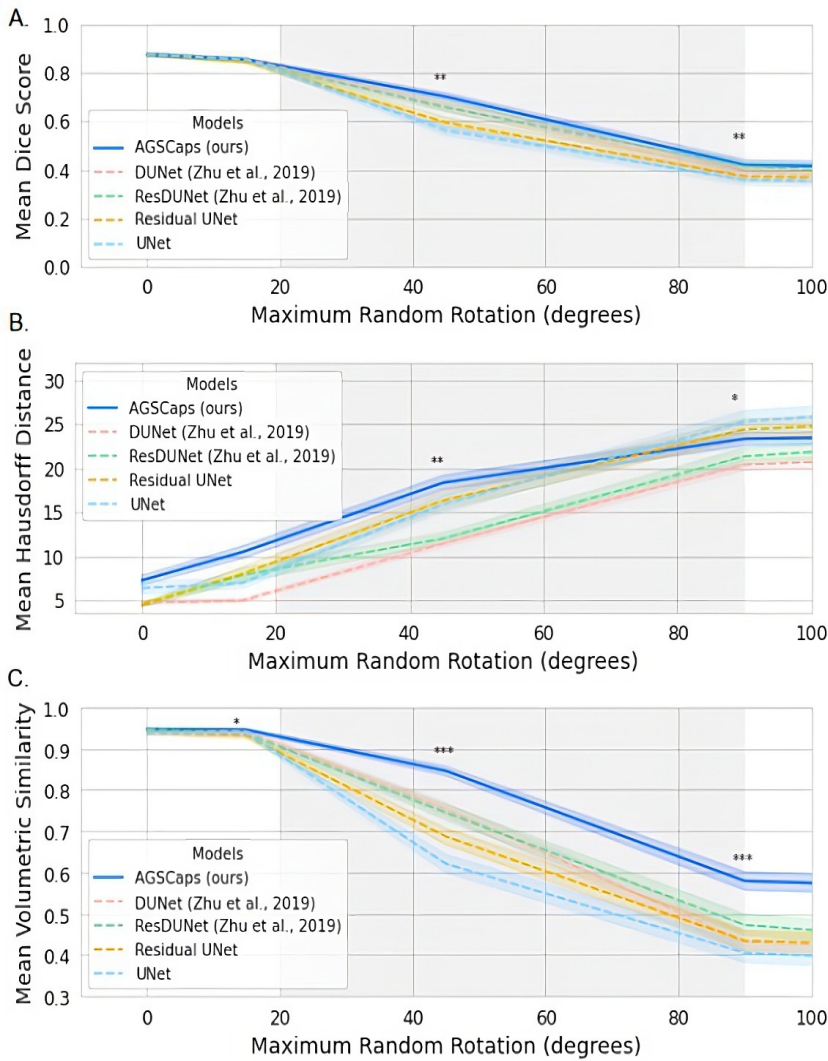


Figure 6: Mean evolution of segmentation quality with respect to an increasing degree of random rotations. p -values are indicated for each metrics (A. Dice coefficient, B. Hausdorff Distance, and C. Volumetric Similarity), where $*$ = $p < 0.05$, $**$ = $p < 0.01$, and $***$ = $p < 0.001$. The gray background roughly indicates ranges of naturally occurring (partial) Incomplete Hippocampal Inversion. p -values are corrected using a Benjamini-Hochberg False Discovery Rate.

3.1.2 A Fast and Robust Hippocampal Subfields Segmentation: HSF Revealing Lifespan Volumetric Dynamics

Clement Poiret, Antoine Bouyeure, Sandesh Patil, Antoine Grigis, Edouard Duchesnay, Matthieu Faillot, Michel Bottlaender, Frederic Lemaitre, and Marion Noulhiane (2023). A fast and robust hippocampal subfields segmentation: HSF revealing lifespan volumetric dynamics. *Frontiers in Neuroinformatics*. <https://doi.org/10.3389/fninf.2023.1130845>.

We found CapsNets to be promising for segmenting the hippocampal subfields, as they can account for anatomical variations and possibly other factors relative to image acquisition. They work by representing visual entities such as hippocampal subfields as “semantic” mathematical objects encoding their instantiation parameters, which effectively made them more robust to adversarial perturbations (Figure 3.7). However, we found CapsNets to not scale well, needing large amounts of memory and compute time because of computations happening on vectors instead of scalars. With environmental concerns and the need for efficient clinical tools, scalability is key.

Therefore, we integrated the best aspects of CapsNets into a more efficient convolutional neural network (CNN) architecture for HSF. We added our Attention Gates that help the network focus on the salient regions, reused the SwitchNorm layers (Luo et al., 2019), and deeply worked on hardware optimizations, capitalizing on pruning and quantization to compress the models.

Abstract

Goal: We aimed to develop an automated segmentation tool for the hippocampal subfields and apply it to study lifespan volumetric trajectories. We introduced HSF, an end-to-end deep learning pipeline that uses the latest advances in computer vision, along with bagging and human feedbacks.

Material and Methods: HSF was trained on manually segmented datasets and validated against ASHS, HIPS, and HippUnfold. We applied HSF to segment subfields in 3,750 HCP subjects aged 5-100+ years. Volumetric trajectories were modeled using natural cubic splines.

Results: HSF demonstrated superior overlap, fewer outliers, and greater volumetric similarity versus other tools. Analysis of HCP revealed distinct nonlinear subfields trajectories. The DG exhibited the strongest age correlation. Sex differences were found, with more pronounced growth and decay in men.

Conclusions: HSF enables fast and accurate subfields segmentation. The application to lifespan data clarified differential subfields dynamics and sex effects. HSF can facilitate the research on hippocampal subfields in large datasets.



OPEN ACCESS

EDITED BY

Xiaohao Cai,
University of Southampton, United Kingdom

REVIEWED BY

Christian Rummel,
University of Bern, Switzerland
Chao Wang,
Southern University of Science and
Technology, China

*CORRESPONDENCE

Marion Noulhiane
✉ marion.noulhiane@cea.fr

RECEIVED 23 December 2022

ACCEPTED 22 May 2023

PUBLISHED 15 June 2023

CITATION

Poiret C, Bouyeure A, Patil S, Grigis A,
Duchesnay E, Faillot M, Bottlaender M,
Lemaitre F and Noulhiane M (2023) A fast and
robust hippocampal subfields segmentation:
HSF revealing lifespan volumetric dynamics.
Front. Neuroinform. 17:1130845.
doi: 10.3389/fninf.2023.1130845

COPYRIGHT

© 2023 Poiret, Bouyeure, Patil, Grigis,
Duchesnay, Faillot, Bottlaender, Lemaitre and
Noulhiane. This is an open-access article
distributed under the terms of the [Creative
Commons Attribution License \(CC BY\)](#). The use,
distribution or reproduction in other forums is
permitted, provided the original author(s) and
the copyright owner(s) are credited and that
the original publication in this journal is cited, in
accordance with accepted academic practice.
No use, distribution or reproduction is
permitted which does not comply with these
terms.

A fast and robust hippocampal subfields segmentation: HSF revealing lifespan volumetric dynamics

Clement Poiret^{1,2,3}, Antoine Bouyeure^{1,2,3}, Sandesh Patil^{1,2,3},
Antoine Grigis^{2,3}, Edouard Duchesnay^{2,3}, Matthieu Faillot^{2,4},
Michel Bottlaender^{2,4}, Frederic Lemaitre^{5,6} and
Marion Noulhiane^{1,2,3*}

¹UNIACT, NeuroSpin, CEA Paris-Saclay, Frederic Joliot Institute, Gif-sur-Yvette, France, ²NeuroSpin, CEA Paris-Saclay, Frederic Joliot Institute, Gif-sur-Yvette, France, ³InDEV, NeuroDiderot, Université Paris Cité, Inserm, Paris, France, ⁴BioMaps, Service Hospitalier Frédéric Joliot, CNRS, Inserm, Université Paris-Saclay, Orsay, France, ⁵CETAPS EA 3832, Université de Rouen, Rouen, France, ⁶CRILOBE, UAR 3278, CNRS-EPHE-UPVD, Mooréa, France

The hippocampal subfields, pivotal to episodic memory, are distinct both in terms of cyto- and myeloarchitecture. Studying the structure of hippocampal subfields *in vivo* is crucial to understand volumetric trajectories across the lifespan, from the emergence of episodic memory during early childhood to memory impairments found in older adults. However, segmenting hippocampal subfields on conventional MRI sequences is challenging because of their small size. Furthermore, there is to date no unified segmentation protocol for the hippocampal subfields, which limits comparisons between studies. Therefore, we introduced a novel segmentation tool called HSF short for hippocampal segmentation factory, which leverages an end-to-end deep learning pipeline. First, we validated HSF against currently used tools (ASHS, HIPS, and HippUnfold). Then, we used HSF on 3,750 subjects from the HCP development, young adults, and aging datasets to study the effect of age and sex on hippocampal subfields volumes. Firstly, we showed HSF to be closer to manual segmentation than other currently used tools ($p < 0.001$), regarding the Dice Coefficient, Hausdorff Distance, and Volumetric Similarity. Then, we showed differential maturation and aging across subfields, with the dentate gyrus being the most affected by age. We also found faster growth and decay in men than in women for most hippocampal subfields. Thus, while we introduced a new, fast and robust end-to-end segmentation tool, our neuroanatomical results concerning the lifespan trajectories of the hippocampal subfields reconcile previous conflicting results.

KEYWORDS

deep learning, semantic segmentation, MRI, development, aging

1. Introduction

Episodic memory, the memory of specific episodes with spatiotemporal details, is critically underpinned by the hippocampal subfields, namely the dentate gyrus (DG), cornu ammonis from 1 to 3 (CA1/2/3), and the subiculum. Each subfield presents a distinct myelo- and cyto-architecture, and plays a critical role in episodic memory functions. For example, the DG and CA3 are involved in pattern separation, which allows the storage and retrieval of similar but distinct events (Yassa and Stark, 2011). CA1 and subiculum are necessary for pattern completion, i.e., the reconstruction of a full memory from partial elements. Since episodic memory performance correlates with variations in hippocampal subfields volume

(Palombo et al., 2018), we hypothesize that hippocampal subfields volumetric trajectories are associated with the evolution of episodic memory performance across the lifespan.

Analyzing hippocampal subfields' dynamics implies delineating their boundaries, boundaries often defined at a microscopic scale. Unfortunately, Magnetic Resonance Imaging (MRI) cannot study the unique myelo- and cyto-architectures of subfields, because structures such as CA1 and the Subiculum have the same contrast (Yushkevich et al., 2015a). Numerous efforts have been made to use geometrical heuristics to map histological features to MRI, thereby providing manual segmentation guidelines (Berron et al., 2017; Dalton et al., 2017). Manual segmentation with these protocols is now considered the gold standard for studying the hippocampal subfields *in vivo*. However, it is a complex, time-consuming, and subjective task which makes it error-prone and limits reproducibility. MRI segmentation of hippocampal subfields faces multiple difficulties, mainly caused by a lack of resolution, tissue ambiguity (notably in the head and the tail of the hippocampus), and noise. This problem is amplified by the lack of standardized segmentation protocols. For example, some protocols merge CA1, 2, and 3, sometimes delineating a separate CA4 or even excluding the hippocampal head or tail. This leads to multiple divergent protocols, inducing a lot of variabilities, notably in the boundary between DG and CA3, 4, and the boundary between CA1 and the subiculum with inter-protocol differences of almost 2 mm (Yushkevich et al., 2015a).

Recent efforts have been made to uniformize and automatize the hippocampal subfields segmentation task (Yushkevich et al., 2015a; Wisse et al., 2017). New hippocampal subfields' segmentation tools have recently been developed, such as ASHS (Yushkevich et al., 2015b), HIPS (Romero et al., 2017), or even more recently HippUnfold (DeKraker et al., 2021). They provide better segmentations, closer to manual segmentation, but neither of them implements state-of-the-art end-to-end deep learning which has been proven to be more fault-tolerant and adaptable to new observations, especially on complex and non-linear tasks (O'Mahony et al., 2020). Recent studies highlighted the possible gains of end-to-end deep learning for hippocampal segmentation (Qiu et al., 2019; Zhu et al., 2019; Yang et al., 2020), promising fast inference time (less than a minute per subject against several hours for FreeSurfer), higher accuracy, and higher robustness to anatomical variations. Unfortunately, most deep learning solutions are currently provided as a proof-of-concept, with either no public implementation, no pre-trained models, or are trained on small and specific datasets limiting generalizability. The current literature lacks an end-to-end deep learning segmentation protocol trained on a heterogeneous database to ensure segmentation quality across (i) contrast, (ii) magnetic field intensity, (iii) age range, or (iv) health condition.

Even though segmentation protocols still need to be uniformized, there is a disparity of available segmentation tools for the hippocampal subfields. The current understanding of the effect of age and sex on volumetric changes in hippocampal subfields across the lifespan is based on manual or (semi-)automatic segmentation studies. Uematsu et al. (2012) found that the total hippocampal volume is increasing until early adulthood. Another study showed a differential maturation between the posterior and anterior hippocampal portions (Gogtay et al., 2006). Regarding

sex difference, Suzuki (2004) showed that the myelination process, which is thought to contribute to the increase in volume during adolescence, takes place earlier in women (i.e., before the age of 18) than in men (i.e., after the age of 20), with a potentially more pronounced developmental dynamic in men than in women. Ziegler et al. (2012) noted an increase of gray matter volume during adulthood in the hippocampus up to 41 years old, with a maximum at 62 years old for the DG and CA, followed by fast atrophy. This is in accordance with Yang et al. (2013), who identified a quadratic relationship between the overall volume of the hippocampus and age, with an inflection point at 63 years old, followed by a strong negative correlation between volume and age.

Non-human primate studies (e.g., 20) have shown that subfields such as the DG, CA2/3, and the subiculum (but not the pre- or para-subiculum) are growing asynchronously until adulthood. However, this question has only been recently addressed in human children and adolescents with inconsistent results. According to Ellis et al. (2021), the DG exhibits a very rapid growth in infants, doubling in size, associated with an increase in CA1 and CA3 volumes during development (8–14 years old). This contrasts with a stable or a slight linear decrease in subicular volumes (Ziegler et al., 2012; Lee et al., 2014). Concerning normal aging, data suggest a volumetric decrease of all subfields which predominates in the DG (de Flores et al., 2015; Foster et al., 2019). While the literature suggests a differential maturation and aging of hippocampal subfields, there is currently a lack of accurate automated segmentation tools which hinders the use of large datasets to study trajectories across the lifespan.

Here, we offer the first end-to-end deep learning pipeline to segment the hippocampal subfields. Hippocampal segmentation factory (HSF) is an open-source tool that leverages new computer vision segmentation methods. It was trained on a heterogeneous database comprising all public datasets with manually segmented hippocampal subfields and new manually segmented observations to ensure generalization. We hypothesized (i) that HSF provides a better overlap (dice coefficient), fewer outliers (Hausdorff distance), and a better volumetric similarity than currently available tools abiding Barron's protocol (Berron et al., 2017); (ii) that subfields such as DG and CA1 exhibit differential lifespan dynamics which can be divided into three periods (a. growth, b. stability, and c. decay); (iii) a fast decay for all subfields starting from 60 to 65 years old; (iv) finally that there are sex differences in the volumetric trajectories, with volumetric variations being more intense in men than in women.

2. Method

This section aims at describing (i) the technical details of HSF development in terms of computational architecture, training regime, and inference peculiarities, (ii) how it differs from other state-of-the-art tools addressing the same segmentation problem, and (iii) how we leveraged the potential of HSF to study hippocampal subfields volumetric trajectories in large healthy individuals datasets which covers the lifespan (5–100+ years old). Please note that we conducted a speed test comparing the tools included in our benchmark. While FreeSurfer, one of the most used neuroimaging tools, possesses modules for hippocampal subfields

segmentation, we chose not to compare it. Although FreeSurfer (Iglesias et al., 2015) is still considered a classic neuroimaging tool, it has recently incorporated deep learning-based approaches. Because it has useful automated features outside the scope of this study, it produces many outputs leading to a long computing time, which makes it inconvenient for scientific studies interested in a single substructure of the human brain: its inference time of approximately 10 h per subject is slower than manual segmentation of the hippocampal subfields. While previous studies found the segmentation quality of FreeSurfer to be good enough to study the hippocampal subfields (Schmidt et al., 2018), others have demonstrated that FreeSurfer has poorer segmentation quality in comparison to the tools included in our benchmark (de Flores et al., 2015; DeKraker et al., 2022), with segmentations that are in a mismatch with known anatomical boundaries leading to a significantly different volumetry, especially in the head and the tail of the hippocampus (Wisse et al., 2014). Thus, as we are only interested in fast tools only tackling hippocampal subfields segmentation, we only used FreeSurfer as a benchmark for speed comparison.

2.1. HSF: description of the hippocampal segmentation factory

HSF is designed to be a fully customizable end-to-end pipeline, handling tasks from the preprocessing of raw anatomical images, to the segmentation of the hippocampal subfields through specialized and highly efficient deep learning models comprised in a “Model Hub” on any hardware acceleration platform such as CUDA, TensorRT, or OpenVINO. HSF also supports the DeepSparse compute engine to benefit from the AVX512 (VNNI) vector instruction set. HSF is distributed under the MIT license at <https://github.com/clementpoiret/HSF>.

2.1.1. Datasets description

The key strength of HSF lies in its training database, which consists of 12 datasets of manually segmented hippocampi by individual expert raters (Table 1), totaling 411 subjects.

2.1.2. Internal information processing

The HSF pipeline consists of three main steps: 1/a preprocessing step handled by ROIloc (a standalone by-product of HSF available at <https://github.com/clementpoiret/ROIloc>) to extract the hippocampi from a given MRI (Figure 1), 2/a augmentation pipeline, and 3/a segmentation by multiple expert models in order to produce both the segmentation and an uncertainty map (Figure 2).

In order to limit the computational impact of HSF, we used a preprocessing step to extract the hippocampi from the MRI. To do so, ROIloc registers the MNI152 09c Sym template (Fonov et al., 2009) to the T1w or T2w input MRI. Utilizing the CerebrA atlas (Manera et al., 2020), the registration process facilitated the inference of approximate coordinates of the hippocampus in native space. ROIloc then crops the MRI

into two volumes corresponding to the right and left hippocampi from head to tail, with an arbitrary safety margin. To finish the preprocessing, the resulting crops are Z-normalized and padded to obtain shapes that are multiple of 8 to satisfy hardware acceleration constraints.

HSF provides a “Model Hub” offering multiple pre-trained models that can handle preprocessed hippocampi. Our built-in models are 3D Residual UNets of depth 4, with ResNet building blocks (Zhang et al., 2018) and transposed convolutions as the upsampling method. We have replaced the additive skip connections with a self-attention mechanism inspired by the one introduced for 2D images by Oktay et al. (2018), with BatchNorm layers replaced by SwitchNorm layers (Luo et al., 2019). Each segmentation model has its efficient counterpart that can benefit from the AVX512-VNNI instruction set due to pruning (at 70%) and int8-Quantization through NeuralMagic’s SparseML.

2.1.3. Training methodology

To augment the quality of the segmentation, we employed the widely used technique called bagging. We trained five “weak-learner” models, each of which was generated by random sampling, with replacement, N samples from the original training set, which contained 822 hippocampi. The bagging technique then amalgamated each weak learner into a strong learner, which displayed a superior accuracy of prediction compared to each weak learner on its own. Bagging outperforms the conventional random split because it introduces more variability (i.e., some subjects can be observed multiple times during a single epoch), thereby enhancing the prediction of the strong learner (Opitz and Maclin, 1999). Each model is trained with an AdamW optimizer, a one-cycle learning rate scheduler, and stochastic weight averaging for 512 epochs with a batch size of 1 to handle heterogeneous input volumes. int8-Quantized models are trained with quantization-aware training.

Given an input x , our segmentation loss L is defined with TP and TN the true positives and negatives, FP and FN the false positives and negatives, and $\alpha = 0.3$, $\beta = 0.7$, $\gamma = \frac{3}{4}$ such as:

$$L = \left(1 - \frac{TP}{TP + \beta FN + \alpha FP}\right)^\gamma$$

While the base loss function is a focal Tversky, the loss function was modulated for each observation to handle different segmentation protocols. As HSF predicts CA1, CA2, and CA3, we merged classes (e.g., CA2 and CA3) at training time to learn from observations that do not distinguish them. For segmentation protocols having a separate head or tail class, all predictions are merged to form a single ‘hippocampus’ class so that predicting any subfield outside the ‘head’ or ‘tail’ class is penalized but not inside of them.

2.1.4. Inference

To further enhance the segmentation pipeline, test-time augmentation is natively implemented, augmenting each hippocampus with random horizontal flips, and with affine and elastic deformations. The final segmentation is computed

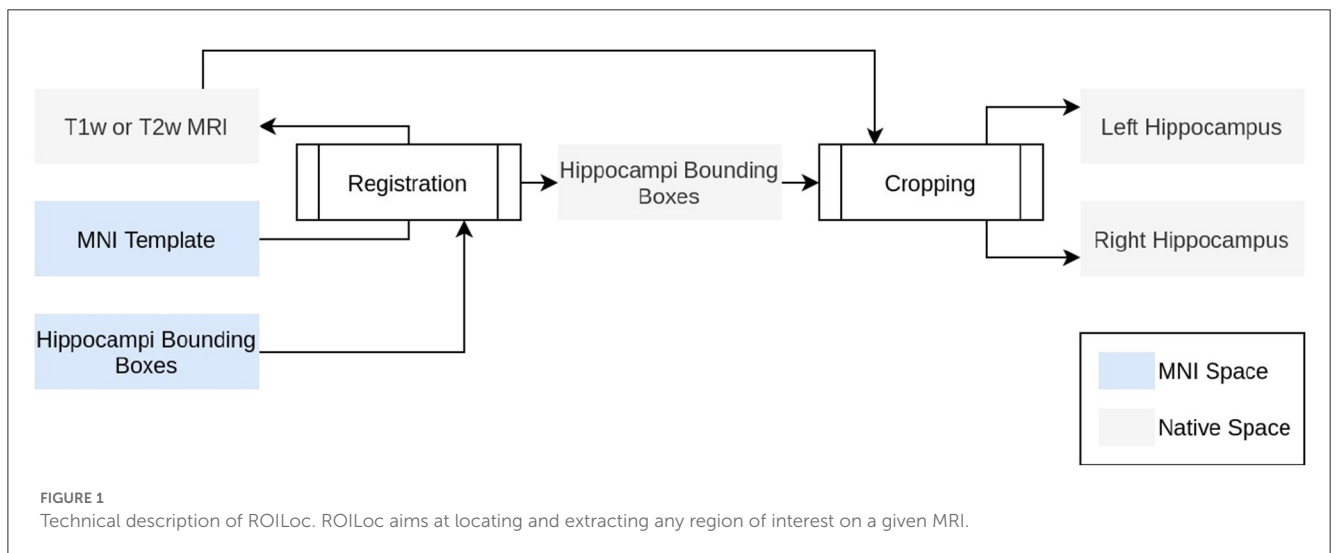
TABLE 1 Training datasets used for HSF.

Dataset	Contrasts	Subfields	Field	Age	Condition
Winterburn et al. (2013)	T1 and T2	DG/CA/Sub	3T	29–57	-
Kulaga-Yoskovitz et al. (2015)	T1 and T2	DG/CA/Sub	3T	21–53	-
Yushkevich et al. (2015b)	T1 and T2	DG/CA1/CA2-3/Sub	3T		MCI
Hindy et al. (2016)	T2	DG/CA1/CA2/CA3/Sub	3T	18–30	-
Bouyeure et al. (2021)	T1 and T2	DG/CA1/CA2-3/Sub	3T	4–12	-
Yushkevich et al. (2010)	T1 and T2	DG/CA1/CA2/CA3/Sub	4T	38–82	MCI/AD
HIPlay7*	T1 and T2	DG/CA1/CA2/CA3/Sub	7T	12–21	TLE
Wisse et al. (2016)	T2	DG/CA1/CA2/CA3/Sub	7T	50–68	-
Berron et al. (2017)	T1 and T2	DG/CA1/CA2/CA3/Sub	7T	19–32	-
Haeger et al. (2020)**	T2	DG/CA1/CA2/CA3/Sub	7T	50–70	-
Shaw et al. (2020)**	T1 and T2	DG/CA1/CA2/CA3/Sub	7T	23–29	-
Lagarde et al. (2021)**	T2	DG/CA1/CA2/CA3/Sub	7T	50–84	SC/MCI/AD

Description of the training database, alongside their manually segmented hippocampal subfields, namely the dentate gyrus (DG), the cornu ammonis (CA)1, CA2, and CA3. Included participants are either healthy, exhibiting mild cognitive impairments (MCI), Alzheimer’s disease (AD), hippocampal sclerosis (SC), or temporal lobe epilepsy (TLE).

*In-house dataset, ANR-16-NEUC-0001-01; Manual Segmentation on 23 controls and 4 temporal lobe epilepsies; 1 mm T1w and 0.125*0.125*1.2 mm T2w MRIs.

**Manual segmentations on 7 subjects per dataset performed by the authors (CP, SP, MF, MB, and MN), following Berron et al. (2017).



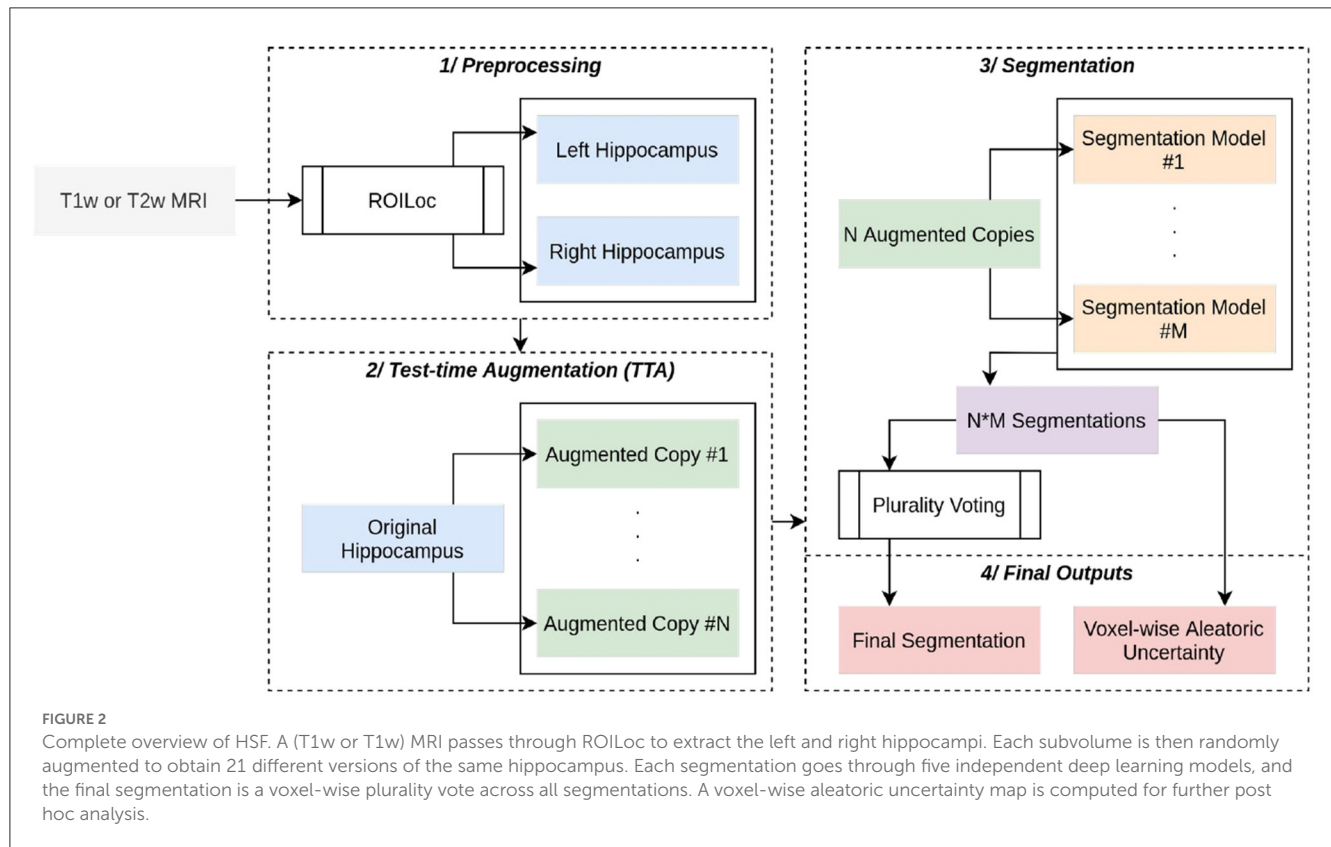
as a voxel-wise plurality vote, assigning to a given voxel the most frequent class. For the sake of further *post hoc* analysis of the segmentation quality, a voxel-wise aleatoric uncertainty $H(Y^i \vee X)$ is also computed (Wang et al., 2019). Given a set Y of i predictions, in HSF:

$$H(Y^i \vee X) \approx - \sum_{m=1}^M \hat{p}_m^i \ln \hat{p}_m^i \tag{1}$$

where \hat{p}_m^i is the frequency of the m^{th} unique value in Y^i .

2.2. Benchmarking HSF against ASHS, HIPS, and HippUnfold

HSF has been assessed against the most recent and widespread tools for hippocampal segmentation: ASHS (Yushkevich et al., 2015b), HIPS (Romero et al., 2017), and HippUnfold (DeKraker et al., 2021). To compare it with manual segmentations, CP, AB, SP, and MF randomly segmented 25 subjects who were excluded from our training set from 5 different datasets: HiPlay7, MemoDev (Bouyeure et al., 2021) (Table 1), as well as HCP-Development (HCP-D), HCP-Young Adults (HCP-YA), and HCP-Aging (HCP-A). This segmentation process took approximately 5 h per hippocampus. In relation to an earlier study on MemoDev, an assessment was conducted by Bouyeure et al. (2021) to determine the reliability of the manual segmentations. This evaluation involved the computation of an inter-rater reliability index,



specifically the dice coefficient, between two individual tracers, who followed the same segmentation protocol. Furthermore, it is worth noting that both raters had no prior knowledge of the participants’ age, sex, or memory performance. The obtained inter-rater reliability indices were notably high at 0.77 and 0.79 for the right and left hippocampi, respectively. Segmentations are compared on three metrics:

- the dice coefficient (DC), an overlap metric ranging from 0 (no overlap) to 1 (full overlap) defined as $DC = \frac{2|y_m \cap y_p|}{|y_m| + |y_p|}$,
- the Hausdorff distance (HD), a metric of surface distance ranging from 0 to +inf. With the directed Hausdorff distance between two point sets X and Y such as $hd(X, Y) = \max_{x \in X} \min_{y \in Y} \|x - y\|_2$, the HD is defined as $HD(y_m, y_p) = \max(hd(y_m, y_p), hd(y_p, y_m))$,
- and the volumetric similarity (VS), a comparison between volumes of two segmentations ranging from 0 (complete dissimilarity between volumes) to 1 (exact match between volumes). With the volume of a region S , it is defined as $VS = 2 \frac{|S_m \cap S_p|}{|S_m + S_p|} \cdot 100\%$.

As both T1w and T2w images can be segmented by HSF, we conducted an additional analysis to evaluate any discrepancies in quality across these contrasts using the same metrics. Given the strong correlation between contrast and resolution (e.g., an isometric millimetric MPRAGE 3D T1w and anisotropic 2D Coro-T2w), we limited our study to only 15 subjects from our test set sourced from the HCP databases, where T1w and T2w MRIs are in the same space and at the same resolution.

Owing to the presence of either heteroscedasticity or non-normal distributions of scores, we compared segmentations utilizing non-parametric Kruskal–Wallis or pairwise Wilcoxon–Mann–Whitney tests, with p-values corrected using the Benjamini–Hochberg false discovery rate.

2.3. HSF: analyzing the Human Connectome Project

The following sections are specifically dedicated to explaining how we used HSF (process and inference) to study hippocampal subfields trajectories across the lifespan in the HCP datasets (HCP-D, HCP-YA, and HCP-A).

2.3.1. Datasets descriptions

All databases are acquired on a 3T Siemens Prisma (Skyra for HCP-YA) scanner:

- **HCP-D:** HCP-D contains 1350 healthy children, adolescents, and young adults aged from 5 to 21 years. T1w and T2w MRIs are acquired at an isotropic resolution of 0.8 mm across four sites (Somerville et al., 2018),
- **HCP-YA:** HCP-YA includes 1,200 subjects with ages ranging from 22 to 35 years. T1w and T2w MRIs have been acquired on a single site at an isotropic resolution of 0.7 mm,
- **HCP-A:** HCP-A comprises 1,200 subjects from 36 to 100+ years old. T1w and T2w MRIs are acquired at an isotropic

resolution of 0.8 mm across four different sites (Bookheimer et al., 2019).

The HCP datasets were provided in part by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research and by the McDonnell Center for Systems Neuroscience at Washington University.

2.3.2. MRI segmentation

Prior to HCP's datasets' segmentation and after the HSF validation, we retrained HSF's models with the manually segmented observations coming from the previous section (see Section 2.2.) including observations from the HCP's datasets. We thereby improved the reliability of segmentations by including new and HCP-specific observations to ensure there was no mismatch between our training set's distribution and HCP's distribution of observations. All segmentations are performed on T2w images, with ROIloc's location algorithm using the 'Affine' registration and a margin of 16 voxels in all directions to ensure that whole hippocampi are included in their boxes.

2.3.3. Lifespan modeling

The whole hippocampus and each subfield were modeled for each sex as a natural cubic spline (NCS) regression between age and volume, a flexible, simple, and efficient model to describe trends (Greenland, 1995; Elhakeem et al., 2022). Cubic models have been validated to study developmental trajectories of the amygdala and the whole hippocampus (Uematsu et al., 2012; Bussy et al., 2021). NCS allowed us to model the growth and decay of hippocampal subfields by fitting a set of piecewise polynomial regressions smoothly joining at points called knots, with a linearity constraint at the extremity of the curve. Significance and goodness of fit for the NCS are computed similarly to linear regressions because NCS are fitted using an ordinary least-squares algorithm. We chose the number of degrees of freedom by minimizing an Akaike Information Criterion. Then, inflection points in the volumetric trajectories of the ROIs were detected as suggested by Satopaa et al. (2011). Finally, we computed an anteroposterior evolution of the subfield's volume on a per-slice basis averaged across every subject.

2.3.4. Statistical analysis

Although lifespan dynamics of the hippocampus and its subfields are thought to be non-linear (e.g., 15,18,34), we assume that within a single period, defined as the uninterrupted period between two distinct inflection points (e.g., young adults), the relationship between age and volume is linear. Therefore, for each lifespan period, we tested (i) the relationship between age and volume, (ii) the relationship between sexes, and (iii) the interaction between these two independent variables using an ordinary least-squares regression. *P*-values are corrected using a Benjamini-Hochberg false discovery rate.

3. Results

3.1. Benchmarking HSF against ASHS, HIPS, and HippUnfold

First, we validated HSF against three state-of-the-art hippocampal subfields segmentation tools: ASHS, HIPS, and HippUnfold (Figure 3). While manual segmentation may require up to 5 h per subject, FreeSurfer 7 may take even longer, exceeding 10 h due to its all-inclusive pipeline, encompassing whole-brain segmentation and cortical morphometry. As we were interested solely in hippocampal subfields segmentation, we have compared only the specialized tools, which were, therefore, much faster: HIPS, ASHS, and HippUnfold can segment a new subject in under an hour. HSF is even faster, taking only minutes to segment a new subject from the HCP. While HIPS requires the use of the volBrain service and can take up to a day to complete due to queueing, HSF is much quicker. In its most accurate mode, HSF takes only 5 mi on a CPU and 90 s on an NVIDIA A100 GPU (Table 2). In its fast mode, HSF can segment a new subject in only 15 s on both CPU and GPU, with the main speed bottleneck being the registration tool ANTs, which is used to localize the hippocampus (ROIloc).

We used dice coefficient, Hausdorff distance, and volumetric similarity (Figure 4) with manual segmentations as benchmarking metrics. We found HSF to exhibit a significantly better DC than ASHS ($p = 4e - 6$; hedge's $g = 1.636$), HIPS ($p = 7e - 9$; hedge's $g = 4.934$), and HippUnfold ($p = 7e - 9$; hedge's $g = 5.440$), with no differences between HippUnfold and HIPS.

Regarding HD, which is sensitive to outlier voxels in the segmentation, we found HSF performing on par with HIPS, but being better than HippUnfold ($p = 7e - 8$; hedge's $g = -1.184$). Importantly, ASHS mainly penalized by poor segmentation results in a few observations although estimation statistics may suggest a difference between the two tools (Figure 4). Our statistical tests failed to reject the null hypothesis.

With respect to the VS, all three methods had similar volumes, but HSF was the closest to manual segmentations (VS = 0.862), better than ASHS ($p = 2e - 4$; hedge's $g = 1.210$), HIPS ($p = 9e - 9$; hedge's $g = 3.391$), and HippUnfold ($p = 8e - 9$, hedge's $g = 3.550$). We found no differences between HIPS and HippUnfold.

After an extensive evaluation, we analyzed the disparities in segmentation quality compared to the T1w and T2w images on a subset of our test set where both contrasts were acquired using the same resolution, as outlined in Table 3. While the effect sizes were negligible, we found that T2w images tend to exhibit a slight inclination, with HSF producing segmentations closer to the manual ones, especially on the smallest regions, CA1, 2, and 3 (DC increased by 0.045, HD decreased by 2.386, and VS increased by 0.035).

3.2. Human Connectome Project

3.2.1. Lifespan development dynamics

After the HSF's retraining including new HCP subjects to ensure segmentation quality, we established lifespan trajectories (Figure 5) consisting of Natural Cubic Splines, from which we inferred inflection points reflecting lifespan critical periods. DG

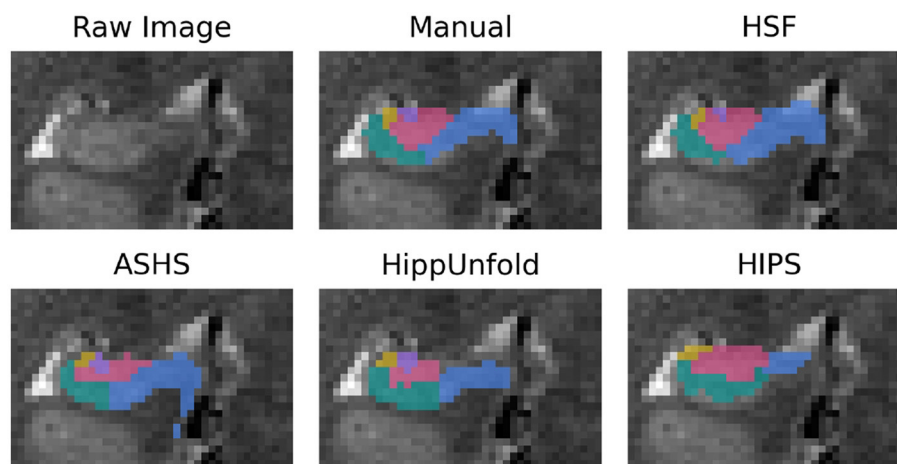


FIGURE 3 Segmentation example from a random subject. The dentate gyrus is in red, CA1/2/3 are in green, yellow, and purple, and the subiculum is in blue.

TABLE 2 Segmentation time of reference software vs. manual segmentation.

	Manual	FreeSurfer 7.3	HIPS	ASHS	HippUnfold	HSF
Segmentation time (min)	~300	678.58 (baseline)	N/A	30.10 ± 1.31	29.52 ± 2.35	1.64 ± 0.27

FreeSurfer segmentation time is computed as a reference point for a single HCP-aging subject (0.8 mm iso.). HIPS, ASHS, and HSF timings are mean ±std of segmentations on our complete test set of 25 subjects. Computations are conducted on a machine with an Intel(R) Xeon(R) Gold 6226R CPU @ 2.90 GHz, and an Nvidia A100 GPU.

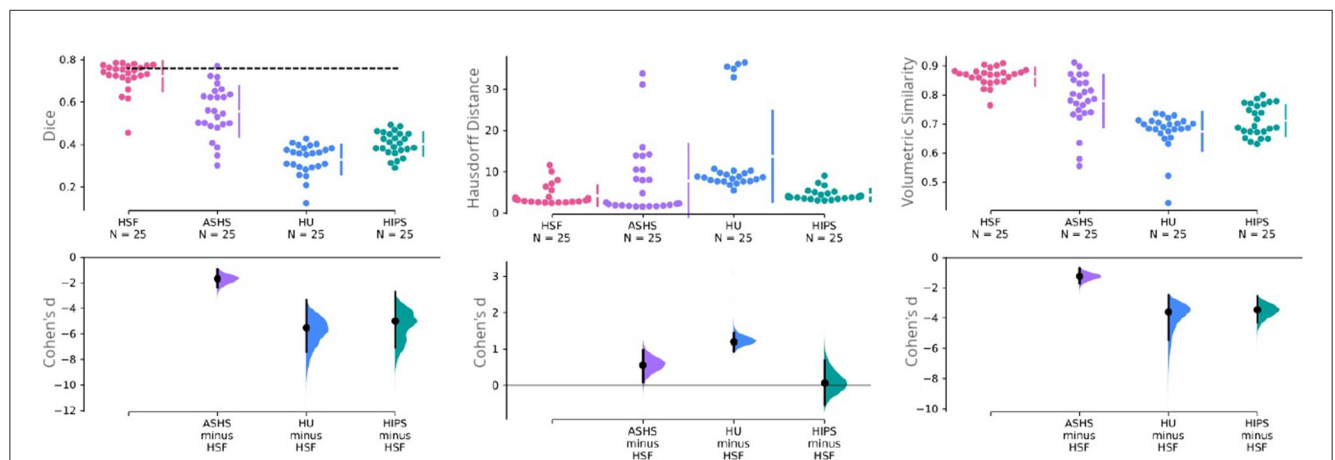


FIGURE 4 Cumming estimation plots comparing HSF (T2w) against ASHS (T1w and T2w), HIPS (T1w and T2w), and HippUnfold (HU) (T1w and T2w). The first row illustrates three performance metrics—the dice coefficient (higher is better), the Hausdorff distance (lower is better), and the volumetric similarity (higher is better). The vertical bars in this row represent the mean ±std for each metric group. The dashed line in this row represents the inter-rater reliability for manual segmentation obtained in the earlier study of Bouyeure et al. (2021). As this earlier study only computed the inter-rater comparison as the dice coefficient, it is not available for the other two metrics. The second row depicts the mean effect size (Cohen's d) with a black dot to facilitate statistical comparison between the groups. The black bars in this row represent 95% CIs for variability estimations. The 95% CIs are obtained through non-parametric bootstrap resampling to generate distributions of all possible effect sizes.

was the subfield whose developmental trajectory was the most correlated with age ($p = 0.005$). Total hippocampal volume was negatively correlated with age for both sexes starting from 70 years old ($p = 0.03$), which is also reflected in the subiculum ($p = 2e - 8$). In addition to significant differences in volumes between sexes mostly during the “stable adulthood” period, except for CA2/3 ($p = 0.120$), we found differences between men and women during

the “development” period in the DG ($p = 0.01$), and CA2/3 ($p = 0.01$), and during the aging period for the DG ($p = 0.015$) and CA1 ($p = 0.04$). Interestingly, we found differences in trajectories between men and women (i.e., interaction between age and sex), for the development period of CA2/3 ($p = 0.017$), for the aging period of the DG ($p = 0.04$), and before 60/70 years for the subiculum ($p = 0.016$).

TABLE 3 Comparison of the segmentations produced on T1w and T2w MRIs.

Label	DC			HD			VS		
	T1w	T2w	Delta	T1w	T2w	Delta	T1w	T2w	Delta
DG	0.850	0.900	0.050	4.880	2.540	-2.340	0.960	0.980	0.010
CA1	0.819	0.868	0.049	5.170	2.448	-2.722	0.907	0.952	0.045
CA2	0.781	0.828	0.047	4.400	1.952	-2.448	0.852	0.905	0.053
CA3	0.796	0.849	0.053	4.534	2.767	-1.767	0.868	0.924	0.056
Sub	0.830	0.859	0.029	5.492	2.787	-2.705	0.921	0.932	0.010

MRIs are coming from HCP-development, HCP- young adults, and HCP-aging. Those 15 subjects are a subset of our test set (N = 25). Those subjects are special cases where T1w and T2w MRIs are in the same space, with the same resolution. Deltas in bold denote significant differences at a p-value of < 0.05.

3.2.2. From head to tail: subfields' distribution

Delineating the subfields in the head and the tail of the hippocampus is a complex task, with some protocols not even delineating subfields in the tail. Due to the peculiar training methods, we trained HSF to segment the head and the tail even when there was no ground truth subfield segmentation in these regions. Using HSF, we created an overall normalized anteroposterior distribution of subfields across all three HCP datasets (Figure 6). We found no anatomical differences between lifespan periods and sexes.

According to HSF, the hippocampal head starts mostly with CA1, quickly followed by the subiculum and then the DG before the hippocampal body. After the body, CA2 and CA3 start to disappear and then followed by the DG. The tail comprises mostly subiculum, CA1, and a small portion of DG which disappears near the middle of the tail.

