Université Paris Cité

École doctorale 474 : Frontières de l'Innovation en Recherche et Éducation
Laboratoire Jean Perrin

Restricted Boltzmann Machines as generic, generative, and interpretable models of the larval zebrafish brain

Par Mattéo DOMMANGET-KOTT

Thèse de doctorat de Neurosciences et Troubles Neuronaux

Dirigée par Volker BORMUTH
Et par Georges DEBRÉGEAS

Présentée et soutenue publiquement le 16/10/2025

Devant un jury composé de :

Lilach AVITAN, SRLECT	Hebrew University of Jerusalem	Rapportrice
Ruben PORTUGUES, PR	Cornell University	Rapporteur
Brice BATHELLIER, DR	Université Paris Cité	Examinateur
N. Alex CAYCO GAJIC, PR	École Normale Supérieure	Examinatrice
Volker BORMUTH, MCF-HDR	Sorbonne Université	Directeur de thèse
Georges DEBRÉGEAS, DR-HDR	CNRS	Directeur de thèse

Résumé

Titre: Développement de Machines de Boltzmann restreintes comme modèles génériques, génératifs, et interprétable du cerveau de la larve de poisson zèbre.

Mots clefs: poisson-zèbre; imagerie calcique; activité neuronale spontanée; Modèles de Markov; Machines de Boltzmann Restreintes

Résumé: Une question centrale en neurosciences est de comprendre comment l'activité cérébrale se généralise d'un individu à l'autre. Les paradigmes classiques basés sur des tâches ou des stimuli s'appuient sur l'alignement anatomique et temporel des enregistrements cérébraux, ainsi que sur le moyennage des données. Ces stratégies se révèlent toutefois inefficaces pour étudier l'activité cérébrale spontanée, car aucun repère temporel commun ne relie les individus.

Dans cette thèse, nous proposons des modèles statistiques permettant de comparer l'activité neuronale spontanée chez la larve de poisson-zèbre. À l'aide de la microscopie fonctionnelle à feuille de lumière, nous enregistrons l'activité de l'ensemble du cerveau à résolution cellulaire et nous utilisons des Modèles de Markov ainsi que des Machines de Boltzmann Restreintes (Restricted Boltzmann Machines, RBM) pour générer des représentations interprétables du comportement et des dynamiques neuronales.

Nous commençons par modéliser la dynamique de réorientation de la larve de poisson-zèbre pendant la nage libre avec un Modèle de Markov caché (Hidden Markov Model, HMM) à trois états moteurs : avant, gauche et droite. Ce même HMM décrit l'activité de la région ARTR (Anterior Rhombencephalic Turning Region) et relie directement la dynamique neuronale au comportement. Nos résultats suggèrent qu'une activation bilatérale équilibrée de l'ARTR pourrait expliquer l'état "avant".

Nous développons ensuite deux procédures pour entraîner des RBMs capables de projeter l'activité neuronale de plusieurs poissons dans un espace latent commun. La première entraîne une RBM unique à partir de l'activité voxélisée et concaténée de plusieurs poissons. La seconde est un paradigme "enseignant-élève" où chaque RBM est entraînée sur un poisson unique, mais sont contraintes entre elles de manière à partager leurs unités cachées. Ces approches produisent une représentation commune qui permet de transférer fidèlement des motifs d'activité d'un individu à l'autre et révèle des régularités conservées entre cerveaux.

Nous exploitons enfin cet espace latent pour segmenter l'activité spontanée en états cérébraux discrets, et mesurer leurs transitions markoviennes. Nous constatons que les séquences d'états suivent des règles stéréotypées d'un animal à l'autre, ce qui confirme que l'activité spontanée encode des priors intrinsèques du calcul neural.

En conclusion, nos résultats montrent comment des modèles génératifs probabilistes peuvent surmonter la variabilité individuelle, et révéler des principes d'organisation communs entres cerveaux chez les vertébrés. La thèse fournit une boîte à outils pour comparer l'activité spontanée entre sujets et souligne la valeur interprétative des RBM et des HMM.

Abstract

Title: Restricted Boltzmann Machines as generic, generative, and interpretable models of the larval zebrafish brain.

Keywords: Zebrafish ; Calcium Imaging ; Hidden Markov Models ; Restricted Boltzmann Machines ; Spontaneous Neural Activity

Abstract: Understanding how brain activity generalises across individuals remains a central challenge in neuroscience. Classical task- or stimulus-driven paradigms align recordings by trial-averaging and registration, but these techniques break down for spontaneous activity, where no common temporal reference exists.