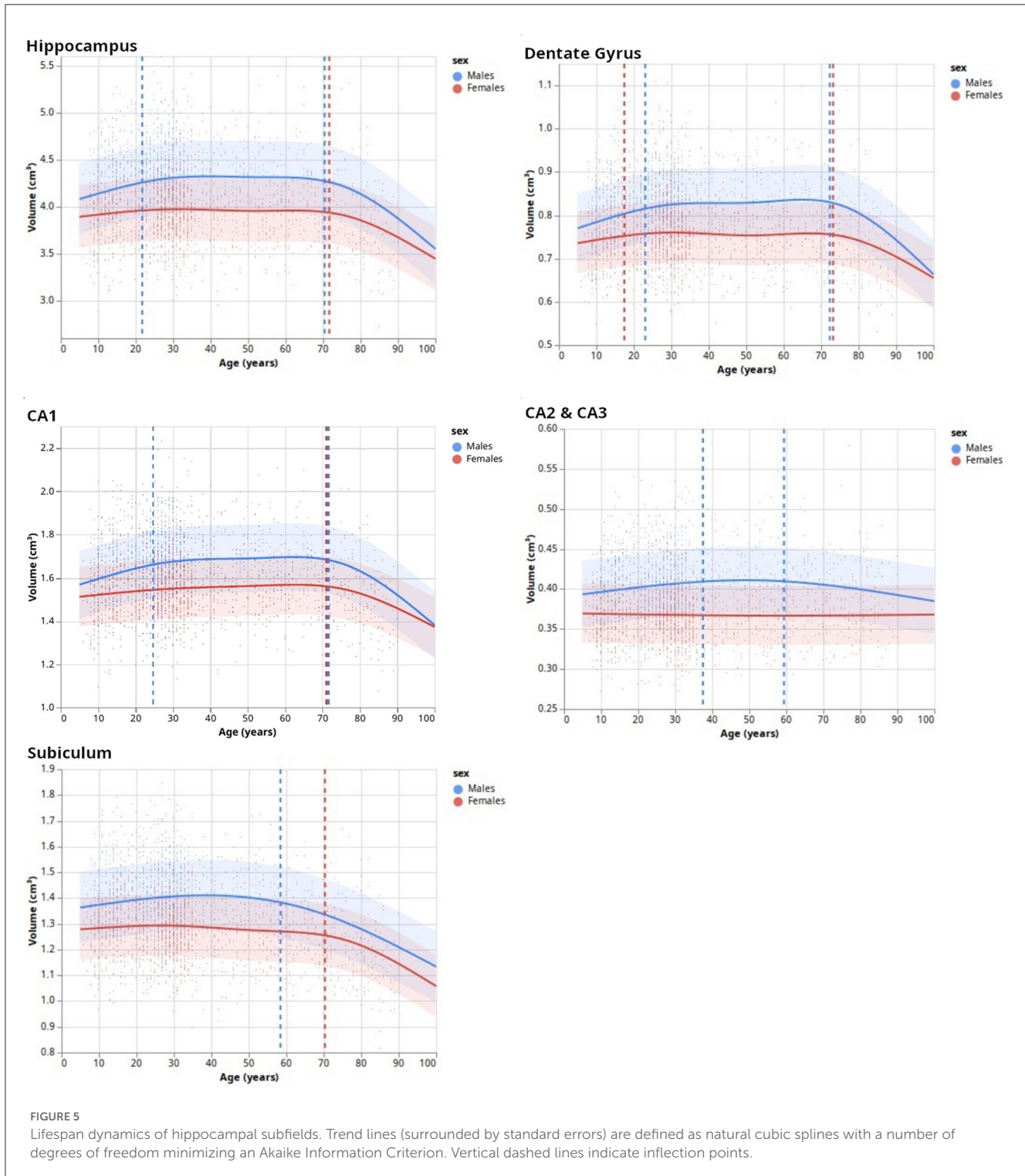
4. Discussion

This study had two main goals: 1/to introduce a new segmentation tool for the hippocampal subfield based on machine learning named hippocampal segmentation factor (HSF), which leverages the latest advances in computer vision, and 2/to study lifespan volumetric trajectories of hippocampal subfields in healthy individuals using the proposed tool. We developed and validated HSF, and demonstrated that it is faster than all previous tools while offering a better segmentation quality closer to manual segmentation. Then, applying our tool to data from 3,750 individuals (HCP-development, HCP-young adults, and HCP-aging), we show that hippocampal subfields have different volumetric trajectories across the lifespan. These trajectories are non-linear, and inflection points differ between males and females in accordance with prior literature (16).

First of all, we validated HSF in comparison to ASHS, HIPS, and HippUnfold. When looking at the DC, it has to be noted that, even in the absence of histological ground truth, HSF matches the inter-rater agreement (Figure 4). Moreover, its scalability benefits out-of-the-box from the latest advances in computing due to the open neural network exchange (ONNX) ecosystem and NeuralMagic's DeepSparse inference engine. HSF shows an unprecedented segmentation speed which makes it particularly suited to the processing of big datasets such as the

HCP. The bootstrap aggregation strategy, coupled with the test-time augmentation, makes HSF more robust than ASHS and HippUnfold as suggested by our results, with a lower variance with respect to the DC, HD, and VS (Figure 4). One feature of interest is the ability of HSF to segment both T1w and T2w images. Our investigation yielded superior quality segmentations through the utilization of T2w images—a result that aligns with the existing literature. It is important to note, however, that our dataset contained a larger quantity of T2w images compared to T1w images. Therefore, we are unable to definitively conclude whether the observed disparities in quality are a direct result of superior T2w contrast or a potential bias within our dataset. However, because each tool was trained using data segmented with different protocols, it is difficult to compare their accuracy, especially regarding the boundary between CA1 and the Subiculum (Yushkevich et al., 2015a). As HSF learned from multiple datasets, we interpret its segmentation as following a consensus between multiple segmentation guidelines, even if our results show it is very close to Barron's protocol (Berron et al., 2017). All tools segment the head and the body of the hippocampus in a similar manner, except HIPS which after manual verification, did not seem to respect the hippocampal subfields' boundaries visible to the naked eye. HippUnfold underperforms compared to HSF and ASHS because it overrepresents CA2 and CA3 in the tail. The way HSF learned to segment the hippocampal tail (Figure 6) is very similar to the histology-based tail segmentation proposed by Dalton et al. (2017), Flores et al. (2020), which both differ from Barron's protocol. There is no histological ground truth to support the superiority of HSF over HippUnfold regarding tail segmentation. If HSF was to be proved wrong regarding this particular point, future investigators could easily add new deep learning models to HSF's Model Hub in a plug-and-play fashion. Ever since the most recent launch of FreeSurfer 7, the original authors (Iglesias et al., 2015) have been endeavoring to enhance their segmentation pipeline of the hippocampal subfields. Due to the fact that this updated version is still untested and limited, it has not been integrated into our benchmark because of the current limitation to low-resolution T1 images. Thus, we highly suggest that future studies thoroughly examine this novel update as soon as it exits the beta stage.

After validating HSF, we segmented and analyzed hippocampal ROIs obtained from the HCP-development, HCP-young adults, and HCP-aging datasets. This allowed us to study the developmental trajectories of hippocampal subfields during the lifespan with a bigger age range than previous studies [e.g., (Yang



et al., 2013; Bookheimer et al., 2019)]. Our model selection of NCS based on AIC found three main patterns. The first pattern, as expected, divided the hippocampus developmental trajectory into three main periods: growth, stabilization, and decay (GSD). This is the overall developmental pattern of the hippocampus, showing a maximal volume at approximately 20 to 25 years old, which is lower than some previous studies [e.g., (Yang et al., 2013)] but this may be due to the finer resolution of our model, thus

allowing the observation of three distinct trends. After the stable period, we found a significantly negative correlation between hippocampal volume and age from 70 years old onwards, which is approximately 8 years later than previously found (Ziegler et al., 2012; Yang et al., 2013; de Flores et al., 2015). As previously, this may be caused by modeling artifacts, survivor bias, or inclusion bias in the used datasets (inclusion of “super-healthy” individuals with better aging than the general population). This GSD trajectory

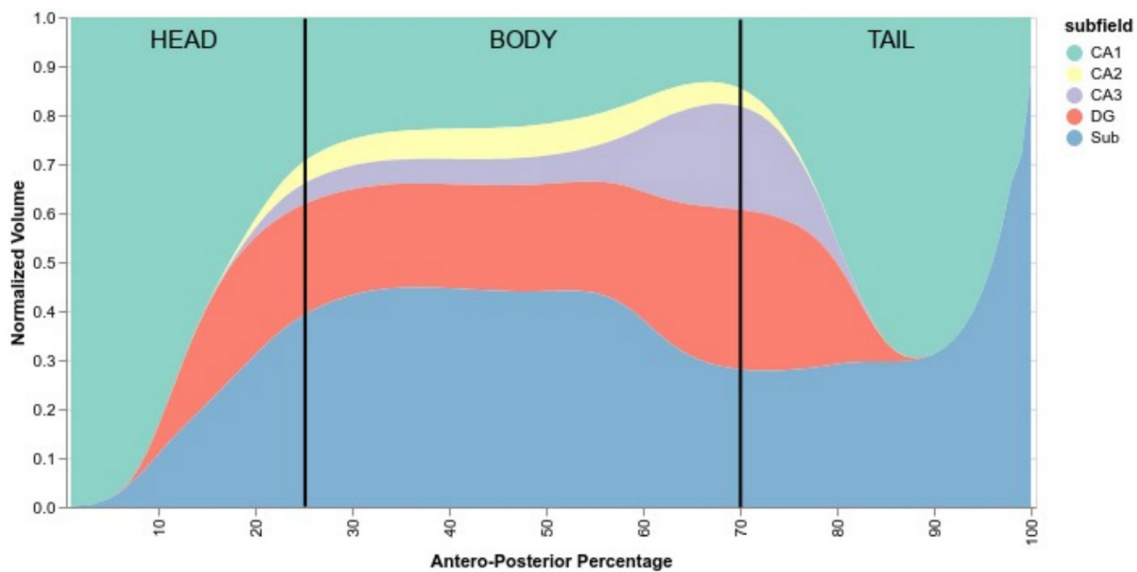


FIGURE 6

Normalized anteroposterior composition of subfields, going from 0% of the hippocampus (head) to 100% (tail). Vertical black lines are approximate delimiters of the head, body, and tail of the hippocampus.

was observed in DG and CA1, which is consistent with previous studies showing growth during infancy and childhood (Lavenex and Banta Lavenex, 2013; Lee et al., 2014; Ellis et al., 2021), [up to a 2-fold increase in size for DG (Bachevalier, 2013)]. Moreover, the inflection points of DG and CA1 were very similar to those of the total hippocampus (Figure 5). However, we observed different trajectories for CA2/3 and the subiculum. Although the literature suggested a volumetric increase of CA2/3 (Lavenex and Banta Lavenex, 2013; Lee et al., 2014), we found this structure to be the most stable across the lifespan with no clear trend. This may be due to an insufficient resolution, forcing us to merge CA2 and CA3, thus averaging their dynamics. Another possible factor might be a too-noisy segmentation because of partial volumes resulting in a lack of sensibility to detect fine changes in these small and complex regions. Finally, our results for the subiculum are consistent with the literature: mostly flat (i.e., absence of correlation of volume with age) or a slight quasi-linear negative correlation between age and volume (Ziegler et al., 2012; Lee et al., 2014; de Flores et al., 2015; Foster et al., 2019). Our bigger age range and finer model allow us to refine those characteristics: by examining our results, we found a plateau, no correlation between age and volume, until the age of 60~70 years after which a fast decay happens similar to other subfields. Overall, this suggests that the DG, followed by CA1, is the most affected by development and aging. Most of the development of the subiculum appears to happen before the age of 5, which would relate to mnemonic developments (Bouyeure and Noulhiane, 2021). While the subicular volume is positively correlated with the learnings of the when, where, and what components of episodic memory (Chi et al., 2022), prior studies found correlations between episodic memory and subiculum only up to 5 years old, which might be caused by the earlier maturation of the monosynaptic pathway (Canada, 2020). If the subiculum appears to mature earlier, it also decays earlier

than others, which suggests that it might be a relevant biomarker for the early identification of age-related cognitive impairments. Furthermore, given that our findings are largely consistent with prior research, this serves to strengthen the validity of HSE, our novel segmentation tool.

Finally, besides sexual dimorphism with men having, over the stable part of their life, bigger hippocampal subfields than women, we found differences in developmental trajectories of hippocampal subfields between men and women. These are debated in the literature since some studies did not find interactions between volume, sex, and age (Sullivan et al., 2005; Mueller et al., 2007), while others did (). The present study suggests a complex relationship since we did not find such an interaction for all subfields. We found significant differences only for the growing period of the DG and CA2/3 with a faster growth in men than in women. This may be due to gonadal hormones modulating neurogenesis and increasing adult-born cells' survival in the DG (Galea et al., 2006; Spritzer and Galea, 2007; Hamson et al., 2013). However, this literature suggests that this interaction also exists in CA1 (Leranth, 2004; Islam et al., 2020), which was not the case in our study. Interestingly, we also observed a stronger negative correlation between age and volume for the DG and CA1 in men than in women. Overall, our results add to the literature and reconcile previous results on the lifespan volumetric trajectories of hippocampal subfields.

Our study suffers from several limitations. First, the lack of a standardized protocol to segment the hippocampal subfields negatively affects the way algorithms will learn to segment. This is partly solved by learning from a consensus between guidelines, but we lack a better *in vivo* ground truth than the one provided by manual segmentations. Then, volume might not reflect all the age-related changes in hippocampal structures. Although we found no anteroposterior differences between subjects, we

believe it is critical to go beyond volumetric analysis and assess additional information, such as shape as suggested by Yang et al. (2013), Voineskos et al. (2015), and Lynch et al. (2019) or other complementary measures gathered through diffusion imaging, or even quantitative T1 relaxation maps, a proxy for intracortical myelin (Vos de Wael et al., 2018).

Therefore, while the hippocampal subfields are critical in the physiology of episodic memory, the lack of efficient segmentation tools hinders the use of large datasets to study their role in health and disease. Here, we introduced a new segmentation tool, HSE, robust to changes in populations, and acquisition parameters such as contrast, resolution, or magnetic field intensity. After its validation against other existing tools (ASHS, HIPS, and HippUnfold), we used it to segment large datasets (HCP-development, HCP-young adults, and HCP-aging) in order to model volumetric trajectories of the hippocampal subfields from 5 to 100 years old. Our volumetric analysis has shown that most subfields except the subiculum are positively correlated with age until the early 20s, and that the most correlated subfield is the dentate gyrus. This study also found a major inflection point at approximately 70 years old (even earlier in the subiculum) where a fast and significant volumetric decrease occurs. Our study has yet to be correlated with evaluations of mnemonic performances, which could help to validate subicular volumes as a relevant biomarker for the early diagnosis of age-related cognitive decline.

Data availability statement

The datasets with the exception of HCP datasets presented in this article are not readily available because participants provided consent only for using their data under the supervision of the principal investigator. Requests to access the datasets should be directed to MN.

Ethics statement

The studies involving human participants were reviewed and approved by CPP 2011-A00058-33. Written informed consent to

participate in this study was provided by the participants' legal guardian/next of kin.

Author contributions

CP: conceptualization, methodology, software, formal analysis, data curation, writing—original draft, and visualization. AB: investigation, resources, data curation, writing—review and editing, and funding acquisition. SP and MF: data curation and writing—review and editing. AG: methodology. ED: methodology and writing—review and editing. MB: resources, data curation, and writing—review and editing. FL: writing—review and editing. MN: conceptualization, investigation, resources, writing—original draft, supervision, project administration, and funding acquisition. All authors contributed to the article and approved the submitted version.

Funding

This study was funded by Fondation de France, Grant/Award Number: 00070721.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

- Bachevalier, J. (2013). The development of memory from a neurocognitive and comparative perspective. In: Bauer PJ, Fivush R, editors. *The Wiley Handbook on the Development of Children's Memory*. Chichester: John Wiley and Sons Ltd (2013). p. 109–25.
- Berron, D., Vieweg, P., Hochkepler, A., Pluta, J. B., Ding, S. L., Maass, A., et al. (2017). A protocol for manual segmentation of medial temporal lobe subregions in 7 Tesla MRI. *NeuroImage Clin.* 15, 466–482. doi: 10.1016/j.nicl.2017.05.022
- Bookheimer, S. Y., Salat, D. H., Terpstra, M., Ances, B. M., Barch, D. M., Buckner, R. L., et al. (2019). The lifespan human connectome project in aging: an overview. *NeuroImage* 185, 335–348. doi: 10.1016/j.neuroimage.2018.10.009
- Bouyeure, A., and Noulhiane, M. (2021). Episodic memory development in normal and adverse environments. In: *Factors Affecting Neurodevelopment*. Elsevier. p. 517–27. Available online at: <https://linkinghub.elsevier.com/retrieve/pii/B9780128179864000444> (accessed October 24, 2022).
- Bouyeure, A., Patil, S., Mauconduit, F., Poiret, C., Isai, D., Noulhiane, M., et al. (2021). Hippocampal subfield volumes and memory discrimination in the developing brain. *Hippocampus* 31, 1202–1214. doi: 10.1002/hipo.23385
- Bussy, A., Plitman, E., Patel, R., Tullo, S., Salaciak, A., Bedford, S. A., et al. (2021). Hippocampal subfield volumes across the healthy lifespan and the effects of MR sequence on estimates. *NeuroImage*. 233, 117931. doi: 10.1016/j.neuroimage.2021.117931
- Canada, K. L. (2020). *Examining the Co-Development of Episodic Memory and Hippocampal Subfields – A Longitudinal Study*. Digital Repository at the University of Maryland. Available online at: <http://drum.lib.umd.edu/handle/1903/26350> (accessed August 31, 2022).
- Chi, C. H., Yang, F. C., and Chang, Y. L. (2022). Age-related volumetric alterations in hippocampal subiculum region are associated with reduced retention of the “when” memory component. *Brain Cogn.* 160, 105877. doi: 10.1016/j.bandc.2022.105877
- Dalton, M. A., Zeidman, P., Barry, D. N., Williams, E., and Maguire, E. A. (2017). Segmenting subregions of the human hippocampus on structural magnetic resonance image scans: an illustrated tutorial. *Brain Neurosci. Adv.* 1:239821281770144. doi: 10.1177/2398212817701448

- de Flores, R., La Joie, R., and Chételat, G. (2015). Structural imaging of hippocampal subfields in healthy aging and Alzheimer's disease. *Neuroscience* 309, 29–50. doi: 10.1016/j.neuroscience.2015.08.033
- DeKraker, J., Haast, R. A., Yousif, M. D., Karat, B., Köhler, S., Khan, A. R., et al. (2021). *HippUnfold: Automated Hippocampal Unfolding, Morphometry, and Subfield Segmentation*. Neuroscience. Available online at: <http://biorxiv.org/lookup/doi/10.1101/12.03.471134> (accessed July 19, 2022).
- DeKraker, J., Haast, R. A., Yousif, M. D., Karat, B., Lau, J. C., Köhler, S., et al. (2022). Automated hippocampal unfolding for morphometry and subfield segmentation with HippUnfold. *Elife*. 11, e77945. doi: 10.7554/eLife.77945
- Elhakeem, A., Hughes, R. A., Tilling, K., Cousminer, D. L., Jackowski, S. A., Cole, T. J., et al. (2022). Using linear and natural cubic splines, SITAR, and latent trajectory models to characterise nonlinear longitudinal growth trajectories in cohort studies. *BMC Med. Res. Methodol.* 22, 68. doi: 10.1186/s12874-022-01542-8
- Ellis, C. T., Skalaban, L. J., Yates, T. S., Bejjanki, V. R., Córdova, N. I., Turk-Browne, N. B., et al. (2021). Evidence of hippocampal learning in human infants. *Curr. Biol.* 31, 3358–3364.e4. doi: 10.1016/j.cub.2021.04.072
- Flores, R., Berron, D., Ding, S., Ittyerah, R., Pluta, J. B., Xie, L., et al. (2020). Characterization of hippocampal subfields using ex vivo MRI and histology data: Lessons for in vivo segmentation. *Hippocampus*. 30, 545–564. doi: 10.1002/hipo.23172
- Fonov, V., Evans, A., McKinstry, R., Almlí, C., and Collins, D. (2009). Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *Neuroimage* 47, S102. doi: 10.1016/S1053-8119(09)70884-5
- Foster, C. M., Kennedy, K. M., Hoagey, D. A., and Rodrigue, K. M. (2019). The role of hippocampal subfield volume and fornix microstructure in episodic memory across the lifespan. *Hippocampus*. 29, 1206–1223. doi: 10.1002/hipo.23133
- Galea, L. A. M., Spritzer, M. D., Barker, J. M., and Pawluski, J. L. (2006). Gonadal hormone modulation of hippocampal neurogenesis in the adult. *Hippocampus* 16, 225–232. doi: 10.1002/hipo.20154
- Gogtay, N., Nugent, T. F., Herman, D. H., Odonez, A., Greenstein, D., Hayashi, K. M., et al. (2006). Dynamic mapping of normal human hippocampal development. *Hippocampus* 16, 664–672. doi: 10.1002/hipo.20193
- Greenland, S. (1995). Avoiding power loss associated with categorization and ordinal scores in dose-response and trend analysis. *Epidemiology*. 6, 450–454. doi: 10.1097/00001648-199507000-00025
- Haeger, A., Mangin, J. F., Vignaud, A., Poupon, C., Grigis, A., Boumezeur, F., et al. (2020). Imaging the aging brain: study design and baseline findings of the SENIOR cohort. *Alz Res Therapy*. 12, 77. doi: 10.1186/s13195-020-00642-1
- Hamson, D. K., Wainwright, S. R., Taylor, J. R., Jones, B. A., Watson, N. V., Galea, L. A. M., et al. (2013). Androgens increase survival of adult-born neurons in the dentate gyrus by an androgen receptor-dependent mechanism in male rats. *Endocrinology* 154, 3294–3304. doi: 10.1210/en.2013-1129
- Hindy, N. C., Ng, F. Y., and Turk-Browne, N. B. (2016). Linking pattern completion in the hippocampus to predictive coding in visual cortex. *Nat. Neurosci.* 19, 665–667. doi: 10.1038/nn.4284
- Iglesias, J. E., Augustinack, J. C., Nguyen, K., Player, C. M., Player, A., Wright, M., et al. (2015). A computational atlas of the hippocampal formation using ex vivo, ultra-high resolution MRI: Application to adaptive segmentation of in vivo MRI. *Neuroimage*. 115, 117–137. doi: 10.1016/j.neuroimage.2015.04.042
- Islam, M. N., Sakimoto, Y., Jahan, M. R., Ishida, M., Tarif, A. M. M., Nozaki, K., et al. (2020). Androgen affects the dynamics of intrinsic plasticity of pyramidal neurons in the CA1 hippocampal subfield in adolescent male rats. *Neuroscience*. 440, 15–29. doi: 10.1016/j.neuroscience.2020.05.025
- Kulaga-Yoskovitz, J., Bernhardt, B. C., Hong, S. J., Mansi, T., Liang, K. E., van der Kouwe, A. J. W., et al. (2015). Multi-contrast submillimetric 3 Tesla hippocampal subfield segmentation protocol and dataset. *Sci Data*. 2, 150059. doi: 10.1038/sdata.2015.59
- Lagarde, J., Olivieri, P., Tonietto, M., Gervais, P., Comtat, C., Caillé, F., et al. (2021). Distinct amyloid and tau PET signatures are associated with diverging clinical and imaging trajectories in patients with amnesic syndrome of the hippocampal type. *Transl Psychiatry*. 11, 498. doi: 10.1038/s41398-021-01628-9
- Lavenex, P., and Banta Lavenex, P. (2013). Building hippocampal circuits to learn and remember: Insights into the development of human memory. *Behav Brain Res*. 254, 8–21. doi: 10.1016/j.bbr.2013.02.007
- Lee, J. K., Ekstrom, A. D., and Gheiti, S. (2014). Volume of hippocampal subfields and episodic memory in childhood and adolescence. *Neuroimage* 94, 162–171. doi: 10.1016/j.neuroimage.2014.03.019
- Leranth, C. (2004). Androgens increase spine synapse density in the CA1 hippocampal subfield of ovariectomized female rats. *J. Neurosci.* 24, 495–499. doi: 10.1523/JNEUROSCI.4516-03.2004
- Luo, P., Ren, J., Peng, Z., Zhang, R., and Li, J. (2019). *Differentiable Learning-to-Normalize via Switchable Normalization*. arXiv: 1806.10779. Available online at: <http://arxiv.org/abs/1806.10779> (accessed May 31, 2023).
- Lynch, K. M., Shi, Y., Toga, A. W., and Clark, K. A. (2019). Hippocampal shape maturation in childhood and adolescence. *Cereb. Cortex*. 29, 3651–3665. doi: 10.1093/cercor/bhy244
- Manera, A. L., Dadar, M., Fonov, V., and Collins, D. L. (2020). CerebRA, registration and manual label correction of Mindboggle-101 atlas for MNI-ICBM152 template. *Sci Data*. 7, 237. doi: 10.1038/s41597-020-0557-9
- Mueller, S. G., Stables, L., Du, A. T., Schuff, N., Truran, D., Cashdollar, N., et al. (2007). Measurement of hippocampal subfields and age-related changes with high resolution MRI at 4T. *Neurobiol Aging* 28, 719–726. doi: 10.1016/j.neurobiolaging.2006.03.007
- Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., et al. (2018). Attention U-Net: Learning Where to Look for the Pancreas. arXiv:1804.03999. Available online at: <http://arxiv.org/abs/1804.03999> (accessed May 31, 2023).
- O'Mahony, N., Campbell, S., Carvalho, A., Harapanahalli, S., Hernandez, G. V., Krpalkova, L., et al. (2020). Deep learning vs. traditional computer vision. In: Arai, K., Kapoor, S., editors. *Advances in Computer Vision*. Cham: Springer International Publishing, p. 128–44.
- Opitz, D., and Maclin, R. (1999). Popular ensemble methods: an empirical study. *Jair*. 11, 169–198. doi: 10.1613/jair.614
- Palombo, D. J., Bacopulos, A., Amaral, R. S. C., Olsen, R. K., Todd, R. M., Anderson, A. K., et al. (2018). Episodic autobiographical memory is associated with variation in the size of hippocampal subregions. *Hippocampus*. 28, 69–75. doi: 10.1002/hipo.22818
- Qiu, Q., Gong, G., Wang, L., Duan, J., and Yin, Y. (2019). Feasibility of Automatic Segmentation of Hippocampus Based on Deep Learning in Hippocampus-Sparing Radiotherapy. *Int. J. Radiat. Oncol. Biol. Phys.* 105, E137–E138. doi: 10.1016/j.ijrobp.2019.06.2177
- Romero, J. E., Coupé, P., and Manjón, J. V. (2017). HIPS: A new hippocampus subfield segmentation method. *Neuroimage*. 163, 286–295. doi: 10.1016/j.neuroimage.2017.09.049
- Satopaa, V., Albrecht, J., Irwin, D., and Raghavan, B. (2011). Finding a “Kneedle” in a haystack: detecting knee points in system behavior. In: 2011 31st International Conference on Distributed Computing Systems Workshops. Minneapolis, MN: IEEE (2011). p. 166–71.
- Schmidt, M. F., Storrs, J. M., Freeman, K. B., Jack, C. R., Turner, S. T., Griswold, M. E., et al. (2018). A comparison of manual tracing and FreeSurfer for estimating hippocampal volume over the adult lifespan. *Hum Brain Mapp*. 39, 2500–2513. doi: 10.1002/hbm.24017
- Shaw, T. B., York, A., Barth, M., and Bollmann, S. (2020). Towards optimising MRI characterisation of tissue (TOMCAT) dataset including all longitudinal automatic segmentation of hippocampal subfields (LASHiS) data. *Data Brief* 32, 106043. doi: 10.1016/j.dib.2020.106043
- Somerville, L. H., Bookheimer, S. Y., Buckner, R. L., Burgess, G. C., Curtiss, S. W., Dapretto, M., et al. (2018). The lifespan human connectome project in development: a large-scale study of brain connectivity development in 5–21 year olds. *Neuroimage*. 183, 456–468. doi: 10.1016/j.neuroimage.2018.08.050
- Spritzer, M. D., and Galea, L. A. M. (2007). Testosterone and dihydrotestosterone, but not estradiol, enhance survival of new hippocampal neurons in adult male rats. *Devel Neurobiol*. 67, 1321–1333. doi: 10.1002/dneu.20457
- Sullivan, E. V., Marsh, L., and Pfefferbaum, A. (2005). Preservation of hippocampal volume throughout adulthood in healthy men and women. *Neurobiol Aging*. 26, 1093–1098. doi: 10.1016/j.neurobiolaging.2004.09.015
- Suzuki, M. (2004). Male-specific volume expansion of the human hippocampus during adolescence. *Cereb Cortex*. 15, 187–193. doi: 10.1093/cercor/bhh121
- Uematsu, A., Matsui, M., Tanaka, C., Takahashi, T., Noguchi, K., Suzuki, M., et al. (2012). Developmental trajectories of amygdala and hippocampus from infancy to early adulthood in healthy individuals. Krueger F, editor. *PLoS ONE* 7, e46970. doi: 10.1371/journal.pone.0046970
- Voineskos, A. N., Winterburn, J. L., Felsky, D., Pipitone, J., Rajji, T. K., Mulsant, B. H., et al. (2015). Hippocampal (subfield) volume and shape in relation to cognitive performance across the adult lifespan: hippocampal volume, shape, and age-related cognitive performance. *Hum. Brain Mapp*. 36, 3020–3037. doi: 10.1002/hbm.22825
- Vos de Wael, R., Larivière, S., Caldaïrou, B., Hong, S. J., Margulies, D. S., Jefferies, E., et al. (2018). Anatomical and microstructural determinants of hippocampal subfield functional connectome embedding. *Proc. Natl. Acad. Sci. USA*. 115, 10154–10159. doi: 10.1073/pnas.1803667115
- Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T., et al. (2019). Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*. 338, 34–45. doi: 10.1016/j.neucom.2019.01.103
- Winterburn, J. L., Pruessner, J. C., Chavez, S., Schira, M. M., Lobaugh, N. J., Voineskos, A. N., et al. (2013). A novel in vivo atlas of human hippocampal subfields using high-resolution 3T magnetic resonance imaging. *Neuroimage*. 74, 254–265. doi: 10.1016/j.neuroimage.2013.02.003
- Wisse, L. E. M., Biessels, G. J., and Geerlings, M. I. A. (2014). Critical appraisal of the hippocampal subfield segmentation package in freesurfer. *Front Aging Neurosci*. 6, 261. doi: 10.3389/fnagi.2014.00261

- Wisse, L. E. M., Daugherty, A. M., Olsen, R. K., Berron, D., Carr, V. A., Stark, C. E. L., et al. (2017). A harmonized segmentation protocol for hippocampal and parahippocampal subregions: Why do we need one and what are the key goals?: A harmonized hippocampal subfield protocol: key goals and impact. *Hippocampus*. 27, 3–11. doi: 10.1002/hipo.22671
- Wisse, L. E. M., Kuijf, H. J., Honingh, A. M., Wang, H., Pluta, J. B., Das, S. R., et al. (2016). Automated Hippocampal Subfield Segmentation at 7T MRI. *AJNR. Am. J. Neuroradiol.* 37, 1050–1057. doi: 10.3174/ajnr.A4659
- Yang, X., Goh, A., Chen, S. H. A., and Qiu, A. (2013). Evolution of hippocampal shapes across the human lifespan: hippocampal shapes in aging. *Hum. Brain Mapp.* 34, 3075–3085. doi: 10.1002/hbm.22125
- Yang, Z., Zhuang, X., Mishra, V., Sreenivasan, K., and Cordes, D. (2020). CAST. A multi-scale convolutional neural network based automated hippocampal subfield segmentation toolbox. *Neuroimage* 218, 116947. doi: 10.1016/j.neuroimage.2020.116947
- Yassa, M. A., and Stark, C. E. L. (2011). Pattern separation in the hippocampus. *Trends Neurosci.* 34, 515–525. doi: 10.1016/j.tins.2011.06.006
- Yushkevich, P. A., Amaral, R. S. C., Augustinack, J. C., Bender, A. R., Bernstein, J. D., Boccardi, M., et al. (2015a). Quantitative comparison of 21 protocols for labeling hippocampal subfields and parahippocampal subregions in in vivo MRI: towards a harmonized segmentation protocol. *Neuroimage*. 111, 526–541. doi: 10.1016/j.neuroimage.2015.01.004
- Yushkevich, P. A., Pluta, J. B., Wang, H., Xie, L., Ding, S. L., Gertje, E. C., et al. (2015b). Automated volumetry and regional thickness analysis of hippocampal subfields and medial temporal cortical structures in mild cognitive impairment: automatic morphometry of MTL subfields in MCI. *Hum. Brain Mapp.* 36, 258–287. doi: 10.1002/hbm.22627
- Yushkevich, P. A., Wang, H., Pluta, J., Das, S. R., Craige, C., Avants, B. B., et al. (2010). Nearly automatic segmentation of hippocampal subfields in in vivo focal T2-weighted MRI. *Neuroimage*. 53, 1208–1224. doi: 10.1016/j.neuroimage.2010.06.040
- Zhang, Z., Liu, Q., and Wang, Y. (2018). Road extraction by deep residual U-net. *IEEE Geosci Remote Sensing Lett.* 15, 749–753. doi: 10.1109/LGRS.2018.2802944
- Zhu, H., Shi, F., Wang, L., Hung, S. C., Chen, M. H., Wang, S., et al. (2019). Dilated dense U-Net for infant hippocampus subfield segmentation. *Front Neuroinform.* 13, 30. doi: 10.3389/fninf.2019.00030
- Ziegler, G., Dahnke, R., Jäncke, L., Yotter, R. A., May, A., Gaser, C., et al. (2012). Brain structural trajectories over the adult lifespan. *Hum Brain Mapp.* 33, 2377–2389. doi: 10.1002/hbm.21374

3.1.3 Additional Discussion

Despite the difficulties posed by Capsule Networks, we were able to develop and release HSF using more traditional architectures. While CapsNets were promising for their built-in robustness, their lack of scalability ultimately made them impractical. Improvements to the routing algorithm and capsule design may make them more usable in the future. It should be noted that GLOM (Hinton, 2021) has been introduced to solve many of the issues introduced by the inner workings of CapsNets. GLOM is better than traditional capsules because it avoids pre-allocating neurons to specific object or part types, instead using universal capsules with distributed representations to allow for more sharing of knowledge between similar parts. GLOM also forms islands of identical vectors through an iterative process to represent parse trees rather than relying on dynamic routing between capsules. However, we also encourage the exploration of other works on equivariances such as Gauge Equivariant Neural Networks (Cohen et al., 2019), or more recent work based on spherical harmonics (Diaz et al., 2023).

Concerning HSF, we would like to emphasize a fact that is only briefly mentioned in the original paper. We used human feedback to improve the tool, — manually correcting poor segmentations to focus training on difficult cases. This made HSF robust even to atypical morphologies (e.g., Figure 3.14). We want to highlight the modular "model hub" design, which also enables users to easily incorporate new models for their needs, meaning that anyone can train his own model and use it without any coding skills. If our first principles were shown wrong, anyone can segment new subjects according to new manual guidelines, include more subjects with a specific health condition, or propose a new architecture. New models can be added as in the code example 3.1, which can be made available to all users through a pull request on <https://github.com/clementpoiret/hsf>. Although it is outside the scope of this dissertation, since ROIloc is already working with any brain region, it is worth noting that HSF could easily be ported to work with any brain region.

Listing 3.1 Example of configuration to include any new model into HSF.

```
ca_mode: "1/2/3"
models_path: "~/hsf/models/single/"
models:
  arunet_bag_0.onnx:
    url: "https://zenodo.org/record/6457484/files/arunet_3.0.0_single.onnx?download=1"
    xxh3_64: "71edec9011f7f304"

segmentation:
  test_time_augmentation: True
  test_time_num_aug: 20
```

Using HSF, we segmented more than 3,700 MRI scans from 4 to 100+ years of age. Analysis revealed distinct volumetric trajectories for each subfield across the lifespan. This provides new insight into the maturation and decay of the underlying medial temporal circuits of memory. In general, HSF establishes that efficient deep learning tools can be applied to segment the hippocampal subfields at scale. Accurate segmentation will enable more powerful structure-function studies that relate subfields anatomy to cognition using functional MRI. The ability to quickly and reliably delineate small substructures like hippocampal subfields promises to shed further light on the role of medial temporal lobe circuits in learning, memory, and disease.

3.1 Segmenting the Hippocampal Subfields: Anatomical Trajectories

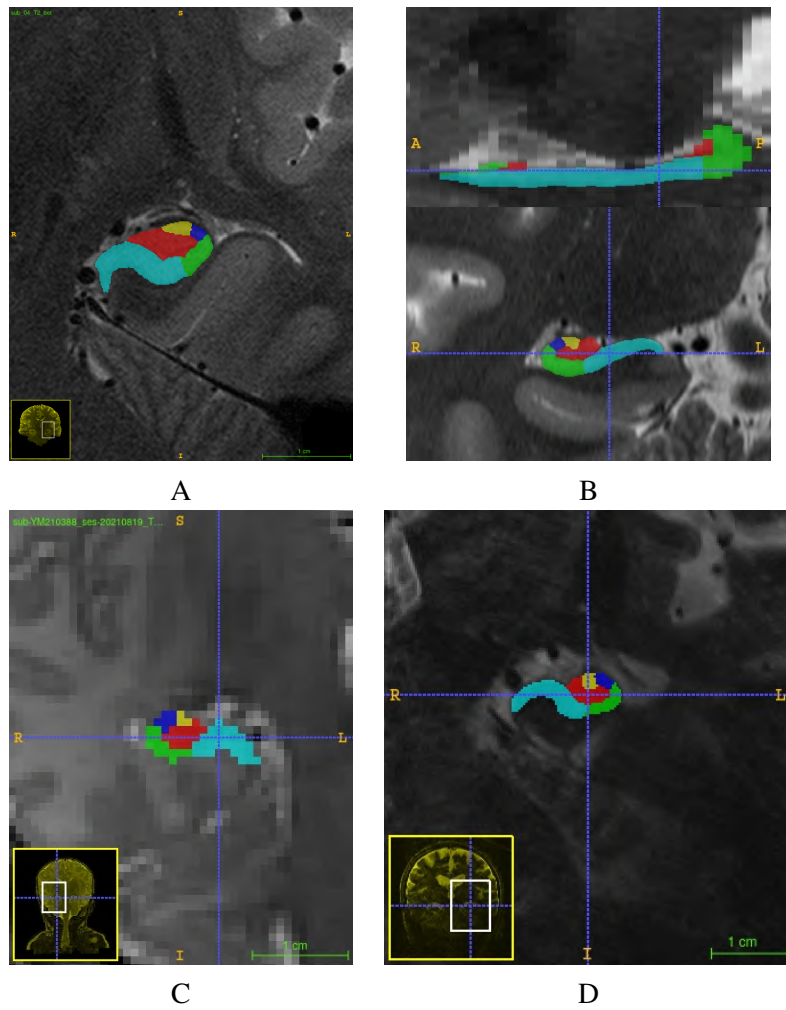


Figure 3.14: HSF results on subjects qualified as “out-of-distribution”. All examples were not included in the training set. A comes from HIPlay7 and show an IHI; B also comes from HIPlay7 and does not have a hippocampal head; C is a preterm infant at the age equivalent to the term of pregnancy from the Robert-Debré hospital database (Elbaz et al., 2023); and D is a subject 80 years of age with heavy motion artifacts from Lagarde et al., 2021.

3.2 Exploring the Functional Maturation of the Hippocampal Subfields

Functional connectivity measured with resting-state fMRI provides insight into the intrinsic networks and interactions between brain regions. Examining connectivity patterns of the hippocampal subfields is of great interest, as this could illuminate their differential involvement in memory, spatial navigation, and neurological disorders. However, accurately mapping the connectivity of the subfields presents significant challenges. The coarse spatial resolution of fMRI combined with the small size and anatomical complexity of subregions like the dentate gyrus easily allow signals to be misattributed. Therefore, a precise delineation of the subfields boundaries is critical to avoid erroneous localization of fMRI timeseries. Our newly developed HSF tool provides accurate automated segmentation of the anatomy of the subfields. Leveraging HSF ensures proper mapping of the subfields as a prerequisite for valid functional connectivity analysis.

Still, even with precise anatomical maps, characterizing subfields functional networks faces difficulties. Coupled to the fact that feature importances derived from machine learning models are not always reliable (Kumar et al., 2020), the multivariate nature of resting-state fMRI connectivity data makes modeling ambiguous, giving rise to the “Rashōmon effect” (Breiman, 2001) where different analytical approaches can produce divergent interpretations of the same data (third green section of Figure 3.15). This is particularly problematic in examining how the connectivity of the hippocampal subfields changes over development, as spurious associations could be inferred. Therefore, we implemented a rigorous methodology that integrates multiple computational models within an explainable AI framework. Specifically, we trained an ensemble of machine learning models relating the connectivity of the subfields to age in a large sample of youths aged 5-21 years from the HCP-Development dataset. Then, we derived consensus between these models to obtain reliable developmental trajectories, validated through out-of-sample testing (fourth green section of Figure 3.15).

This analysis focused on early life connectivity because the hippocampal subfields are known to have prolonged maturation supporting gains in memory (Bouyeure & Noulhiane, 2020; Bouyeure et al., 2021; Poiret, Bouyeure, et al., 2023). Elucidating the dynamics of subfields’ networks during this critical period of development provides new insights into the anatomofunctional properties underpinnings of memory evolution. Overall, our work underscores the need for multimodel explainable AI to handle ambiguities in mapping the functional connectivity changes of the subfields over the lifespan. The framework established here serves as a template for rigorously probing the neural substrates of cognition and behavior through the integration of precise anatomy and robust computational modeling.

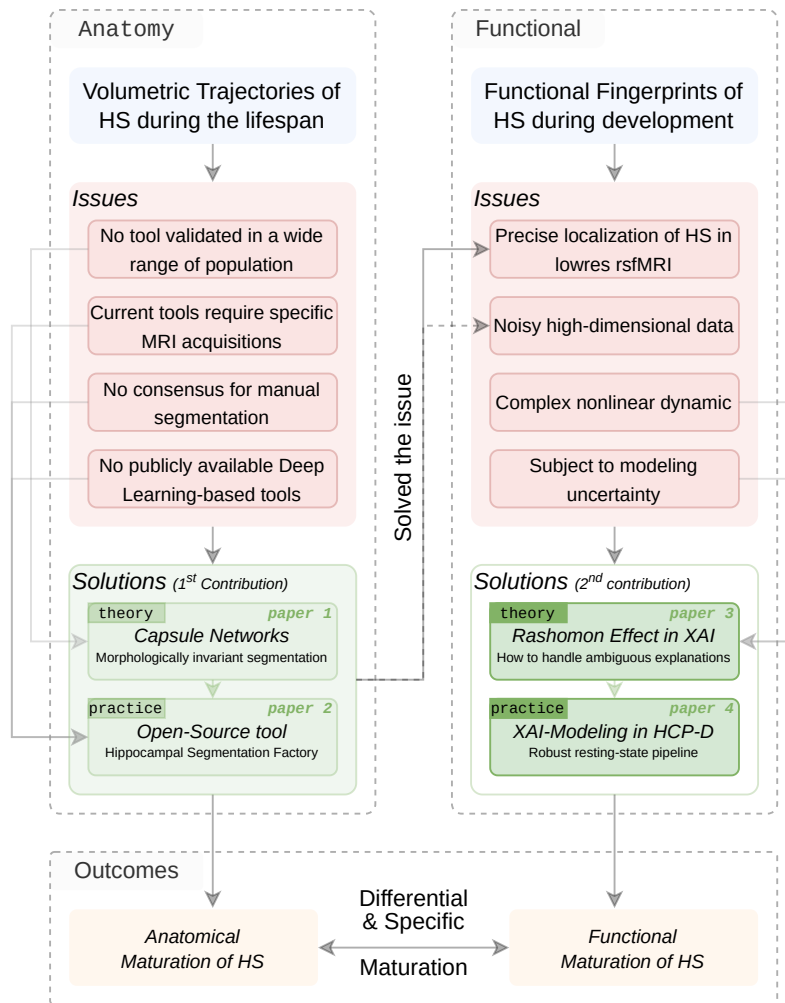


Figure 3.15: Achitecture of the dissertation: Functional Connectivity. This chapter aimed at characterizing functional changes in hippocampal subfields with a specific emphase in childhood and adolescence. The chapter is organized into two main parts with its accompanying paper: (1) the application of this model to explore volumetric developmental trajectories in a large lifespan dataset; (2) the investigation of the maturation of functional connectivity of the subfields using resting-state fMRI and explainable AI modeling. Together, these novel tools and analytical approaches provide insights into the prolonged development of hippocampal circuits supporting gains in memory and cognition.

3.2.1 Can we Agree? On the Rashōmon Effect and the Reliability of Post-Hoc Explainable AI

Clement Poiret, Antoine Grigis, Justin Thomas, and Marion Noulhiane (2023). Can we Agree? On the Rashōmon Effect and the Reliability of Post-Hoc Explainable AI. *arXiv:2308.07247 [cs, stat]*.

As stated earlier, we need a framework for thinking about and handling modeling uncertainties. The framework established in the following paper provides a template to find consensus between competing explanations to obtain reliable information based on SHAP (SHapley Additive exPlanations) values (Lundberg & Lee, 2017) as a post-hoc explanation method. Therefore, we proposed a new methodology that uses traditional machine learning models. Because this issue is transversal to any domain, we used 5 datasets from multiple domains to ensure the general characteristic of our approach. We focused on machine learning and not deep learning, because, although deep learning shows promise for neuroimaging, tabular data is still better handled by gradient boosting and tree-based methods (Grinsztajn et al., 2022).

Abstract

Goal: To examine the influence of sample size on the reliability and agreement of explanations from machine learning models that exhibit the Rashōmon effect. To test our hypothesis, we introduced a new metric, called weighted cosine similarity. We also proposed a consensus method based on mean absolute SHAP values across multiple high-performing models.

Material and Methods: Experiments were carried out using SHAP values as explanations on 5 classification datasets. Multiple model types (linear, tree, ensemble) were trained on varying subsample sizes from 16 to full data. 10-fold cross-validation assessed intramodel agreement through weighted cosine similarity between SHAP values. Intermodel agreement towards consensus was measured by similarity to mean SHAP values of all models on full data.

Results: Weighted cosine similarity increased significantly with sample size for 4 of 5 datasets, indicating convergence of explanations. Explanations from <128 samples showed high variability in cross-validation, limiting reliability. Bagging ensembles often had higher agreement versus individual models.

Conclusions: A sample size of at least 128 is probably required to trust the explanations of the models in a Rashōmon set. Explanation variability at low samples means conclusions may be unreliable without further validation. Although bagging can improve agreement, individual models should also be tested. The approaches explored here demonstrate ways to elicit knowledge from ambiguous machine learning models.

Can we Agree? ON THE RASHŌMON EFFECT AND THE RELIABILITY OF POST-HOC EXPLAINABLE AI

A PREPRINT

Clément Poiret ^{1,2,3,4}, Antoine Grigis ³, Justin Thomas², and Marion Noulhiane ^{1,3,4*}

¹UNIACT, NeuroSpin, CEA Paris-Saclay, Frederic Joliot Institute, Gif-sur-Yvette, France

²Department of Research and Development, Caminov, Paris, France

³NeuroSpin, CEA Paris-Saclay, Frederic Joliot Institute, Gif-sur-Yvette, France

⁴InDEV, NeuroDiderot, Université Paris Cité, Inserm, Paris, France

October 1, 2023

ABSTRACT

The Rashōmon effect — the fact that equally performing models may have different underlying assumptions — poses challenges for deriving reliable knowledge from machine learning models. This study examined the influence of sample size on explanations from models in a Rashōmon set using SHAP. Experiments on 5 public datasets showed that explanations gradually converged as the sample size increased. Explanations from less than a hundred samples exhibited high variability, limiting reliable knowledge extraction. However, agreement between models improved with more data, allowing for consensus. Bagging ensembles often had higher agreement. The results provide guidance on sufficient data to trust explanations. Variability at low samples suggests that conclusions may be unreliable without validation. Further work is needed with more model types, data domains, and explanation methods. Testing convergence in neural networks and with model-specific explanation methods would be impactful. The approaches explored here point towards principled techniques for eliciting knowledge from ambiguous models.

Keywords Machine Learning · Interpretability · Explanations Robustness · Sample size · Guidelines

1 Introduction

In recent years, the widespread integration of AI-powered systems in various domains has highlighted the need for increased transparency and trust in these complex models. Explainable AI (XAI) methodologies have emerged as a means to address this need by providing insights into model training and clarifying predictions through post hoc explanations. As these systems become more integrated into our daily lives and are employed in critical domains such as healthcare (e.g. [1, 2]) and cybersecurity (e.g. [3]), it is crucial to understand the basis of their decision-making processes. By unraveling the inner workings of these complex models, we can ascertain whether their conclusions are based on genuine features or merely rely on spurious correlations and shortcuts. In addition, ethical considerations play a role in the adoption and deployment of these AI systems. As they become increasingly sophisticated, it is essential to ensure that their decision-making processes are in line with ethical guidelines and do not follow biases or discriminate against certain individuals

or communities. Explainable AI methodologies, such as the post hoc explanations provided by XAI techniques, offer a means of scrutinizing and mitigate potential biases and ethical concerns. By shedding light on the decision-making process and revealing the factors that contribute to predictions, XAI not only provides accountability, but also empowers stakeholders to address and rectify any biases or ethical issues that may arise. Beyond ethical concerns, if these models outperform human experts on key metrics, illuminating the black box could lead to new findings that might enhance expertise, expand knowledge, or spark new research ideas.

While there is an active research field dedicated to building explainable models out-of-the-box, those models are not yet widely available or as effective as state-of-the-art black-box models in some domains. Therefore, this current work will focus solely on post-hoc explanations: methods that will link a model’s prediction to contextual information. Unfortunately, multiple models can perform very well on a given task, but there is no guarantee on their use of similar

features to construct their prediction. Multiple models can base their predictions on multiple set of different features, an effect named “the Rashōmon Effect”

1.1 Explainable AI: a Practical Overview

Although interpretability and explainability may be used interchangeably, interpretability refers to the ability to understand the internal workings and mechanisms of a model, whereas explainability, on which this work will focus, refers to the capacity to provide explanations for the model’s predictions or decisions in a way that can be understood by humans. However, as Del Giudice et al., (2022) previously described [4], there exists a Prediction-Explanation Fallacy. It arises when one employs prediction-optimized models for explanatory purposes, without taking into account the delicate balance between explanation and prediction. On the one hand, of the spectrum lie interpretable models, which are either unrealistically simplistic or complex to deploy in real-world situations. On the other hand, the models employ excessively complex structures that are almost impossible to interpret (Figure 1). Thus, if explainability can come from models with built-in explanation mechanisms (e.g. [5]), the easiest approach comes from post hoc explanation methods, as they allow the explanation of any given model, for any given task. Finally, the objectives of explainability are (i) to provide information on the training or generalization of a specific model, or (ii) to clarify its predictions by explaining them in terms of the input of the model [6]. By addressing the question of “*why did this model make this particular prediction?*”, XAI generates a set of features that highlight distinct patterns present in a model’s internal representation of a phenomenon.

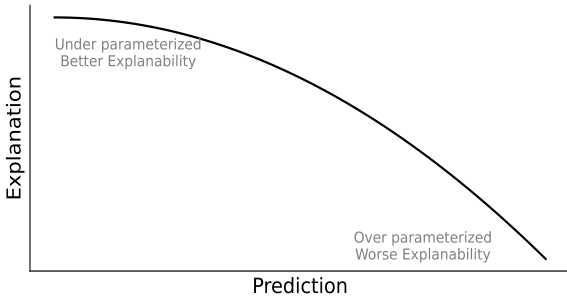


Figure 1: Usual Perception of the Prediction-Explanation trade-off. The prediction-explanation trade-off illustrates the difficulty behind XAI. The more complex a model is, the more predictive power it has, but its use of highly complex structures render its interpretability nearly intractable.

Post-hoc explainability encompasses two main approaches: model-specific and model-agnostic methods. Model-specific methods are tailored to a specific type of model and provide insight into its internal workings. For example, attention maps (e.g. [7]) highlight the important regions in an image that contribute to the model’s prediction, while GradCAM (e.g. [8]) visualizes the regions that are crucial

for a specific class. On the other hand, model-agnostic methods are applicable to any model and focus on explaining predictions without accessing internal information such as weights or architecture [9]. SHAP (for SHapley Additive exPlanations) [10] - the most popular method of this kind - uses Shapley values to attribute the importance of each input feature in making the final prediction. For example, in an image classification task, SHAP can identify the pixels that contribute the most to the predicted class. Shapley values are a way to fairly distribute the “credit” for a particular outcome among the different factors that contributed to it. In the context of Machine and Deep Learning, “credit” refers to the importance of each input feature (e.g. pixels in an image) in making the final prediction. SHAP computes the Shapley value for each feature by considering all possible combinations of features and measuring the change in prediction.

In addition to the increase in trust and transparency, explaining a model gives the possibility of detecting inconsistencies in its modeling quality. Some models may demonstrate equivalent predictive accuracy, but their biases may be distinct, implying the “Rashōmon effect”, — i.e. conflicting interpretations of the underlying phenomenon [11].

1.2 On the Rashōmon Effect

The tension between prediction and explanation is correlated to the Rashōmon Effect. According to Anderson (2016), the Rashōmon Effect “is the naming of an epistemological framework—or ways of thinking, knowing, and remembering—required for understanding the complex and ambiguous on both the small and large scale” [12]. It originates from a Japanese film of the same name, by Akira Kurosawa, which portrays a crime from four different perspectives, each with a different interpretation of the event caused by the subjectivity of human perception and memory. Thus, Breiman (2001) imported the concept into statistics and Machine Learning, where the concept refers to the phenomenon in which different near-optimal models trained on the same task may actually base their prediction on different sets of features [11]. Different models that perform similarly may have different underlying structures and assumptions.

Those models fall into what we call a Rashōmon set: a set of models showing the same predictive power, i.e., models whose training loss is below a specific threshold (Figure 2). Given a loss function L and a model class F , the Rashōmon set R can be written as [13]:

$$R(F, f^*, \epsilon) = \{f \in F \text{ such that } L(f) \leq L(f^*) + \epsilon\}$$

where f^* is an optimal model, ϵ is a threshold named the Rashōmon parameter. While multiple models can be found at $\epsilon = 0$, we usually use an ϵ -level set where $\epsilon > 0$ [14] because a low ϵ could result in an empty set. It has to be noted that computing the Rashōmon set is NP-hard and requires brute-force methods by sampling the hypothesis space [13].

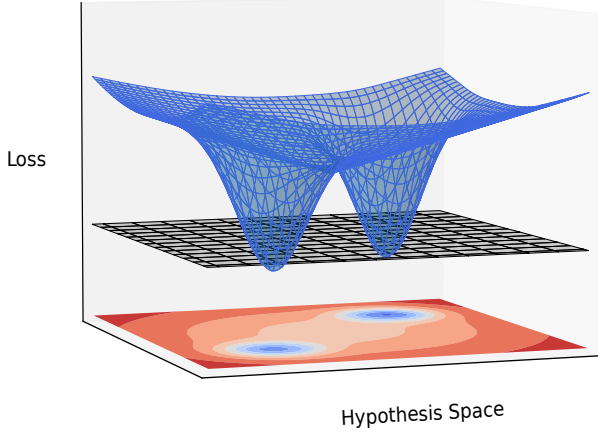


Figure 2: A simplified example of a Rashōmon set in a two-dimensional hypothesis space. The loss of each model in the space is in blue, while the black plan represents the value of the Rashōmon parameter, ϵ . As a result, the Rashōmon set is illustrated in blue in the projection at the bottom of the figure. We observe two distinct bassins, meaning we have at least two models with equals performances with different underlying hypotheses.

Let H be the hypothesis space that contains all possible models M , and let f_1 and f_2 be two near-optimal models in $R(F, f^*, \epsilon)$ of H . The Rashōmon Effect refers to the phenomenon where f_1 and f_2 produce similar predictions on the same input x , despite showing different feature importances. Formally, with w_1 and w_2 be the feature importances of f_1 and f_2 respectively, let r_1 and r_2 be the ranks of each feature. Intuitively, we can say that we observe a Rashōmon effect when:

$$r_1 \neq r_2$$

The Rashōmon Effect, in this case, can be problematic because it can lead to different interpretations of the same dataset and make it difficult to identify the most important features for a given task, or even to derive reliable knowledge from this interpretation.

1.3 Finding a Consensus

As formulated by Teney et al. (2022), predicting is not understanding [15]. The presence of the Rashōmon effect can be seen as a manifestation of underspecification, where a model fails to capture all the underlying patterns in the data accurately. To address this issue, Teney et al. propose training multiple models that are compatible with the data. Del Giudice (2022) suggests several ways to mitigate this problem, notably: (i) seeking consensus among models with different assumptions or biases, and (ii) noting that the severity of this effect tends to decrease when working with larger datasets [4].

Thus, we hypothesize that:

1. if computing the Rashōmon set without brute force is challenging, it is still possible to identify models that belong to the set,
2. the robustness against small input perturbations increases as the sample size grows,
3. the agreement among explanations from diverse models in the Rashōmon set converges, allowing for a consensus when the sample size is sufficiently large.

To further enhance the exploration of the Rashōmon effect, we included a bootstrap aggregation (bagging) strategy above all top models, a common method that enhances overall performance by alleviating fluctuations in predictions from multiple models, as we can suppose that they do so also by minimizing the variability of feature importances. To investigate the said hypothesis, the work has been organized into three main contributions on five distinct public datasets.

1. we proposed a methodology to assess an intra-model agreement, in which the explainability is disturbed by a varying input dataset through cross-validation,
2. we proposed a methodology to assess an inter-model agreement, and its convergence toward a consensus,
3. we analyzed the behavior of a bagging strategy to assess its behavior with respect to intra- and inter-model agreement.

The intent of this paper is not to benchmark explanation methods, but instead presents a novel approach to evaluating the robustness of explanations, notably with respect to sample size. The general goal of the present work is to propose a new framework for computing explanations of machine learning models (Figure 3). Its aim is to emphasize the fact that explaining a model is not enough, but we need to assess the convergence and robustness of feature importances.

2 Methodology

The method (Figure 3) can be divided into four parts: (i) a pre-selection of models based on their performances on the training set (Table 2) followed by their hyperparameter tuning to ensure we ended up in a Rashōmon set; (ii) a 10-fold cross-validation on subsets of the training set with varying size of each model, to ensure that the models are generalizing well; (iii) a bagging strategy on all the selected models, which is a technique that combines the predictions of multiple models to improve the overall accuracy; and (iv) an inference on a holdout test set followed by an explanation using the SHAP methodology. While we assume feature independence, it has to be noted that, when possible, we computed the exact SHAP values (e.g., with the exact and tree explainers) to avoid issues related

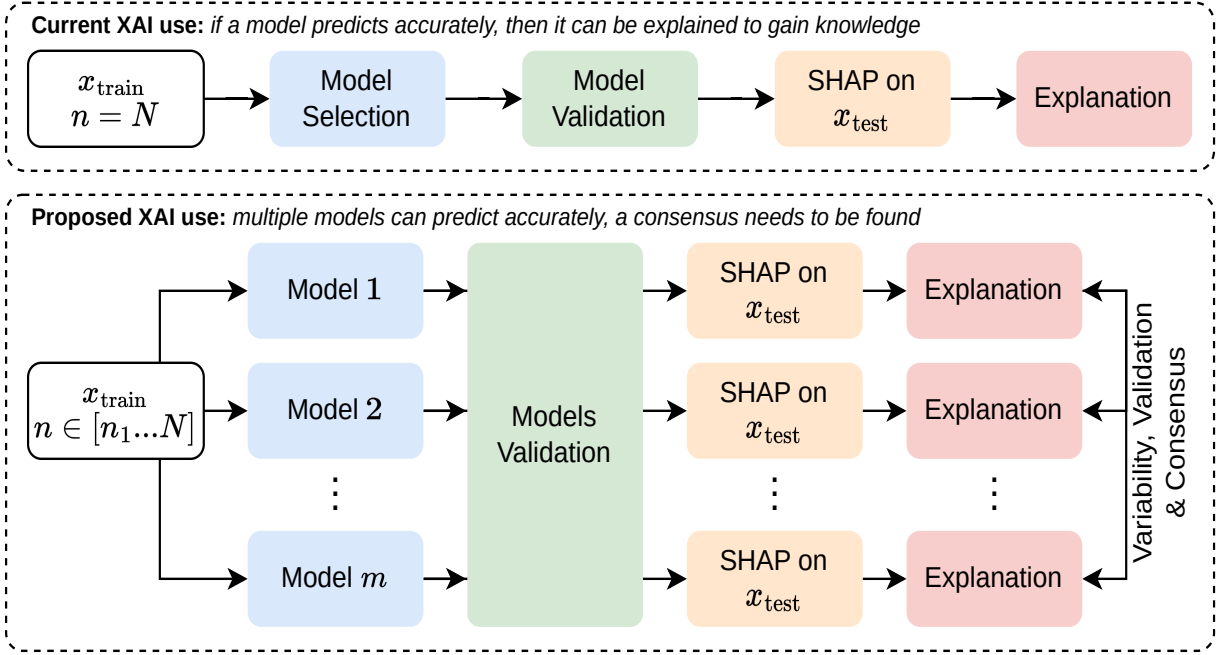


Figure 3: Proposed framework for enhancing explainability through model validation and consensus finding. It illustrates the proposed framework to improve the reliability of explanations from machine learning models. In contrast to the typical use of XAI methods where a single predictive model is explained, the proposed approach first validates multiple high-performing models on a dataset. SHAP explanations are generated for each model in the validation stage. The resulting explanations are then analyzed to find consensus and reduce variability, leading to more trustworthy explanations.

to multi-collinearity. All training, inference, and explanation steps are summarized in algorithm 1. As sample size might have an impact on hyperparameters, we conducted a simple hyperparameter tuning using a random search strategy for each instance of a model, to ensure that the best hyperparameters are chosen for each model.

2.1 Datasets

To test our hypothesis under various realistic conditions, we conducted our experiments on five datasets (Table 1), of various sizes, dimensionality, and with or without class imbalance (e.g. sex or site imbalances). Most of them come from the OpenXAI framework, an open initiative for the robust and repeatable evaluation of XAI methodologies [16].

2.2 Exploring the Rashōmon Set

The first step of this methodology is to select the models that we will explain. As the exploration of the Rashōmon set is NP-hard, its space has been divided into categories of models M (see Table 2). Using the PyCaret library [21], each of these models has been benchmarked with multiple configurations using 10-fold cross-validation, and the three configurations with the highest Cohen’s Kappa coefficient have been selected. The Cohen’s Kappa coefficient [22] is a statistical measure of the interrater agreement taking into account the possibility of agreement occurring by chance.

This is particularly important in cases where the classes are imbalanced. In addition to these top three models, a bagging strategy has been implemented by combining the predictions of the aforementioned models to further improve the overall comparison.

2.3 Shapley Additive Explanations (SHAP)

After model selection and training, the core of our hypothesis relies on feature importances that have to be comparable across all models. As all datasets are already given by the original authors as separate train and a test sets, each model inferred on the latter. Each model state - i.e. each cross-validation step at each sample size - went through the SHAP pipeline [10]. We chose the SHAP method for its ease of use, widespread adoption, and its axiomatic superiority to its counterpart. Shapley-based explanation methods provide a way to assign each feature an importance in a model-agnostic way [10], allowing the comparison between very different models that could have been previously selected. It has to be noted that, while SHAP is a methodological choice, the method we introduce could have been used with any other explanation framework. Following the definition of [23], let $\phi_{f_i}^{SHAP}$ be the SHAP values obtained from a machine learning model f_i , for a set of N instances and their labels y in a dataset X of K features. $\phi_{f_i}^{SHAP}$ is then a matrix of size (N, K) containing the feature importances of each prediction. In simpler terms, SHAP values are used to explain predictions made

Dataset	Short name	N	N Features	Balanced
Framingham Heart study [17]	Framingham	4240	16	No
German Credit [18]	German	1000	20	No
Pima-Indians Diabetes [19]	Diabetes	768	9	No
COMPAS [20]	Compas	18876	7	No
Student [16]	Student	100	30	Yes

Table 1: Datasets used to study the Rashōmon effect.

Algorithm 1 Model Selection and Explanations

Input: x_{train}, x_{test} (training and test sets)
Output: sim_{intra}, sim_{inter} (similarities between explanations)

```

1: procedure SELECT( $M$ )
2:    $r = []$  ▷ Array of Kappa scores
3:   for  $f \in M$  do
4:     Train( $f, x_{train}$ )
5:      $\hat{y} = f(x_{test})$ 
6:      $\mathcal{K} = \text{Kappa}(\hat{y}, y)$ 
7:      $r.append(\mathcal{K})$ 
8:   end for
9:    $top_3 = \text{ArgSort}(\text{Rank}(r)) [3]$ 
10: end procedure
11: procedure EXPLAIN( $top_3$ )
12:    $S = \{2^4, 2^5, \dots, N\}$  ▷ Varying sample sizes
13:    $\Phi = []$  ▷ Array of explanations
14:   for  $s \in S$  do
15:      $x = x_{train}[:s]$ 
16:     for  $f \in top_3$  do ▷ Repeated as 10-Fold CV
17:       Train( $f, x$ )
18:        $\hat{y} = f(x_{test})$ 
19:        $\phi_f^{SHAP} = \text{Explain}(\hat{y})$ 
20:        $\Phi.append(\phi_f^{SHAP})$ 
21:     end for
22:   end for
23:    $sim_{intra} = \text{similarity}(\Phi|s)$  ▷ For each model,
computes the similarity between cross-validation folds
24:    $sim_{inter} = \text{similarity}(\Phi|f)$  ▷ For each sample
size, computes the similarity between all the models
25: end procedure

```

by a model at an individual level. Because SHAP values can be positive or negative, and because we only care about the absolute importance of a feature, transitioning from the individual level to the population level can be achieved by taking the mean of the absolute values of the vector $\phi_{f_i}^{SHAP}$ across all instances:

$$\frac{1}{N} \sum_{n=1}^N |\phi_{f_i}^{SHAP}_n|$$

Model	Acronym	Linearity
Logistic Regression	lr	Linear
Linear Discriminant analysis	lda	Linear
Ridge	ridge	Linear
Gaussian Naive Bayes	nb	Linear
Singular Vector Machine	svm	Linear
Singular Vector Machine (RBF)	rbfsvm	Non-linear
Decision Tree	dt	Non-linear
Quadratic Discriminant analysis	qda	Non-linear
Random Forest	rf	Non-linear
AdaBoost	ada	Non-linear
Extra Trees	et	Non-linear
Gradient Boosting	gbc	Non-linear
K-Nearest Neighbors	knn	Non-linear
Gaussian Process	gp	Non-linear
CatBoost	catboost	Non-linear
LightGBM	lightgbm	Non-linear
XGBoost	xgboost	Non-linear

Table 2: List of all models included in our experiments. For each dataset, all models are trained and tuned, and the best ones are kept for further analysis.

2.4 Evaluating the similarity between multiple explanations

As our main goals were to compare the SHAP values given by (i) the same models with variations in the training set induced by cross-validation, and (ii) different near-optimal models on the same training set, we defined two metrics. First, since not all features may be of interest, we used a top_j similarity as introduced in [24]. Then, we used a weighted cosine similarity, which allows us to compare the similarity between multiple vectors while taking into account that features that are not important are possibly randomly ranked by a given model f_i .

2.4.1 Metrics

top_j Similarity

The top_j similarity metric is a useful tool for comparing the similarity between two sets of features selected by different models. This metric is based on the idea that not all features may be of interest, and therefore, we can focus on the top j features selected by each model. When comparing the overlap between these top j features, we can obtain a measure of similarity between the two models. This metric is particularly useful when comparing models with different feature selection methods or when only a subset of features is relevant to the problem at hand. The top_j

similarity metric is a simple, yet effective way to compare the interpretability of different machine learning models and can provide valuable insight into the behavior of these models. To begin with, let us define the ranking function R that maps the SHAP values to the ranked features as follows:

$$R(\phi_{f_i}^{SHAP}) = \text{argsort}(|\phi_{f_i}^{SHAP}|)[::- -1][: j] \quad (1)$$

where $\text{argsort}(a)$ returns the indices that would sort the array a in ascending order, and the indexing operation $[::- -1]$ reverses the order of the sorted indices to obtain the descending order. The final indexing operation $[: j]$ selects the top j indices based on this ranking.

Using this ranking function, we can obtain the sets of top j features selected by models f_1 and f_2 for the n -th instance as $S_n^{f_1} = R(\phi_n^{SHAP_{f_1}})$ and $S_n^{f_2} = R(\phi_n^{SHAP_{f_2}})$. The top_j similarity metric is the ratio of the average number of common features between the two models and j , where j is the number of features selected. Thus, the top_j similarity metric can be expressed as:

$$top_j \text{ similarity} = \frac{1}{j} \sum_{i=1}^j \frac{|S_i^{f_1} \cap S_i^{f_2}|}{|S_i^{f_2}|}$$

Here, the numerator represents the number of common features between the two models, and the denominator represents the number of features selected by the second model. As we compare M models, the final metric top_j^M computes the mean top_j similarity between all possible combinations of two different models using their respective SHAP values, such as:

$$top_j^M \text{ similarity} = \frac{1}{\binom{M}{2}} \sum_{a,b \in \binom{[1,M]}{2}} top_j \text{ similarity}(a,b) \quad (2)$$

where $\binom{M}{2}$ is the number of possible combinations of two different models, and $\binom{[1,M]}{2}$ is the set of all such combinations.

Weighted Cosine Similarity

We introduced the weighted cosine similarity as a useful metric to compare similarity between multiple vectors while considering that features that are not important are possibly randomly ranked by a given model f_i . First, with K features, we compute the mean absolute SHAP value for each feature k across all samples as:

$$mas_k = \frac{1}{K} \sum_{k=1}^K |\phi_{f_i,k}^{SHAP}|$$

Then, we compute the weighted SHAP values for each feature k as:

$$w_{f_i,k} = |\phi_{f_i,k}^{SHAP}| \cdot mas_k$$

where $k \in \{1, K\}$. Let us define the weighted cosine similarity $W_{cos_{sim}}$ between models f_1 and f_2 such as:

$$W_{cos_{sim}}(w_{f_1}, w_{f_2}) = \frac{w_{f_1} \cdot w_{f_2}}{\|w_{f_1}\|_2 \|w_{f_2}\|_2}$$

where \cdot denotes the dot product and $\|\cdot\|_2$ denotes the Euclidean norm. Similarly to top_j similarity, we can finally define $W_{cos_{sim}}^M$ as the weighted cosine similarity between all combinations of two different models:

$$W_{cos_{sim}}^M = \frac{1}{\binom{m}{2}} \sum_{a,b \in \binom{[1,M]}{2}} W_{cos_{sim}}(a,b) \quad (3)$$

We used the weighted cosine similarity to compare the similarity of models' SHAP values while accounting for the randomness of unimportant feature ranking.

2.5 Assessing the convergence towards a near-optimal consensus

The mean of the absolute SHAP values for the largest sample size N available from all models f_i can be expressed as:

$$\bar{\phi}_{consensus}^{SHAP} = \frac{1}{MN} \sum_{i=1}^M \sum_{j=1}^N |\phi_{f_i,j}^{SHAP}| \quad (4)$$

where $\bar{\phi}_{consensus}^{SHAP}$ represents the mean absolute SHAP values for the consensus of all models, M is the total number of models, N represents the largest sample size available across all models, and $|\phi_{f_i,j}^{SHAP}|$ represents the absolute SHAP value for the j -th instance of the i -th model. This formulation calculates the mean of the absolute SHAP values across all instances and models for the largest available sample size N , providing a baseline for comparison with the evolution of the explanations of individual models over time.

2.6 Statistical Analysis

To assess the intra-model agreement, we used Spearman correlations between sample sizes and the weighted cosine similarities of each individual model on the SHAP values computed at each step of the cross-validation. Spearman correlations are used to account for the monotonic nonlinear nature of the relationship. A similar analysis is performed for the inter-model analysis, where we computed Spearman correlations between sample sizes and weighted cosine similarities between individual models and the consensus $\bar{\phi}_{consensus}^{SHAP}$. To account for multiple comparisons, p -values are corrected using the Benjamini-Hochberg false discovery rate (FDR) [25].

3 Results

Having selected the top three models that maximize the Kappa coefficient on the training set, we then computed their performances on the test set. Subsequently, we compared both their internal variability (i.e., the variability caused by the cross-validation) and their overall relative variability (i.e., the variability caused by a switch of near-optimal models).

3.1 Models performances

We implemented a 10-fold cross-validation strategy to assess the accuracy (ACC), F1, and Matthews correlation coefficients (MCC) of various models on the Framingham, German, Diabetes, Compas, and Student datasets, the results of which are illustrated in Table 3. The findings showed the existence of a Rashōmon set for the Framingham, Compas, Diabetes, and Student datasets, where the precision, F1 score, and MCC of the proposed models were similar. An examination of the learning curve revealed that most of the models reached a convergence state for the Framingham, Diabetes, Compas, and Student datasets, as shown in Figure 4. Interestingly, it can be noted that none of the models successfully converged on the German dataset, which is characterized by high dimensionality and exhibits a pronounced class imbalance, in contrast to the Student dataset where optimal performance levels were achieved early in the learning progression, as depicted in Figure 4.

3.1.1 Intra-model agreement

In our analyses of the Compas, Diabetes, and Framingham datasets, we found a significant correlation between sample size and intramodel agreement, as shown in Table 4. However, this correlation was absent in the context of the Student dataset where, remarkably, even with smaller sample sizes, we recorded great levels of predictive precision and harmony, as evidenced in Table 3. Conversely, the German dataset exhibited persistently poor predictive accuracy throughout our experimental endeavors, a deficiency mirrored in the intra-model agreement. Overall, the top_j similarity demonstrated a propensity for the top_j features to converge toward uniformity, particularly for the bagging models (Figure 5). These models displayed a higher internal agreement on the Framingham, German, and Diabetes datasets. An intriguing insight was the ability to attain a high degree of agreement with smaller samples, a level that initially receded but subsequently increased as the sample size grown.

3.1.2 Inter-model agreement and consensus

The outcomes of the convergence towards an optimal consensus is similar to the intramodel agreement. Nearly each individual model converges towards the near-optimal consensus $\phi_{consensus}^{SHAP}$ (Table 5 and Table 6), except for the student model, which demonstrates a significantly high level of similarity even for small sample sizes. As for the

	ACC	F1	MCC
A.			
baseline	0.663 ±0.001	0.231 ±0.001	0.200 ±0.002
nb	0.833 ±0.001	0.268 ±0.002	0.196 ±0.002
qda	0.828 ±0.006	0.318 ±0.007	0.228 ±0.002
lda	0.808 ±0.001	0.354 ±0.001	0.242 ±0.001
bagging	0.841 ±0.001	0.24 ±0.004	0.19 ±0.004
B.			
baseline	0.535 ±0.060	0.600 ±0.073	-0.320 ±0.000
gpc	0.595 ±0.001	0.714 ±0.001	0.022 ±0.001
et	0.65 ±0.001	0.783 ±0.001	-0.072 ±0.001
nb	0.5 ±0.000	0.59 ±0.000	-0.005 ±0.000
bagging	0.504 ±0.012	0.603 ±0.019	-0.026 ±0.005
C.			
baseline	0.721 ±0.001	0.581 ±0.001	0.393 ±0.000
rbfsvm	0.759 ±0.003	0.622 ±0.018	0.449 ±0.009
et	0.759 ±0.006	0.622 ±0.028	0.451 ±0.019
gpc	0.739 ±0.001	0.583 ±0.0	0.397 ±0.001
bagging	0.748 ±0.001	0.598 ±0.001	0.418 ±0.003
D.			
baseline	0.850 ±0.000	0.912 ±0.000	0.413 ±0.000
catboost	0.853 ±0.002	0.914 ±0.001	0.432 ±0.010
gbc	0.851 ±0.005	0.913 ±0.003	0.427 ±0.019
ada	0.851 ±0.003	0.913 ±0.002	0.429 ±0.007
bagging	0.854 ±0.001	0.915 ±0.001	0.438 ±0.005
E.			
baseline	0.960 ±0.000	0.947 ±0.002	0.918 ±0.000
xgboost	0.980 ±0.001	0.976 ±0.001	0.959 ±0.001
catboost	1.000 ±0.000	1.000 ±0.000	1.000 ±0.000
dt	0.960 ±0.001	0.947 ±0.001	0.919 ±0.001
bagging	0.980 ±0.001	0.974 ±0.001	0.959 ±0.001

Table 3: Test-set performances of the sampled models on accuracy (ACC), F1, and Matthews Correlation Coefficients (MCC) on the Framingham (A), German (B), Diabetes (C), Compas (D), and Student (E) datasets. Bold results indicate best models. The baseline is a simple logistic regression.

	p	p_{cor}	r	power
compas	0.000	0.000	0.889	1.000
diabetes	0.025	0.061	0.458	0.635
german	0.664	0.664	0.080	0.072
framingham	0.045	0.075	0.412	0.534
student	0.097	0.122	0.298	0.389

Table 4: Correlation between sample size and intra-model correlation. Results are Spearman correlations between similarities of cross-validation explanations and sample size.

German model, it follows a U-shaped pattern with respect to the sample size (Figure 6).

4 Discussion

The problem of the existence of multiple high-performing models relying on different features — the Rashōmon effect — poses significant challenges to derive reliable

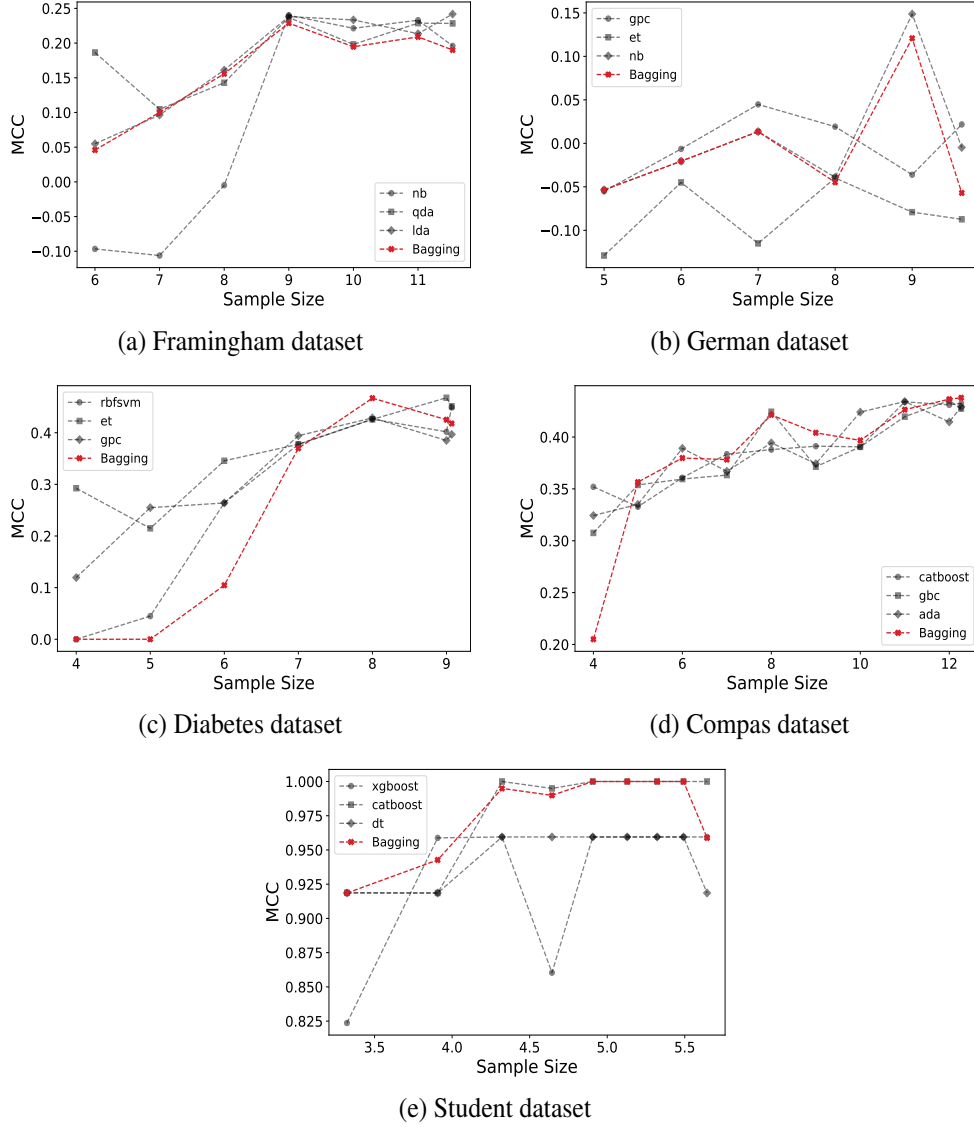


Figure 4: Impact of the sample size on model performance (Matthews Correlation Coefficient, MCC). MCC is computed on the test set during a 10-fold cross-validation. The gray lines represent individual models, and the red line represents the result when a bagging strategy is used on all models. Sample sizes are on a log2 basis.

	p	p_{cor}	r	power
compas	0.000	0.000	0.766	0.999
diabetes	0.000	0.000	0.862	0.999
german	0.111	0.139	0.334	0.366
framingham	0.000	0.000	0.875	1.000
student	0.654	0.654	-0.096	0.073

Table 5: Correlation between sample size and inter-model correlation convergence towards the best consensus. Results are Spearman correlations between similarities of all models and $\bar{\phi}_{consensus}^{SHAP}$ and sample size.

	p	p_{cor}	r	power
compas	0.067	0.084	0.633	0.483
diabetes	0.000	0.000	1.000	1.000
german	0.037	0.061	0.738	0.604
framingham	0.329	0.329	0.486	0.171
student	0.002	0.005	0.905	0.936

Table 6: Correlation between sample size and bootstrap aggregation convergence towards the best consensus. Results are Spearman correlations between similarities of a bagging strategy of all top_3 models and $\bar{\phi}_{consensus}^{SHAP}$ and sample size.

knowledge from machine learning systems. In this study, our aim was to examine the influence of sample size on models within a Rashōmon set using a model-agnostic explainability technique. Our research revolved around

two main aspects: (i) the enhancement of explainability in relation to sample size, and (ii) the convergence of explanations from diverse models in the Rashōmon set, leading

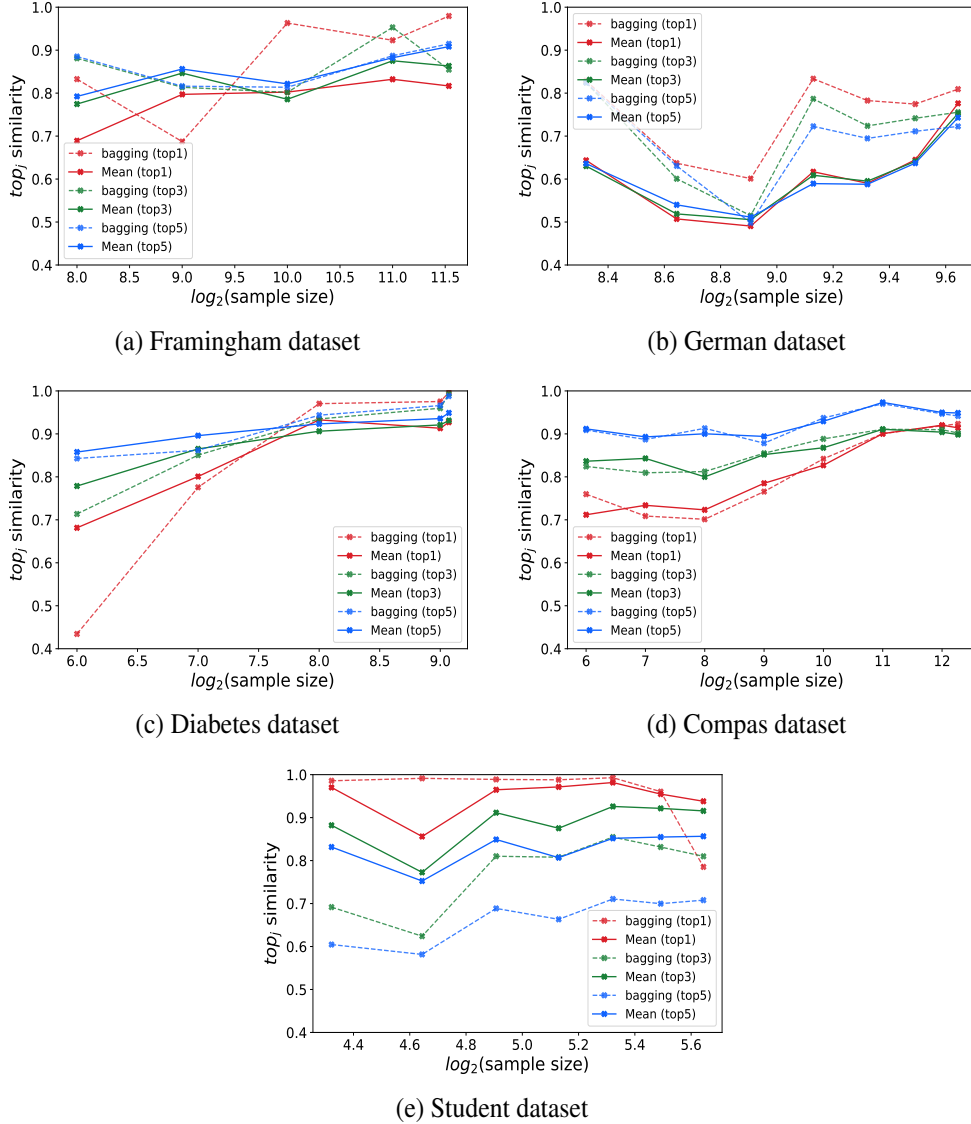


Figure 5: Impact of the sample size on top_j similarity between model explanations. Explanations given for each model are computed using a 10-fold cross-validation. The dashed lines in light colors represent individual models, the solid lines represent the mean top_j for $j = 1, j = 3$, and $j = 5$ across all models, and their lighter counterparts represent a bagging strategy on all models.

to a consensus when the sample size reaches a sufficient size. Additionally, we explored the characteristics of bagging strategies within the same scenarios to gain additional information. The key findings of this study demonstrate the substantial impact of sample size on the reliability and agreement of explanations from machine learning models that exhibit a Rashōmon effect (Table 4, Table 5, and Table 6). Our experiments in five public data sets revealed that explanations derived from subsets with fewer than 128 samples showed high variability in cross-validation, indicating spuriousness (Figure 5 and Figure 6). This effectively limits the actionable knowledge that can be extracted from any single model’s interpretations. However, as the sample size increased, the variance in similarity diminished and models converged towards a unified explanation.

Our first experiment was primarily devoted to the study of model robustness when the training set is perturbed through cross-validation, a property we named internal agreement. All models, except the one used for the German dataset, demonstrated convergence towards an acceptable level of classification accuracy (Figure 4 and Table 3), while manifesting a strong correlation between sample size and weighted cosine similarity (Table 4). However, it is important to highlight two exceptions. The first pertains to the methodologies employed on the German dataset, which were unable to identify suitable models within the Rashōmon set, implying that no models were successful in determining a valid solution, resulting in a lack of convergence in internal agreement. The second exception is linked to the Student dataset, where convergence was observed remarkably early in the learning trajectory (Fig-

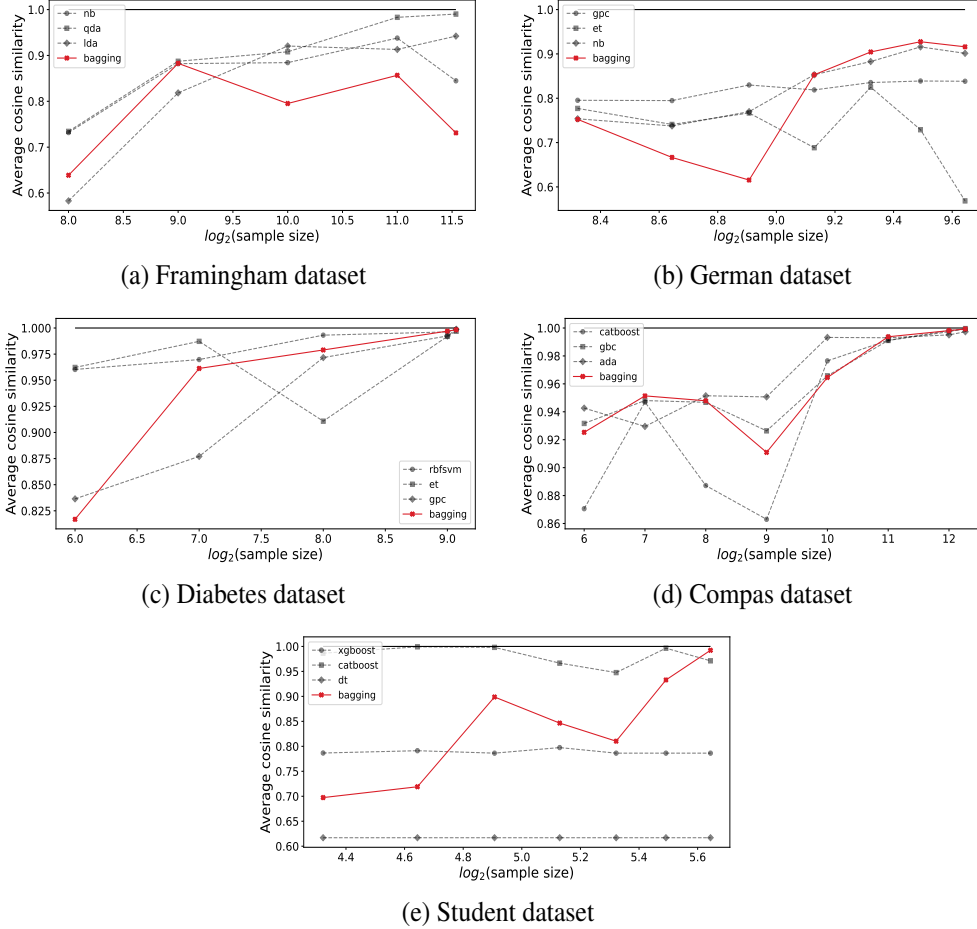


Figure 6: Impact of the convergence of models towards the best consensus available. Explanations given for each model are computed using 10-fold cross-validation. Dashed lines represent individual models; the black line represents the best consensus available (the mean of the absolute SHAP values for the biggest sample size available). The red line represents the result when using a bagging strategy on all models.

Figure 4e), indicating that a high degree of internal agreement was achieved at the beginning of the experiment itself. An interesting insight was the nonlinear association between the sample size and the intra-model consensus. At small sample sizes, high similarity could be achieved (Figure 5b) despite a poor predictive capacity (Figure 4b), which subsequently experienced a decline, only to ascend once again with the addition of more data. This aligns with previous research indicating false confidence in underspecified models [26], drawing a parallel to the Dunning-Kruger effect observed in human learning. Although the agreement on the German dataset seems to increase, this suggests that more data is needed to perform our analysis on our dataset. The second important point that this experiment highlights is the regular superiority of bagging approaches, where they showed higher internal agreement at large sample sizes for 3 of our 5 datasets (Figure 5), but this was not always the case at low sample sizes, especially for the Diabetes dataset. This first experiment allows us to say that while sample size has a positive impact on a model’s internal agreement, performance is not always correlated to agreement in explanations, highlighting the need to per-

form additional analysis before the explanation of machine learning models. Moreover, contrary to popular belief, bagging is not always the best strategy — especially at low sample sizes — regarding explainability.

The supplementary experiment engaged in an in-depth exploration concerning the coherence among various models and their convergence towards a near-optimal consensus, — the average explanation of high-performing models at the highest sample size. Once again, we observed a positive correlation between the sample size and the agreement among the high-performance models (Table 5 and Table 6). Analogous to the intra-model agreement, the experiment failed to exhibit any correlation within the Student dataset due to the previously delineated complications. The convergence of inter-model explanations reaffirms the hypothesis that large datasets attenuate the Rashōmon effect and facilitate reliable knowledge extraction [4]. Despite the fact that the bagging ensembles did not always converged better than the individual models (Figure 6a, c, and d), they converged towards the optimal solution, while some models remained stuck within their respective basins (Figure 6e). This secondary experiment allows the conclusion that the

learning process functions herein as a refinement of feature importance, a process whereby an increase in sample size triggers a convergence towards a singular, coherent explanation of the predictions. However, it is interesting to observe the phenomenon that underscores the disparities between a consensus derived from explanations coming from diverse models, and the explanation of a bagging procedure over the exact same models (e.g. Figure 6a). Although it was generally not possible to differentiate between the predictive performance of individual models and the bagging procedure (Figure 4), this research does not offer the means to draw a definitive conclusion about which is superior: should we favor the explanation derived from an aggregation of multiple models, or a consensus predicated upon the explanations of the individual models? However, the conclusion we can draw is that both were similar for 4 datasets in the 5 we studied, meaning that our method should be appropriate in both cases.

However, there are some limitations to the generalization of these findings. First, this study limited itself to certain types of models, such as linear models or random forests. Testing other complex models, such as neural networks, could reveal different behaviors and challenges. Extending this research to neural network models presents challenges due to their complexity, but also opportunities to mitigate the Rashōmon effect through transfer learning. As shown by [27], different initializations of the same architecture can learn distinct solutions. Transfer learning provides more consistent representations and could align explanations between models, because the learning on the specific task starts in an already nonrandom state. Questions such as “*does the surrogate task impact explanations?*” or “*are explanations found through transfer learning better than those of models trained from scratch?*” remain to be answered. Testing the sample size effects shown here in deep learning scenarios with and without transfer learning would be an impactful extension.

Second, while the datasets covered various sizes and domains, more extensive experiments are needed on diverse real-world problems, which could include regression problems. Regarding those datasets, the Student task was definitely too simple for the correct assessment of our methodology, even if it allowed us to draw interesting conclusions. However, the German dataset clearly requested a modified methodology, certainly a dimensionality reduction, such as PCA, or a class-imbalance correction, such as SMOTE. We intentionally chose not to adopt this modified methodology to stay consistent with the other analysis. Also, PCA would have complexified our SHAP workflow as we would have had to interpret its principal components. However, this raises a relevant link between the curse of dimensionality and overconfidence issues in machine learning models as described in other works (e.g. [28]).

Finally, while this study used SHAP, a model-agnostic approach, further research could explore how model-specific explanation methods such as attention maps or Grad-CAM behave in the context of the Rashōmon effect. As model-

specific techniques provide insights into a model’s internal representations, they may reveal divergences between architectures that model-agnostic methods cannot access. Comparing the convergence patterns of both approaches could offer additional validation and confidence in the explanations. Furthermore, the evaluation of other post-hoc explanation techniques such as LIME could reveal different insights. For example, LIME perturbs inputs and trains interpretable local models, which may show higher variability at low sample sizes. Testing multiple perturbation and explanation strategies could indicate which are the most robust for reliable explanations from underspecified models.

Despite these constraints, this work has meaningful practical implications. The results suggest that sample sizes below 100 can lead to unreliable explanations that lack consensus between equally performing models. We argue that our methodology provides a kind of power analysis to determine sufficient data to trust explanations. Furthermore, the variability at low sizes indicates that conclusions drawn from small datasets could be spurious and should be validated. In general, these findings can guide machine learning practitioners in selecting appropriate data volumes, models, and explanation techniques for their applications.

Future work should explore ensembles such as bagging for explanation robustness across broader model types, data domains, and explanation methods. Testing the Rashōmon effect in online learning settings where models are continuously retrained would also have a great impact. As interpretations become increasingly critical for trustworthy AI, developing a rigorous understanding of how to evaluate, improve, and trust model explanations remains an essential challenge. The approaches explored here — leveraging sample size, ensembles, consensus-finding, and variability quantification — point towards principled pathways for eliciting knowledge from ambiguous models. This study provides an initial data-driven perspective on this important problem.

5 Conclusion

This study examined how sample size impacts the explainability of models that exhibit Rashōmon effect. Experiments on multiple datasets revealed explanations become more consistent as sample size grows, with variability limiting reliability below 100 samples. Key takeaways for practitioners include: 1) larger data volumes attenuate the Rashōmon effect and improve explanation consensus, 2) explanations derived from limited data may be spurious and require validation, 3) bagging ensembles can enhance agreement between models. Overall, these findings provide guidance on selecting appropriate sample sizes, models, and explanation techniques when interpreting machine learning systems to ensure credible and actionable knowledge.

References

- [1] Andrzej Grzybowski and Piotr Brona. Approval and Certification of Ophthalmic AI Devices in the European Union. *Ophthalmology and Therapy*, 12(2): 633–638, April 2023. ISSN 2193-8245, 2193-6528. doi: 10.1007/s40123-023-00652-w.
- [2] Mohit Pandey, Michael Fernandez, Francesco Gentile, Olexandr Isayev, Alexander Tropsha, Abraham C. Stern, and Artem Cherkasov. The transformational role of GPU computing and deep learning in drug discovery. *Nature Machine Intelligence*, 4(3):211–221, March 2022. ISSN 2522-5839. doi: 10.1038/s42256-022-00463-x.
- [3] Priyanka Dixit and Sanjay Silakari. Deep Learning Algorithms for Cybersecurity Applications: A Technological and Status Review. *Computer Science Review*, 39:100317, February 2021. ISSN 15740137. doi: 10.1016/j.cosrev.2020.100317.
- [4] Marco Del Giudice. The Prediction-Explanation Fallacy: A Pervasive Problem in Scientific Applications of Machine Learning. preprint, PsyArXiv, December 2021. URL <https://osf.io/4vq8f>.
- [5] Xiaoxiao Li, Yuan Zhou, Nicha Dvornek, Muhan Zhang, Siyuan Gao, Juntang Zhuang, Dustin Scheinost, Lawrence H Staib, Pamela Ventola, and James S Duncan. BrainGNN: Interpretable brain graph neural network for fMRI analysis. *Med. Image Anal.*, 74(102233):102233, December 2021.
- [6] Gabrielle Ras, Ning Xie, Marcel Van Gerven, and Derek Doran. Explainable Deep Learning: A Field Guide for the Uninitiated. *Journal of Artificial Intelligence Research*, 73:329–397, January 2022. ISSN 1076-9757. doi: 10.1613/jair.1.13200.
- [7] Hiroshi Fukui, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi. Attention branch network: Learning of attention mechanism for visual explanation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [8] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *International Journal of Computer Vision*, 128(2):336–359, February 2020. ISSN 0920-5691, 1573-1405. doi: 10.1007/s11263-019-01228-7.
- [9] Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso. Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics*, 8(8): 832, July 2019. ISSN 2079-9292. doi: 10.3390/electronics8080832. URL <https://www.mdpi.com/2079-9292/8/8/832>.
- [10] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.
- [11] Leo Breiman. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3), August 2001. ISSN 0883-4237. doi: 10.1214/ss/1009213726.
- [12] Robert Anderson. The Rashomon Effect and Communication. *Canadian Journal of Communication*, 41(2):249–270, May 2016. ISSN 0705-3657, 1499-6642. doi: 10.22230/cjc.2016v41n2a3068. URL <https://cjc.utpjournals.press/doi/10.22230/cjc.2016v41n2a3068>.
- [13] Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges, July 2021. URL <http://arxiv.org/abs/2103.11251>. arXiv:2103.11251 [cs, stat].
- [14] Charles T. Marx, Flavio du Pin Calmon, and Berk Ustun. Predictive Multiplicity in Classification, September 2020. URL <http://arxiv.org/abs/1909.06677>. arXiv:1909.06677 [cs, stat].
- [15] Damien Teney, Maxime Peyrard, and Ehsan Abbasnejad. Predicting is not Understanding: Recognizing and Addressing Underspecification in Machine Learning, July 2022. URL <http://arxiv.org/abs/2207.02598>. arXiv:2207.02598 [cs].
- [16] Chirag Agarwal, Satyapriya Krishna, Eshika Saxena, Martin Pawelczyk, Nari Johnson, Isha Puri, Marinka Zitnik, and Himabindu Lakkaraju. OpenXAI: Towards a Transparent Evaluation of Model Explanations, January 2023.
- [17] Ashish Bhardwaj. Framingham heart study dataset.
- [18] Hans Hofmann. Statlog (German Credit Data), 1994.
- [19] Jack W. Smith, J.E. Everhart, W.C. Dickson, W.C. Knowler, and R.S. Johannes. Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus. pages 261–265, 1988.
- [20] Kareem L. Jordan and Tina L. Freiburger. The Effect of Race/Ethnicity on Sentencing: Examining Sentence Type, Jail Length, and Prison Length. *Journal of Ethnicity in Criminal Justice*, 13(3):179–196, July 2015. ISSN 1537-7938, 1537-7946. doi: 10.1080/15377938.2014.984045.
- [21] Moez Ali. *PyCaret: An open source, low-code machine learning library in Python*, April 2020. URL <https://www.pycaret.org>. PyCaret version 1.0.
- [22] Jacob Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46, April 1960. ISSN 0013-1644, 1552-3888. doi: 10.1177/001316446002000104.

- [23] Pål Vegard Johnsen, Inga Strümke, Mette Langaas, Andrew Thomas DeWan, and Signe Riemer-Sørensen. Inferring feature importance with uncertainties with application to large genotype data. *PLOS Computational Biology*, 19(3):e1010963, March 2023. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1010963.
- [24] Andreas Messalas, Yiannis Kanellopoulos, and Christos Makris. Model-Agnostic Interpretability with Shapley Values. In *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*, pages 1–7, PATRAS, Greece, July 2019. IEEE. ISBN 978-1-72814-959-2. doi: 10.1109/IISA.2019.8900669.
- [25] Yosef Hochberg. A sharper Bonferroni procedure for multiple tests of significance. page 3, 1988.
- [26] Yoav Wald, Amir Feder, Daniel Greenfeld, and Uri Shalit. On Calibration and Out-of-domain Generalization, January 2022.
- [27] Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang. What is being transferred in transfer learning?, January 2021.
- [28] Dr. Mehmet Cem Çatalbaş. An Investigation into the Relationship between Curse of Dimensionality and Dunning-Kruger Effect. *Sakarya University Journal of Computer and Information Sciences*, 3(2):121–130, August 2020. ISSN 2636-8129. doi: 10.35377/saucis.03.02.727032.

Additional Discussion

The study on the Rashōmon effect in machine learning models provides valuable insights into assessing model robustness and reliability prior to extracting knowledge. The approach bears similarities to power analysis in conventional statistics, quantifying the amount of data needed for stable model explanations. The inability to find agreement on the German dataset highlights the need for such sanity checks, indicating that more data may be required for reliable modeling. Overall, this methodology offers a principled way to validate machine learning models before deriving conclusions, analogous to confirming model assumptions in classical statistics.

These techniques have exciting potential for application in neuroimaging studies using machine learning. In particular, resting state fMRI of the hippocampus poses challenges due to small structure size and high-dimensional connectivity data. Functional connections between hippocampal subfields and other brain regions can provide development biomarkers. However, reliably extracting knowledge requires evidence that models have converged on robust explanations, given the potential for Rashōmon effects with limited samples. The proposed methods offer ways to assess the consistency of predictive models applied to resting state data, helping determine if explanations reflect genuine development patterns versus spurious correlations. This work highlights the need for validation before interpreting model feature importance, providing confidence in discovered biomarkers. Overall, the study of Rashōmon effects delivers a framework to enhance model trustworthiness in exploratory neuroscience applications.

3.2.2 Charting Hippocampal Development with Robust Explainable AI and Resting-State fMRI

Clement Poiret, and Marion Noulhiane (2023). Charting Hippocampal Development with Robust Explainable AI and Resting-State fMRI. *ArXiv preprint*.

The previous study demonstrated a robust methodology for eliciting reliable knowledge from machine learning models exhibiting the Rashōmon effect. By quantifying agreement between explanations from an ensemble of high-performing models, it was possible to distinguish consistent predictive features from unstable ones (Poiret, Grigis, et al., 2023). To further validate this approach and study the functional maturation of the hippocampal subfields, we applied this methodology on the functional connections from the hippocampus and the rest of the brain. Functional connectivity captured via resting-state fMRI provides a window into coordinated neural circuit dynamics supporting cognition. However, deriving reliable biomarkers is challenging due to noise and high-dimensionality. We hypothesized that our multi-model explainable AI approach could overcome these obstacles to chart typical hippocampal subfields development. The hippocampus plays a critical role in learning and memory, with distinct subfields showing differential maturation. By training an ensemble of models on hippocampal subfields connectivity and aggregating explanations, we hoped to extract robust developmental fingerprints despite noise. Thus, we tested whether the consensus methodology established for mitigating the Rashōmon effect could elucidate the complex interactive specialization of hippocampal circuits from multidimensional resting-state fMRI data.

Abstract

Goal: This study aimed to apply a robust multi-model explainable AI methodology to elucidate developmental changes in hippocampal subfields functional connectivity from resting-state fMRI data.

Material and Methods: Functional connectivity between hippocampal subfields and 360 cortical regions was computed from resting-state fMRI data on 588 healthy youths ages 5-21 from the Human Connectome Project. Subfields were segmented from T2 MRI scans. An ensemble of machine learning models was trained to predict age from connectivity features. Model explanations were generated using SHAP values and aggregated to obtain a consensus on core developmental fingerprints.

Results: The models achieved high accuracy in predicting age solely from hippocampal subfields connectivity patterns. Explanations identified 10 connections, including the interactions of the subfields with medial temporal, frontal, and parietal cortices, that decreased in strength with age. Explanation agreement improved with model accuracy.

Conclusion: The results suggest interactive specialization and pruning of hippocampal circuits unrelated to memory. Decreasing connectivity likely reflects maturation supporting relational binding, pattern separation, and spatial navigation. The study demonstrates that with model convergence, explainable AI can chart typical hippocampal development from resting-state fMRI despite noise and high-dimensionality.

CHARTING HIPPOCAMPAL DEVELOPMENT WITH ROBUST EXPLAINABLE AI AND RESTING-STATE fMRI

A PREPRINT

Clément Poiret ^{1,2,3} and Marion Noulhiane ^{1,2,3*}

¹UNIACT, NeuroSpin, CEA Paris-Saclay, Frederic Joliot Institute, Gif-sur-Yvette, France

²NeuroSpin, CEA Paris-Saclay, Frederic Joliot Institute, Gif-sur-Yvette, France

³InDEV, NeuroDiderot, Université Paris Cité, Inserm, Paris, France

October 1, 2023

ABSTRACT

The hippocampus plays a central role in memory and spatial navigation. It contains anatomically distinct subfields with differential functional roles. This study employed a robust multi-model explainable AI (XAI) approach to analyze resting-state fMRI data, elucidating developmental changes in the functional connectivity of hippocampal subfields. 588 healthy youths (ages 5-21) from the Human Connectome Project Development dataset were analyzed. Individual subfields were segmented from T2 MRI scans. Functional connectivity between subfields and 360 cortical regions was computed from resting fMRI data. An ensemble of machine learning models was trained to predict age from connectivity features. Bayesian regression achieved high accuracy (mean absolute error = 2.16 years) using only hippocampal subfields connectivity. Model explanations based on SHAP values revealed 10 core connections where strength decreased with age, including subfields interactions with medial temporal, frontal, and parietal cortical areas. Explanation agreement improved with model accuracy, distinguishing reliable developmental “fingerprints” from unstable features. The findings suggest interactive specialization of hippocampal circuits, with pruning of connections unrelated to memory. Decreasing connectivity may reflect neural optimization supporting maturation of relational binding, pattern separation, spatial navigation and memory retrieval. Overall, the study demonstrates that a robust multi-model XAI approach applied to resting-state fMRI can chart typical hippocampal development. With thoughtful analysis and model convergence, AI methods can provide robust biomarkers characterizing neurodevelopment despite noise and dimensionality challenges in connectivity data. The approach could be extended to assess developmental disorders.

Keywords Connectome · Neurodevelopment · Biomarker · Functional connectivity · Rashomon Effect

1 Introduction

The hippocampus is a critical region for learning, memory, and spatial navigation. Given its central role in key cognitive functions, understanding hippocampal development provides fundamental insights into the maturation of memory systems. The hippocampus can be anatomically divided into distinct subfields including the dentate gyrus (DG), and cornu ammonis (CA) regions CA1, CA2, CA3, and the subiculum. Each subfield plays differential roles in memory encoding, consolidation, and retrieval. Examining the functional connectivity of hippocampal subfields in resting-state fMRI data provides a powerful approach to elucidate the maturation of hippocampal neural circuits. Resting-state functional connectivity captures coordinated

spontaneous activity between brain regions, providing insight into networks supporting cognition without requiring active tasks. By applying advanced analytics like machine learning to resting-state functional connectivity patterns, it is possible to chart the typical developmental trajectories of hippocampal subfields and their interactions with distributed cortical regions involved in memory and navigation. This enables sensitive detection of connectivity biomarkers that could be extended to assess developmental disorders.

1.1 Functional Connectivity

Functional connectivity captures coordinated activity between spatially remote regions of the brain. Formally de-

defined as “statistical dependencies among remote neurophysiological events” [1], functional connectivity is widely studied using resting-state fMRI. This data-driven approach measures spontaneous signal fluctuations when a subject is at rest, not performing any explicit task [2]. Functional connectivity is then typically assessed by computing the correlation of time series between different brain regions.

Extensive research has delineated the functional properties of the hippocampus and its differential contributions to various aspects of memory [3]. While the hippocampus undergoes protracted structural maturation extending into early adulthood [4, 5, 6, 7], studies have shown that it forms distinct functional connections with other regions relatively early in development [8]. These links support the integration of information across domains and are thought to reflect interactive specialization, marked by organizational changes that support memory development [9, 10].

In particular, the anterior and posterior hippocampal segments exhibit differential patterns of connectivity and functional specialization. The anterior hippocampus shows preferential connectivity with the anterior and medial temporal lobe regions involved in semantic memory, while the posterior hippocampus connects more strongly with the posterior cortical areas that support spatial navigation and imagery [11, 12, 13]. Connectivity strengthens over development between the hippocampus and critical cortical hubs such as the precuneus and posterior cingulate, regions of the default mode network [8, 12]. On the contrary, connectivity weakens between the hippocampus and regions less related to memory, such as the inferior frontal gyrus [12]. Anterior versus posterior hippocampal connectivity also shifts dynamically with distinct prefrontal subregions that support executive control processes [14].

Overall, hippocampal functional connectivity undergoes interactive specialization with age, marked by organizational changes supporting memory development [12]. The differentiation enables efficient binding of multimodal information, providing a neural basis for episodic memory formation [3]. As connectivity patterns mature, hippocampal-cortical interactions become increasingly optimized to support the consolidation and retrieval of lifelong memories. Although the literature suggests that the hippocampus becomes more functionally integrated with memory-related cortical regions and segregated from non-memory regions with age, related to memory improvements, the precise developmental timecourse and functional significance of hippocampal subfields connectivity remain poorly understood. Elucidating subfield-specific maturation is key to deciphering hippocampal contributions to memory development. However, the precise nature and functional significance of these connections remain incompletely understood, even more incomplete at the scale of the hippocampal subfields.

1.2 Machine Learning

Machine learning techniques provide promising and powerful opportunities to extract predictive biomarkers from fMRI data. However, realizing this potential requires care-

ful consideration of the unique properties of connectome data. Specifically, connectome data possesses complex topological properties distinct from grid-like image data, which presents both advantages and challenges for applying machine learning methods [15]. On the one hand, the rich relational information in connectivity patterns enables extraction of predictive features at multiple interrelated scales, from individual edges and nodes to local subgraphs and full connectomes. However, these complex topological properties also require adapting and optimizing machine learning techniques for connectivity data. A key challenge is that commonly used univariate fMRI measures, such as regional averages, often lack sufficient test-retest reliability for biomarker discovery or brain-behavior mapping [16]. This highlights the need for developing optimized multivariate techniques tailored for connectivity data. Encouragingly, fMRI can demonstrate high test-retest reliability if leveraging appropriate multivariate methods and connectivity-based measures beyond individual region averages [17]. Reliability is thus not an inherent fixed property, but rather depends on the specific measures and analyses employed. Overall, machine learning provides a promising approach to extract generalized biomarkers from fMRI data, but achieving its full potential requires thoughtful application adapted for connectome data properties. If applied judiciously, machine learning could enable fMRI to yield reliable and clinically useful biomarkers for diverse applications. The complex topological properties of connectomes raise challenges but also provide rich opportunities if addressed carefully using optimized machine learning techniques.

The number of studies applying machine learning techniques to connectome data is growing rapidly, with a predominant focus on adult populations and conditions like Alzheimer’s disease. Prediction tasks have included identifying patients with various neurological and psychological disorders, as well as predicting individual physiological attributes [15]. Machine learning has been successfully applied for biomarker discovery in conditions such as autism [18], schizophrenia [19, 20], and even identifying individuals [21]. Several studies have also analyzed developmental populations using fMRI. For example, multivariate pattern analysis of resting state fMRI data could detect widespread connectivity differences in preterm versus term infants at term age. These machine learning methods can also estimate gestational age at birth and may predict neurodevelopmental outcomes at the individual level [22]. Other work has predicted brain age using functional networks and machine learning [23, 24, 25]. To date, to our knowledge, such developmental applications have not been explored in hippocampal subfields, although they could illuminate their development. The small size of the hippocampal subfields produces noisy signals, which pose challenges in the application of machine learning. In general, machine learning shows promise for the detection of connectome biomarkers, but applications remain limited thus far.

Transitioning from conventional statistical methods to AI-driven analytics presents exciting opportunities, as such

approaches are not paradigm-bound [26]. As [27] argues, improved predictive models can yield better theories, even without theoretical frameworks, although model limitations profoundly impact conclusions. Given machine learning’s demonstrated efficacy in elucidating complex relationships in large neuroimaging datasets, these techniques could illuminate hippocampal subfields functional connectivity. However, capturing dynamic interactions between regions requires specialized methods. Although linear modeling approaches [15] can effectively model connectivity, the high dimensionality of connectomes relative to sample sizes requires careful modeling pipelines such as feature selection and dimensionality reduction. As proposed in [28], our approach represents a principled solution to overcome ambiguities and validate insights using explainable AI for resting-state fMRI. Overall, machine learning provides opportunities to elucidate subfields connectivity, but realizing its full potential requires adapting techniques to address the complexity, dimensionality, and dynamics of connectome data.

In summary, machine learning techniques show considerable promise for deriving predictive biomarkers from fMRI connectome data, but realizing this potential in hippocampal subfields poses both opportunities and challenges. While machine learning has been fruitfully applied to extract biomarkers in various neurological conditions using whole-brain connectomes, subfield-specific applications remain limited thus far. We hypothesize that thoughtfully tailored machine learning techniques could give robust insights into the complex dynamic connectivity patterns of the hippocampal subfields. This could enable detection of sensitive developmental biomarkers based on subfields interactions with other brain regions. However, capturing subfields connectivity dynamics requires adapting analytical approaches to address noise and high-dimensionality. Building on encouraging applications of explainable AI to validate insights from resting-state fMRI, we propose specialized methods leveraging consensus of multiple high-performing predictive models. We aim to demonstrate that with a judicious application of machine learning, calibrated to the intricacies of subfields connectivity data, it is possible to derive sensitive developmental biomarkers from hippocampal subfields. Although difficult, elucidating subfields connectivity trajectories could provide key insights into typical and atypical development.

2 Methodology

The following section describes the pipeline used to compute the functional connectivity of the hippocampal subfields, our age-modeling with Machine Learning, and the resulting consensual explanations. The whole process is summarized in Figure 1.

2.1 Dataset

Data used in this study were obtained from the Human Connectome Project Development (HCP-Development)

dataset, which includes functional and structural data from 652 healthy young subjects (5-21 years) acquired on a Siemens 3T Prisma scanner [29, 30]. rs-fMRI data were acquired for 21-26 minutes using a multiband echoplanar imaging protocol (TR = 800 ms, TE = 37 ms, flip angle = 52°, multiband factor = 8, voxel size = 2 mm isotropic). High-resolution T1-weighted sMRI images were also obtained using a 3D MPRAGE sequence (TR = 2400 ms, TE = 2.14 ms, TI = 1000 ms, flip angle = 8°, voxel size = 0.8 mm isotropic). For the T2w scan, other parameters were TR/TE = 3200/564 ms, turbo factor = 314, and up to 25TRs allowed for motion-induced reacquisition. The rs-fMRI data from the HCP-Development dataset were pre-processed according to the minimal preprocessing pipeline by the Human Connectome Project, using the MSMAll multimodal surface registration. This included motion correction, registration to structural space, spatial normalization, global intensity normalization, and artifact/distortion correction [29, 30].

2.2 ROI Extraction

The rs-fMRI data were parcellated using both hippocampal subfields segmentations from the HSF tool [7], and the 360-region Glasser atlas. The Glasser atlas provides a state-of-the-art, comprehensive brain parcellation for exploring functional connectivity within the Human Connectome Project (HCP) dataset. Unlike traditional atlases based on cytoarchitecture or anatomical landmarks, the Glasser atlas integrates multimodal imaging data, inclusive of resting-state functional connectivity, to segment 180 functionally distinct sections per hemisphere [31]. This data-driven, functionally-informed approach resonates with the objectives of the present functional connectivity research. Specifically, the HCP Development dataset employs the Glasser atlas to assess alterations in functional connectivity throughout childhood and adolescence. Although the Glasser atlas parcellates the entire cortex, it lacks the granularity required to accurately outline minute subcortical structures such as the hippocampal subfields. To counteract this limitation, we augmented the Glasser atlas with high-resolution segmentation of the hippocampal subfields obtained from each participant’s T2-weighted structural MRI scan using HSF. This combined strategy leverages the power of the Glasser atlas for comprehensive brain parcellation while concurrently capturing fine-grained individual variability. Owing to the overly coarse resolution, we opted to merge CA2-3, given that CA2 occupies a minimal number of voxels at such a resolution. Specifically, HSF provided masks — for DG, CA1, CA2-3, and Subiculum — which superseded the “hippocampus” and “presubiculum” classes in the Glasser Atlas (Figure 2).

2.3 XAI Modeling

To model the relationship between functional connectivity and age, PyCaret [32] was first used to select optimal regression models. The model library included linear and nonlinear models (Table 1). The models were

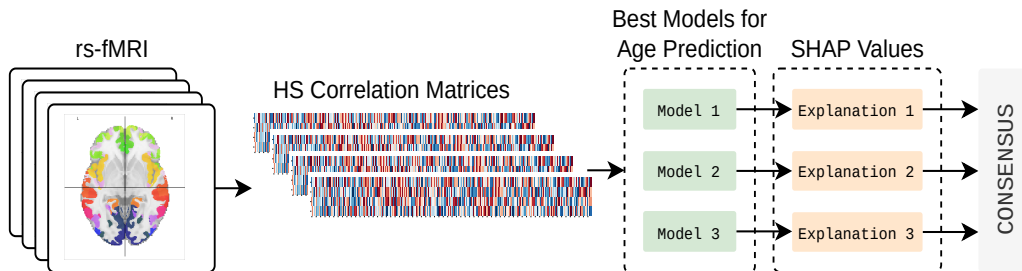


Figure 1: Overview of the pipeline for modeling age from hippocampal subfields (HS) connectivity. Resting-state fMRI data is used to derive correlation matrices capturing functional connectivity between hippocampal subfields and other brain regions. An ensemble of machine learning models is trained to predict age from connectivity features. Model explanations based on SHAP values are aggregated across models to obtain a consensus on developmental fingerprints characterizing the influence of age on specific connections.

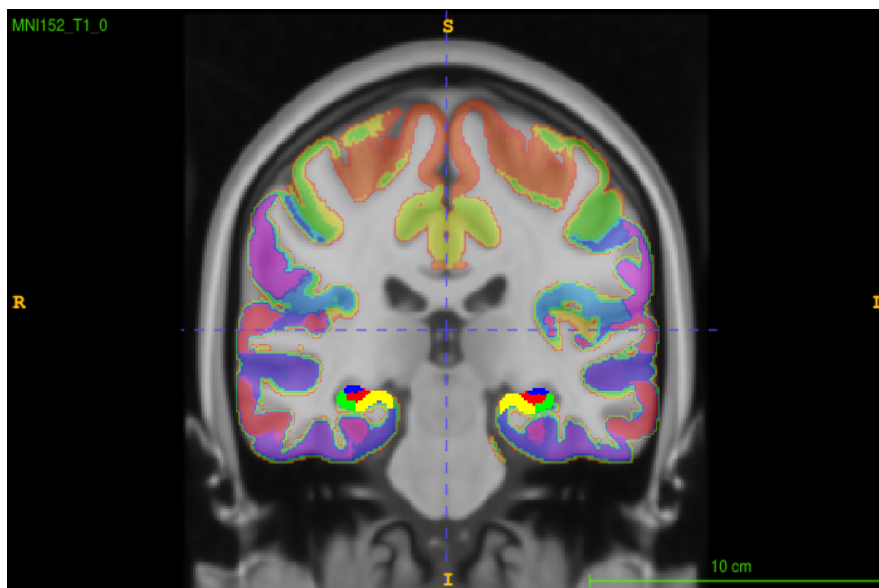


Figure 2: Illustrative example of a superposition between the Glasser Atlas and HSF's ROIs. The Glasser atlas is shown with transparency above the MNI152 template. ROIs from HSF are displayed with full opacity above the hippocampus, with the DG (red), CA1 (green), CA2-3 (blue), and Subiculum (yellow). Note that while the MNI template is displayed here, hippocampal subfields are calculated on a per-subject basis.

evaluated using a 10-fold cross-validation on the training data. For model selection and hyperparameter tuning, we used PyCaret's built-in support for grid search cross-validation. For each model, hyperparameters were tuned by grid search over reasonable ranges. We selected the top models based on their cross-validated RMSLE performance across the full sample size range. Models with the lowest prediction error and highest generalization were chosen. These top models were then re-trained using nested cross-validation with varying sample sizes from 16 to 588 samples. For each sample size and fold, the top 3 models were trained on the training data, and hyperparameters were tuned using grid search. The models were evaluated in test folds using the mean absolute error (MAE), the mean squared error (MSE). After retraining with nested CV, models performances were also evaluated on an independent holdout set, a test fold left out for the

final evaluation containing 60 carefully curated individuals, representing a uniform age distribution and gender parity.

To generate explanations for model predictions, the SHapley Additive Explanations (SHAP) values were calculated on the holdout set for each model at each sample size and fold. SHAP is a model interpretation method based on game theory and Shapley values [33]. It explains each prediction by computing the contribution of each feature. The intuition is that features that contribute most to changing the model output have the highest impact on predictions. SHAP connects to several local explanation techniques like LIME but with guarantees of consistency and accuracy. SHAP provides a framework for attributing importance values to each feature, in this case, the functional connections between the hippocampal subfields and other brain regions. By using SHAP, we can identify the specific connections that contribute most significantly to the predictive

Model	Acronym	Linearity
Logistic Regression	lr	Linear
Linear Discriminant analysis	lda	Linear
Ridge	ridge	Linear
Gaussian Naive Bayes	nb	Linear
Singular Vector Machine	svm	Linear
Singular Vector Machine (RBF)	rbfsvm	Non-linear
Decision Tree	dt	Non-linear
Quadratic Discriminant analysis	qda	Non-linear
Random Forest	rf	Non-linear
AdaBoost	ada	Non-linear
Extra Trees	et	Non-linear
Gradient Boosting	gbc	Non-linear
K-Nearest Neighbors	knn	Non-linear
Gaussian Process	gp	Non-linear
CatBoost	catboost	Non-linear
LightGBM	lightgbm	Non-linear
XGBoost	xgboost	Non-linear

Table 1: List of all models included in our experiments. For each dataset, all models are trained and tuned, and the best ones are kept for further analysis.

models, thereby gaining insights into the key drivers of the developmental changes in the functional connectivity of the hippocampal subfields during childhood. The mean absolute SHAP values were extracted as explanations of feature importance. Because of the multi-collinearity of the rs-fMRI timeseries, we computed the SHAP values using exact explainers or tree explainers when it was possible as they do not assume feature independence. The resulting SHAP values across models, sample sizes, and CV folds were then analyzed to assess agreement between explanations, as detailed in [28]. More specifically, with SHAP values for a model f_i noted as $\phi_{f_i}^{SHAP}$, we used the top j similarity and the weighted cosine similarity defined in [28]. With $S_n^{f_1}$ and $S_n^{f_2}$ the sets of the top j features selected by the models f_1 and f_2 for the n -th instance:

$$top_j \text{ similarity} = \frac{1}{j} \sum_{i=1}^j \frac{|S_i^{f_1} \cap S_i^{f_2}|}{|S_i^{f_2}|}$$

Similarly, with K features and mas_k the mean absolute SHAP value for each feature k , the weighted cosine similarity $w_{f_i,k}$ is defined as:

$$w_{f_i,k} = |\phi_{f_i,k}^{SHAP}| \cdot mas_k$$

The optimal number of features to consider for model interpretation was determined by computing the elbow point on the curve of top_j similarity versus j using the kneed library [34]. The inflection point indicates the number of top features, where additional features are showing a substantially decreasing the agreement between models, meaning they are too randomly distributed. This cutoff balances interpretability with stability of selected features across models.

3 Results

Machine learning analysis revealed valuable insights into how functional connectivity between hippocampal subfields and other cortical regions changes over development. Cross-validated model performance provided a global picture of how predictive connectivity patterns are of age, while model explanations based on SHAP values highlight the specific connections most influential in driving predictions. By aggregating explanations across models, a consensus emerged on the key developmental ‘‘fingerprints’’ that characterize the maturation of the hippocampal subfields.

3.1 Models performances

The top performing models achieved strong predictive accuracy, with an MAE of 2.156 years (Table 2). Model performance improved with larger sample sizes ($p = 0.021$), plateauing around 512 subjects, with a plateau as soon as 256 subjects for the best model (Figure 3). The bayesian regression model performed best, followed by gradient boosting regression, and support vector regression. All models showed significantly higher accuracy than chance (Table 2). Robust cross-validated performance demonstrates that functional connectivity contains rich information about the neurodevelopmental stage.

Model	MAE	MSE
<i>br</i>	2.156 ±0.000	6.570 ±0.000
<i>gbr</i>	2.459 ±0.004	9.022 ±0.060
<i>svr</i>	2.325 ±0.000	8.034 ±0.000
<i>bagging</i>	2.211 ±0.001	7.085 ±0.001
<i>dummy</i>	5.389 ±1.161	19.410 ±2.210

Table 2: Predictive performance of machine learning models for age prediction. Mean absolute error (MAE) and mean squared error (MSE) for age prediction from hippocampal subfields connectivity using bayesian regression (*br*), gradient boosting regression (*gbr*), support vector regression (*svr*). Errors are reported as mean ± std over 10-fold nested cross-validation.

3.2 Convergence and consensus of the explanations

Explanations from the machine learning models showed strong convergence, providing a robust consensus on the core developmental fingerprints. The agreement between models steadily improved as predictive accuracy increased (Figure 4, $p = 0.002$). However, the knee point analysis (Figure 5) showed a strong decrease in feature important agreement after 10 features. This indicates that, beyond a consistent set of 10 connections with the highest explanatory importance across all models, other features were inconsistently taken into account for age prediction.

3.3 Most important features for age-modeling

When considering average feature importances, the consensus set included a strong influence of the connections

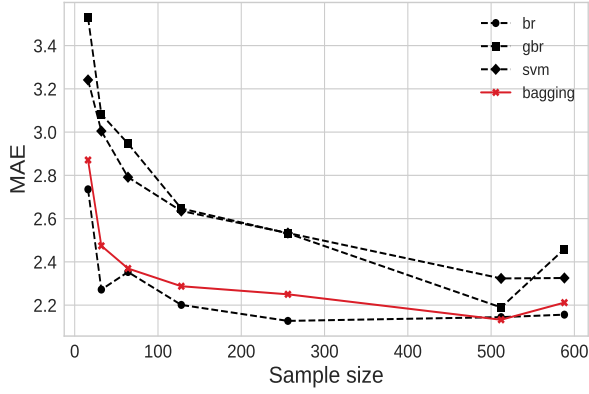


Figure 3: Learning curves depicting the mean absolute error (MAE) on age prediction as a function of sample size for the best performing machine learning models and their bagging ensemble. Performance generally improves with more training data before plateauing around $n=512$.

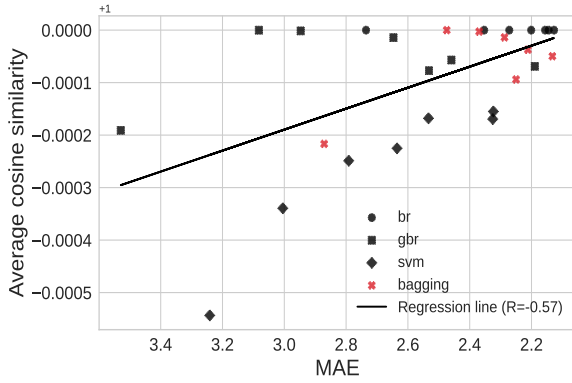


Figure 4: Model error decreases as the agreement between explanations increases. Each point represents a model trained in cross-validation at a given sample size. Agreement is quantified as the average cosine similarity of SHAP values between a model and the consensus.

implying the DG, followed by the Subiculum, CA1, and finally CA2-3. Interestingly, all ten predictive features were found to be negatively correlated with age. The ventral visual, dorsal attention, limbic, and somatomotor networks showed the strongest developmental changes in connectivity to the hippocampal subfields. In particular, the perirhinal cortex, frontal eye fields, parietal regions such as POS2, and areas of the lateral frontal cortex, including pOFC and OFC, showed highly predictive connectivity patterns, suggesting their functional coordination with hippocampal subfields matures over age.

4 Discussion

The results demonstrate that diverse machine learning models can accurately predict age based on patterns of functional connectivity between hippocampal subfields and distributed cortical regions. The best performing models

Rank	Node 1	Node 2
1	DG	Perirhinal Ectorhinal Cortex (L)
2	CA1	posterior OFC Complex (R)
3	CA1	Orbital Frontal Complex (R)
4	Sub	ParaHippocampal Area 1 (L)
5	Sub	Frontal Eye Field (R)
6	DG	posterior OFC Complex (R)
7	CA1	Parieto-Occipal Sulcus Area 2 (L)
8	CA2-3	Perirhinal Ectorhinal Cortex (L)
9	DG	Parieto-Occipital Sulcus Area 2 (R)
10	DG	Area 23c

Table 3: The 10 most predictive functional connections between hippocampal subfields and cortical regions for modeling age, based on model explanation agreement.

achieved a mean absolute error of only 2.156 years in predicting the chronological age of individuals based solely on functional connectivity of the hippocampal subfields. Performance improved with larger sample sizes before plateauing around 512 subjects. Explanations from the models converged on a consistent set of core developmental “fingerprints” characterized by a decrease in connectivity strength between hippocampal subfields and regions of the visual, dorsal attention, limbic, and somatomotor networks. Despite variability across modeling approaches, the convergence towards a consensus set of predictive features helps resolve the Rashōmon Effect wherein different equally performing models can yield divergent explanations. By quantifying agreement through similarity metrics and aggregating SHAP values across models, our analysis extracts stable insights from the machine learning explanations while accounting for instability across poorly performing models. This principled consensus approach provides greater confidence in the identified developmental fingerprints most influential in predicting individuals’ brain maturity from hippocampal subfields interactions. The convergence and consistency of explanations highlight the potential of using AI methods like explainable machine learning to chart typical and atypical neurodevelopment through predictive modeling of brain connectivity patterns.

The learning curve (Figure 3) showed prediction accuracy steadily improved with larger sample sizes, plateauing around 512 subjects. The best performing model was Bayesian Ridge regression, achieving a mean absolute error of 2.156 years and a mean squared error of 6.570 in predicting age (Table 2). The predictive performance of this linear model aligns with prior evidence that linear techniques effectively capture patterns in fMRI data [15]. The accuracy is notable given the focus on only a small subset of the full connectome, indicating our approach may work as a feature selection, using solely hippocampal subfields connectivity. Critically, model explanations showed stronger consensus with lower error (Figure 4), with high agreement for the top 10 features ($p = 0.002$). However, beyond 10 features, agreement dropped substantially, likely reflecting the curse of dimensionality where in

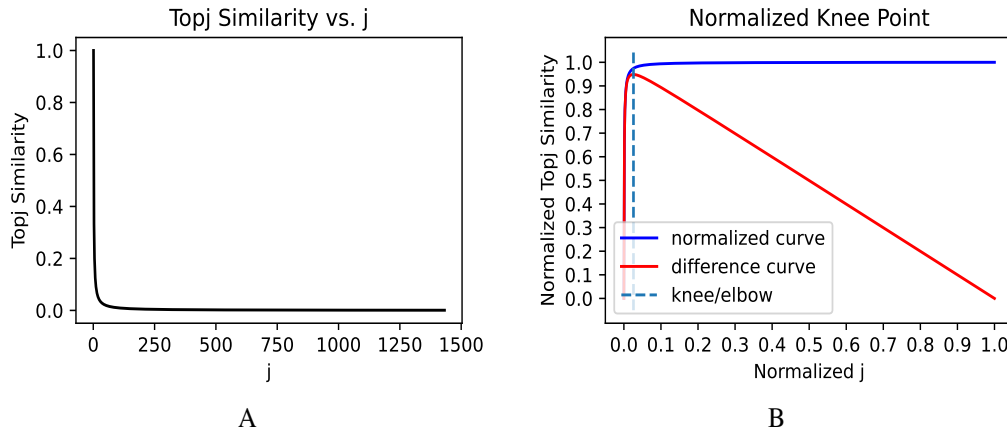


Figure 5: Analysis of model explanation agreement when considering an increasing number of top features. (A) Top- j similarity levels off after $\tilde{10}$ features. (B) Analysis of the knee point indicates the optimal cut-off point for the number of features.

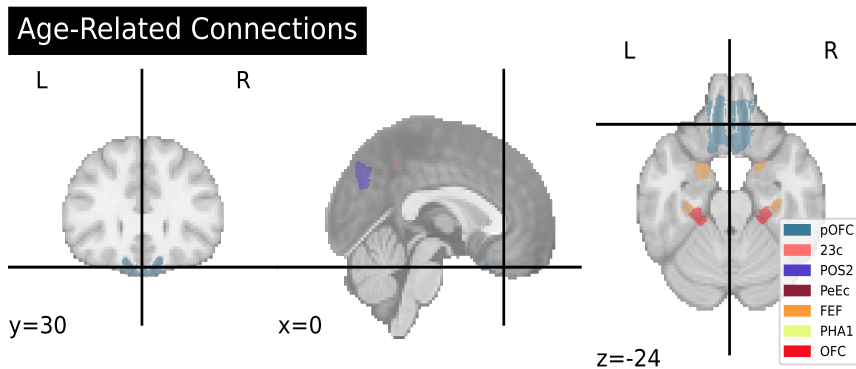


Figure 6: Cortical regions whose functional connectivity with hippocampal subfields is most predictive of age, based on the top connections in Table 3. Predictive connections originate from both left (L) and right (R) hemispheres. Abbreviations: POS2: Parieto-Occipital Sulcus Area 2; PeEc: Perirhinal Ectorhinal Cortex; pOFC: posterior Orbital Frontal Cortex; 23c: Area 23c; FEF: Frontal Eye Field; PHA1: Parahippocampal Area 1; OFC: Orbital Frontal Cortex.

high-dimensional data, most features are too unstable for consistent selection across models (Figure 5). Interestingly enough, we obtain nearly the same feature importances curve as in [21]. By quantifying explanation agreement, our analysis distinguishes reproducible developmental fingerprints from sporadic connections reflecting noise or model instability. In general, the sample size learning curve and the relationship between accuracy and explanation convergence validate the reliability of the identified core set of functional connectivity patterns that characterize the maturation of the hippocampal subfields.

The findings align with known maturation of hippocampal-cortical interactions supporting memory development, and the identified developmental fingerprints provide insights into hippocampal subfields interactions supporting memory maturation. The ranking of subfields importance provides further insights. The prominence of DG aligns with its role in pattern separation, which may mature later as perceptual discrimination abilities require refinement. The decreasing connectivity between the DG and medial tem-

poral cortical regions like perirhinal/ectorhinal cortex suggests pruning of inputs not necessary for perceptual discrimination that matures earlier. The importance of CA1 connectivity with prefrontal areas like orbital frontal cortex likely supports relational memory binding and retrieval, as suggested in [35] and [36]. The role of subiculum connections with its involvement in spatial navigation circuits as these skills improve [37]. Subiculum interactions with parahippocampal cortex and frontal eye fields likely underlie navigational circuit maturation. CA2-3's minimal importance fits with earlier maturation of its social memory functions. Overall, the findings suggest specialization and pruning of hippocampal subfields connectivity enables maturation of perceptual discrimination, executive control over memories, and spatial navigation. The details of predictive connectivity inform theories of hippocampal circuit dynamics during this critical window of neurocognitive growth. Decreasing connectivity strength likely reflects interactive specialization and neural pruning as the hippocampus becomes more optimized for episodic memory processing.

However, functional connectivity provides only an indirect measure of interactions. Developmental patterns warrant validation using task-based fMRI. Small hippocampal structures suffer from low signal-to-noise ratio in fMRI, likely restricting detected connectivity patterns. Exploring ultra-high field MRI could improve signal. Longitudinal data would elucidate trajectories. Studying clinical populations could assess connectivity disturbances reflecting developmental arrest. Another possible limitation is that functional connectivity differences could be influenced by changes related to cortical development [38, 39, 40], but the convergence of machine learning explanations on reproducible biomarkers highlights the potential of AI methods to chart typical and atypical neurodevelopment using predictive modeling of brain connectivity.

5 Conclusion

This study demonstrates that machine learning techniques can be valuably applied to resting-state fMRI data to reveal insights into hippocampal subfields development. By training an ensemble of predictive models on functional connectivity patterns, we were able to accurately estimate individuals' neurodevelopmental maturity. Explanations from the models converged on a set of core connectivity biomarkers that reliably track maturation. These developmental "fingerprints" align with known interactive specialization of hippocampal circuits and suggest pruning of unnecessary connections enabling optimization of memory functions. The multi-model explainable AI approach illustrates how, with careful analysis, machine learning can overcome challenges posed by noise and high-dimensionality to extract generalized patterns from functional connectivity data. The identified hippocampal subfields biomarkers could be extended to assess developmental disorders or the effects of interventions. Overall, the study highlights the potential of AI methods to chart typical and abnormal neurodevelopment using predictive modeling of brain connectivity.

References

- [1] Karl J. Friston. Functional and effective connectivity in neuroimaging: A synthesis. 2(1-2): 56–78, 1994. ISSN 10659471. doi: 10.1002/hbm.460020107. URL <https://onlinelibrary.wiley.com/doi/10.1002/hbm.460020107>.
- [2] Stephen M. Smith, Karla L. Miller, Gholamreza Salimi-Khorshidi, Matthew Webster, Christian F. Beckmann, Thomas E. Nichols, Joseph D. Ramsey, and Mark W. Woolrich. Network modelling methods for FMRI. 54(2):875–891, 2011. ISSN 10538119. doi: 10.1016/j.neuroimage.2010.08.063. URL <https://linkinghub.elsevier.com/retrieve/pii/S1053811910011602>.
- [3] Jordan Poppenk, Hallvard R. Evensmoen, Morris Moscovitch, and Lynn Nadel. Long-axis specialization of the human hippocampus. 17(5):230–240, 2013. ISSN 13646613. doi: 10.1016/j.tics.2013.03.005. URL <https://linkinghub.elsevier.com/retrieve/pii/S1364661313000673>.
- [4] Ylva Østby, Christian K. Tamnes, Anders M. Fjell, Lars T. Westlye, Paulina Due-Tønnessen, and Kristine B. Walhovd. Heterogeneity in Subcortical Brain Development: A Structural Magnetic Resonance Imaging Study of Brain Maturation from 8 to 30 Years. 29(38):11772–11782, 2009. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.1242-09.2009. URL <https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.1242-09.2009>.
- [5] Akiko Uematsu, Mie Matsui, Chiaki Tanaka, Tsutomu Takahashi, Kyo Noguchi, Michio Suzuki, and Hisao Nishijo. Developmental Trajectories of Amygdala and Hippocampus from Infancy to Early Adulthood in Healthy Individuals. *PLoS ONE*, 7(10): e46970, October 2012. ISSN 1932-6203. doi: 10.1371/journal.pone.0046970.
- [6] Gabriel Ziegler, Robert Dahnke, Lutz Jäncke, Rachel Aine Yotter, Arne May, and Christian Gaser. Brain structural trajectories over the adult lifespan. *Human Brain Mapping*, 33(10):2377–2389, October 2012. ISSN 10659471. doi: 10.1002/hbm.21374.
- [7] Clement Poiret, Antoine Bouyeure, Sandesh Patil, Antoine Grigis, Edouard Duchesnay, Matthieu Faillot, Michel Bottlaender, Frederic Lemaitre, and Marion Noulhiane. A fast and robust hippocampal subfields segmentation: HSF revealing lifespan volumetric dynamics. 17:1130845, 2023. ISSN 1662-5196. doi: 10.3389/fninf.2023.1130845. URL <https://www.frontiersin.org/articles/10.3389/fninf.2023.1130845/full>.
- [8] Sarah L. Blankenship, Elizabeth Redcay, Lea R. Dougherty, and Tracy Riggins. Development of hippocampal functional connectivity during childhood. 38(1):182–201, 2017. ISSN 1065-9471, 1097-0193. doi: 10.1002/hbm.23353. URL <https://onlinelibrary.wiley.com/doi/10.1002/hbm.23353>.
- [9] Mark H. Johnson. Functional brain development in humans. 2(7):475–483, 2001. ISSN 1471-003X, 1471-0048. doi: 10.1038/35081509. URL <https://www.nature.com/articles/35081509>.
- [10] Howard Eichenbaum. A cortical–hippocampal system for declarative memory. 1(1):41–50, 2000. ISSN 1471-003X, 1471-0048. doi: 10.1038/35036213. URL <https://www.nature.com/articles/35036213>.
- [11] Laura A. Libby, Arne D. Ekstrom, J. Daniel Ragland, and Charan Ranganath. Differential Connectivity of Perirhinal and Parahippocampal Cortices within Human Hippocampal Subregions Revealed by High-Resolution Functional Imaging. 32(19): 6550–6560, 2012. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.3711-11.2012. URL

- <https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.3711-11.2012>.
- [12] Tracy Riggins, Fengji Geng, Sarah L. Blankenship, and Elizabeth Redcay. Hippocampal functional connectivity and episodic memory in early childhood. 19:58–69, 2016. ISSN 18789293. doi: 10.1016/j.dcn.2016.02.002. URL <https://linkinghub.elsevier.com/retrieve/pii/S1878929315300827>.
- [13] Jessica S. Damoiseaux, Raymond P. Viviano, Peng Yuan, and Naftali Raz. Differential effect of age on posterior and anterior hippocampal functional connectivity. *NeuroImage*, 133:468–476, June 2016. ISSN 10538119. doi: 10.1016/j.neuroimage.2016.03.047.
- [14] Lingfei Tang, Patrick J. Pruitt, Qijing Yu, Roya Homayouni, Ana M. Daugherty, Jessica S. Damoiseaux, and Noa Ofen. Differential Functional Connectivity in Anterior and Posterior Hippocampus Supporting the Development of Memory Formation. 14:204, 2020. ISSN 1662-5161. doi: 10.3389/fnhum.2020.00204. URL <https://www.frontiersin.org/article/10.3389/fnhum.2020.00204/full>.
- [15] Colin J. Brown and Ghassan Hamarneh. Machine Learning on Human Connectome Data from MRI. 2016. URL <http://arxiv.org/abs/1611.08699>.
- [16] Maxwell L. Elliott, Annchen R. Knodt, David Ireland, Meriwether L. Morris, Richie Poulton, Sandhya Ramrakha, Maria L. Sison, Terrie E. Moffitt, Avshalom Caspi, and Ahmad R. Hariri. What Is the Test-Retest Reliability of Common Task-Functional MRI Measures? New Empirical Evidence and a Meta-Analysis. 31(7):792–806, 2020. ISSN 0956-7976, 1467-9280. doi: 10.1177/0956797620916786. URL <http://journals.sagepub.com/doi/10.1177/0956797620916786>.
- [17] Philip A. Kragel, Xiaochun Han, Thomas Edward Kraynak, Peter J. Gianaros, and Tor D Wager. fMRI can be highly reliable, but it depends on what you measure. 2020. doi: 10.31234/osf.io/9eaxk. URL <https://osf.io/9eaxk>.
- [18] Alexandre Abraham, Michael P. Milham, Adriana Di Martino, R. Cameron Craddock, Dimitris Samaras, Bertrand Thirion, and Gaël Varoquaux. Deriving reproducible biomarkers from multi-site resting-state data: An Autism-based example. 147:736–745, 2017. ISSN 10538119. doi: 10.1016/j.neuroimage.2016.10.045. URL <https://linkinghub.elsevier.com/retrieve/pii/S1053811916305924>.
- [19] Mohammad R. Arbabshirani, Kent A. Kiehl, Godfrey D. Pearlson, and Vince D. Calhoun. Classification of schizophrenia patients based on resting-state functional network connectivity. 7, 2013. ISSN 1662-453X. doi: 10.3389/fnins.2013.00133. URL <http://journal.frontiersin.org/article/10.3389/fnins.2013.00133/abstract>.
- [20] Danielle S. Bassett, Brent G. Nelson, Bryon A. Mueller, Jazmin Camchong, and Kelvin O. Lim. Altered resting state complexity in schizophrenia. 59(3):2196–2207, 2012. ISSN 10538119. doi: 10.1016/j.neuroimage.2011.10.002. URL <https://linkinghub.elsevier.com/retrieve/pii/S1053811911011633>.
- [21] Andrew Hannum, Mario A. Lopez, Saúl A. Blanco, and Richard F. Betzel. High-accuracy machine learning techniques for functional connectome fingerprinting and cognitive state decoding. page hbm.26423, 2023. ISSN 1065-9471, 1097-0193. doi: 10.1002/hbm.26423. URL <https://onlinelibrary.wiley.com/doi/10.1002/hbm.26423>.
- [22] Christopher D. Smyser, Nico U.F. Dosenbach, Tara A. Smyser, Abraham Z. Snyder, Cynthia E. Rogers, Terrie E. Inder, Bradley L. Schlaggar, and Jeffrey J. Neil. Prediction of brain maturity in infants using machine-learning algorithms. 136:1–9, 2016. ISSN 10538119. doi: 10.1016/j.neuroimage.2016.05.029. URL <https://linkinghub.elsevier.com/retrieve/pii/S1053811916301483>.
- [23] Anqi Qiu, Annie Lee, Mingzhen Tan, and Moo K. Chung. Manifold learning on brain functional networks in aging. 20(1):52–60, 2015. ISSN 13618415. doi: 10.1016/j.media.2014.10.006. URL <https://linkinghub.elsevier.com/retrieve/pii/S1361841514001522>.
- [24] Franziskus Liem, Gaël Varoquaux, Jana Kynast, Frauke Beyer, Shahrzad Kharabian Masouleh, Julia M. Huntenburg, Leonie Lampe, Mehdi Rahim, Alexandre Abraham, R. Cameron Craddock, Steffi Riedel-Heller, Tobias Luck, Markus Loeffler, Matthias L. Schroeter, Anja Veronica Witte, Arno Villringer, and Daniel S. Margulies. Predicting brain-age from multimodal imaging data captures cognitive impairment. 148:179–188, 2017. ISSN 10538119. doi: 10.1016/j.neuroimage.2016.11.005. URL <https://linkinghub.elsevier.com/retrieve/pii/S1053811916306103>.
- [25] Shi Gu, Theodore D. Satterthwaite, John D. Medaglia, Muzhi Yang, Raquel E. Gur, Ruben C. Gur, and Danielle S. Bassett. Emergence of system roles in normative neurodevelopment. 112(44):13681–13686, 2015. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1502829112. URL <https://pnas.org/doi/full/10.1073/pnas.1502829112>.
- [26] Gaël Varoquaux and Russell A Poldrack. Predictive models avoid excessive reductionism in cognitive neuroimaging. 55:1–6, 2019. ISSN 09594388. doi: 10.1016/j.conb.2018.11.002. URL <https://linkinghub.elsevier.com/retrieve/pii/S0959438818301089>.
- [27] Gaël Varoquaux. AI as statistical methods for imperfect theories. 2021.

- [28] Clement Poirer, Antoine Grigis, Justin Thomas, and Marion Noulhiane. Can we Agree? On the Rashomon Effect and the Reliability of Post-Hoc Explainable AI. URL <http://arxiv.org/abs/2308.07247>.
- [29] Michael P. Harms, Leah H. Somerville, Beau M. Ances, Jesper Andersson, Deanna M. Barch, Matteo Bastiani, Susan Y. Bookheimer, Timothy B. Brown, Randy L. Buckner, Gregory C. Burgess, Timothy S. Coalson, Michael A. Chappell, Mirella Dapretto, Gwenaëlle Douaud, Bruce Fischl, Matthew F. Glasser, Douglas N. Greve, Cynthia Hodge, Keith W. Jamison, Saad Jbabdi, Sridhar Kandala, Xiufeng Li, Ross W. Mair, Silvia Mangia, Daniel Marcus, Daniele Mascali, Steen Moeller, Thomas E. Nichols, Emma C. Robinson, David H. Salat, Stephen M. Smith, Stamatis N. Sotiropoulos, Melissa Terpstra, Kathleen M. Thomas, M. Dylan Tisdall, Kamil Ugurbil, Andre Van Der Kouwe, Roger P. Woods, Lilla Zöllei, David C. Van Essen, and Essa Yacoub. Extending the Human Connectome Project across ages: Imaging protocols for the Lifespan Development and Aging projects. *NeuroImage*, 183: 972–984, December 2018. ISSN 10538119. doi: 10.1016/j.neuroimage.2018.09.060.
- [30] Leah H. Somerville, Susan Y. Bookheimer, Randy L. Buckner, Gregory C. Burgess, Sandra W. Curtiss, Mirella Dapretto, Jennifer Stine Elam, Michael S. Gaffrey, Michael P. Harms, Cynthia Hodge, Sridhar Kandala, Erik K. Kastman, Thomas E. Nichols, Bradley L. Schlaggar, Stephen M. Smith, Kathleen M. Thomas, Essa Yacoub, David C. Van Essen, and Deanna M. Barch. The Lifespan Human Connectome Project in Development: A large-scale study of brain connectivity development in 5–21 year olds. *NeuroImage*, 183:456–468, December 2018. ISSN 10538119. doi: 10.1016/j.neuroimage.2018.08.050.
- [31] Matthew F. Glasser, Timothy S. Coalson, Emma C. Robinson, Carl D. Hacker, John Harwell, Essa Yacoub, Kamil Ugurbil, Jesper Andersson, Christian F. Beckmann, Mark Jenkinson, Stephen M. Smith, and David C. Van Essen. A multi-modal parcellation of human cerebral cortex. 536(7615):171–178, 2016. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature18933. URL <https://www.nature.com/articles/nature18933>.
- [32] Moez Ali. *PyCaret: An open source, low-code machine learning library in Python*, April 2020. URL <https://www.pycaret.org>. PyCaret version 1.0.0.
- [33] Scott Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. 2017. URL <http://arxiv.org/abs/1705.07874>.
- [34] Ville Satopaa, Jeannie Albrecht, David Irwin, and Barath Raghavan. Finding a "Kneedle" in a Haystack: Detecting Knee Points in System Behavior. In *2011 31st International Conference on Distributed Computing Systems Workshops*, pages 166–171. IEEE, 2011. ISBN 978-1-4577-0384-3. doi: 10.1109/ICDCSW.2011.20. URL <http://ieeexplore.ieee.org/document/5961514/>.
- [35] Myrcea A. De S. Tilger, Renan B. Gaiardo, and Suzete M. Cerutti. Inactivation of the dorsal CA1 hippocampus impairs the consolidation of discriminative avoidance memory by modulating the intrinsic and extrinsic hippocampal circuitry. 128:102209, 2023. ISSN 08910618. doi: 10.1016/j.jchemneu.2022.102209. URL <https://linkinghub.elsevier.com/retrieve/pii/S0891061822001399>.
- [36] Krubeal Danieli, Alice Guyon, and Ingrid Bethus. Episodic Memory formation: A review of complex Hippocampus input pathways. 126:110757, 2023. ISSN 02785846. doi: 10.1016/j.pnpbp.2023.110757. URL <https://linkinghub.elsevier.com/retrieve/pii/S027858462300043X>.
- [37] Jean Simonnet and Michael Brecht. Burst Firing and Spatial Coding in Subicular Principal Cells. 39 (19):3651–3662, 2019. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.1656-18.2019. URL <https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.1656-18.2019>.
- [38] Nitin Gogtay, Jay N. Giedd, Leslie Lusk, Kira M. Hayashi, Deanna Greenstein, A. Catherine Vaituzis, Tom F. Nugent, David H. Herman, Liv S. Clasen, Arthur W. Toga, Judith L. Rapoport, and Paul M. Thompson. Dynamic mapping of human cortical development during childhood through early adulthood. 101(21):8174–8179, 2004. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0402680101. URL <https://pnas.org/doi/full/10.1073/pnas.0402680101>.
- [39] Tonya White, Shu Su, Marcus Schmidt, Chiu-Yen Kao, and Guillermo Sapiro. The development of gyri-fication in childhood and adolescence. 72(1):36–45, 2010. ISSN 02782626. doi: 10.1016/j.bandc.2009.10.009. URL <https://linkinghub.elsevier.com/retrieve/pii/S0278262609002012>.
- [40] Sophia Mueller, Danhong Wang, Michael D. Fox, B.T. Thomas Yeo, Jorge Sepulcre, Mert R. Sabuncu, Rebecca Shafee, Jie Lu, and Hesheng Liu. Individual Variability in Functional Connectivity Architecture of the Human Brain. 77(3):586–595, 2013. ISSN 08966273. doi: 10.1016/j.neuron.2012.12.028. URL <https://linkinghub.elsevier.com/retrieve/pii/S0896627313000044>.

4 Discussion

The present dissertation made several key advances that enabled the detailed characterization of the anatomy and functional connectivity of the subfields of the hippocampus throughout the lifespan. To begin, we developed novel deep learning architectures for automated subfields segmentation in structural MRI. As a first step, we explored Capsule Networks (Section 3.1.1, 1st paper), which can inherently model anatomical variations through built-in equivariances. However, scalability challenges led us to integrate the strengths of Capsule Networks into more efficient convolutional architectures. This resulted in HSF (Section 3.1.2, 2nd paper), which leverages a heterogeneous training database, learned from human feedbacks, and employed state-of-the-art computer vision techniques such as visual attention to efficiently and accurately delineate subfields boundaries. HSF is open source and publicly accessible at <https://hsf.rtf.io>.

Enabled by the development of HSF, we conducted an unprecedented large-scale analysis of hippocampal subfields volumes across the lifespan, hypothesizing distinct developmental patterns and asynchronous maturation of subfields aligned with their cognitive roles (2nd paper). This revealed several key anatomical findings:

- Application of HSF to 3,750 MRI scans aged 4-100+ years provided unprecedented insights into hippocampal subfields volumetric development (Section 3.1.2, Figure 3.1),
- Modeling revealed differential nonlinear trajectories for each subfield, rather than treating the hippocampus as a developmentally homogeneous structure (Poiret, Bouyeure, et al., 2023), The prolonged development of the subfields parallels behavioral refinements in learning and memory over childhood and adolescence (Lavenex & Banta Lavenex, 2013; Lee et al., 2016; Shing & Lindenberger, 2011),
- For example, the dentate gyrus, pivotal for pattern separation and neurogenesis, exhibited the strongest age-related volume changes and rapid growth extending into the second decade of life,
- In contrast, the subiculum showed early volumetric maturation by age 5, followed by susceptibility to atrophy from 60-70 years,
- These differential patterns likely arise from asynchronous cytoarchitectonic and myeloarchitectonic development.

To reliably model the functional maturation of the hippocampal subfields, we implemented a rigorous computational approach using explainable AI (Section 3.2.1, 3rd paper) based on SHappley Additive exPlanations. We first introduced a framework to find consensus between competing explanations, addressing the Rashōmon effect, where different models yield divergent explanations for the same prediction (Breiman, 2001). This approach was designed to overcome noise and high-dimensionality in resting-state fMRI connectivity data to chart typical development. The key functional findings enabled by this approach include:

- Application in 588 youths revealed decreasing interactions between hippocampal subfields and sensory and attentional cortical regions (Section 3.2.2, Figure 4.1, 4th paper). The perirhinal cortex, frontal eye fields, posterior parietal cortex, and lateral frontal areas showed particularly predictive developmental changes in coordination with the subfields,
- Convergence of explanations from the model ensemble highlights ten core connections most influential in predicting age,

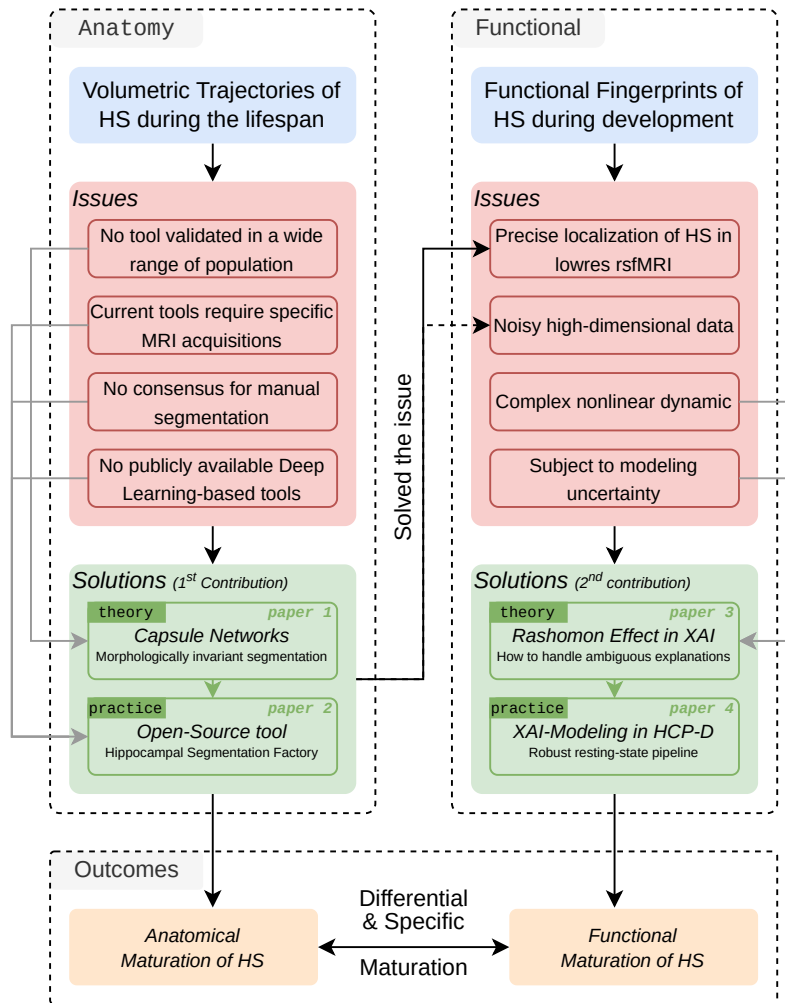


Figure 4.1: Architecture of the Ph.D. dissertation aimed at characterizing anatomical and functional changes in hippocampal subfields across the lifespan. The work was organized into four main parts with its accompanying paper: (1) a methodological exploration of innovative and disruptive segmentation methods; (2) the development of an automated deep learning tool for hippocampal subfields segmentation from structural MRI; (3) the application of this model to explore volumetric developmental trajectories in a large lifespan dataset; (4) the investigation of the maturation of functional connectivity of the subfields using resting-state fMRI and explainable AI modeling. Together, these novel tools and analytical approaches provided insights into the prolonged development of hippocampal circuits supporting gains in memory and cognition.

-
- The dentate gyrus showed the greatest change in connectivity, consistent with its role in pattern separation and prolonged structural development,
 - Decreased connectivity likely reflects optimization of subfields interactions with cortical networks as hippocampal circuits become optimized for episodic memory and spatial navigation processing,
 - Finding consensus between competing computational explanations mitigates the Rashōmon effect and provides confidence in identifying reproducible developmental biomarkers of brain maturity.

Methodologically, this work makes several key contributions that can guide future hippocampal subfields research, provides tools that can be ported to other domains, and guide machine learning practitioners in their use of XAI methods:

- Development of HSF, an efficient and modular deep learning pipeline for hippocampal subfields segmentation that is evolvable and generalizable, and that can be ported to other brain regions with minimal work,
- Introduction of a rigorous explainable AI framework to elicit reliable knowledge from multivariate neuroimaging data and handle model ambiguity. Finding consensus between competing explanations mitigates issues like the Rashōmon effect,
- Validation approaches including quantification of agreement between model explanations to overcome instability and ensure reproducibility of discovered biomarkers,
- Guidelines for the responsible development and deployment of AI tools in the medical field, emphasizing the need for transparency, efficiency, and public availability to realize benefits,
- Demonstration of how machine learning can scale studies relating subfields anatomy to cognition, aging, and disease by enabling large-scale efficient segmentation.

Together, the complementary analyses provide unmatched insight into hippocampal subfields development. Tight correspondence exists between structural maturation timecourses and functional connectivity changes. For example, the dentate gyrus, with its protracted structural development, exhibits the greatest alterations in functional connectivity. On the contrary, CA2-3 was the least age-related region, both anatomically and functionally. Decreasing connectivity likely reflects optimization of subfields interactions with cortical networks as memory abilities mature. Overall, modeling subfields heterogeneity reveals coordinated developmental pathways between anatomy and function that strengthen episodic memory.

In summary, the dissertation provides innovative tools and integrative modeling approaches to elucidate hippocampal subfields structure-function relationships over the lifespan. The findings reconcile disconnected results and underscore the need to model subfields diversity. This framework serves as a springboard for large-scale inquiries into hippocampal subfields contributions in health and disease.

Significance and implications

In an initial exploration, we implemented Capsule Networks for hippocampal subfields segmentation given their inherent robustness from built-in equivariances (Sabour et al., 2017). We developed a novel 3D Capsule Network architecture with attention gating (3D-AGSCaps, <https://github.com/clementpoiret/3d-agscaps>) and compared performance to 3D convolutional neural networks (CNNs) using three manually labeled datasets. Under typical conditions, 3D-AGSCaps performed comparably to CNNs, validating its feasibility for subfields segmentation. However, a representative use-case is the incomplete hippocampal inversion (IHI). Normally during brain development, the hippocampus undergoes a process called “inversion”, where it changes from its original position and shape to its final “inverted” form. While this process is generally completed before birth, in some individuals, this inversion process is not fully completed, leading to what is termed as “incomplete hippocampal inversion”. IHI is not necessarily a pathological condition. It is often considered a variant of normal brain anatomy, but causes segmentation issues, and therefore we decided to use this condition as a proof-of-concept for our methodology. With increasing rotational perturbations simulating anatomical variations such as incomplete hippocampal inversion, 3D-AGSCaps demonstrated significantly higher overlap (Dice similarity 0.460 vs 0.331 for CNNs at 90° rotation), lower surface distance (Hausdorff distance 25.868 vs 29.960 voxels), and greater volumetric similarity (0.660 vs 0.501). This greater robustness likely arises from capsules’ equivariance, enabling recognition despite transformations. We made seminal progress in the field with the open-sourcing of the first 3D implementation of capsules, the SMSquash function which enables the segmentation of multiple regions of interest, or the attention between capsules, which were later reused in other articles (e.g., Liu et al., 2023; Viriyasaranon and Choi, 2022) that we used as a way to route the information between the capsules to replace the original Routing-by-Agreement algorithm.

The current trend is to empower deep learning neural networks with ever more expressive powers. Although the current trend concentrates on unconstrained — or unbiased — architectures such as Transformers (Dosovitskiy et al., 2021; Vaswani et al., 2017), resulting in incrementally powerful AI models at the expense of data requirements and energy consumption, we are staunch advocates for the notion that deep learning neural networks stand to gain from encapsulating such advanced capabilities within architectures that exhibit them in an inherent manner. Capsule Networks were proposed with this very specific goal in mind. If their computational demands currently limit widespread usage, as summarized by Haq et al., 2023, other alternatives are yet to be explored in neuroimaging, such as Gauge Equivariant Neural Networks (Cohen et al., 2019), or more recent work based on spherical harmonics (Diaz et al., 2023). Future work could explore architectural improvements, such as employing self-routing (Hahn et al., 2019), or dot product routing (Tsai et al., 2020) instead of original dynamic routing, to improve efficiency. Finally, since the main drawback is the vector representation that causes high RAM usage, we suggest the exploration of pruning and quantization methods, as explored in Costa et al., 2022 and Marchisio et al., 2022. As CapsNets become more optimized, they may find utility in handling unusual morphologies.

To develop a more efficient and scalable solution while retaining robustness benefits, we transitioned back from CapsNets to convolutional architectures, ultimately resulting in HSF. We integrated the strengths of CapsNets — including attention gating and switchnorm layers — and model compression techniques thanks to our close collaboration with NeuralMagic and their SparseML library, to which we contributed (Kurtz et al., 2020). Compared to competing tools, HSF achieved superior

overlap, lower outliers, and greater volume similarity. We designed HSF to be as modular and future-proof as possible, embracing open-source and transparency standards. To our knowledge, we were the first to augment training with human feedback on difficult cases, improving generalization. Although we did not use reinforcement learning as in Bai et al., 2022, we adopted an iterative approach, voluntarily segmenting notoriously difficult hippocampi presenting abnormalities such as heavy motion artifacts, anatomical anomalies, or heavy hippocampal sclerosis (Haeger et al., 2020; Lagarde et al., 2021). The model hub enables easy incorporation of advances, ensuring evolvability. Future work could explore the incorporation of geometric deep learning to capture finer anatomical details or even try completely different approaches such as SAM-based approaches (Kirillov et al., 2023). Generalizability testing is warranted across scanners, resolutions, and clinical populations. As datasets grow exponentially, optimizing efficiency without sacrificing accuracy remains an important aim. Overall, HSF provides an accurate, efficient, and scalable solution to explore the hippocampal subfields across diverse studies and populations.

Application of HSF to 3,750 MRI scans aged 4-100+ years provided unprecedented insights into hippocampal subfields volumetric development. Modeling revealed differential nonlinear trajectories for each subfield, rather than treating the hippocampus as a developmentally homogeneous structure (Poiret, Bouyeure, et al., 2023). The prolonged development of the subfields parallels behavioral refinements in learning and memory over childhood and adolescence (Lavenex & Banta Lavenex, 2013; Lee et al., 2016; Shing & Lindenberger, 2011). For example, the DG, pivotal for pattern separation and place of adult neurogenesis, exhibited the strongest age-related volume changes and a rapid growth extending into the second decade of life. In contrast, the subiculum showed early volumetric maturation by age 5, followed by susceptibility to atrophy in later decades. These differential patterns likely arise from asynchronous cytoarchitectonic and myeloarchitectonic development. Precise segmentation enables relating subfields volumes to cognitive abilities and disorders that affect the medial temporal lobe (Bouyeure et al., 2021). Future work should examine additional measures such as cortical thickness, quantitative T1 mapping, and shape morphometry (Lynch et al., 2019). Relating such quantitative markers to in-vivo cytoarchitecture could better elucidate typical and disturbed developmental pathways. Longitudinal analyzes are also needed to disentangle aging effects from cohort variations. Overall, characterizing subfields heterogeneity provides insights into typical and atypical neurodevelopment underlying memory. Although HSF enables efficient subfields segmentation at scale, several frontiers remain to enhance anatomical modeling. Critically, joint analysis of multimodal MRI, coupling structural, diffusion, and functional data, would enable a more complete perspective on hippocampal maturation. If the cytoarchitectony is not accessible in MRI, and because each subfield has its function, it should be possible to refine the segmentation process by jointly modeling both structural and functional data. Ultimately, optimized segmentation will help unlock the hippocampal subfields' contributions to cognition, development, and disease.

The resting state functional connectivity analysis revealed that models could accurately predict an individual's age based solely on patterns of hippocampal subfields connectivity, achieving a mean absolute error of only 2.156 years. The Rashōmon framework we introduced is key to extracting reliable insights from noisy high-dimensional data by quantifying model agreement. This helped overcome instability and ambiguity in machine learning explanations. Explanations from various models showed convergence, identifying a consistent developmental “fingerprint” — distinctive patterns that reliably distinguish individuals based on their age — characterized by decreasing connectivity strength between hippocampal subfields and distributed cortical regions. This aligns with the known maturation of hippocampal-cortical interactions (Riggins et al., 2016) and suggests

long specializations up until early adulthood that optimize episodic memory and spatial navigation. The DG showed the greatest change in connectivity, consistent with its role in pattern separation (Leal & Yassa, 2018), which probably continues refinement in this age range. The decreased connectivity of CA1 with the prefrontal and parietal areas supports the maturation of relational binding and retrieval (Schlichting et al., 2014), while the minimal importance of CA2-3 — implied in social memory functions (Hitti & Siegelbaum, 2014) — could suggest that they mature earlier than 4 years, leaving their maturational dynamics outside of the scope of our datasets. However, since the analysis relied on functional connectivity, more validation is needed using task-based fMRI. Here again, longitudinal data would clarify developmental trajectories, and studying clinical groups with memory disorders could reveal altered connectivity.

The results have implications for linking connectivity to cognition and behavior. Identified “fingerprints” probably reflect the optimization of the hippocampus circuit enabling episodic memory. DG and CA1 connectivity suggest a protracted maturation of pattern separation and relational binding into adulthood. Subicular connectivity potentially reflects improving spatial navigation and memory retrieval. Together, this elucidates neural mechanisms that support the development of learning, memory, and spatial abilities. Studying clinical groups with conditions such as neonatal hypoxia, temporal lobe epilepsy, or Alzheimer’s disease, would help assess whether altered subfields connectivity disrupts cognition and behavior. Longitudinal tracking of connectivity biomarkers could aid in early diagnosis and monitoring. The approach demonstrates the promise of using explainable AI to derive sensitive developmental indices. However, rigorous validation is required to confirm that connectivity changes translate to memory proficiency. Multimodal imaging and cognitive testing are needed to directly relate functional biomarkers to behavior and structure. Tasks probing pattern separation and completion would help assess the significance of identified DG and CA1 connectivity shifts, and navigation tasks could validate implications of subicular network changes.

Overall, the observed lifespan trajectories of hippocampal subfields volumes and functional connectivity align with known synaptic dynamics. In the same age range as in our study, overall number of synapses has been shown to decrease, reflecting synaptic pruning (Huttenlocher & Dabholkar, 1997). This pruning is related to previously reported changes in functional connectivity within the brain (Fair et al., 2009). During early development, there is a rapid increase in synapses, leading to a highly interconnected network with high functional connectivity. However, not all of these connections are efficient or necessary. Synaptic pruning helps refine these connections, removing weaker or less efficient synapses and strengthening the more active ones (Dosenbach et al., 2010). This leads to a decrease in overall connectivity, but an increase in the efficiency and specificity of the remaining connections. Consistent with this pruning, we found a decrease in functional connectivity between the hippocampal subfields and the parahippocampal cortices with age. The hippocampus becomes less densely connected but more specialized and efficient, with stronger connections between regions that frequently interact with its subfields. While we conjecture that synaptic pruning leads to decreased connectivity, it is an essential developmental process to form a more specialized neural network. Abnormalities in this process, leading to too much or too little pruning, have been associated with several neurological and psychiatric disorders, such as schizophrenia and autism (Sakai, 2020). Overall, the converging anatomical and functional evidence underscores the importance of synaptic refinement in optimizing hippocampal circuits that support memory over the lifespan.

As summarized in the Figure 4.1, this dissertation successfully met its core objectives of studying the anatomical and functional development of the hippocampal subfields. Through innovative tools like HSF and rigorous computational modeling, the prolonged maturation and coordinated specialization of subfield structure and interactions were characterized from early childhood into older adulthood. The tight correspondence between subfield-specific volumetric trajectories and changes in intrinsic connectivity provides critical insights into the development of hippocampal circuits supporting gains in memory and cognition. While important contributions were made towards these aims, several limitations should be acknowledged, along with exciting future research directions to build on the present work.

Limitations and future directions

Although this work made important contributions towards characterizing hippocampal subfields anatomy and function, several limitations must be acknowledged along with exciting future research directions. In addition to the already mentioned and discussed limitations of Capsule Networks and HSF, manual protocols still require further standardization to enable clearer cross-study comparisons and to facilitate the training process of tools similar to ours. Fortunately, initiatives such as the Hippocampal Subfields Group (<https://hippocampalsubfields.com>) strive to achieve consensus in this intricate field of research.

The developmental and aging trajectories were examined in cross-sectional data. While the focus was on group-level developmental patterns, longitudinal studies tracking the same individuals over time would provide stronger evidence of within-subject changes. Also, the participants in the HCP datasets, while large in number, lacked diversity in ethnicity, socioeconomic status, etc. Studies in more representative populations are needed. Furthermore, only volumetric features of the subfields were considered, but we believe that volumetry has to be coupled to shape analyses to provide additional and complementary information (Lynch et al., 2019). Finally, examining quantitative MRI measures such as T1 relaxation time could give insight into intracortical myelination (Vos de Wael et al., 2018).

For functional connectivity modeling, establishing generalizability across an array of datasets and explanation methods, including LIME, remains a crucial undertaking. Likewise, longitudinal data are required to robustly establish developmental timecourses. Additionally, we introduced a weighted cosine similarity metric, which has been shown to be artificially high on very high-dimensional datasets where many features have a mean absolute SHAP value near zero, as is the case for functional connectivity analyses. Improved metrics in high-dimensional data could enhance the reliability of the method. Lastly, correlations between BOLD time series, though easy to interpret, may oversimplify the “actual” functional connectivity. Exploring causality through Granger modeling, or simply other customary metrics such as partial correlations or correlations in tangent space, could provide additional evidence concerning the observed connectivity patterns. Although only resting-state fMRI was used, incorporating task fMRI could help determine the functional significance of the connectivity fingerprints. Additionally, the spatial resolution of the fMRI limits the functional segregation of the subfields, but ultra-high field MRI at 7T or beyond, improving signal-to-noise ratios and enhancing contrasts, could help go further in the analysis. Another possible investigation path is the multimodal integration with EEG/MEG which could

give richer characterization of subfields temporal dynamics. Lastly, only typical development was characterized. Comparing to clinical groups could identify disturbed connectivity related to memory disorders.

A fundamental limitation is the inability of structural MRI to directly examine cellular-level factors that influence subfields trajectories, such as neurogenesis, apoptosis, or synaptic pruning. Delving into how genetic and epigenetic factors shape the anatomy of the subfields would yield profound insights. Methodologically, transfer learning techniques could take advantage of animal MRIs to enhance model generalization. Broadening HSF to encompass animal models could yield preclinical biomarkers, and modeling hippocampal subfields in a cross-species context could shed light on their evolutions over time.

In a nutshell, this study underscores the necessity to integrate multimodal data spanning from molecular to cognitive levels. Open-source tools like HSF aspire to catalyze the next wave of hippocampal research by enabling large-scale studies correlating structure, function, and behavior, notwithstanding the fact that HSF could easily be ported to other brain regions without major issues.

Key methodological contributions

The experimental work highlights several key methodological insights regarding the use of machine learning and deep learning for hippocampal subfields segmentation and functional connectivity mapping. First, HSF emphasizes the importance of building models that can generalize to diverse data. Overly homogenizing datasets by excluding atypical observations risks leading models to fail when presented with real-world variability. Instead, retaining heterogeneity during training enables better generalization, as evidenced by HSF's robustness across ages, health conditions, imaging protocols, and even atypical anatomies. This suggests future work could validate HSF in patient populations by collaborating with medical centers to acquire data. Second, given the potential for the Rashōmon effect in multivariate neuroimaging analyses, rigorous workflows such as the framework introduced here are essential. Training an ensemble of high-performing models helps identify consensus between competing explanations to extract reliable insights. Cross-validation and out-of-sample testing further validate the results. These principles help overcome ambiguities when mapping complex dynamics like hippocampal subfields functional connectivity.

Looking ahead to clinical usage, the need to ensure model robustness to data heterogeneity and ease of use cannot be forgotten. Robustness to variations in scanners, operators, subjects, and protocols will be key for adoption (Lekadir et al., 2021). Efficient inference that reduces diagnostic time versus manual segmentation will increase utility. To spur adoption despite privacy limitations on data sharing, providing open-source code and pretrained models with ample documentation is pivotal.

Overall, the experimental work reveals important methodological lessons regarding the judicious application of machine learning in hippocampal subfields analysis and beyond. The need for heterogeneous training data, explainable AI to find consensus, and publicly available tools to enable clinical translation represent key takeaways that can guide future efforts to unlock the full potential of AI in the medical field.

Impact on Clinical Practice

The methodologies introduced in this dissertation have substantial potential to advance clinical care by enabling earlier and more accurate diagnosis, guiding personalized treatments, and accelerating biomedical discovery.

In particular, we showed that HSF is reliable across a wide range of population, scanners, and sites (Figure 3.14). The ability to efficiently and reliably quantify volumes and characterize functional connectivity patterns of the hippocampal subfields could aid the prognosis and guide interventions across a range of conditions affecting memory and cognition, an additional step towards individualized medicine. In temporal lobe epilepsy, the degree of sclerosis in vulnerable subfields such as CA1 correlates with the severity of the seizures (Blümcke et al., 2013). Relating subfields abnormalities to outcomes has the potential to optimize surgical planning and response prediction. Beyond surgical interventions, tracking the trajectories of the subfields through the lens of normative analysis (Bethlehem et al., 2022) could help to target early interventions to foster healthy development, especially since in dementias such as Alzheimer's disease, hippocampal subfields atrophy is one of the earliest structural biomarkers of pathology (Barnes et al., 2009). Detecting initial subicular and CA1 loss could enable earlier diagnosis and monitoring of disease progression. Finally, the use of such tools at a larger scale can help better describe and maybe prevent the consequences of adverse events. For example, preterm birth has been linked to an impaired episodic memory caused by anatomical and functional alterations of the hippocampus (Nosarti & Froudust-Walsh, 2016), while neonatal hypoxia has been shown to alter the hippocampal volume and reduce adult neurogenesis (Takada et al., 2016).

However, for AI systems such as HSF to be responsibly implemented in clinical settings, key prerequisites must be met. As outlined in a report by the European Parliament on AI in healthcare (European Parliament. Directorate General for Parliamentary Research Services., 2022), robustness to diverse conditions is paramount (Lekadir et al., 2021). For such tools to transition responsibly from research to clinical practice, they must demonstrate robust performance in various real-world conditions. Rigorous validation on heterogeneous data from multiple sites using various protocols is essential to establish appropriate use cases and limitations, preventing patient harm from improper application. As models are validated on more diverse data, transparency about performance gaps must be clearly communicated to set proper expectations among clinicians. Additionally, privacy and consent issues around data collection and sharing must be handled ethically through de-identification — if de-identification is possible — and opt-in procedures. It is also critical that multidisciplinary teams including clinicians, engineers, and domain experts oversee tool development to consider diverse needs and perspectives, not just technological capabilities. Clinician education programs will then be key to facilitate appropriate integration of AI systems into clinical workflows. Finally, policymakers have an important role to play in enacting tailored regulations that encourage innovation while adequately managing risks. Overall, achieving successful clinical implementation requires addressing data gaps, integration challenges, and managerial obstacles, not just improving algorithmic accuracy. A collaborative, holistic approach focused on people is vital as AI moves from bench to bedside.

While introducing a robust, future-proof, fast, and automated hippocampal subfields segmentation represents progress, fully realizing the potential of AI in hippocampal research — and more broadly, neurosciences — will require ongoing efforts. As emphasized by the European Parliament report, external validation from independent entities, comprehensive reporting, and continuous monitoring

are vital as tools transition to practice. Incorporating multimodal data, such as genetics and histology, could strengthen biological validity. Longitudinal and multicenter studies are needed to replicate findings across populations. Moving forward, the methodological lessons from this work can inform the development of robust and transparent AI systems that augment clinicians in delivering enhanced care while putting people first. With responsible design and application, AI promises to unlock new frontiers in medicine, but as demonstrated by the Rashōmon Effect, having high-performing models is not a guarantee on their underlying mechanisms, and systematic explanatory methods are necessary.

On the ethic and epistemology of AI in the medical field

Recent advances in AI have enabled remarkable progress in medical imaging capabilities, sometimes surpassing human experts (e.g. Esteva et al., 2017). When developed responsibly, AI systems can enable earlier and more precise diagnoses, reduce the need for invasive procedures, and accelerate discovery through large-scale data analysis (Ienca & Ignatiadis, 2020). However, as promising as these technologies may be, there are important ethical considerations regarding their development and deployment that must be addressed. Ultimately, to reach a level of usable intelligence, we need (1) to learn from prior data, (2) to extract knowledge, (3) to generalize, (4) to fight the curse of dimensionality, and (5) to disentangle the underlying explanatory factors of the data (Bengio et al., 2013; Holzinger et al., 2017). Therefore, XAI is relevant because it validates the enumerated needs. Even if post hoc explanations, as employed in this work, may present risks such as perpetuating bad practices (Rudin, 2019), we argue that this approach is a necessary step before the emergence of high-performing interpretable models. In this context, models must come with transparency standards and be committed to fairness (Holzinger et al., 2017). But, as we showed, XAI should be taken with a grain of salt. Explanation methods, without further validation, can be highly variable, with only a small valid subset representing true model behavior. Blind trust in AI explanations is unwise, as issues can still arise from “inconclusive evidence” — probabilistic conclusions are not infallible — “inscrutable evidence” — misunderstood models are hard to control systems — and “misguided evidence” — the reliability is limited by the underlying data (Mittelstadt et al., 2016).

We also want to highlight the concern that much AI research is done in proprietary settings, limiting public availability of models and data, even when scientific publications are available. As AI has demonstrated substantial benefits for medical care, maintaining open access to these critical resources should be a priority, — closed-source AI cannot become the default standard. Similarly to open access to code bases, open and responsible data sharing is crucial. As an example, this entire dissertation would not have been possible without publicly available datasets. Unfortunately, only a small percentage of scientific investigators in biomedicine currently share data openly, — the majority of investigators remain relatively reluctant to make their data available for reuse and repurposing (Scruggs et al., 2015).

There are also ecological motivations for developing efficient AI that do not require massive computational resources for training and inference. With computations required for deep learning research doubling every few months since 2012, making efficiency a more common evaluation criterion — e.g. reporting FLOPs count, number of parameters, or elapsed real time — for AI

research is a necessity (Schwartz et al., 2019). This ever-growing computational cost goes beyond ecology, as currently only large tech companies have the capacity to produce state-of-the-art models, disadvantaging academics and small companies without such resources.

Finally, an open epistemological question remains regarding the extent to which AI models build understanding as scientists do. As models may rely on different explanatory factors than domain experts, interpretability merits analysis, but as AI research rapidly evolves, how do we know when to reject an old model entirely as opposed to seeking incremental improvements? Realizing the full potential of AI in medicine requires grappling with complex trade-offs and unanswered questions. The present dissertation aimed to contribute evidence to guide the responsible and beneficial adoption of these transformative technologies.

Conclusion

This dissertation presented several key contributions towards understanding the anatomical and functional development of the hippocampal subfields throughout the lifespan. A major methodological advance was the creation of HSF, an efficient deep learning tool for automated segmentation of subfields boundaries from structural MRI data. HSF achieved superior performance compared to existing methods by leveraging state-of-the-art computer vision techniques and incorporating diverse manually labeled training data. The modular architecture enables continued evolution by incorporating new models. Critical to the success of HSF was the aggregation of publicly available manual segmentation datasets to maximize heterogeneity during training. This likely contributed to the tool's robustness across ages, conditions, and even atypical anatomies. The public release as an open-source tool means HSF can be widely adopted to catalyze large-scale studies relating subfields anatomy to cognition and disease.

The application of HSF in over 3700 individuals aged 4 to 100+ years provided unprecedented characterization of hippocampal subfields volumetric trajectories across the lifespan. Modeling revealed distinct non-linear development patterns for each subfield, underscoring the importance of not treating the hippocampus as a homogeneous entity. Prolonged volumetric maturation was observed for the dentate gyrus, which aligns with its role in adult neurogenesis and pattern separation. The stable volume of CA2-3 mirrors its early maturation supporting social memory. In contrast, the subiculum exhibited early development followed by susceptibility to age-related atrophy. Differences likely arise from asynchronous cytoarchitectonic and myeloarchitectonic timelines. Overall, delineating subfields heterogeneity elucidated coordinated maturation supporting memory improvements.

Functional connectivity analysis using rigorous explainable AI techniques revealed decreasing interactions between hippocampal subfields and distributed cortical regions during childhood and adolescence. Convergence between computational models highlighted robust connectivity biomarkers despite the ambiguity of high-dimensional resting-state fMRI data. The strongest effects were found for the dentate gyrus, consistent with its protracted structural development. Decreasing connectivity likely reflects refinement of hippocampal-cortical circuits optimizing episodic memory and spatial processing. There was a tight correspondence between subfield-specific anatomical and functional trajectories over development.

Together, the findings provide critical insights into the prolonged maturation and coordinated specialization of hippocampal subfields anatomy and interactions that strengthen memory and cognition. By enabling large-scale segmentation, HSF can significantly accelerate studies that examine the role of the structure and function of the subfields from normal development to degenerative conditions. Moving forward, integrating multimodal data and increasing sample diversity will be important to enhance understanding of these complex neurobiological mechanisms. Ultimately, characterizing

4 Discussion

subfields heterogeneity promises to enlight typical and atypical developmental pathways, aid in the early diagnosis of memory disorders, and guide interventions to promote lifelong cognitive health.

In summary, this dissertation introduced innovative tools and modeling approaches to chart the coordinated maturation of the structure and function of the hippocampal subfields across the lifespan. Experimental contributions establish a foundation to study the hippocampal subfield contributions in health and disease at an unprecedented scale.

Bibliography

- Aggleton, J. P. (2012). Multiple anatomical systems embedded within the primate medial temporal lobe: Implications for hippocampal function. *Neuroscience & Biobehavioral Reviews*, 36(7), 1579–1596. <https://doi.org/10.1016/j.neubiorev.2011.09.005> (cit. on p. 20)
- Aggleton, J. P., & Christiansen, K. (2015). The subiculum. In *Progress in Brain Research* (pp. 65–82). Elsevier. <https://doi.org/10.1016/bs.pbr.2015.03.003>. (Cit. on p. 20)
- Aimone, J. B., Wiles, J., & Gage, F. H. (2006). Potential role for adult neurogenesis in the encoding of time in new memories. *Nature Neuroscience*, 9(6), 723–727. <https://doi.org/10.1038/nn1707> (cit. on p. 15)
- Bai, Y., Jones, A., Ndousse, K., Askill, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., . . . Kaplan, J. (2022). *Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback*. arXiv: 2204.05862 [cs]. Retrieved August 9, 2023, from <http://arxiv.org/abs/2204.05862>. (Cit. on p. 113)
- Barnes, J., Bartlett, J. W., Van De Pol, L. A., Loy, C. T., Scahill, R. I., Frost, C., Thompson, P., & Fox, N. C. (2009). A meta-analysis of hippocampal atrophy rates in Alzheimer’s disease. *Neurobiology of Aging*, 30(11), 1711–1723. <https://doi.org/10.1016/j.neurobiolaging.2008.01.010> (cit. on p. 117)
- Bender, A. R., Keresztes, A., Bodammer, N. C., Shing, Y. L., Werkle-Bergner, M., Daugherty, A. M., Yu, Q., Kühn, S., Lindenberger, U., & Raz, N. (2018). Optimization and validation of automated hippocampal subfield segmentation across the lifespan. *Human Brain Mapping*, 39(2), 916–931. <https://doi.org/10.1002/hbm.23891> (cit. on p. 22)
- Bengio, Y., Courville, A., & Vincent, P. (2013). Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8), 1798–1828. <https://doi.org/10.1109/TPAMI.2013.50> (cit. on p. 118)
- Berron, D., Vieweg, P., Hochkeppeler, A., Pluta, J., Ding, S.-L., Maass, A., Luther, A., Xie, L., Das, S., Wolk, D., Wolbers, T., Yushkevich, P., Düzel, E., & Wisse, L. (2017). A protocol for manual segmentation of medial temporal lobe subregions in 7 Tesla MRI. *NeuroImage: Clinical*, 15, 466–482. <https://doi.org/10.1016/j.nicl.2017.05.022> (cit. on pp. 16, 21, 32, 34)
- Bethlehem, R. A. I., Seidlitz, J., White, S. R., Vogel, J. W., Anderson, K. M., Adamson, C., Adler, S., Alexopoulos, G. S., Anagnostou, E., Areces-Gonzalez, A., Astle, D. E., Auyeung, B., Ayub, M., Bae, J., Ball, G., Baron-Cohen, S., Beare, R., Bedford, S. A., Benegal, V., . . . Alexander-Bloch, A. F. (2022). Brain charts for the human lifespan. *Nature*, 604(7906), 525–533. <https://doi.org/10.1038/s41586-022-04554-y> (cit. on p. 117)
- Bijsterbosch, J. (2017). *Introduction to resting state fMRI functional connectivity* (First edition). Oxford University Press. (Cit. on pp. 23, 24).
- Blackstad, T. W. (1956). Commissural connections of the hippocampal region in the rat, with special reference to their mode of termination. *The Journal of Comparative Neurology*, 105(3), 417–537. <https://doi.org/10.1002/cne.901050305> (cit. on p. 16)

- Blümcke, I., Thom, M., Aronica, E., Armstrong, D.D., Bartolomei, F., Bernasconi, A., Bernasconi, N., Bien, C. G., Cendes, F., Coras, R., Cross, J. H., Jacques, T. S., Kahane, P., Mathern, G. W., Miyata, H., Moshé, S. L., Oz, B., Özkara, Ç., Perucca, E., . . . Spreafico, R. (2013). International consensus classification of hippocampal sclerosis in temporal lobe epilepsy: A Task Force report from the ILAE Commission on Diagnostic Methods. *Epilepsia*, *54*(7), 1315–1329. <https://doi.org/10.1111/epi.12220> (cit. on p. 117)
- Bouyeure, A., & Noulhiane, M. (2020). Memory: Normative development of memory systems. In *Handbook of Clinical Neurology* (pp. 201–213). Elsevier. <https://doi.org/10.1016/B978-0-444-64150-2.00018-6>. (Cit. on pp. 15, 80)
- Bouyeure, A., Patil, S., Mauconduit, F., Poiret, C., Isai, D., & Noulhiane, M. (2021). Hippocampal subfield volumes and memory discrimination in the developing brain. *Hippocampus*, hipo.23385. <https://doi.org/10.1002/hipo.23385> (cit. on pp. 18, 32, 33, 80, 113)
- Braun, U., Plichta, M. M., Esslinger, C., Sauer, C., Haddad, L., Grimm, O., Mier, D., Mohnke, S., Heinz, A., Erk, S., Walter, H., Seiferth, N., Kirsch, P., & Meyer-Lindenberg, A. (2012). Test–retest reliability of resting-state connectivity network characteristics using fMRI and graph theoretical measures. *NeuroImage*, *59*(2), 1404–1412. <https://doi.org/10.1016/j.neuroimage.2011.08.044> (cit. on p. 24)
- Breiman, L. (2001). Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3). <https://doi.org/10.1214/ss/1009213726> (cit. on pp. 80, 109)
- Burgess, N., Maguire, E. A., & O’Keefe, J. (2002). The Human Hippocampus and Spatial and Episodic Memory. *Neuron*, *35*(4), 625–641. [https://doi.org/10.1016/S0896-6273\(02\)00830-9](https://doi.org/10.1016/S0896-6273(02)00830-9) (cit. on p. 15)
- Canada, K. L., Saifullah, S., Gardner, J. C., Sutton, B. P., Fabiani, M., Gratton, G., Raz, N., & Daugherty, A. M. (2023). Development and validation of a quality control procedure for automatic segmentation of hippocampal subfields. *Hippocampus*, hipo.23552. <https://doi.org/10.1002/hipo.23552> (cit. on p. 21)
- Cansino, S. (2009). Episodic memory decay along the adult lifespan: A review of behavioral and neurophysiological evidence. *International Journal of Psychophysiology*, *71*(1), 64–69. <https://doi.org/10.1016/j.ijpsycho.2008.07.005> (cit. on p. 15)
- Cendes, F. (2005). Progressive hippocampal and extrahippocampal atrophy in drug resistant epilepsy: Review. *Current Opinion in Neurology*, *18*(2), 173–177. <https://doi.org/10.1097/01.wco.0000162860.49842.90> (cit. on p. 15)
- Chen, Y., Gan, H., Chen, H., Zeng, Y., Xu, L., Heidari, A. A., Zhu, X., & Liu, Y. (2023). Accurate iris segmentation and recognition using an end-to-end unified framework based on MADNet and DSANet. *Neurocomputing*, *517*, 264–278. <https://doi.org/10.1016/j.neucom.2022.10.064> (cit. on p. 39)
- Chen, Y., & Liu, Y. (2021). Automatic Segmentation of Hippocampal Subfields MRI Based on FPN-DenseVoxNet. *2021 Asia-Pacific Conference on Communications Technology and Computer Science (ACCTCS)*, 58–62. <https://doi.org/10.1109/ACCTCS52002.2021.00020> (cit. on pp. 22, 38)
- Chételat, G., Fouquet, M., Kalpouzos, G., Denghien, I., De La Sayette, V., Viader, F., Mézenge, F., Landeau, B., Baron, J., Eustache, F., & Desgranges, B. (2008). Three-dimensional surface mapping of hippocampal atrophy progression from MCI to AD and over normal aging as assessed using voxel-based morphometry. *Neuropsychologia*, *46*(6), 1721–1731. <https://doi.org/10.1016/j.neuropsychologia.2007.11.037> (cit. on p. 18)

- Chiappiniello, A., Tarducci, R., Muscio, C., Bruzzone, M. G., Bozzali, M., Tiraboschi, P., Nigri, A., Ambrosi, C., Chipi, E., Ferraro, S., Festari, C., Gasparotti, R., Gianeri, R., Giulietti, G., Mascaro, L., Montanucci, C., Nicolosi, V., Rosazza, C., Serra, L., . . . Jovicich, J. (2021). Automatic multispectral MRI segmentation of human hippocampal subfields: An evaluation of multicentric test–retest reproducibility. *Brain Structure and Function*, 226(1), 137–150. <https://doi.org/10.1007/s00429-020-02172-w> (cit. on p. 22)
- Cohen, T. S., Weiler, M., Kicanaoglu, B., & Welling, M. (2019). Gauge Equivariant Convolutional Networks and the Icosahedral CNN. *arXiv:1902.04615 [cs, stat]* (cit. on pp. 78, 112).
- Costa, M., Costa, D., Gomes, T., & Pinto, S. (2022). Shifting Capsule Networks from the Cloud to the Deep Edge. *ACM Transactions on Intelligent Systems and Technology*, 13(6), 1–25. <https://doi.org/10.1145/3544562> (cit. on p. 112)
- Dalton, M. A., McCormick, C., & Maguire, E. A. (2019). Differences in functional connectivity along the anterior-posterior axis of human hippocampal subfields. *NeuroImage*, 192, 38–51. <https://doi.org/10.1016/j.neuroimage.2019.02.066> (cit. on pp. 20, 25)
- Dalton, M. A., Zeidman, P., Barry, D. N., Williams, E., & Maguire, E. A. (2017). Segmenting subregions of the human hippocampus on structural magnetic resonance image scans: An illustrated tutorial. *Brain and Neuroscience Advances*, 1, 239821281770144. <https://doi.org/10.1177/2398212817701448> (cit. on pp. 17, 21)
- Damoiseaux, J. S., Viviano, R. P., Yuan, P., & Raz, N. (2016). Differential effect of age on posterior and anterior hippocampal functional connectivity. *NeuroImage*, 133, 468–476. <https://doi.org/10.1016/j.neuroimage.2016.03.047> (cit. on p. 25)
- Das, S. R., Pluta, J., Mancuso, L., Kliot, D., Orozco, S., Dickerson, B. C., Yushkevich, P. A., & Wolk, D. A. (2013). Increased functional connectivity within medial temporal lobe in mild cognitive impairment. *Hippocampus*, 23(1), 1–6. <https://doi.org/10.1002/hipo.22051> (cit. on p. 25)
- de Flores, R., La Joie, R., & Chételat, G. (2015). Structural imaging of hippocampal subfields in healthy aging and Alzheimer’s disease. *Neuroscience*, 309, 29–50. <https://doi.org/10.1016/j.neuroscience.2015.08.033> (cit. on pp. 16, 18, 22)
- De Flores, R., Mutlu, J., Bejanin, A., Gonneaud, J., Landeau, B., Tomadesso, C., Mézengé, F., De La Sayette, V., Eustache, F., & Chételat, G. (2017). Intrinsic connectivity of hippocampal subfields in normal elderly and mild cognitive impairment patients. *Human Brain Mapping*, 38(10), 4922–4932. <https://doi.org/10.1002/hbm.23704> (cit. on p. 25)
- De Raad, K., Van Garderen, K., Smits, M., Van Der Voort, S., Incekara, F., Oei, E., Hirvasniemi, J., Klein, S., & Starmans, M. (2021). The Effect of Preprocessing on Convolutional Neural Networks for Medical Image Segmentation. *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 655–658. <https://doi.org/10.1109/ISBI48211.2021.9433952> (cit. on p. 33)
- Deco, G., Jirsa, V. K., & McIntosh, A. R. (2011). Emerging concepts for the dynamical organization of resting-state activity in the brain. *Nature Reviews Neuroscience*, 12(1), 43–56. <https://doi.org/10.1038/nrn2961> (cit. on p. 24)
- DeKraker, J., Haast, R. A., Yousif, M. D., Karat, B., Köhler, S., & Khan, A. R. (2021). *HippUnfold: Automated hippocampal unfolding, morphometry, and subfield segmentation* (Preprint). Neuroscience. <https://doi.org/10.1101/2021.12.03.471134>. (Cit. on p. 22)
- DeKraker, J., Haast, R. A., Yousif, M. D., Karat, B., Lau, J. C., Köhler, S., & Khan, A. R. (2022). Automated hippocampal unfolding for morphometry and subfield segmentation with HippUnfold. *eLife*, 11, e77945. <https://doi.org/10.7554/eLife.77945> (cit. on pp. 22, 38)

- Diaz, I., Geiger, M., & McKinley, R. I. (2023). An end-to-end SE(3)-equivariant segmentation network. (Cit. on pp. 78, 112).
- Ding, S.-L., Royall, J. J., Sunkin, S. M., Ng, L., Facer, B. A., Lesnar, P., Guillozet-Bongaarts, A., McMurray, B., Szafer, A., Dolbeare, T. A., Stevens, A., Tirrell, L., Benner, T., Caldejon, S., Dalley, R. A., Dee, N., Lau, C., Nyhus, J., Reding, M., . . . Lein, E. S. (2016). Comprehensive cellular-resolution atlas of the adult human brain. *Journal of Comparative Neurology*, 524(16), 3127–3481. <https://doi.org/10.1002/cne.24080> (cit. on p. 34)
- Ding, S.-L., & Van Hoesen, G. W. (2015). Organization and detailed parcellation of human hippocampal head and body regions based on a combined analysis of Cyto- and chemoarchitecture: Detailed subfields in human hippocampal head. *Journal of Comparative Neurology*, 523(15), 2233–2253. <https://doi.org/10.1002/cne.23786> (cit. on p. 34)
- Dong, C., Xu, S., & Li, Z. (2022). A novel end-to-end deep learning solution for coronary artery segmentation from CCTA. *Medical Physics*, 49(11), 6945–6959. <https://doi.org/10.1002/mp.15842> (cit. on p. 39)
- Dosenbach, N. U. F., Nardos, B., Cohen, A. L., Fair, D. A., Power, J. D., Church, J. A., Nelson, S. M., Wig, G. S., Vogel, A. C., Lessov-Schlaggar, C. N., Barnes, K. A., Dubis, J. W., Feczko, E., Coalson, R. S., Pruett, J. R., Barch, D. M., Petersen, S. E., & Schlaggar, B. L. (2010). Prediction of Individual Brain Maturity Using fMRI. *Science*, 329(5997), 1358–1361. <https://doi.org/10.1126/science.1194144> (cit. on p. 114)
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Hounsfield, N. (2021). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. arXiv: 2010.11929 [cs]. Retrieved August 15, 2023, from <http://arxiv.org/abs/2010.11929>. (Cit. on p. 112)
- Douglas, R. J. (1967). The hippocampus and behavior. *Psychological Bulletin*, 67(6), 416–442. <https://doi.org/10.1037/h0024599> (cit. on p. 15)
- Duvernoy, H. M., Cattin, F., & Risold, P.-Y. (2013). *The Human Hippocampus*. Springer Berlin Heidelberg. <https://doi.org/10.1007/978-3-642-33603-4>. (Cit. on pp. 16, 19)
- Eichenbaum, H., & Cohen, N. J. (2004). *From conditioning to conscious recollection: Memory systems of the brain* (1. issued as paperback). Oxford Univ. Press. (Cit. on p. 15).
- Eichenbaum, H., Dudchenko, P., Wood, E., Shapiro, M., & Tanila, H. (1999). The Hippocampus, Memory, and Place Cells. *Neuron*, 23(2), 209–226. [https://doi.org/10.1016/S0896-6273\(00\)80773-4](https://doi.org/10.1016/S0896-6273(00)80773-4) (cit. on p. 15)
- Elbaz, N., Devisscher, L., Ghosland, C., Adibpour, P., Elmaleh, M., Héneau, A., Neumane, S., Hertz-Pannier, L., Biran, V., Dubois, J., & Alison, M. (2023). Evaluation IRM des lésions cérébrales chez le prématuré à terme : Quels facteurs de risque ? (Cit. on p. 79).
- Ellis, C. T., Skalaban, L. J., Yates, T. S., Bejjanki, V. R., Córdova, N. I., & Turk-Browne, N. B. (2021). Evidence of hippocampal learning in human infants. *Current Biology*, 31(15), 3358–3364.e4. <https://doi.org/10.1016/j.cub.2021.04.072> (cit. on p. 18)
- Erickson, K. I., Voss, M. W., Prakash, R. S., Basak, C., Szabo, A., Chaddock, L., Kim, J. S., Heo, S., Alves, H., White, S. M., Wojcicki, T. R., Mailey, E., Vieira, V. J., Martin, S. A., Pence, B. D., Woods, J. A., McAuley, E., & Kramer, A. F. (2011). Exercise training increases size of hippocampus and improves memory. *Proceedings of the National Academy of Sciences*, 108(7), 3017–3022. <https://doi.org/10.1073/pnas.1015950108> (cit. on p. 16)
- Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), 115–118. <https://doi.org/10.1038/nature21056> (cit. on p. 118)

- European Parliament. Directorate General for Parliamentary Research Services. (2022). *Artificial intelligence in healthcare: Applications, risks, and ethical and societal impacts*. Publications Office. Retrieved August 15, 2023, from <https://data.europa.eu/doi/10.2861/568473>. (Cit. on p. 117)
- Fair, D. A., Cohen, A. L., Power, J. D., Dosenbach, N. U. F., Church, J. A., Miezin, F. M., Schlaggar, B. L., & Petersen, S. E. (2009). Functional Brain Networks Develop from a “Local to Distributed” Organization (O. Sporns, Ed.). *PLoS Computational Biology*, 5(5), e1000381. <https://doi.org/10.1371/journal.pcbi.1000381> (cit. on p. 114)
- Foster, C. M., Kennedy, K. M., Hoagey, D. A., & Rodrigue, K. M. (2019). The role of hippocampal subfield volume and fornix microstructure in episodic memory across the lifespan. *Hippocampus*, 29(12), 1206–1223. <https://doi.org/10.1002/hipo.23133> (cit. on p. 18)
- Fragueiro, A., Committeri, G., & Cury, C. (2023). Incomplete hippocampal inversion and hippocampal subfield volumes: Implementation and inter-reliability of automatic segmentation. (Cit. on p. 22).
- Frankland, P. W., Köhler, S., & Josselyn, S. A. (2013). Hippocampal neurogenesis and forgetting. *Trends in Neurosciences*, 36(9), 497–503. <https://doi.org/10.1016/j.tins.2013.05.002> (cit. on p. 15)
- Gallo, S., El-Gazzar, A., Zhutovsky, P., Thomas, R. M., Javaheripour, N., Li, M., Bartova, L., Bathula, D., Dannlowski, U., Davey, C., Frodl, T., Gotlib, I., Grimm, S., Grotegerd, D., Hahn, T., Hamilton, P. J., Harrison, B. J., Jansen, A., Kircher, T., . . . Van Wingen, G. (2023). Functional connectivity signatures of major depressive disorder: Machine learning analysis of two multicenter neuroimaging studies. *Molecular Psychiatry*. <https://doi.org/10.1038/s41380-023-01977-5> (cit. on p. 24)
- Gashler, M., Giraud-Carrier, C., & Martinez, T. (2008). Decision Tree Ensemble: Small Heterogeneous Is Better Than Large Homogeneous. *2008 Seventh International Conference on Machine Learning and Applications*, 900–905. <https://doi.org/10.1109/ICMLA.2008.154> (cit. on p. 31)
- Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C. F., Jenkinson, M., Smith, S. M., & Van Essen, D. C. (2016). A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615), 171–178. <https://doi.org/10.1038/nature18933> (cit. on p. 35)
- Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J. R., Van Essen, D. C., & Jenkinson, M. (2013). The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage*, 80, 105–124. <https://doi.org/10.1016/j.neuroimage.2013.04.127> (cit. on p. 33)
- Gregorians, L., & Spiers, H. J. (2022). Affordances for Spatial Navigation. In Z. Djebbara (Ed.), *Affordances in Everyday Life* (pp. 99–112). Springer International Publishing. https://doi.org/10.1007/978-3-031-08629-8_10. (Cit. on p. 16)
- Greicius, M. D., Krasnow, B., Reiss, A. L., & Menon, V. (2003). Functional connectivity in the resting brain: A network analysis of the default mode hypothesis. *Proceedings of the National Academy of Sciences*, 100(1), 253–258. <https://doi.org/10.1073/pnas.0135058100> (cit. on p. 23)
- Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). *Why do tree-based models still outperform deep learning on tabular data?* arXiv: 2207.08815 [cs, stat]. Retrieved August 6, 2023, from <http://arxiv.org/abs/2207.08815>. (Cit. on p. 82)

- Haeger, A., Mangin, J.-F., Vignaud, A., Poupon, C., Grigis, A., Boumezbeur, F., Frouin, V., Deverre, J.-R., Sarazin, M., Hertz-Pannier, L., Bottlaender, M., the SENIOR team, Baron, C., Berland, V., Blancho, N., Desmidt, S., Doublé, C., Ginisty, C., Joly-Testault, V., . . . Vuilleumard, C. (2020). Imaging the aging brain: Study design and baseline findings of the SENIOR cohort. *Alzheimer's Research & Therapy*, *12*(1), 77. <https://doi.org/10.1186/s13195-020-00642-1> (cit. on pp. 32, 33, 113)
- Hahn, T., Pyeon, M., & Kim, G. (2019). Self-routing capsule networks. *Advances in neural information processing systems*, *32* (cit. on p. 112).
- Hao, Z. Y., Zhong, Y., Ma, Z. J., Xu, H. Z., Kong, J. Y., Wu, Z., Wu, Y., Li, J., Lu, X., Zhang, N., & Wang, C. (2020). Abnormal resting-state functional connectivity of hippocampal subfields in patients with major depressive disorder. *BMC Psychiatry*, *20*(1), 71. <https://doi.org/10.1186/s12888-020-02490-7> (cit. on p. 25)
- Haq, M. U., Sethi, M. A. J., & Rehman, A. U. (2023). Capsule Network with Its Limitation, Modification, and Applications—A Survey. *Machine Learning and Knowledge Extraction*, *5*(3), 891–921. <https://doi.org/10.3390/make5030047> (cit. on p. 112)
- Hindy, N. C., Ng, F. Y., & Turk-Browne, N. B. (2016). Linking pattern completion in the hippocampus to predictive coding in visual cortex. *Nature Neuroscience*, *19*(5), 665–667. <https://doi.org/10.1038/nn.4284> (cit. on p. 32)
- Hinton, G. (2021). How to represent part-whole hierarchies in a neural network. *arXiv:2102.12627 [cs]* (cit. on p. 78).
- Hitti, F. L., & Siegelbaum, S. A. (2014). The hippocampal CA2 region is essential for social memory. *Nature*, *508*(7494), 88–92. <https://doi.org/10.1038/nature13028> (cit. on p. 114)
- Holt, D. J., Öngür, D., Wright, C. I., Dickerson, B. C., & Rauch, S. L. (2008). Neuroanatomical Systems Relevant to Neuropsychiatric Disorders. In *Massachusetts General Hospital Comprehensive Clinical Psychiatry* (pp. 975–995). Elsevier. <https://doi.org/10.1016/B978-0-323-04743-2.50073-1>. (Cit. on p. 19)
- Holzinger, A., Biemann, C., Pattichis, C. S., & Kell, D. B. (2017). *What do we need to build explainable AI systems for the medical domain?* arXiv: 1712.09923 [cs, stat]. Retrieved August 8, 2023, from <http://arxiv.org/abs/1712.09923>. (Cit. on p. 118)
- Honey, C. J., Sporns, O., Cammoun, L., Gigandet, X., Thiran, J. P., Meuli, R., & Hagmann, P. (2009). Predicting human resting-state functional connectivity from structural connectivity. *Proceedings of the National Academy of Sciences*, *106*(6), 2035–2040. <https://doi.org/10.1073/pnas.0811168106> (cit. on p. 24)
- Honey, C. J., Thivierge, J.-P., & Sporns, O. (2010). Can structure predict function in the human brain? *NeuroImage*, *52*(3), 766–776. <https://doi.org/10.1016/j.neuroimage.2010.01.071> (cit. on p. 24)
- Huttenlocher, P. R., & Dabholkar, A. S. (1997). Regional differences in synaptogenesis in human cerebral cortex. *The Journal of Comparative Neurology*, *387*(2), 167–178. [https://doi.org/10.1002/\(SICI\)1096-9861\(19971020\)387:2<167::AID-CNE1>3.0.CO;2-Z](https://doi.org/10.1002/(SICI)1096-9861(19971020)387:2<167::AID-CNE1>3.0.CO;2-Z) (cit. on p. 114)
- Ienca, M., & Ignatiadis, K. (2020). Artificial Intelligence in Clinical Neuroscience: Methodological and Ethical Challenges. *AJOB Neuroscience*, *11*(2), 77–87. <https://doi.org/10.1080/21507740.2020.1740352> (cit. on p. 118)
- Iglesias, J. E., Augustinack, J. C., Nguyen, K., Player, C. M., Player, A., Wright, M., Roy, N., Frosch, M. P., McKee, A. C., Wald, L. L., Fischl, B., & Van Leemput, K. (2015). A computational atlas of the hippocampal formation using ex vivo , ultra-high resolution MRI: Application to adaptive segmentation of in vivo MRI. *NeuroImage*, *115*, 117–137. <https://doi.org/10.1016/j.neuroimage.2015.04.042> (cit. on pp. 22, 37)

- Jarrard, L. E. (1993). On the role of the hippocampus in learning and memory in the rat. *Behavioral and Neural Biology*, 60(1), 9–26. [https://doi.org/10.1016/0163-1047\(93\)90664-4](https://doi.org/10.1016/0163-1047(93)90664-4) (cit. on p. 15)
- Jarrard, L. E. (1995). What does the hippocampus really do? *Behavioural Brain Research*, 71(1-2), 1–10. [https://doi.org/10.1016/0166-4328\(95\)00034-8](https://doi.org/10.1016/0166-4328(95)00034-8) (cit. on p. 15)
- Ketz, N., Morkonda, S. G., & O'Reilly, R. C. (2013). Theta Coordinated Error-Driven Learning in the Hippocampus (O. Sporns, Ed.). *PLoS Computational Biology*, 9(6), e1003067. <https://doi.org/10.1371/journal.pcbi.1003067> (cit. on p. 19)
- Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., & Girshick, R. (2023). *Segment Anything*. arXiv: 2304.02643 [cs]. Retrieved August 9, 2023, from <http://arxiv.org/abs/2304.02643>. (Cit. on p. 113)
- Koelsch, S. (2014). Brain correlates of music-evoked emotions. *Nature Reviews Neuroscience*, 15(3), 170–180. <https://doi.org/10.1038/nrn3666> (cit. on p. 20)
- Kubie, J. L., Levy, E. R. J., & Fenton, A. A. (2020). Is hippocampal remapping the physiological basis for context? *Hippocampus*, 30(8), 851–864. <https://doi.org/10.1002/hipo.23160> (cit. on p. 16)
- Kulaga-Yoskovitz, J., Bernhardt, B. C., Hong, S.-J., Mansi, T., Liang, K. E., van der Kouwe, A. J., Smallwood, J., Bernasconi, A., & Bernasconi, N. (2015). Multi-contrast submillimetric 3 Tesla hippocampal subfield segmentation protocol and dataset. *Scientific Data*, 2(1), 150059. <https://doi.org/10.1038/sdata.2015.59> (cit. on p. 32)
- Kumar, I. E., Venkatasubramanian, S., Scheidegger, C., & Friedler, S. (2020). Problems with shapley-value-based explanations as feature importance measures. In H. D. III & A. Singh (Eds.), *Proceedings of the 37th international conference on machine learning* (pp. 5491–5500). PMLR. <https://proceedings.mlr.press/v119/kumar20e.html>. (Cit. on p. 80)
- Kurtz, M., Kopinsky, J., Gelashvili, R., Matveev, A., Carr, J., Goin, M., Leiserson, W., Moore, S., Nell, B., Shavit, N., & Alistarh, D. (2020). Inducing and exploiting activation sparsity for fast inference on deep neural networks. In H. D. III & A. Singh (Eds.), *Proceedings of the 37th international conference on machine learning* (pp. 5533–5543). PMLR. <http://proceedings.mlr.press/v119/kurtz20a.html>. (Cit. on p. 112)
- Kwabena Patrick, M., Felix Adekoya, A., Abra Mighty, A., & Edward, B. Y. (2019). Capsule Networks – A survey. *Journal of King Saud University - Computer and Information Sciences*, S1319157819309322. <https://doi.org/10.1016/j.jksuci.2019.09.014> (cit. on p. 39)
- La Joie, R., Fouquet, M., Mézenge, F., Landeau, B., Villain, N., Mevel, K., Pélerin, A., Eustache, F., Desgranges, B., & Chételat, G. (2010). Differential effect of age on hippocampal subfields assessed using a new high-resolution 3T MR sequence. *NeuroImage*, 53(2), 506–514. <https://doi.org/10.1016/j.neuroimage.2010.06.024> (cit. on pp. 18, 21)
- La Joie, R., Perrotin, A., De La Sayette, V., Egret, S., Doeuvre, L., Belliard, S., Eustache, F., Desgranges, B., & Chételat, G. (2013). Hippocampal subfield volumetry in mild cognitive impairment, Alzheimer's disease and semantic dementia. *NeuroImage: Clinical*, 3, 155–162. <https://doi.org/10.1016/j.nicl.2013.08.007> (cit. on p. 16)
- Lace, G., Savva, G. M., Forster, G., De Silva, R., Brayne, C., Matthews, F. E., Barclay, J. J., Dakin, L., Ince, P. G., & Wharton, S. B. (2009). Hippocampal tau pathology is related to neuroanatomical connections: An ageing population-based study. *Brain*, 132(5), 1324–1334. <https://doi.org/10.1093/brain/awp059> (cit. on p. 18)

- Lagarde, J., Olivieri, P., Tonietto, M., Gervais, P., Comtat, C., Caillé, F., Bottlaender, M., & Sarazin, M. (2021). Distinct amyloid and tau PET signatures are associated with diverging clinical and imaging trajectories in patients with amnesic syndrome of the hippocampal type. *Translational Psychiatry*, *11*(1), 498. <https://doi.org/10.1038/s41398-021-01628-9> (cit. on pp. 32, 33, 79, 113)
- LaLonde, R., Xu, Z., Irmakci, I., Jain, S., & Bagci, U. (2020). Capsules for biomedical image segmentation. *Medical Image Analysis*, *68*, 101889. <https://doi.org/10.1016/j.media.2020.101889> (cit. on pp. 39, 40)
- Lavenex, P., & Amaral, D. G. (2000). Hippocampal-neocortical interaction: A hierarchy of associativity. *Hippocampus*, *10*(4), 420–430. [https://doi.org/10.1002/1098-1063\(2000\)10:4<\\$>420::AID-HIPO8\\$>3.0.CO;2-5](https://doi.org/10.1002/1098-1063(2000)10:4<$>420::AID-HIPO8$>3.0.CO;2-5) (cit. on p. 20)
- Lavenex, P., & Banta Lavenex, P. (2013). Building hippocampal circuits to learn and remember: Insights into the development of human memory. *Behavioural Brain Research*, *254*, 8–21. <https://doi.org/10.1016/j.bbr.2013.02.007> (cit. on pp. 18, 109, 113)
- Leal, S. L., & Yassa, M. A. (2018). Integrating new findings and examining clinical applications of pattern separation. *Nature Neuroscience*, *21*(2), 163–173. <https://doi.org/10.1038/s41593-017-0065-1> (cit. on p. 114)
- Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (Nov./1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324. <https://doi.org/10.1109/5.726791> (cit. on p. 39)
- Lee, J. K., Ekstrom, A. D., & Ghetti, S. (2014). Volume of hippocampal subfields and episodic memory in childhood and adolescence. *NeuroImage*, *94*, 162–171. <https://doi.org/10.1016/j.neuroimage.2014.03.019> (cit. on p. 18)
- Lee, J. K., Wendelken, C., Bunge, S. A., & Ghetti, S. (2016). A Time and Place for Everything: Developmental Differences in the Building Blocks of Episodic Memory. *Child Development*, *87*(1), 194–210. <https://doi.org/10.1111/cdev.12447> (cit. on pp. 109, 113)
- Lekadir, K., Osuala, R., Gallin, C., Lazrak, N., Kushibar, K., Tsakou, G., Aussó, S., Alberich, L. C., Marias, K., Tsiknakis, M., Colantonio, S., Papanikolaou, N., Salahuddin, Z., Woodruff, H. C., Lambin, P., & Martí-Bonmatí, L. (2021). *FUTURE-AI: Guiding Principles and Consensus Recommendations for Trustworthy Artificial Intelligence in Medical Imaging*. arXiv: 2109.09658 [cs]. Retrieved August 8, 2023, from <http://arxiv.org/abs/2109.09658>. (Cit. on pp. 116, 117)
- Liu, S., Wang, Z., An, Y., Zhao, J., Zhao, Y., & Zhang, Y.-D. (2023). EEG emotion recognition based on the attention mechanism and pre-trained convolution capsule network. *Knowledge-Based Systems*, *265*, 110372. <https://doi.org/10.1016/j.knosys.2023.110372> (cit. on p. 112)
- Lundberg, S., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions*. Retrieved February 19, 2020, from <http://arxiv.org/abs/1705.07874>. (Cit. on p. 82)
- Luo, P., Ren, J., Peng, Z., Zhang, R., & Li, J. (2019). Differentiable Learning-to-Normalize via Switchable Normalization. *arXiv:1806.10779 [cs]* (cit. on p. 64).
- Lynch, K. M., Shi, Y., Toga, A. W., Clark, K. A., & Pediatric Imaging, Neurocognition and Genetics Study. (2019). Hippocampal Shape Maturation in Childhood and Adolescence. *Cerebral Cortex*, *29*(9), 3651–3665. <https://doi.org/10.1093/cercor/bhy244> (cit. on pp. 113, 115)
- Maass, A., Schütze, H., Speck, O., Yonelinas, A., Tempelmann, C., Heinze, H.-J., Berron, D., Cardenas-Blanco, A., Brodersen, K. H., Enno Stephan, K., & Düzel, E. (2014). Laminar activity in the hippocampus and entorhinal cortex related to novelty and episodic encoding. *Nature Communications*, *5*(1), 5547. <https://doi.org/10.1038/ncomms6547> (cit. on p. 25)

- Marchisio, A., Bussolino, B., Salvati, E., Martina, M., Masera, G., & Shafique, M. (2022). Enabling Capsule Networks at the Edge through Approximate Softmax and Squash Operations. *Proceedings of the ACM/IEEE International Symposium on Low Power Electronics and Design*, 1–6. <https://doi.org/10.1145/3531437.3539717> (cit. on p. 112)
- Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2), 205395171667967. <https://doi.org/10.1177/2053951716679679> (cit. on p. 118)
- Mueller, S., Chao, L., Berman, B., & Weiner, M. (2011). Evidence for functional specialization of hippocampal subfields detected by MR subfield volumetry on high resolution images at 4T. *NeuroImage*, 56(3), 851–857. <https://doi.org/10.1016/j.neuroimage.2011.03.028> (cit. on pp. 19, 20)
- Mueller, S., Stables, L., Du, A., Schuff, N., Truran, D., Cashdollar, N., & Weiner, M. (2007). Measurement of hippocampal subfields and age-related changes with high resolution MRI at 4T. *Neurobiology of Aging*, 28(5), 719–726. <https://doi.org/10.1016/j.neurobiolaging.2006.03.007> (cit. on p. 21)
- Neves, G., Cooke, S. F., & Bliss, T. V. P. (2008). Synaptic plasticity, memory and the hippocampus: A neural network approach to causality. *Nature Reviews Neuroscience*, 9(1), 65–75. <https://doi.org/10.1038/nrn2303> (cit. on p. 19)
- Ngo, C. T., Michelmann, S., Olson, I. R., & Newcombe, N. S. (2021). Pattern separation and pattern completion: Behaviorally separable processes? *Memory & Cognition*, 49(1), 193–205. <https://doi.org/10.3758/s13421-020-01072-y> (cit. on p. 20)
- Norman, K. A., & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: A complementary-learning-systems approach. *Psychological Review*, 110(4), 611–646. <https://doi.org/10.1037/0033-295X.110.4.611> (cit. on p. 19)
- Nosarti, C., & Froudust-Walsh, S. (2016). Alterations in development of hippocampal and cortical memory mechanisms following very preterm birth. *Developmental Medicine & Child Neurology*, 58, 35–45. <https://doi.org/10.1111/dmcn.13042> (cit. on p. 117)
- O'Mahony, N., Campbell, S., Carvalho, A., Harapanahalli, S., Hernandez, G. V., Krpalkova, L., Riordan, D., & Walsh, J. (2020). Deep Learning vs. Traditional Computer Vision. In K. Arai & S. Kapoor (Eds.), *Advances in Computer Vision* (pp. 128–144). Springer International Publishing. https://doi.org/10.1007/978-3-030-17795-9_10. (Cit. on p. 22)
- Pessoa, L. (2017). A Network Model of the Emotional Brain. *Trends in Cognitive Sciences*, 21(5), 357–371. <https://doi.org/10.1016/j.tics.2017.03.002> (cit. on p. 20)
- Pluta, J., Yushkevich, P., Das, S., & Wolk, D. (2012). In vivo Analysis of Hippocampal Subfield Atrophy in Mild Cognitive Impairment via Semi-Automatic Segmentation of T2-Weighted MRI. *Journal of Alzheimer's Disease*, 31(1), 85–99. <https://doi.org/10.3233/JAD-2012-111931> (cit. on p. 16)
- Poiret, C., Bouyeure, A., Patil, S., Grigis, A., Duchesnay, E., Faillot, M., Bottlaender, M., Lemaitre, F., & Noulhiane, M. (2023). A fast and robust hippocampal subfields segmentation: HSF revealing lifespan volumetric dynamics. *Frontiers in Neuroinformatics*, 17, 1130845. <https://doi.org/10.3389/fninf.2023.1130845> (cit. on pp. 37, 80, 109, 113)
- Poiret, C., Grigis, A., Thomas, J., & Noulhiane, M. (2023, August 14). *Can we Agree? On the Rashomon Effect and the Reliability of Post-Hoc Explainable AI*. arXiv: 2308.07247 [cs, stat]. Retrieved August 15, 2023, from <http://arxiv.org/abs/2308.07247>. (Cit. on p. 97)

- Qiu, Q., Gong, G., Wang, L., Duan, J., & Yin, Y. (2019). Feasibility of Automatic Segmentation of Hippocampus Based on Deep Learning in Hippocampus-Sparing Radiotherapy. *International Journal of Radiation Oncology*Biophysics*, 105(1), E137–E138. <https://doi.org/10.1016/j.ijrobp.2019.06.2177> (cit. on pp. 22, 38)
- Qiu, T., Zeng, Q., Zhang, Y., Luo, X., Xu, X., Li, X., Shen, Z., Li, K., Wang, C., Huang, P., Zhang, M., Dai, S., Xie, F., & for the Alzheimer’s Disease Neuroimaging Initiative. (2022). Altered functional connectivity pattern of hippocampal subfields in individuals with objectively-defined subtle cognitive decline and its association with cognition and cerebrospinal fluid biomarkers. *European Journal of Neuroscience*, 56(12), 6227–6238. <https://doi.org/10.1111/ejn.15860> (cit. on p. 25)
- Rao, Y. L., Ganaraja, B., Murlimanju, B. V., Joy, T., Krishnamurthy, A., & Agrawal, A. (2022). Hippocampus and its involvement in Alzheimer’s disease: A review. *3 Biotech*, 12(2), 55. <https://doi.org/10.1007/s13205-022-03123-4> (cit. on p. 15)
- Richter-Levin, G., & Akirav, I. (2000). Amygdala-Hippocampus Dynamic Interaction in Relation to Memory. *Molecular Neurobiology*, 22(1-3), 011–020. <https://doi.org/10.1385/MN:22:1-3:011> (cit. on p. 15)
- Riggins, T., Geng, F., Blankenship, S. L., & Redcay, E. (2016). Hippocampal functional connectivity and episodic memory in early childhood. *Developmental Cognitive Neuroscience*, 19, 58–69. <https://doi.org/10.1016/j.dcn.2016.02.002> (cit. on p. 113)
- Robin, J., & Moscovitch, M. (2017). Details, gist and schema: Hippocampal–neocortical interactions underlying recent and remote episodic and spatial memory. *Current Opinion in Behavioral Sciences*, 17, 114–123. <https://doi.org/10.1016/j.cobeha.2017.07.016> (cit. on p. 20)
- Rodriguez, P. F. (2009). Neural decoding of goal locations in spatial navigation in humans with fMRI. *Human Brain Mapping*, NA–NA. <https://doi.org/10.1002/hbm.20873> (cit. on p. 23)
- Rolls, E. T. (2016). Pattern separation, completion, and categorisation in the hippocampus and neocortex. *Neurobiology of Learning and Memory*, 129, 4–28. <https://doi.org/10.1016/j.nlm.2015.07.008> (cit. on p. 15)
- Romero, J. E., Coupé, P., & Manjón, J. V. (2017). HIPS: A new hippocampus subfield segmentation method. *NeuroImage*, 163, 286–295. <https://doi.org/10.1016/j.neuroimage.2017.09.049> (cit. on pp. 22, 38)
- Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv:1505.04597 [cs]* (cit. on p. 38).
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215. <https://doi.org/10.1038/s42256-019-0048-x> (cit. on p. 118)
- Sabour, S., Frosst, N., & Hinton, G. E. (2017). Dynamic Routing Between Capsules. *arXiv:1710.09829 [cs]* (cit. on pp. 39, 40, 112).
- Sakai, J. (2020). How synaptic pruning shapes neural wiring during development and, possibly, in disease. *Proceedings of the National Academy of Sciences*, 117(28), 16096–16099. <https://doi.org/10.1073/pnas.2010281117> (cit. on p. 114)
- Sanders, A. F. P., Harms, M. P., Kandala, S., Marek, S., Somerville, L. H., Bookheimer, S. Y., Dapretto, M., Thomas, K. M., Van Essen, D. C., Yacoub, E., & Barch, D. M. (2023). Age-related differences in resting-state functional connectivity from childhood to adolescence. *Cerebral Cortex*, 33(11), 6928–6942. <https://doi.org/10.1093/cercor/bhad011> (cit. on p. 25)

- Sandner, M., Zeier, P., Lois, G., & Wessa, M. (2021). Cognitive emotion regulation withstands the stress test: An fMRI study on the effect of acute stress on distraction and reappraisal. *Neuropsychologia*, *157*, 107876. <https://doi.org/10.1016/j.neuropsychologia.2021.107876> (cit. on p. 23)
- Schapiro, A. C., Turk-Browne, N. B., Botvinick, M. M., & Norman, K. A. (2017). Complementary learning systems within the hippocampus: A neural network modelling approach to reconciling episodic memory with statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1711), 20160049. <https://doi.org/10.1098/rstb.2016.0049> (cit. on p. 19)
- Schlichting, M. L., Zeithamova, D., & Preston, A. R. (2014). CA₁ subfield contributions to memory integration and inference: Memory Integration in CA₁. *Hippocampus*, *24*(10), 1248–1260. <https://doi.org/10.1002/hipo.22310> (cit. on p. 114)
- Schmidt, M. F., Storrs, J. M., Freeman, K. B., Jack, C. R., Turner, S. T., Griswold, M. E., & Mosley, T. H. (2018). A comparison of manual tracing and FreeSurfer for estimating hippocampal volume over the adult lifespan. *Human Brain Mapping*, *39*(6), 2500–2513. <https://doi.org/10.1002/hbm.24017> (cit. on p. 22)
- Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2019). *Green AI*. arXiv: 1907.10597 [cs, stat]. Retrieved August 8, 2023, from <http://arxiv.org/abs/1907.10597>. (Cit. on p. 119)
- Scoville, W. B., & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery & Psychiatry*, *20*(1), 11–21. <https://doi.org/10.1136/jnnp.20.1.11> (cit. on p. 15)
- Scruggs, S. B., Watson, K., Su, A. I., Hermjakob, H., Yates, J. R., Lindsey, M. L., & Ping, P. (2015). Harnessing the Heart of Big Data. *Circulation Research*, *116*(7), 1115–1119. <https://doi.org/10.1161/CIRCRESAHA.115.306013> (cit. on p. 118)
- Sekeres, M. J., Winocur, G., & Moscovitch, M. (2018). The hippocampus and related neocortical structures in memory transformation. *Neuroscience Letters*, *680*, 39–53. <https://doi.org/10.1016/j.neulet.2018.05.006> (cit. on p. 20)
- Shaw, T. B., York, A., Barth, M., & Bollmann, S. (2020). Towards Optimising MRI Characterisation of Tissue (TOMCAT) Dataset including all Longitudinal Automatic Segmentation of Hippocampal Subfields (LASHiS) data. *Data in Brief*, *32*, 106043. <https://doi.org/10.1016/j.dib.2020.106043> (cit. on p. 32)
- Sheline, Y. I., Sanghavi, M., Mintun, M. A., & Gado, M. H. (1999). Depression Duration But Not Age Predicts Hippocampal Volume Loss in Medically Healthy Women with Recurrent Major Depression. *The Journal of Neuroscience*, *19*(12), 5034–5043. <https://doi.org/10.1523/JNEUROSCI.19-12-05034.1999> (cit. on p. 15)
- Shin, L. M. (2006). Amygdala, Medial Prefrontal Cortex, and Hippocampal Function in PTSD. *Annals of the New York Academy of Sciences*, *1071*(1), 67–79. <https://doi.org/10.1196/annals.1364.007> (cit. on p. 15)
- Shing, Y. L., & Lindenberger, U. (2011). The Development of Episodic Memory: Lifespan Lessons: Episodic Memory Across the Lifespan. *Child Development Perspectives*, *5*(2), 148–155. <https://doi.org/10.1111/j.1750-8606.2011.00170.x> (cit. on pp. 109, 113)
- Simis, G., Kostovis, I., Winblad, B., & Bogdanovis, N. (1997). Volume and number of neurons of the human hippocampal formation in normal aging and Alzheimer's disease. *The Journal of Comparative Neurology*, *379*(4), 482–494. [https://doi.org/10.1002/\(SICI\)1096-9861\(19970324\)379:4<482::AID-CNE23.0.CO;2-Z](https://doi.org/10.1002/(SICI)1096-9861(19970324)379:4<482::AID-CNE23.0.CO;2-Z) (cit. on p. 18)

- Simpson, A. L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., van Ginneken, B., Kopp-Schneider, A., Landman, B. A., Litjens, G., Menze, B., Ronneberger, O., Summers, R. M., Bilic, P., Christ, P. F., Do, R. K. G., Gollub, M., Golia-Pernicka, J., Heckers, S. H., Jarnagin, W. R., . . . Cardoso, M. J. (2019). A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv:1902.09063 [cs, eess]* (cit. on p. 32).
- Small, S. A., Schobel, S. A., Buxton, R. B., Witter, M. P., & Barnes, C. A. (2011). A pathophysiological framework of hippocampal dysfunction in ageing and disease. *Nature Reviews Neuroscience*, *12*(10), 585–601. <https://doi.org/10.1038/nrn3085> (cit. on p. 37)
- Smith, S. M., Beckmann, C. F., Andersson, J., Auerbach, E. J., Bijsterbosch, J., Douaud, G., Duff, E., Feinberg, D. A., Griffanti, L., Harms, M. P., Kelly, M., Laumann, T., Miller, K. L., Moeller, S., Petersen, S., Power, J., Salimi-Khorshidi, G., Snyder, A. Z., Vu, A. T., . . . Glasser, M. F. (2013). Resting-state fMRI in the Human Connectome Project. *NeuroImage*, *80*, 144–168. <https://doi.org/10.1016/j.neuroimage.2013.05.039> (cit. on p. 33)
- Smith, S. M., Vidaurre, D., Beckmann, C. F., Glasser, M. F., Jenkinson, M., Miller, K. L., Nichols, T. E., Robinson, E. C., Salimi-Khorshidi, G., Woolrich, M. W., Barch, D. M., Uğurbil, K., & Van Essen, D. C. (2013). Functional connectomics from resting-state fMRI. *Trends in Cognitive Sciences*, *17*(12), 666–682. <https://doi.org/10.1016/j.tics.2013.09.016> (cit. on p. 24)
- Stark, S. M., Frithsen, A., & Stark, C. E. (2021). Age-related alterations in functional connectivity along the longitudinal axis of the hippocampus and its subfields. *Hippocampus*, *31*(1), 11–27. <https://doi.org/10.1002/hipo.23259> (cit. on p. 25)
- Takada, S. H., Motta-Teixeira, L. C., Machado-Nils, A. V., Lee, V. Y., Sampaio, C. A., Polli, R. S., Malheiros, J. M., Takase, L. F., Kihara, A. H., Covolan, L., Xavier, G. F., & Nogueira, M. I. (2016). Impact of neonatal anoxia on adult rat hippocampal volume, neurogenesis and behavior. *Behavioural Brain Research*, *296*, 331–338. <https://doi.org/10.1016/j.bbr.2015.08.039> (cit. on p. 117)
- Tang, P., Yang, P., Nie, D., Wu, X., Zhou, J., & Wang, Y. (2022). Unified medical image segmentation by learning from uncertainty in an end-to-end manner. *Knowledge-Based Systems*, *241*, 108215. <https://doi.org/10.1016/j.knosys.2022.108215> (cit. on p. 39)
- Tejwani, R., Liska, A., You, H., Reinen, J., & Das, P. (2017). Autism Classification Using Brain Functional Connectivity Dynamics and Machine Learning. (Cit. on p. 24).
- Teyler, T. J., & DiScenna, P. (1986). The hippocampal memory indexing theory. *Behavioral Neuroscience*, *100*(2), 147–154. <https://doi.org/10.1037/0735-7044.100.2.147> (cit. on p. 20)
- Teyler, T. J., & Rudy, J. W. (2007). The hippocampal indexing theory and episodic memory: Updating the index. *Hippocampus*, *17*(12), 1158–1169. <https://doi.org/10.1002/hipo.20350> (cit. on p. 20)
- Tsai, Y.-H. H., Srivastava, N., Goh, H., & Salakhutdinov, R. (2020). Capsules with inverted dot-product attention routing. *arXiv preprint arXiv:2002.04764* (cit. on p. 112).
- Tulving, E. (1972). Episodic and Semantic Memory. *Organization of memory* (cit. on p. 15).
- Uematsu, A., Matsui, M., Tanaka, C., Takahashi, T., Noguchi, K., Suzuki, M., & Nishijo, H. (2012). Developmental Trajectories of Amygdala and Hippocampus from Infancy to Early Adulthood in Healthy Individuals (F. Krueger, Ed.). *PLoS ONE*, *7*(10), e46970. <https://doi.org/10.1371/journal.pone.0046970> (cit. on p. 18)

- van den Heuvel, M. P., & Hulshoff Pol, H. E. (2010). Exploring the brain network: A review on resting-state fMRI functional connectivity. *European Neuropsychopharmacology*, *20*(8), 519–534. <https://doi.org/10.1016/j.euroneuro.2010.03.008> (cit. on p. 24)
- Van Erp, T. G. M., Hibar, D. P., Rasmussen, J. M., Glahn, D. C., Pearlson, G. D., Andreassen, O. A., Agartz, I., Westlye, L. T., Haukvik, U. K., Dale, A. M., Melle, I., Hartberg, C. B., Gruber, O., Kraemer, B., Zilles, D., Donohoe, G., Kelly, S., McDonald, C., Morris, D. W., . . . Turner, J. A. (2016). Subcortical brain volume abnormalities in 2028 individuals with schizophrenia and 2540 healthy controls via the ENIGMA consortium. *Molecular Psychiatry*, *21*(4), 547–553. <https://doi.org/10.1038/mp.2015.63> (cit. on p. 16)
- Van Leemput, K., Bakkour, A., Benner, T., Wiggins, G., Wald, L. L., Augustinack, J., Dickerson, B. C., Golland, P., & Fischl, B. (2009). Automated segmentation of hippocampal subfields from ultra-high resolution in vivo MRI. *Hippocampus*, *19*(6), 549–557. <https://doi.org/10.1002/hipo.20615> (cit. on p. 21)
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need*. Retrieved December 11, 2020, from <http://arxiv.org/abs/1706.03762>. (Cit. on p. 112)
- Viriyasaranon, T., & Choi, J.-H. (2022). Object detectors involving a NAS-gate convolutional module and capsule attention module. *Scientific Reports*, *12*(1), 3916. <https://doi.org/10.1038/s41598-022-07898-7> (cit. on p. 112)
- Vos de Wael, R., Larivière, S., Caldaïrou, B., Hong, S.-J., Margulies, D. S., Jefferies, E., Bernasconi, A., Smallwood, J., Bernasconi, N., & Bernhardt, B. C. (2018). Anatomical and microstructural determinants of hippocampal subfield functional connectome embedding. *Proceedings of the National Academy of Sciences*, *115*(40), 10154–10159. <https://doi.org/10.1073/pnas.1803667115> (cit. on pp. 16, 115)
- Whitlock, J. R. (2006). Learning Induces Long-Term Potentiation in the Hippocampus. *Science*, *313*(5790), 1093–1097. <https://doi.org/10.1126/science.1128134> (cit. on p. 15)
- Winocur, G., Moscovitch, M., & Bontempi, B. (2010). Memory formation and long-term retention in humans and animals: Convergence towards a transformation account of hippocampal–neocortical interactions. *Neuropsychologia*, *48*(8), 2339–2356. <https://doi.org/10.1016/j.neuropsychologia.2010.04.016> (cit. on p. 20)
- Winterburn, J. L., Pruessner, J. C., Chavez, S., Schira, M. M., Lobaugh, N. J., Voineskos, A. N., & Chakravarty, M. M. (2013). A novel in vivo atlas of human hippocampal subfields using high-resolution 3T magnetic resonance imaging. *NeuroImage*, *74*, 254–265. <https://doi.org/10.1016/j.neuroimage.2013.02.003> (cit. on p. 32)
- Wisse, L. E. M., Chételat, G., Daugherty, A. M., Flores, R., Joie, R., Mueller, S. G., Stark, C. E. L., Wang, L., Yushkevich, P. A., Berron, D., Raz, N., Bakker, A., Olsen, R. K., & Carr, V. A. (2021). Hippocampal subfield volumetry from structural isotropic 1 mm³ MRI scans: A note of caution. *Human Brain Mapping*, *42*(2), 539–550. <https://doi.org/10.1002/hbm.25234> (cit. on p. 21)
- Wisse, L. E., Daugherty, A. M., Olsen, R. K., Berron, D., Carr, V. A., Stark, C. E., Amaral, R. S., Amunts, K., Augustinack, J. C., Bender, A. R., Bernstein, J. D., Boccardi, M., Bocchetta, M., Burggren, A., Chakravarty, M. M., Chupin, M., Ekstrom, A., de Flores, R., Insausti, R., . . . for the Hippocampal Subfields Group. (2017). A harmonized segmentation protocol for hippocampal and parahippocampal subregions: Why do we need one and what are the key goals?: A HARMONIZED HIPPOCAMPAL SUBFIELD PROTOCOL: KEY GOALS AND IMPACT. *Hippocampus*, *27*(1), 3–11. <https://doi.org/10.1002/hipo.22671> (cit. on p. 21)

- Wisse, L., Gerritsen, L., Zwanenburg, J., Kuijf, H., Luijten, P., Biessels, G., & Geerlings, M. (2012). Subfields of the hippocampal formation at 7T MRI: In vivo volumetric assessment. *NeuroImage*, *61*(4), 1043–1049. <https://doi.org/10.1016/j.neuroimage.2012.03.023> (cit. on pp. 21, 34)
- Wisse, L., Kuijf, H., Honingh, A., Wang, H., Pluta, J., Das, S., Wolk, D., Zwanenburg, J., Yushkevich, P., & Geerlings, M. (2016). Automated Hippocampal Subfield Segmentation at 7T MRI. *American Journal of Neuroradiology*, *37*(6), 1050–1057. <https://doi.org/10.3174/ajnr.A4659> (cit. on p. 32)
- Yan, B., Xu, X., Liu, M., Zheng, K., Liu, J., Li, J., Wei, L., Zhang, B., Lu, H., & Li, B. (2020). Quantitative Identification of Major Depression Based on Resting-State Dynamic Functional Connectivity: A Machine Learning Approach. *Frontiers in Neuroscience*, *14*, 191. <https://doi.org/10.3389/fnins.2020.00191> (cit. on p. 24)
- Yang, J., Gohel, S., & Vachha, B. (2020). Current methods and new directions in resting state fMRI. *Clinical Imaging*, *65*, 47–53. <https://doi.org/10.1016/j.clinimag.2020.04.004> (cit. on p. 24)
- Yang, X., Goh, A., Chen, S.-H. A., & Qiu, A. (2013). Evolution of hippocampal shapes across the human lifespan: Hippocampal Shapes in Aging. *Human Brain Mapping*, *34*(11), 3075–3085. <https://doi.org/10.1002/hbm.22125> (cit. on p. 18)
- Yang, Z., Zhuang, X., Mishra, V., Sreenivasan, K., & Cordes, D. (2020). CAST: A multi-scale convolutional neural network based automated hippocampal subfield segmentation toolbox. *NeuroImage*, *218*, 116947. <https://doi.org/10.1016/j.neuroimage.2020.116947> (cit. on pp. 22, 38)
- Yassa, M. A., & Stark, C. E. (2011). Pattern separation in the hippocampus. *Trends in Neurosciences*, *34*(10), 515–525. <https://doi.org/10.1016/j.tins.2011.06.006> (cit. on p. 20)
- Yassa, M. A., Stark, S. M., Bakker, A., Albert, M. S., Gallagher, M., & Stark, C. E. (2010). High-resolution structural and functional MRI of hippocampal CA3 and dentate gyrus in patients with amnesic Mild Cognitive Impairment. *NeuroImage*, *51*(3), 1242–1252. <https://doi.org/10.1016/j.neuroimage.2010.03.040> (cit. on p. 25)
- Yushkevich, P. A., Amaral, R. S., Augustinack, J. C., Bender, A. R., Bernstein, J. D., Boccardi, M., Bocchetta, M., Burggren, A. C., Carr, V. A., Chakravarty, M. M., Chételat, G., Daugherty, A. M., Davachi, L., Ding, S.-L., Ekstrom, A., Geerlings, M. I., Hassan, A., Huang, Y., Iglesias, J. E., . . . Zeineh, M. M. (2015). Quantitative comparison of 21 protocols for labeling hippocampal subfields and parahippocampal subregions in in vivo MRI: Towards a harmonized segmentation protocol. *NeuroImage*, *111*, 526–541. <https://doi.org/10.1016/j.neuroimage.2015.01.004> (cit. on p. 21)
- Yushkevich, P. A., Pluta, J. B., Wang, H., Xie, L., Ding, S.-L., Gertje, E. C., Mancuso, L., Kliot, D., Das, S. R., & Wolk, D. A. (2015). Automated volumetry and regional thickness analysis of hippocampal subfields and medial temporal cortical structures in mild cognitive impairment: Automatic Morphometry of MTL Subfields in MCI. *Human Brain Mapping*, *36*(1), 258–287. <https://doi.org/10.1002/hbm.22627> (cit. on pp. 22, 32, 37)
- Yushkevich, P. A., Wang, H., Pluta, J., Das, S. R., Craige, C., Avants, B. B., Weiner, M. W., & Mueller, S. (2010). Nearly automatic segmentation of hippocampal subfields in in vivo focal T2-weighted MRI. *NeuroImage*, *53*(4), 1208–1224. <https://doi.org/10.1016/j.neuroimage.2010.06.040> (cit. on p. 32)
- Zhu, H., Shi, F., Wang, L., Hung, S.-C., Chen, M.-H., Wang, S., Lin, W., & Shen, D. (2019). Dilated Dense U-Net for Infant Hippocampus Subfield Segmentation. *Frontiers in Neuroinformatics*, *13*, 30. <https://doi.org/10.3389/fninf.2019.00030> (cit. on pp. 22, 38)

Ziegler, G., Dahnke, R., Jäncke, L., Yotter, R. A., May, A., & Gaser, C. (2012). Brain structural trajectories over the adult lifespan. *Human Brain Mapping*, 33(10), 2377–2389. <https://doi.org/10.1002/hbm.21374> (cit. on p. 18)

A Appendix 1: Hippocampal Subfield Volumes and Memory Discrimination in the Developing Brain

Antoine Bouyeure, Sandesh Patil, Franck Mauconduit, Clément Poiret, Damien Isai, Marion Noulhiane (2021). Hippocampal subfield volumes and memory discrimination in the developing brain. *Hippocampus*. <https://doi.org/10.1002/hipo.23385>.

Abstract




Goal: Pattern separation, supported by hippocampal subfields like dentate gyrus (DG) and CA3, is important for episodic memory development. We aimed to examine hippocampal subfield volumes and memory discrimination, a proxy for pattern separation, in children.

Methods: 26 children (5-12 years) underwent MRI scanning and a memory discrimination task. Hippocampal subfields (DG, CA1, CA2/3, subiculum) were manually segmented. Associations between subfield volumes and memory discrimination were assessed.

Results: Memory discrimination improved with age. CA1 and subiculum volumes increased with age, but not DG or CA2/3. CA2/3 volume positively correlated with memory discrimination. A negative association between subiculum and memory discrimination in younger children shifted to positive in older children.

Conclusions: During childhood, CA1 and subiculum volumes increased, while CA2/3 related to memory discrimination. Subiculum-memory links changed with age. Our findings clarify hippocampal subfield contributions to developing pattern separation.

Hippocampal subfield volumes and memory discrimination in the developing brain

Antoine Bouyeure^{1,2}  | Sandesh Patil^{1,2} | Franck Mauconduit³  |
Clément Poiret^{1,2} | Damien Isai^{1,2} | Marion Noulhiane^{1,2} 

¹UNIACT, NeuroSpin, CEA, Université Paris-Saclay, Gif-sur-Yvette, France

²UMR1141, Inserm, Université de Paris, Paris, France

³BAOBAB, NeuroSpin, CEA, CNRS, Université Paris-Saclay, Gif-sur-Yvette, France

Correspondence

Marion Noulhiane, UNIACT, NeuroSpin, CEA, Université Paris-Saclay, Gif-sur-Yvette 91191, France.

Email: marion.noulhiane@u-paris.fr

Funding information

Fondation de France, Grant/Award Number: 00070721; Fondation Mustela

Abstract

The ability to keep distinct memories of similar events is underpinned by a type of neural computation called pattern separation (PS). Children typically report coarse-grained memories narratives lacking specificity and detail. This lack of memory specificity is illustrative of an immature or impaired PS. Despite its importance for the ontogeny of memory, data regarding the maturation of PS during childhood is still scarce. PS is known to rely on the hippocampus, particularly on hippocampal subfields DG and CA3. In this study, we used a memory discrimination task, a behavioral proxy for PS, and manually segmented hippocampal subfields volumes in the hippocampal body in a cohort of 26 children aged from 5 to 12 years. We examined the association between subfields volumes and memory discrimination performance. The main results were: (1) we showed age-related differences of memory discrimination suggesting a continuous increase of memory performance during early to late childhood. (2) We evidenced distinct associations between age and the volumes of hippocampal subfield, suggesting distinct developmental trajectories. (3) We showed a relationship between memory discrimination performance and the volumes of CA3 and subiculum. Our results further confirm the role of CA3 in memory discrimination, and suggest to scrutinize more closely the role of the subiculum. Overall, we showed that hippocampal subfields contribute distinctively to PS during development.

KEYWORDS

development, episodic memory, hippocampus, pattern separation, segmentation, subfields

1 | INTRODUCTION

A key aspect of episodic memory (EM) is the formation of memory representations of events without these representations interfering with each other. This interference is more likely to occur if the represented events are highly similar. For example, two distinct but ordinary school days share a certain amount of common features. The resulting feature overlap could lead to memory interference, which would impair the quality and specificity of recall. A type of neural computation, called pattern separation (PS), was theorized decades ago as the putative mechanism by which similar representations could be discriminated in memory (Marr &

Brindley, 1971); Complementary Learning Systems Theory: (Norman & O'Reilly, 2003). PS is the process by which distinct neural activation patterns that do not overlap are assigned to similar memory representations (Norman & O'Reilly, 2003; Yassa & Stark, 2011). In other words, orthogonal memory representations are created from similar inputs, reducing memory interference. Strong evidence now suggests that PS is indeed the mechanism by which similar memory representations are discriminated and that its neural substrate lies in the hippocampus, specifically hippocampal subfields dentate gyrus (DG) and *cornu Ammonis* area 3 (CA3; Bakker, Kirwan, et al., 2008; Mankin et al., 2015; Nakashiba et al., 2012; Yassa & Stark, 2011).

At the behavioral level, PS is usually assessed with the Mnemonic Similarity Task (MST; (Stark et al., 2019; Yassa, Mattfeld, et al., 2011). The MST was conceived as a behavioral proxy for PS by eliciting discrimination judgments between highly similar items. The ability to discriminate between identical and similar representations of previously presented items, memory discrimination, is thus thought to tap on PS-dependent processes. The direct involvement of DG and CA3 in memory discrimination, and by extension in PS, have been shown in humans by functional magnetic resonance imaging (fMRI) studies, as previously suggested by computational models of hippocampal function (Bakker, Kirwan, et al., 2008; Berron et al., 2016; Leutgeb et al., 2007; Myers & Scharfman, 2009; Neunuebel & Knierim, 2014; Schmidt et al., 2012); Yassa, Lacy, et al., 2011; Yassa, Mattfeld, et al., 2011). Structural magnetic resonance imaging (sMRI) studies also showed associations between in DG and CA3 volumes and memory discrimination performance (Doxey & Kirwan, 2015; Stark & Stark, 2017).

Memory narratives reported by young children are coarse-grained and lack specificity and detail; this could thus suggest that “immature” memory discrimination competence during childhood could play a key role in the ontogeny of EM (Canada et al., 2019; Ramsaran et al., 2019). Despite this, the development of PS during childhood and the relationship between PS development and hippocampal subfields maturation is poorly known. The scarcity of available data is mainly because this question emerged as an object of study only in recent years (e.g., Benear et al., 2020; Canada et al., 2019; Hassevoort et al., 2020; Keresztes et al., 2017; Ngo et al., 2018, 2019). The current study aims to contribute to our understanding of the relationship between memory discrimination and hippocampal subfields during childhood by examining the association between memory discrimination performance and the volumes of manually segmented subfields in children aged from 5 to 12 years. As the acquisition of fMRI data in young children is particularly challenging, this correlational sMRI approach is particularly suited to assess the relationship between memory competence and hippocampal subfields in children.

1.1 | Development of memory discrimination

To date, a few studies investigated the development of PS during childhood using the MST. These studies showed age-related differences in memory discrimination performance, but there are discrepancies between the suggested maturational timelines (Ngo et al., 2018; Rollins & Cloude, 2018). For example, Ngo et al. (2018) suggested an early maturation of memory discrimination, with adult-like performance reached around 6 years of age (Ngo et al., 2018), while Rollins and Cloude (2018) suggested a more protracted maturation, with age-related differences observed until 9–10 years of age. The development of PS early in life thus needs to be further investigated. An earlier or later maturation of PS would induce different interpretations regarding the relationship between the development of memory discrimination and the structural and functional maturation of its neural substrates (i.e., the hippocampal subfields).

1.2 | Maturation of hippocampal subfields

The maturation of hippocampal subfields during childhood is protracted. An initial phase of rapid maturational changes during the first 2–3 years of life (Lavenex & Banta Lavenex, 2013; Olson & Newcombe, 2013; Utsunomiya et al., 1999), is followed by a phase of more modest age-related changes (e.g., volumetric increases or decreases) which extends into adulthood (Krogsrud et al., 2014; Lee et al., 2014; Riggins et al., 2018; Tamnes et al., 2018). These age-related volumetric differences could be related to several causes, including neurogenesis, synaptogenesis, synaptic pruning, or myelination, among others (Lenroot & Giedd, 2006; Riggins et al., 2018). Importantly, age-differences in hippocampal subfields volumes during childhood and adolescence have been associated to age-differences in memory performance (e.g., Lee et al., 2014; Riggins et al., 2018; Tamnes et al., 2018), showing their functional significance.

To date, two studies have directly examined the association between differences in hippocampal subfields volumes and differences in memory discrimination performance in the developing brain. Canada et al. (2019) examined the individual contribution of hippocampal subfields' volumes to memory discrimination performance in 4–8 years old children. They showed an age-mediated association between combined DG/CA3 volume and memory discrimination, highlighting the pivotal role of these subfields in PS. However, as these two subfields were combined in a single ROI in the aforementioned study, a separate assessment of the roles of DG and CA3 is still lacking. Another study, Keresztes et al. (2017), used a multivariate approach describing the shared variance between age and all hippocampal subfields' volume in a single latent variable, which was positively correlated to memory discrimination performance in 6–14 years old children, and young adults. This suggested that hippocampal subfield maturation as a whole is related to memory discrimination performance, suggesting inter-dependency in the maturational processes of each subfield. Indeed, while most of the literature points to the privileged role of DG and CA3 in PS, there is also data suggesting the contribution of other subfields, for example, the subiculum (Potvin et al., 2009) or CA1 (Hanert et al., 2019). Therefore, more investigations are necessary to disentangle the association between hippocampal subfields and PS in the developing brain.

1.3 | Current study

Here, we aimed to contribute to the understanding of the relationship between hippocampal maturation and PS during development. We assessed memory discrimination performance as a proxy for PS and manually segmented hippocampal subfields on the MRI images of 26 children aged from 5 to 12 years old. This allowed us to examine the association between hippocampal subfields' volumes and memory discrimination performance to investigate how PS is related to hippocampal subfields in the developing brain. Our hypotheses were the following: (1) we expected to observe a positive correlation between age and memory discrimination performance. (2) We expected to

observe associations between age and hippocampal subfields volumes, with specific associations for each subfield, suggesting developmental trajectories specific to each subfield. (3) We hypothesized that differences in memory discrimination performance would be associated with differences in hippocampal subfields volumes, particularly for DG and/or CA2-3, which are known to be the main neural correlates of PS, but not necessarily restricted to them (see Keresztes et al., 2017).

2 | METHODS

2.1 | Participants

Fifty children aged from 4–12 years old (mean: 8.27 years, standard deviation: 2.3 years) participated in this study as part of a larger study on the neural correlates of EM during development. Fifty-five percent of the participants were males and 45% females. Among our 50 participants, 11 children had no or incomplete data, resulting a sample of 39 children with neuroimaging data. Data acquisition was performed under the regulations of an appropriate Ethical Committee board (CPP 2011-A00058-33).

2.2 | MRI acquisition

Imaging data were collected at the NeuroSpin research center, CEA, Gif-sur-Yvette, France. Children first followed an MRI training session on a mock scanner set in a children-friendly environment. They were told a compelling story, making them astronauts on a mission to understand the brain, taking aboard a spaceship (the scanner), and wearing a space helmet (the head coil). For the mission to succeed, children were told to try staying still as much as possible, for the scanner to take accurate pictures of their brains. Once the children were familiarized with the sonic and visual environment of the scanner, the acquisition begun. Images were acquired on a Siemens PRISMA 3T scanner (Siemens Medical Solutions, Erlangen, Germany) with a 64-channel head coil. The animation movie *Wall-E* (Pixar Animation Studios) was shown to children during the scanning sessions to bolster engagement and reduce head motion caused by intolerance to noise and sensation of boredom.

We first acquired a T1-weighted MPRAGE volume (TR = 300 ms, TE = 2.98 ms, 0.9 mm isotropic resolution, 175 slices, acceleration factor GRAPPA2). The resulting image was used to localize the hippocampus in a subsequent oblique coronal T2-weighted structural sequence, which was acquired perpendicular to the main axis of the hippocampus (interleaved TR = 3970 ms, TE = 89 ms, FOV 173 mm, 0.45 × 0.45 mm in-plane resolution, 2.1 mm through-plane resolution, 46 slices). Two T2w images were acquired for each subject and interleaved to produce the full T2w scan.

2.3 | MRI preprocessing

T1w data was corrected for B1 bias and skull-stripped, using the fsl anat pipeline (fsl.fmrib.ox.ac.uk/fsl/fslwiki/fsl_anat). Each of the two

T2w sequences were 2D aligned using fsl FLIRT (3 degrees of freedom). This ensures that the images are correctly aligned on the x and y-axis (in-plane resolution), but remain at their respective through-plane location. The co-registered images were then interleaved to obtain a single T2w sequence for each participant and skull-stripped. The skull-stripped T1 image was registered to the skull-stripped T2 image using freesurfer's (Zöllei et al., 2020) MRI robust register command with the following parameters: 6 degrees of freedom (rigid body alignment), normalized mutual information cost function, and no prior initialization. After visual inspection of the quality of registration, the orientation of the registered T1 image was swapped to match that of the T2 image, that is, to have an oblique orientation.

2.4 | Segmentation of hippocampal subfields

Visual inspection of T2w data prior to segmentation showed that several subjects had poor data quality due to excessive head motion. This was mainly due to the fact that the T2w sequence acquisition was performed at the end of a 45 min-long neuroimaging protocol. Thus, among the 39 subjects with neuroimaging data, 11 were excluded because of insufficient data quality to perform a reliable segmentation. Exclusion of poor data was made by experimenters with expertise in structural segmentation of hippocampal subfields, based on the careful visual inspection of each MR image. Images where identification of subfields boundaries was compromised by motion, resulting in blurred data, were thus excluded from our sample. Overall, this resulted in a final segmentation sample of 28 subjects, giving 56 data points (2 per hemisphere) for each subfield.

We manually segmented the 28 retained images using ITK-SNAP (Yushkevich et al., 2006). Only subfields inside the hippocampal body were segmented, as subfields in the hippocampal body have particularly identifiable anatomical landmarks. This ensured a reliable segmentation given our resolution (for similar approaches, see (Lee et al., 2014; Mueller et al., 2011; Neylan et al., 2010; Yushkevich et al., 2010). The anterior limit of the body was identified based on the presence of the uncal apex: the body section began one slice posterior to the uncal apex (Bernasconi et al., 2003). The posterior limit of the body was identified one slice anterior to the coronal slice at which the colliculi disappeared (Bernasconi et al., 2003).

Hippocampal subfields were then manually segmented in the hippocampal body following relevant anatomical landmarks following the protocol of Dalton et al. (2017). This protocol was chosen because of its precision and exhaustivity in terms of segmentation procedure details, because it allows to segment separately the hippocampal body from the hippocampal head and tail, and because it was conceived as a synthesis of several widely used segmentation protocols (see Dalton et al., 2017). Each subfield per slice of the hippocampal body was segmented following the methodology described by Dalton et al. (2017) in this order: DG/CA4, CA2-3, CA1, and subiculum. Table 1 shows the boundary landmarks used for manually segmenting the hippocampal subfields. As CA4 is often considered as a part of the DG (often called the hilar region), DG/CA4 will be referred to as the DG in the manuscript for simplicity.

The reliability of the obtained manual segmentations was assessed by computing an inter-rater reliability index between two independent tracers. Of the 28 subjects with usable data manually segmented by one rater (S.P.), 10 subjects have been re-segmented by another rater (D.I.) using the same segmentation protocol to assess inter-rater reliability. Each rater was blind to age, sex, and memory performance of participants. According to Bartko (1991), it has been agreed that an inter-rater reliability (as measured through Dice's index [Dice, 1949]) ≥ 0.7 represents good to excellent spatial agreement. For the left hippocampus, inter-rater reliability was 0.83 for CA1, 0.68 for CA2-3, 0.88 for DG, and 0.77 for subiculum. For the right hippocampus, inter-rater reliability was 0.79 for CA1, 0.67 for CA2-3, 0.84 for DG, and 0.77 for the subiculum. With mean inter-reliability of 0.79 and 0.77 for the left and right hippocampus, our results are consistent with inter-rater reliability found in previous studies (e.g., Palombo et al., 2013), and deemed satisfactory.

The obtained ROI (CA1, CA2-3, DG, and subiculum) volumes (Figure 1) were then corrected for intracranial volume (ICV) using a covariance approach (Raz et al., 2005). Correcting for ICV is necessary to account for the fact that differences in ROI volumes are related to differences in head size, as estimated by ICV. ICV was computed as the volume of a mask representing the whole brain, which was

estimated by combining three brain extraction methods: Freesurfer "recon-all" (Desikan et al., 2006), ANTs "BrainExtraction" (Avants et al., 2009) with OASIS template and SPM8 (Ashburner & Friston, 2005). The brain masks resulting from these 3 methods were averaged, and manually corrected when necessary (for similar approaches, see Bender et al., 2018; Canada et al., 2019; Keresztes et al., 2017; Riggins et al., 2018). To adjust ROI volumes for ICV, we computed the slope of the linear regression between each ROI volume (including the total hippocampal body volume) and ICV (β_{ICV}) to determine the statistical relationship between ICV and ROI volume. Then, we multiplied each β_{ICV} slope by each subject's mean-centered ICV and subtracted this product from raw ROI volumes (see Schlichting et al., 2019, for a similar approach). This removes the statistical relationship between ROI volumes and ICV volume. Thus, the corrected subfields volumes were obtained as follow:

$$Volume_{corrected} = Volume_{Raw\ i} - \beta(ICV_{Raw\ i} * ICV_{Mean\ i})$$

To verify that the reported statistical effects described in this study were not the sole product of this adjustment procedure, we conducted analyses on raw volumes first and then on corrected volumes. Only the latter analyses are reported. The age-related trajectories of raw (unadjusted) volumes are presented in Figure S1.

TABLE 1 Landmarks used for segmentation of hippocampal subfields

Subfield	Location	Landmark
Cornu Ammonis 1 (CA1)	Lateral to CA2-3 and the DG	Lateral border—started at the inferior point of the dorsomedial border and drew the ventromedial border of the CA1 mask by following the VHS until we reached the center of the DG/CA4 mask. Dorsolateral border—we create a straight line as the inferior border and draw a line following the dorsolateral wall of the hippocampus until the starting point.
Cornu Ammonis 2-3 (CA2-3)	Dorsal to the DG	Lateral border with CA1—straight diagonal line from the dorsal portion of the VHS where it begins to turns ventrally to the dorsolateral corner of the superior wall of the hippocampus. Medial border—follow the anatomical limit of the superior wall in a medial direction toward the medial extent of the lateral external digitation until reached to the point where we started.
Dentate Gyrus (DG)/CA4	Center portion of the hippocampus	Ventral, lateral, and dorsal border—traced the dark line of the VHS by beginning on the lateral extent of the uncus sulcus until we reached the point above at which we started. Medial border—draw a line from the dorsal limit of the VHS to the ventral direction until we reach the point where we started.
Subiculum	Ventral portion of the hippocampus, medial to CA1	The inferior boundary of the subiculum from the parahippocampal cortex was demarcated at the nadir of the concavity in the medial wall between the collateral sulcus and hippocampus. The boundary between the CA1 and subiculum was based on the CA1 mask. Indeed, the lateral border of the subiculum is the ventromedial border of the previously created CA1 mask.

Note: These landmarks are adapted from the segmentation protocol described in Dalton et al. (2017).

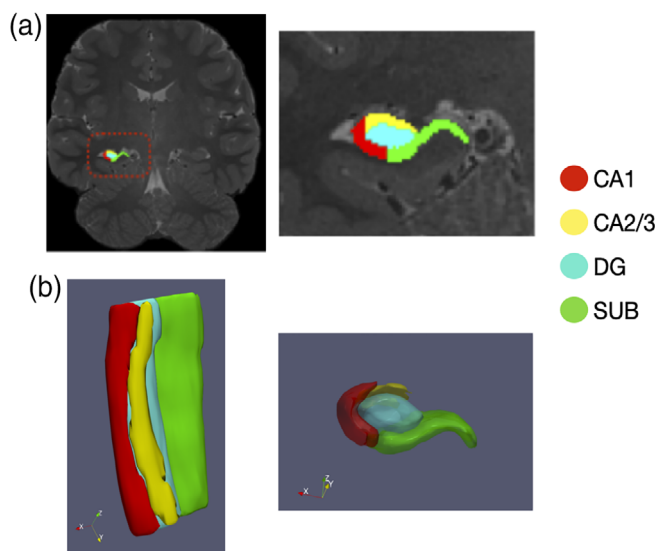


FIGURE 1 (a) Manually segmented subfields of one subject, superimposed on its T2w scan. (b) Three-dimensional surface reconstruction of the hippocampus of another subject, in dorsal and coronal views. CA, cornu ammonis; DG, dentate gyrus; SUB, subiculum

2.5 | Behavioral assessments

After MRI acquisitions, children were given a battery of cognitive tests. This included the MST (Stark et al., 2019), and the children's version of Raven's Colored Progressive Matrices task (PM47; Raven et al., 2003), a measurement of fluid intelligence.

An incidental encoding version of the MST, as described in Ngo et al. (Ngo et al., 2018), was used to assess memory discrimination as a behavioral proxy for PS. One hundred images of objects were selected from Craig Stark's Database specifically designed for the MST task (<http://faculty.sites.uci.edu/starklab/mnemonic-similarity-task-mst/>). Images were chosen for their appeal and familiarity to children (e.g., toys and food). The MST consisted in an incidental encoding phase where the participant had to perform indoors/outdoors judgments, followed by a test phase evaluating memory discrimination. In an initial incidental encoding phase, 60 pictures of objects were displayed one by one in a randomized order. The participant looked at each picture for 3 s. The picture then disappeared in order to control for duration of stimuli exposure. From that moment, the participant had to state whether the object seen in the picture was something used outdoors or indoors. The participant had 3 s to provide orally its answer, which was recorded by the experimenter, after which the experiment proceeded automatically to the next trial. In the subsequent test phase, 60 pictures were displayed one by one in a randomized order. Out of these 60 pictures, 20 were already presented during the incidental encoding phase ("target" trials); 20 were similar, but not identical, to the pictures presented during the incidental encoding phase ("lure" trials); and 20 were totally new ("foil" trials). For each picture, the participant had to make "old," "new," or "similar" judgments, in order to correctly identify the target, lure, or foil trials, respectively. The discrimination phase was preceded

by six training trials (two training trials per type of response). The responses were given orally by the participant and recorded by the examiner.

Following previous studies (e.g., Ngo et al., 2018), PS was assessed through memory discrimination, which is the percentage of correct "old" responses from which was subtracted the percentage of responses were subjects incorrectly gave "old" responses to "lure" items. Additionally, we also computed a measurement of item memory as the percentage of correct "old" responses from which was subtracted the percentage of responses were subjects incorrectly gave "new" answers to "target" items. While memory discrimination is a proxy for PS, using item memory as an additional measurement of a different memory function allowed us to control for the specificity of our findings, following previous studies (Canada et al., 2019; Ngo et al., 2018).

2.6 | Statistical analyses

Out of the 28 subjects with usable segmentation data, 2 were excluded from the analyses because of lack of compliance during the behavioral assessments, resulting in lacking or unusable MST data. Therefore, 26 subjects were included in our final sample (mean age: 7.95, median age: 7.45, age standard deviation: 2.25, 16 males). Because we had no hypotheses regarding hemispherical differences between subfields, data points were collapsed across hemispheres, resulting in total bilateral subfield volume of each subfield (for a similar approach, see Canada et al., 2019).

2.6.1 | Age-related differences of memory discrimination and item memory

We examined potential age-related differences in memory discrimination and item memory by conducting regressions models predicting item memory and memory discrimination with age. To look for potential nonlinear relations, we also tested models included quadratic and cubic age terms. The best fitting model was chosen with a hierarchical linear regression approach: we used an ANOVA test to assess if the model including more variables predicted memory discrimination or item memory above and beyond the contribution of a single age predictor. We controlled for multiple comparisons by adjusting raw p -values with an FDR procedure (Benjamini & Hochberg, 1995).

2.6.2 | Age-related differences of hippocampal subfields' volumes

We examined age-related differences in hippocampal subfields' volumes by conducting regressions models predicting the volume of each hippocampal subfields with age. Because the development of hippocampal subfields is heterogeneous (Østby et al., 2009), we do not necessarily

expect a linear relation between subfields' volume and age. Therefore, we also tested models including quadratic and cubic age terms. Model selection was performed with a hierarchical regression approach similarly to the previous section. Sex was added as a covariate in a second step to control for potential sex-related differences of hippocampal subfields volumes. We controlled for multiple comparisons by adjusting raw p -values with an FDR procedure across models (Benjamini & Hochberg, 1995).

2.6.3 | Association between hippocampal subfields' volumes and memory discrimination

Finally, we examined the association between memory discrimination performance and the volumes of hippocampal subfields. We used a multilinear regression model per subfield ($N = 4$), predicting memory discrimination with the volume of each subfield. Age, sex, and Raven's matrix standard scores, were added in the models as covariates to control for their potential confounding effects. To assess the significance of the association between memory discrimination performance and hippocampal subfields, we used a hierarchical linear regression approach. In the first step, we used an ANOVA test to assess if the model including the tested hippocampal subfield volume predicted memory discrimination above and beyond the contribution of the control variables (age, sex, and Raven's matrix standard scores). The significance of the model comparison ANOVA was deemed to express the significant contribution of hippocampal subfields' volumes for explaining the variance of memory discrimination performance. In the second step, we added an interaction term between the tested subfield's volume and age, to test for potential interactions between age and subfields volumes in relation to memory discrimination performance (see Canada et al., 2019, for a similar approach). We used an ANOVA test to assess if the model including the interaction term predicted memory discrimination above and beyond the contribution of the model without the interaction term.

These analyses were also performed with item memory and Raven's matrix standard scores as the dependent variables, in order to assess the specificity of our findings. The correlogram showing correlations between memory measures and subfields volumes is presented in Figure S2.

The p -values of all tested models were adjusted with FDR to adjust for multiple comparisons. Moreover, p -values of the predictive variables inside each model were corrected with FDR separately for each model. An alpha value of .05 was used for all analyses.

3 | RESULTS

3.1 | Age-related differences of memory discrimination and item memory

Memory discrimination was positively associated to age using a simple linear model ($F = 9.18$, $R^2 = 0.28$, $p = .011$). Models including quadratic and cubic age terms did not explained memory discrimination variance above and beyond the variance explained by the linear model

(quadratic model vs. linear model: $F = 0.58$, $p = .45$; cubic vs. linear model: $F = 1.14$, $p = .33$). Item memory was not correlated to age, in all tested models (linear model: $F = 0.06$, $R^2 = 0.003$, $p = .79$). Figure 2 illustrates the association between age and these two behavioral measurements.

Memory discrimination and item memory performances were not correlated, when controlling for sex ($r = -.08$, $p = .71$) or sex and age ($r = -.11$, $p = .58$).

Memory discrimination and Raven's matrix standard scores were not correlated, when controlling for sex ($r = -0.10$, $p = .59$), or sex and age ($r = -0.01$, $p = .94$). Similarly, item memory and Raven's matrix standard scores were not correlated, controlling for sex ($r = -0.18$, $p = .37$), or sex and age ($r = -0.16$, $p = .41$).

3.2 | Age-related differences of hippocampal subfields' volumes

We examined age-related differences of hippocampal subfields volumes using regressions models, which were compared with a hierarchical regression approach. The plots illustrating the fitting models for each subfield are shown in Figure 3.

The linear model predicting CA1 with a simple age term was significant ($F = 8.32$, $R^2 = 0.26$, $p = .02$). Adding sex as a covariate in the model, sex was not a significant predictor of CA1 volume, while age was still significant ($t = 2.8$, $p = .01$). The model with a quadratic age term did not predict the variance of CA1 volume above and beyond than the variance predicted by the linear model ($F = 0.29$, $p = .59$), as well as the model including a cubic age term ($F = 0.30$, $p = .73$).

For CA2-3, the linear model was not significant ($F = 1.34$, $R^2 = 0.05$, $p = .33$). Adding quadratic and cubic terms did not explain the variance of CA2-3 volume above and beyond that explained by the linear model (quadratic vs. linear: $F = 1.24$, $p = .27$; cubic vs. linear: $F = 0.62$, $p = .54$). Sex was not a significant predictor of CA2-3 volume and adding sex as a covariate did not change the non-significance of age to predict CA2-3 volume.

For DG, the linear model was not significant ($F = 0.09$, $R^2 = 0.004$, $p = .75$). Neither adding a quadratic age term ($F = 1.36$, $p = .25$), nor a cubic age term ($F = 1.59$, $p = .22$) explained DG variance above and beyond that explained by the linear model. Sex was not a significant predictor of CA2-3 volume and adding sex as a covariate did not change the nonsignificance of age.

For the subiculum, the linear model was significant ($F = 6.37$, $R^2 = 0.21$, $p = .031$). Adding sex as a covariate in the model, sex was not a significant predictor of CA1 volume, while age was still significant ($t = 2.43$, $p = .02$). Neither adding a quadratic age term ($F = 0.65$, $p = .41$), nor a cubic age term ($F = 0.33$, $p = .72$) explained subiculum variance above and beyond that explained by the linear model.

Finally, we examined the association between total volume of the hippocampal body with age, controlling for sex. We observed a positive relationship between age and hippocampal body volume in the linear model ($F = 9.95$, $R^2 = 0.29$, $p = .02$). Adding quadratic ($F = 0.03$, $p = .84$) or cubic ($F = 0.16$, $p = .84$) age terms did not

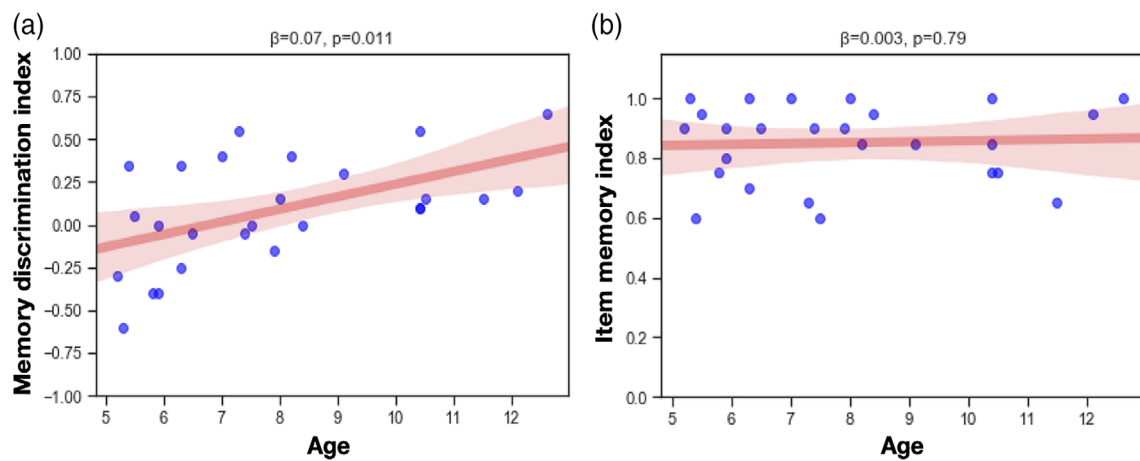


FIGURE 2 (a) Regression between age and memory discrimination; (b) regression between age and item memory

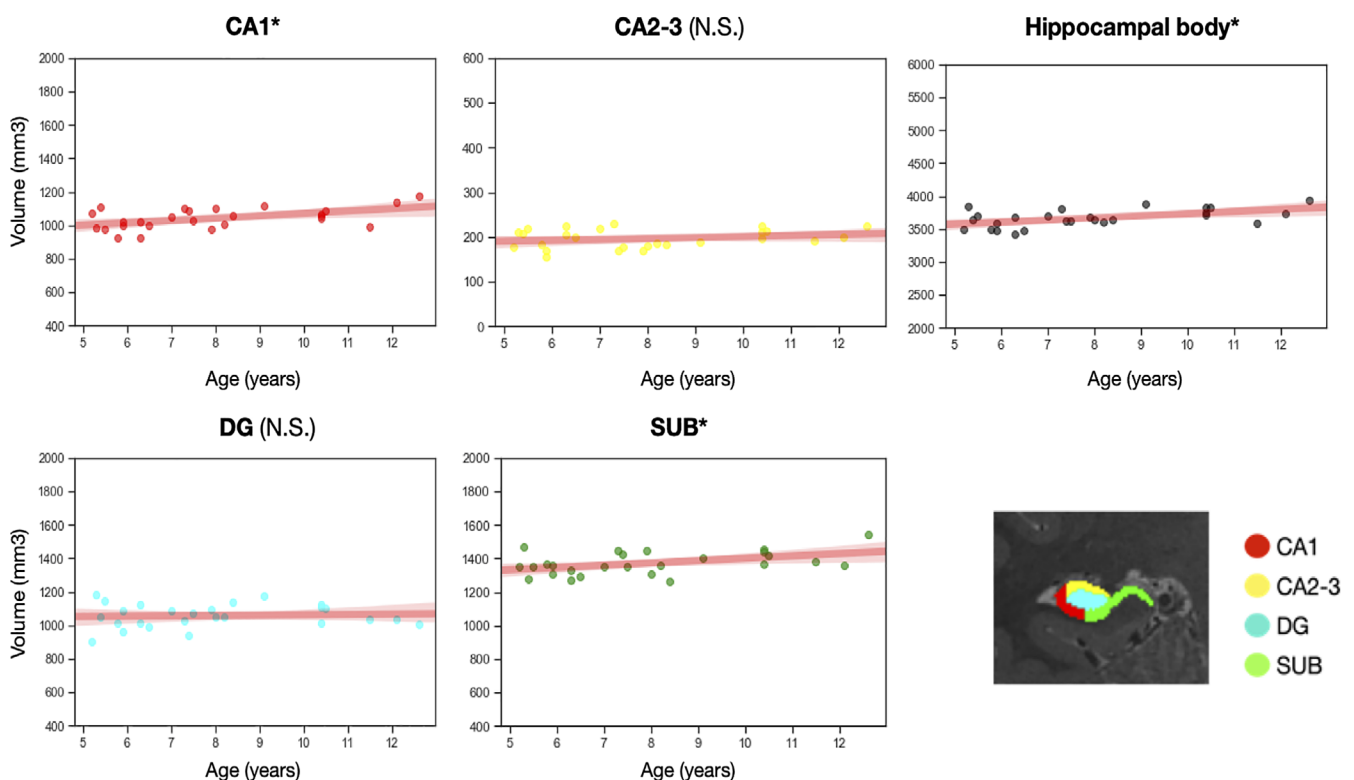


FIGURE 3 Plots of the regressions between adjusted bilateral hippocampal subfields' volumes and adjusted total hippocampal body volume, with age. CA, cornu ammonis; DG, dentate gyrus; SUB, subiculum. Subfields are shown on a coronal slice of the hippocampus on the bottom right. *, corrected $p < .05$, N.S, not significant

explain hippocampal body volume variance above and beyond the linear model. Sex was not a significant predictor of CA2-3 volume and adding sex as a covariate did not change the nonsignificance of age.

3.3 | Association between hippocampal subfields' volumes and memory discrimination

We used a two-stepped hierarchical linear regression approach to assess the association between subfield's volumes and memory discrimination. In a first step, we used four regression models (one per

subfield) including control variables and the tested subfield's volumes to assess if they predicted the variance of memory discrimination performance above and beyond a model including only control variables (age, sex, and Raven's matrix scores). In a second step, we added an interaction term between age and the tested subfield's volume in all models to assess if this model predicted the variance of memory discrimination performance above and beyond the model of the first step. p -values were adjusted for multiple comparisons by correcting the p -values of all models with a FDR procedure. Moreover, p -values of the predictive variables inside each model were corrected with FDR separately for each model. To verify that our results were not

impacted by collinearity, we computed variance inflation factors (VIFs) for all models. VIFs ranged from 1.03 to 1.49, which was lower than the traditionally retained thresholds of 5 or 10 (James et al., 2017; Vittinghoff et al., 2012).

TABLE 2 Summary of the full model predicting memory discrimination with control variables and CA2-3 volume

Variable	β	t-value	Corrected <i>p</i> -value
Intercept	-1.56	-2.72	.048
Sex	-0.03	-0.29	.91
Age	0.059	2.50	.048
Raven's matrix score	-0.003	-0.11	.91
CA2-3	0.006	2.34	.048

Note: Dependent variable: Memory discrimination. Bold shows significance at $p < 0.05$ (corrected). Full model: $F = 4.29$, $R^2 = 0.45$, corrected p -value = .04.

Abbreviation: CA, cornu ammonis.

TABLE 3 Summary of the full model predicting memory discrimination with control variables, subiculum volume, and subiculum volume*age interaction

Variable	β	t-value	Corrected <i>p</i> -value
Intercept	11.35	2.80	.022
Sex	-0.24	-2.18	.049
Age	-0.6633	-2.66	.022
Raven's matrix score	0.026	0.861	.4
SUB	-0.008	-2.91	.01
SUB*age	0.001	2.83	.022

Note: Dependent variable: Memory discrimination. Bold shows significance at $p < 0.05$ (corrected). Full model: $F = 4.21$, $R^2 = 0.51$, corrected p -value = .04.

Abbreviation: SUB, subiculum.

The model with CA1 volume did not explain the variance of memory performance above and beyond the model comprising only control variables (model comparison: $F = 1.87$, $p = .18$). The model including an interaction term did not explain the variance of memory discrimination above and beyond the variance explained by the model without the interaction term (model comparison: $F = 0.06$, $p = .80$).

The model with CA2-3 volume explained the variance of memory performance above and beyond the model comprising only control variables (model comparison: $F = 5.5$, $p = .029$) and explained 45% of the variance of memory discrimination ($F = 4.29$, $R^2 = 0.45$, $p = .04$; Table 2). CA2-3 was a significant predictor of memory discrimination performance ($t = 2.34$, $p = .04$). Adding an interaction term did not explain the variance of memory discrimination above and beyond the variance explained by the model without the interaction term (model comparison: $F = 0.02$, $p = .88$).

The model with DG volume did not explain the variance of memory performance above and beyond the model comprising only control variables (model comparison: $F = 0.41$, $p = .52$). The model including an interaction term did not explain the variance of memory discrimination above and beyond the variance explained by the model without the interaction term (model comparison: $F = 0.11$, $p = .73$).

The model with subiculum volume did not explain the variance of memory performance above and beyond the model comprising only control variables (model comparison: $F = 0.36$, $p = .55$). However, the model including an interaction term explained the variance of memory discrimination above and beyond the variance explained by the model without the interaction term (model comparison: $F = 8.01$, $p = .01$). The model with the interaction term explained 51% of the variance of memory discrimination ($F = 4.21$, $R^2 = 0.51$, $p = .04$) (Table 3). In this model, subiculum volume, and the interaction term between subiculum volume and age were significant predictors of memory

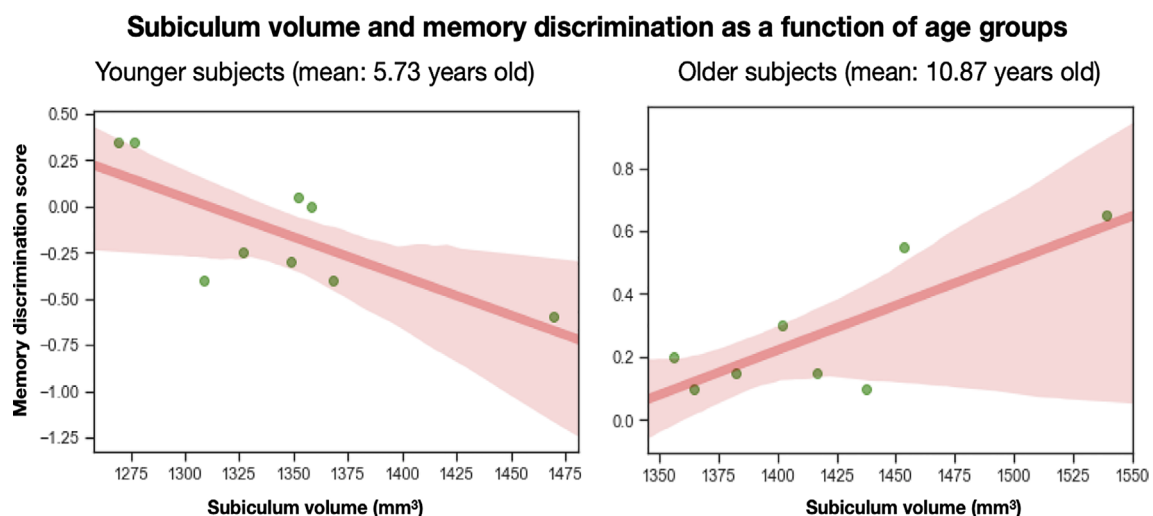


FIGURE 4 Association between memory discrimination and subiculum volume as a function of age. Left: for younger subjects (younger than the median age minus half of its standard deviation), the correlation between memory discrimination and subiculum volume is negative. Right: for older subjects (older than the median age plus half of its standard deviation), the correlation between memory discrimination and subiculum volume is positive

discrimination performance (subiculum volume: $t = -2.91, p = .01$; interaction term: $t = 2.83, p = .022$).

The significance of the interaction term between subiculum and age for predicting memory discrimination suggests an age-moderated association between subiculum and memory discrimination. To illustrate the moderating effect of age, we plotted memory discrimination scores in relation to subiculum volumes separately for younger subjects (younger than 0.5 standard deviation below the median age: i.e., from the youngest subject to 6.33 years old) and for older subjects (older than 0.5 standard deviation above the median age: i.e., from 8.57 years old to the oldest subject; Figure 4). These separate subiculum-memory discrimination plots as a function of age showed that for younger subjects, the relation between subiculum volume and memory discrimination is negative, while for older subjects, the relation between subiculum volume and memory discrimination is positive.

Raven's matrix standard scores were not a significant predictor of memory discrimination in all models described above. Performing the analyses without Raven's matrix standard scores did not significantly change the results. Moreover, hippocampal subfields volumes were not significant predictors of Raven's matrix standard scores (Table S1).

To verify for a possible confounding effect of the variance of total hippocampal body size, we also performed the same regression analyses with hippocampal body volume as a covariate, which did not significantly change the results. Combining the volumes of CA2-3 and DG in a single ROI, combined DG/CA2-3 was not a significant predictor of memory performance in a model without the interaction term (model comparison: $F = 0.001, p = .96$) and with the interaction term (model comparison: $F = 0.22, p = .63$).

Last, to examine the specificity of the association between hippocampal subfields and memory discrimination, we also ran models predicting item memory with control variables and subfields' volumes. None of the model predicting item memory with hippocampal subfields' volumes were significant.

4 | DISCUSSION

We aimed to assess the association between hippocampal subfields' volumes and memory discrimination, a behavioral proxy for PS, during development. Our main results were: (1) we observed age-related differences of memory discrimination performance in our age span; (2) we highlighted distinct age-related differences of hippocampal subfields' volumes, suggesting distinct developmental trajectories for each subfield; and (3) we showed an association between hippocampal subfields' volumes and memory discrimination performance. CA2-3 and subiculum were significant predictors of memory discrimination. Furthermore, the association between memory discrimination and subiculum was moderated by age. Subfields' volumes were not predictors of item memory, showing that the reported association between memory and subfields' volumes is specific to memory discrimination.

4.1 | Memory discrimination, but not item memory, is subject to age-related differences

Memory discrimination was positively correlated with age, suggesting memory discrimination could continue to improve continuously until late childhood. This result is in agreement with the results reported by Rollins and Cloude (2018). The authors found that 8–9 years old children memory discrimination performance was worse than adults, while 11–12 years old children performed similarly to adults. Combined with ours, these results suggest that memory discrimination performance could continue to improve during childhood until approximately 10 years of age. However, as our study did not include adult subjects, we cannot test this hypothesis directly. This contradicts the finding that memory discrimination performance of 6 years old children at a MST task was similar to that of adults' by Ngo et al. (2018). Task difficulty, relative to children's familiarity to the presented stimuli, has been shown to influence memory discrimination performance (Benear et al., 2020). Thus, these differences might be explained by the possibility that some items included in our study or the study of Rollins and Cloude were too unfamiliar to elicit a plateau of children's performance, compared to the items used by Ngo and colleagues. Still, the reasons explaining these disagreements remain elusive. Further studies will have to examine the precise developmental trajectory of memory discrimination from early childhood to adulthood while considering variables that might impact performance (e.g., item familiarity). Knowing the precise developmental timeline of memory discrimination is essential to understand how memory discrimination relates to EM development as a whole.

By contrast to memory discrimination, item memory performance was not associated with age, as previously reported (e.g., Ngo et al., 2018). We also did not find an association between memory discrimination performance and item memory performance. Hence, we further highlight that memory discrimination and item memory are two independent memory processes, as they likely rely on distinct neural correlates. While memory discrimination is mainly associated with the hippocampus, item memory has been shown to rely on medial temporal lobe regions, such as the perirhinal cortex (e.g., Davachi, 2006).

4.2 | Age-related differences of hippocampal subfields volumes suggest distinct developmental trajectories

We observed age-related differences of hippocampal subfields volumes from early to late childhood. Specifically, we found that CA1 and subiculum volumes were linearly and positively associated with age, suggesting continuous volumetric increases of these two subfields during childhood. CA2/3 and DG volumes were not associated with age. These findings echo the results of studies that previously examined the developmental trajectories of hippocampal subfields in the hippocampal body (Lee et al., 2014; Riggins et al., 2018). Lee et al. (2014) reported positive associations of CA1 of and

CA3/DG volumes with age, but not of the subiculum, in an 8–14 years old cohort. Riggins et al. (2018) reported age-related differences of subfields volumes in the hippocampal head, but not in the hippocampal body, in a 4–9 years old cohort. Our results are thus partly in agreement with theirs, which are also only in partial mutual agreement. Several factors, such as differences in segmentation protocols, sample size, or studied age span, could contribute to study-specific findings. Besides the hippocampal body, several studies have described the developmental trajectories of hippocampal subfields in the whole hippocampus, with differing results. Canada et al. (2019) found a positive association between age and subiculum volume in a 4–9 years old cohort, as well as a nonlinear association for CA1. Krogsrud et al. (2014) found positive associations between age and the volumes of the subiculum, CA1, DG, and CA2-3, in a 4–22 years old cohort. In a longitudinal study including subjects from 8 to 28 years old, Tamnes et al. (2018) found linear increases of CA1 and the subiculum, and linear CA2-3 and DG decreases. Overall, most of the aforementioned studies reported linear positive associations between age and CA1 and subiculum volumes, as we did in the present work.

We thus suggest that CA1 and the subiculum volumes likely undergo age-related volumetric increases during childhood in the hippocampal body, as it might be the case in the hippocampus as a whole. Myelination processes in the subiculum, occurring until adulthood (Krogsrud et al., 2014; van Praag et al., 2005), could explain the positive association between age and subiculum volume reported here. Age-related differences of CA1 volume could be related to increased connectivity and synaptogenesis between the pyramidal cells of the CA1 and the other subfields, or the entorhinal cortex. The size growth of CA1 and the subiculum in the hippocampal body suggested here could be the contributors of the larger hippocampal body size in adults, compared to children (DeMaster et al., 2014).

4.3 | Association between hippocampal subfields' volumes and memory discrimination

PS plays a crucial role in forming episodic memories by ensuring that similar representations are kept distinct from each other, reducing memory interference (see Keresztes et al., 2018 for a discussion). We hypothesized that the DG and/or CA3 volume would be associated with memory discrimination performance, a behavioral proxy for PS (Canada et al., 2019; Yassa & Stark, 2011). Partly in accordance with this hypothesis, we found that CA2-3 but not DG volume was associated with memory discrimination performance. We hence further confirm the association between CA2-3 and memory discrimination (e.g., Yassa & Stark, 2011). The positive correlation between memory discrimination suggests that, at a given age, larger CA2-3 in the hippocampal body is associated to better memory discrimination performance. Specifically, as we controlled for the volumes of other subfields, a relative larger CA2-3 size than the size of other hippocampal subfields could contribute to better memory discrimination. Similar relations between CA2-3 size in the body and memory performance

(using other tasks than memory discrimination) were found by former studies (Daugherty et al., 2016; Lee et al., 2014; Riggins et al., 2018; Tamnes et al., 2014). A larger CA3 could be related to several factors, such as increasing connectivity and synaptogenesis between CA3 pyramidal cells and DG granule cells.

As the DG is frequently highlighted as the main neural correlate of PS (e.g., Berron et al., 2017; Yassa & Stark, 2011) we would have expected to observe an association between memory discrimination and the volume of this subfield. This absence of relation is thus somewhat surprising. As we limited our study to the hippocampal body, it is possible that this relation is not observed, or less important, in the hippocampal body compared to the whole hippocampus. Indeed, a relationship between memory discrimination performance and the combined volumes of DG and CA2-4 (total hippocampus) during development was found by a previous study (Canada et al., 2019). However, the study from Canada and colleagues combined DG and CA2-4 in a single region of interest. This approach prevented assessing if the association between subfields volumes and memory discrimination was shared by both subfields, or only driven by one. Future studies will have to more precisely assess associations during development between PS, CA3, and DG, in the whole hippocampus and separately for the hippocampal head, body, and tail.

We found an association between the volume of the subiculum and memory discrimination. The subiculum is not classically associated with PS, but instead with pattern completion (e.g., Bakker, Kirwan, et al., 2008). However, some previous findings suggested that the subiculum takes part, in some cases, in PS. A study conducted in rats found that impairments of the dorsal subiculum were associated with impaired PS performance (Potvin et al., 2009). An ultra-high-field MRI study in adults found subiculum's role in scene discrimination (Hodgetts et al., 2017). A study from Lee et al. (2014), while not directly using a measurement of memory discrimination, showed an association in children and adolescents between subiculum volume and item false alarm rates, which could rely on discrimination processes partly dependent on PS. Therefore, the subiculum can also be involved in, or related to, PS, which we also suggest here. The subiculum is not part of the trisynaptic circuit, a loop connecting CA1 to the DG and to CA2-3 classically associated with PS. However, the subiculum is a major output of the trisynaptic circuit through connections via the fornix. A possible explanation for the relation between subiculum volume and memory discrimination performance could be that the subiculum volume partly expresses the efficacy at which information is transmitted between the bilateral hippocampi or to other cortical or subcortical regions through the fornix. It is possible that, to some extent, the subiculum behaves similarly to CA3, in the sense that it can take part in both PS and completion processes.

The relationship between subiculum volume and memory discrimination was moderated by age. Visualization of the relationship between subiculum volume and memory discrimination (Figure 4) as a function of age showed that for younger subjects, the association between subiculum and memory discrimination was negative, and positive for older subjects. A similar age-moderated relation between subiculum size and memory performance was found by Riggins

et al. (2018) in the hippocampal head, albeit in the opposite direction (positive correlation for younger children, negative correlation for older children). The head and body of the hippocampus are subjects to distinct maturational trajectories during childhood (e.g., DeMaster et al., 2014; Riggins et al., 2018). Distinct types of age-moderated associations, as a function of hippocampal subregions (head or body) and memory function, are thus likely to be observed. Here, the age-moderated relationship between memory discrimination and subiculum volume could suggest that the subiculum is differently related to memory discrimination as children approach puberty. For example, increased myelination processes in the subiculum in later childhood and adolescence could explain the positive association found in older children. This increased myelination could conduct information from the DG and CA3 subfields to cortical output regions through the subiculum and the fornix, contributing to PS performance. Even if the nature of this age-moderated association should be interpreted with caution, this nevertheless shows the overall relation between subiculum volume and memory discrimination performance during development.

Overall, we further confirm an association between CA2-3 and memory discrimination (Yassa & Stark, 2011), and suggest an association between memory discrimination and the subiculum. We provided evidence regarding the specificity of these findings as hippocampal subfields volumes were not correlated to item memory or to Raven's matrix standard scores.

4.4 | Limitations and future directions

Our study is subject to several limitations. First, although our initial sample comprised 50 subjects, which can be deemed reasonable in a developmental study, our final sample was limited to 26 data points. This important loss of data was mainly caused by the sensitivity to motion of the high-resolution T2w sequence used for the segmentation of the hippocampal subfields. A relatively low sample size thus limits the reach of our conclusions, despite satisfactory statistical effect size and correction for multiple comparisons. Low sample size was the consequence of our choice to keep only totally satisfying data to perform segmentation in order to maintain high accuracy. Second, another caveat of our study is that our design was cross-sectional rather than longitudinal. Longitudinal studies are better endowed to capture developmental trajectories by examining intra-subject rather than inter-subject variability. Third, the approach used here is indirect and correlational, only suggesting an association between subfields and memory discrimination by using subfields volumes as a proxy for hippocampal function. This type of approach could be completed by investigations of the role of hippocampal subfields in memory discrimination through more direct means, for example, using functional activation studies (e.g., Benear et al., 2020). Finally, we restricted our segmentation to the hippocampus' body to ensure reliable segmentations, but this limits our conclusion to this subregion rather than to the hippocampus as a whole. As subfields' developmental trajectories vary along the anteroposterior axis (e.g., Riggins et al., 2018), it is

relevant to assess subfields separately for the hippocampal head and body. Segmenting the head of the hippocampus is more complex than segmenting the body, but several protocols allow to do so (e.g., Joie et al., 2020). Still, we provide the first examination of the relation between memory discrimination and the main hippocampal subfields (separating DG and CA2-3) in the context of development. Future directions could include extending this investigation to the hippocampus' head, to verify if the relationship between memory discrimination and the subiculum is also found in the hippocampal head. Our results also invite, more generally, to scrutinize more closely the putative role of the subiculum in PS.

5 | CONCLUSION

We showed that memory discrimination performance is associated to age from early to late childhood. We highlighted distinct age-related association between age and hippocampal subfields in the hippocampal body, and showed that volumes of CA2-3 and subiculum were associated with memory discrimination performance. Our results confirm the role of CA2-3 in PS during childhood, and suggest an involvement of the subiculum (at least in the hippocampus' body) in memory discrimination. These results stress the need to further investigate the different contributions of hippocampal subfields to memory discrimination, and thus to PS, during development.

ACKNOWLEDGMENTS

This work was supported by Fondation de France (grant n°00070721, P.I. Marion Noulhiane), Fondation Mustela (Bourses de Recherche 2017 to Antoine Bouyeure), and Paris University (Bourse ministérielle de doctorat to Antoine Bouyeure). The authors would like to thank the nurses and radio manipulators for their help in this study. Special thanks to Chantal Ginisty.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Antoine Bouyeure  <https://orcid.org/0000-0002-0689-6878>

Franck Mauconduit  <https://orcid.org/0000-0002-0128-061X>

Marion Noulhiane  <https://orcid.org/0000-0003-2832-0332>

REFERENCES

- Ashburner, J., & Friston, K. J. (2005). Unified segmentation. *NeuroImage*, 26(3), 839–851. <https://doi.org/10.1016/j.neuroimage.2005.02.018>
- Avants, B. B., Tustison, N., & Song, G. (2009). Advanced normalization tools (ANTS). *Insight Journal*, 2(365), 1–35. <https://scicomp.ethz.ch/public/manual/ants/2.x/ants2.pdf>
- Bakker, A., Kirwan, C. B., Miller, M., & Stark, C. E. (2008). Pattern separation in the human hippocampal CA3 and dentate gyrus. *Science*, 319(5870), 1640–1642.
- Bartko, J. J. (1991). Measurement and reliability: Statistical thinking considerations. *Schizophrenia Bulletin*, 17(3), 483–489. <https://doi.org/10.1093/schbul/17.3.483>

- Bender, A. R., Keresztes, A., Bodammer, N. C., Shing, Y. L., Werkle-Bergner, M., Daugherty, A. M., Yu, Q., Kühn, S., Lindenberger, U., & Raz, N. (2018). Optimization and validation of automated hippocampal subfield segmentation across the lifespan. *Human Brain Mapping, 39*(2), 916–931.
- Beneat, S. L., Horwath, E. A., Cowan, E., Camacho, M. C., Ngo, C., Newcombe, N., Olson, I. R., Perlman, S. B., & Murty, V. P. (2020). Children show adult-like hippocampal pattern similarity for familiar but not novel events. *PsyArXiv*. <https://doi.org/10.31234/osf.io/3d8tv>
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological), 57*(1), 289–300.
- Bernasconi, N. L., Onai, N., & Lanzavecchia, A. (2003). A role for Toll-like receptors in acquired immunity: Up-regulation of TLR9 by BCR triggering in naive B cells and constitutive expression in memory B cells. *Blood, 101*(11), 4500–4504. <https://doi.org/10.1182/blood-2002-11-3569>
- Berron, D., Schütze, H., Maass, A., Cardenas-Blanco, A., Kuijff, H. J., Kumaran, D., & Düzel, E. (2016). Strong evidence for pattern separation in human dentate gyrus. *Journal of Neuroscience, 36*(29), 7569–7579.
- Berron, D., Vieweg, P., Hochkeppeler, A., Pluta, J. B., Ding, S.-L., Maass, A., Luther, A., Xie, L., Das, S. R., Wolk, D. A., Wolbers, T., Yushkevich, P. A., Düzel, E., & Wisse, L. E. M. (2017). A protocol for manual segmentation of medial temporal lobe subregions in 7Tesla MRI. *NeuroImage: Clinical, 15*, 466–482. <https://doi.org/10.1016/j.nicl.2017.05.022>
- Canada, K. L., Ngo, C. T., Newcombe, N. S., Geng, F., & Riggins, T. (2019). It's all in the details: Relations between young Children's developing pattern separation abilities and hippocampal subfield volumes. *Cerebral Cortex (New York, N.Y.: 1991), 29*(8), 3427–3433. <https://doi.org/10.1093/cercor/bhy211>
- Dalton, M.A., Zeidman, P., Barry, D.N., Williams, E., Maguire, E.A., 2017. Segmenting subregions of the human hippocampus on structural magnetic resonance image scans: An illustrated tutorial. <https://doi.org/10.1177/2398212817701448>
- Daugherty, A. M., Bender, A. R., Raz, N., & Ofen, N. (2016). Age differences in hippocampal subfield volumes from childhood to late adulthood. *Hippocampus, 26*(2), 220–228.
- Davachi, L. (2006). Item, context and relational episodic encoding in humans. *Current Opinion in Neurobiology, 16*(6), 693–700.
- DeMaster, D., Pathman, T., Lee, J. K., & Ghetti, S. (2014). Structural development of the hippocampus and episodic memory: Developmental differences along the anterior/posterior axis. *Cerebral Cortex (New York, N.Y.: 1991), 24*(11), 3036–3045. <https://doi.org/10.1093/cercor/bht160>
- Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., Albert, M. S., & Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage, 31*(3), 968–980. <https://doi.org/10.1016/j.neuroimage.2006.01.021>
- Dice, L. R. (1949). The selection index and its test of significance. *Evolution, 3*, 262–265.
- Doxey, C. R., & Kirwan, C. B. (2015). Structural and functional correlates of behavioral pattern separation in the hippocampus and medial temporal lobe. *Hippocampus, 25*(4), 524–533. <https://doi.org/10.1002/hipo.22389>
- Hanert, A., Pedersen, A., & Bartsch, T. (2019). Transient hippocampal CA1 lesions in humans impair pattern separation performance. *Hippocampus, 29*(8), 736–747. <https://doi.org/10.1002/hipo.23073>
- Hassevoort, K. M., Khan, N. A., Hillman, C. H., & Cohen, N. J. (2020). Differential development of relational memory and pattern separation. *Hippocampus, 30*(3), 210–219. <https://doi.org/10.1002/hipo.23146>
- Hodgetts, C. J., Postans, M., Warne, N., Varnava, A., Lawrence, A. D., & Graham, K. S. (2017). Distinct contributions of the fornix and inferior longitudinal fasciculus to episodic and semantic autobiographical memory. *Cortex, 94*, 1–14. <https://doi.org/10.1016/j.cortex.2017.05.010>
- James, L. M., Christova, P., Engdahl, B. E., Lewis, S. M., Carpenter, A. F., & Georgopoulos, A. P. (2017). Human Leukocyte Antigen (HLA) and Gulf War Illness (GWI): HLA-DRB1*13:02 spares subcortical atrophy in gulf war veterans. *eBioMedicine, 26*, 126–131. <https://doi.org/10.1016/j.ebiom.2017.11.005>
- Joie, R. L., Olsen, R., Berron, D., Amunts, K., Augustinack, J., Bakker, A., Bender, A., Boccardi, M., Bocchetta, M., Chakravarty, M. M., Chetelat, G., de Flores, R., DeKraaker, J., Ding, S.-L., Insausti, R., Kedo, O., Mueller, S. G., Ofen, N., Palombo, D., ... Daugherty, A. M. (2020). The development of a valid, reliable, harmonized segmentation protocol for hippocampal subfields and medial temporal lobe cortices: A progress update. *Alzheimer's & Dementia, 16*(S4), e046652. <https://doi.org/10.1002/alz.046652>
- Keresztes, A., Bender, A. R., Bodammer, N. C., Lindenberger, U., Shing, Y. L., & Werkle-Bergner, M. (2017). Hippocampal maturity promotes memory distinctiveness in childhood and adolescence. *Proceedings of the National Academy of Sciences of the United States of America, 114*(34), 9212–9217. <https://doi.org/10.1073/pnas.1710654114>
- Keresztes, A., Ngo, C. T., Lindenberger, U., Werkle-Bergner, M., & Newcombe, N. S. (2018). Hippocampal maturation drives memory from generalization to specificity. *Trends in Cognitive Sciences, 22*(8), 676–686. <https://doi.org/10.1016/j.tics.2018.05.004>
- Krogsrud, S. K., Tamnes, C. K., Fjell, A. M., Amlien, I., Grydeland, H., Sulutvedt, U., Due-Tønnessen, P., Bjørnerud, A., Søsnes, A. E., Håberg, A. K., Skrane, J., & Walhovd, K. B. (2014). Development of hippocampal subfield volumes from 4 to 22 years. *Human Brain Mapping, 35*(11), 5646–5657. <https://doi.org/10.1002/hbm.22576>
- Lavenex, P., & Banta Lavenex, P. (2013). Building hippocampal circuits to learn and remember: Insights into the development of human memory. *Behavioural Brain Research, 254*, 8–21. <https://doi.org/10.1016/j.bbr.2013.02.007>
- Lee, J. K., Ekstrom, A. D., & Ghetti, S. (2014). Volume of hippocampal subfields and episodic memory in childhood and adolescence. *NeuroImage, 94*, 162–171.
- Lenroot, R. K., & Giedd, J. N. (2006). Brain development in children and adolescents: Insights from anatomical magnetic resonance imaging. *Neuroscience & Biobehavioral Reviews, 30*(6), 718–729. <https://doi.org/10.1016/j.neubiorev.2006.06.001>
- Leutgeb, J. K., Leutgeb, S., Moser, M.-B., & Moser, E. I. (2007). Pattern separation in the dentate gyrus and CA3 of the hippocampus. *Science, 315*(5814), 961–966.
- Mankin, E. A., Diehl, G. W., Sparks, F. T., Leutgeb, S., & Leutgeb, J. K. (2015). Hippocampal CA2 activity patterns change over time to a larger extent than between spatial contexts. *Neuron, 85*(1), 190–201. <https://doi.org/10.1016/j.neuron.2014.12.001>
- Marr, D., & Brindley, G. S. (1971). Simple memory: A theory for archicortex. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences, 262*(841), 23–81. <https://doi.org/10.1098/rstb.1971.0078>
- Mueller, S. G., Chao, L. L., Berman, B., & Weiner, M. W. (2011). Evidence for functional specialization of hippocampal subfields detected by MR subfield volumetry on high resolution images at 4T. *NeuroImage, 56*(3), 851–857. <https://doi.org/10.1016/j.neuroimage.2011.03.028>
- Myers, C. E., & Scharfman, H. E. (2009). A role for hilar cells in pattern separation in the dentate gyrus: A computational approach. *Hippocampus, 19*(4), 321–337.
- Nakashiba, T., Cushman, J. D., Pelkey, K. A., Renaudineau, S., Buhl, D. L., McHugh, T. J., Rodriguez Barrera, V., Chittajallu, R., Iwamoto, K. S., McBain, C. J., Fanselow, M. S., & Tonegawa, S. (2012). Young dentate granule cells mediate pattern separation, whereas old granule cells facilitate pattern completion. *Cell, 149*(1), 188–201. <https://doi.org/10.1016/j.cell.2012.01.046>

- Neunuebel, J. P., & Knierim, J. J. (2014). CA3 retrieves coherent representations from degraded input: Direct evidence for CA3 pattern completion and dentate gyrus pattern separation. *Neuron*, *81*(2), 416–427.
- Neylan, T. C., Mueller, S. G., Wang, Z., Metzler, T. J., Lenoci, M., Truran, D., Marmar, C. R., Weiner, M. W., & Schuff, N. (2010). Insomnia severity is associated with a decreased volume of the CA3/dentate Gyrus hippocampal subfield. *Biological Psychiatry*, *68*(5), 494–496. <https://doi.org/10.1016/j.biopsych.2010.04.035>
- Ngo, C. T., Horner, A. J., Newcombe, N. S., & Olson, I. R. (2019). Development of holistic episodic recollection. *Psychological Science*, *30*(12), 1696–1706. <https://doi.org/10.1177/0956797619879441>
- Ngo, C. T., Newcombe, N. S., & Olson, I. R. (2018). The ontogeny of relational memory and pattern separation. *Developmental Science*, *21*(2), e12556. <https://doi.org/10.1111/desc.12556>
- Norman, K. A., & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: A complementary-learning-systems approach. *Psychological Review*, *110*(4), 611–646. <https://doi.org/10.1037/0033-295X.110.4.611>
- Olson, I. R., & Newcombe, N. S. (2013). Binding together the elements of episodes: Relational memory and the developmental trajectory of the hippocampus. *The Wiley Handbook on the Development of Children's Memory*, *1*, 285–308.
- Østby, Y., Tamnes, C. K., Fjell, A. M., Westlye, L. T., Due-Tønnessen, P., & Walhovd, K. B. (2009). Heterogeneity in subcortical brain development: A structural magnetic resonance imaging study of brain maturation from 8 to 30 years. *Journal of Neuroscience*, *29*(38), 11772–11782. Retrieved from <https://www.jneurosci.org/content/29/38/11772>
- Palombo, D., Williams, L. J., Abdi, H., & Levine, B. (2013). The survey of autobiographical memory (SAM): A novel measure of trait mnemonics in everyday life. *Cortex*, *49*, 1526–1540. <https://doi.org/10.1016/j.cortex.2012.08.023>
- Potvin, O., Doré, F. Y., & Goulet, S. (2009). Lesions of the dorsal subiculum and the dorsal hippocampus impaired pattern separation in a task using distinct and overlapping visual stimuli. *Neurobiology of Learning and Memory*, *91*(3), 287–297. <https://doi.org/10.1016/j.nlm.2008.10.003>
- Ramsaran, A. I., Schlichting, M. L., & Frankland, P. W. (2019). The ontogeny of memory persistence and specificity. *Developmental Cognitive Neuroscience*, *36*, 100591. <https://doi.org/10.1016/j.dcn.2018.09.002>
- Raven, J., Raven, J. C., & Court, J. H. (2003). *Manual for Raven's progressive matrices and vocabulary scales*. Section 3. Harcourt Assessment.
- Raz, N., Lindenberger, U., Rodrigue, K. M., Kennedy, K. M., Head, D., Williamson, A., Dahle, C., Gerstorf, D., & Acker, J. D. (2005). Regional brain changes in aging healthy adults: General trends, individual differences and modifiers. *Cerebral Cortex*, *15*(11), 1676–1689. <https://doi.org/10.1093/cercor/bhi044>
- Riggins, T., Geng, F., Botdorf, M., Canada, K., Cox, L., & Hancock, G. R. (2018). Protracted hippocampal development is associated with age-related improvements in memory during early childhood. *NeuroImage*, *174*, 127–137.
- Rollins, L., & Cloude, E. B. (2018). Development of mnemonic discrimination during childhood. *Learning & Memory*, *25*(6), 294–297. <https://doi.org/10.1101/lm.047142.117>
- Schlichting, M. L., Mack, M. L., Guarino, K. F., & Preston, A. R. (2019). Performance of semi-automated hippocampal subfield segmentation methods across ages in a pediatric sample. *NeuroImage*, *191*, 49–67.
- Schmidt, B., Marrone, D. F., & Markus, E. J. (2012). Disambiguating the similar: The dentate gyrus and pattern separation. *Behavioural Brain Research*, *226*(1), 56–65.
- Stark, S. M., Kirwan, C. B., & Stark, C. E. L. (2019). Mnemonic similarity task: A tool for assessing hippocampal integrity. *Trends in Cognitive Sciences*, *23*(11), 938–951. <https://doi.org/10.1016/j.tics.2019.08.003>
- Stark, S. M., & Stark, C. E. L. (2017). Age-related deficits in the mnemonic similarity task for objects and scenes. *Behavioural Brain Research*, *333*, 109–117. <https://doi.org/10.1016/j.bbr.2017.06.049>
- Tamnes, C. K., Overbye, K., Ferschmann, L., Fjell, A. M., Walhovd, K. B., Blakemore, S.-J., & Dumontheil, I. (2018). Social perspective taking is associated with self-reported prosocial behavior and regional cortical thickness across adolescence. *Developmental Psychology*, *54*(9), 1745–1757. <https://doi.org/10.1037/dev0000541>
- Tamnes, C. K., Walhovd, K. B., Engvig, A., Grydeland, H., Krogsrud, S. K., Østby, Y., Holland, D., Dale, A. M., & Fjell, A. M. (2014). Regional hippocampal volumes and development predict learning and memory. *Developmental Neuroscience*, *36*(3–4), 161–174. <https://doi.org/10.1159/000362445>
- Utsunomiya, H., Takano, K., Okazaki, M., & Mitsudome, A. (1999). Development of the temporal lobe in infants and children: Analysis by MR-based volumetry. *American Journal of Neuroradiology*, *20*(4), 717–723.
- van Praag, H., Shubert, T., Zhao, C., & Gage, F. H. (2005). Exercise enhances learning and hippocampal neurogenesis in aged mice. *Journal of Neuroscience*, *25*(38), 8680–8685. <https://doi.org/10.1523/JNEUROSCI.1731-05.2005>
- Vittinghoff, E., Glidden, D. V., Shiboski, S. C., & McCulloch, C. E. (2012). *Regression methods in biostatistics: Linear, logistic, survival, and repeated measures models* (2nd ed.). Springer-Verlag. <https://doi.org/10.1007/978-1-4614-1353-0>
- Yassa, M. A., Lacy, J. W., Stark, S. M., Albert, M. S., Gallagher, M., & Stark, C. E. (2011). Pattern separation deficits associated with increased hippocampal CA3 and dentate gyrus activity in nondemented older adults. *Hippocampus*, *21*(9), 968–979.
- Yassa, M. A., Mattfeld, A. T., Stark, S. M., & Stark, C. E. L. (2011). Age-related memory deficits linked to circuit-specific disruptions in the hippocampus. *Proceedings of the National Academy of Sciences*, *108*(21), 8873–8878. <https://doi.org/10.1073/pnas.1101567108>
- Yassa, M. A., & Stark, C. E. (2011). Pattern separation in the hippocampus. *Trends in Neurosciences*, *34*(10), 515–525.
- Yushkevich, P. A., Piven, J., Hazlett, H. C., Smith, R. G., Ho, S., Gee, J. C., & Gerig, G. (2006). User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *NeuroImage*, *31*(3), 1116–1128. <https://doi.org/10.1016/j.neuroimage.2006.01.015>
- Yushkevich, P. A., Wang, H., Pluta, J., Das, S. R., Craige, C., Avants, B. B., Weiner, M. W., & Mueller, S. (2010). Nearly automatic segmentation of hippocampal subfields in in vivo focal T2-weighted MRI. *NeuroImage*, *53*(4), 1208–1224. <https://doi.org/10.1016/j.neuroimage.2010.06.040>
- Zöllei, L., Iglesias, J. E., Ou, Y., Grant, P. E., & Fischl, B. (2020). Infant FreeSurfer: An automated segmentation and surface extraction pipeline for T1-weighted neuroimaging data of infants 0–2 years. *NeuroImage*, *218*, 116946. <https://doi.org/10.1016/j.neuroimage.2020.116946>

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Bouyeure, A., Patil, S., Mauconduit, F., Poiret, C., Isai, D., & Noulhiane, M. (2021). Hippocampal subfield volumes and memory discrimination in the developing brain. *Hippocampus*, 1–13. <https://doi.org/10.1002/hipo.23385>

Curriculum Vitae

Education

Institute	Degree & Certification	Year
University of Lyon	Ethique de la Recherche	2021
University of Rouen	Master of Science, Msc Games Analytics and Big Data	2020
DeepLearning.AI	Mentor for the AI for Medicine course, Deep Learning Specialization (TF2.0)	2020
HSE University	Bayesian Methods for Machine Learning	2020
SuperDataScience	Deep Learning and Computer Vision, Tensorflow 2.0 Practical Advanced	2019
Google Flutter Team	Flutter Development with Dart	2019
Alberta Machine Intelligence Institute	Optimizing Machine Learning Model Performance	2019
IBM	Advanced Data Science Specialist	2019
John Hopkins University	Fundamental Neurosciences and “Neurohacking in R”	2019
Santa Fe Institute	Nonlinear Dynamics: Mathematical and Computational Approaches & Fractals and Scaling	2019
HarvardX	Using Python for Research	2018

Publications

- Poiret, et al. Charting Hippocampal Development with Robust Explainable AI and Resting-State fMRI. in prep, 2023.
- Poiret, et al. Can we Agree? On the Rashōmon Effect and the Reliability of Post-Hoc Explainable AI. in prep, 2023.
- Poiret, et al. A Fast and Robust Hippocampal Subfields Segmentation: HSF revealing Lifespan Volumetric Dynamics. Frontiers in Neuroinformatics, 2023.
- Poiret, et al. 3D-AGSCaps: Are CapsNet Any Good for Hippocampal Segmentation? Journal of Medical Imaging, submitted.

- Bouyeure, et al. Hippocampal subfield volumes and memory discrimination in the developing brain. Hippocampus, 2021.
- Poiret, et al. Breath-hold diving strategies to avoid loss of consciousness: speed is the key factor. Sports Biomechanics, 2020.

Contributions to Open Source Softwares

Personal repositories:

- HSF, an end-to-end Deep Learning tool to segment the Hippocampus <https://github.com/clementpoiret/HSF>,
- ROIloc, a preprocessing tool to crop MRIs around any region in the brain <https://github.com/clementpoiret/roiloc>,
- 3D-AGSCaps, the first 3D implementation of Capsule Networks, <https://github.com/clementpoiret/3D-AGSCaps>,
- Perceiver MNIST, a toy example of a multi-modal Deep Learning Architecture, https://github.com/clementpoiret/Perceiver_MNIST,
- Pingouin.js, a Julia implementation of statistical tests such as T-Tests, Correlations, Friedman, etc., <https://github.com/neuralmagic/sparseml>.

External repositories (via Pull Requests):

- Obsidian, a note-taking app <https://github.com/obsidianmd/obsidian-releases>,
- Pytorch Lightning, a wrapper around PyTorch to accelerate and facilitate Deep Learning research <https://github.com/Lightning-Universe/lightning-bolts>,
- SparseML, a library to prune and quantize neural networks, <https://github.com/neuralmagic/sparseml>.

Scientific Communications

- HSF as a Future-proof Solution to Hippocampal Subfields Segmentation, OHBM 2022,
- Viewpoint Equivariance: Are Capsule Networks the Future of Hippocampal Segmentation? OHBM 2021,
- Poster at CJC-SCO 2020.