This thesis develops statistical frameworks that enable direct, cross-individual comparison of spontaneous activity in larval zebrafish. Leveraging whole-brain, single-cell calcium imaging acquired with light-sheet functional microscopy, we use Markov Models and Restricted Boltzmann Machines (RBMs) to build generative, interpretable representations of behavior and neural activity.

First, we characterize spontaneous swimming reorientation with a Hidden Markov Model (HMM) comprising forward, left, and right motor states. The same 3-state architecture also captures the activity of the Anterior Rhombencephalic Turning Region (ARTR), linking neural dynamics to behavior and suggesting that balanced bilateral ARTR activity mediates the forward state.

Next, we introduce two strategies for training RBMs that embed the neural activity of multiple fish into a common latent space. A voxel-based approach trains a single RBM on pooled recordings, while a teacher-student paradigm constrains individual-fish RBMs to shared hidden units. The shared representation allows realistic activity patterns to be transferred between animals, demonstrating that the model captures generic structures conserved across brains.

Finally, we exploit the common latent space to segment spontaneous activity into discrete brain states and quantify their Markovian transition statistics. We find that state-to-state dynamics are remarkably stereotyped across individuals, supporting the hypothesis that spontaneous activity reflects intrinsic priors of neural computation.

In conclusion, our results show how probabilistic generative models can bridge individual variability and reveal shared organisational principles of vertebrate brains. The thesis provides a toolbox for cross-subject comparison of spontaneous activity, and underscores the biological interpretability of probabilistic models.

Résumé détaillé en langue française

Chapitre 1 : Introduction Générale

Cette thèse interroge la nature, l'organisation et la comparabilité inter-individuelle de l'activité cérébrale spontanée (AS). L'hypothèse directrice est double :
(i) la AS reflète des contraintes d'organisation du cerveau, des "priors" internes résultant de la connectivité et de l'histoire sensorimoteur du système. Ces contraintes ayants des bases génétiques, développementales, et environnementales, elles devraient être partagées entre individus.
(ii) ces contraintes peuvent être mises au jour par des modèles génératifs qui synthétisent les motifs observés plutôt que de seulement les décrire.

Nous adoptons pour cela un cadre probabiliste combinant des modèles de Markov pour segmenter et comparer des dynamiques, et des Machines des Boltzmann Restreintes (Restricted Boltzmann Machines, RBM) pour apprendre des espaces latents interprétables et génératifs où les co-activations neuronales sont explicitement modélisées. Cette approche vise à concilier interprétabilité biologique et pouvoir génératif, deux conditions clés pour relier activité spontanée au comportement, mais aussi pour construire des modèles généralisables à plusieurs individus.

Conceptuellement, l'AS n'est pas du "bruit", elle est présente à tous les états de vigilance (durant anesthésie, le sommeil, pendant la réfection et pendant des états moteurs), elle est structurée spatio-temporellement, et impliquée dans le développement, la plasticité et la mémoire. Ses motifs sont proches de l'activité évoquée, et contraignent les réponses aux stimuli. Elle peut être vue comme l'exploration d'un répertoire interne de configurations cérébrales, depuis lequel des événements sensoriels sélectionnent ou réorganisent l'activité. Modéliser l'AS consiste donc à identifier une structure et un vocabulaire récurrent de configurations cérébrales et des règles de transition entre celles-ci. À ce titre, des outils de segmentation (modèles de Markov) et de représentation (RBM) s'avèrent naturels pour passer du neurone à l'organisation large-échelles, puis à des états cérébraux.

Sur le plan expérimental, la larve de poisson-zèbre offre un compromis unique : un cerveau suffisamment compact pour être imagé à résolution cellulaire sur de larges volumes (jusqu'au cerveau entier) grâce à la microscopie fonctionnelle à feuille de lumière, tout en exprimant des comportements variés et des réseaux fonctionnels identifiables. Depuis l'établissement d'enregistrements couvrant jusqu'à $\sim 80\%$ des neurones chez la larve, cette technique s'est imposée pour étudier l'AS à l'échelle du cerveau entier.

Sur le plan méthodologique, la thèse s'articule en trois volets :
(1) lier l'activité cérébrale et le comportement de réorientation de la larve de poisson zèbre, via des Modèles de Markov Cachés (Hidden Markov Model, HMM) tri-états (gauche/avant/droite) appris séparément sur les données comportementales et sur l'activité du circuit ARTR (Anterior Rhombencephalic Turning Region).
(2) la construction d'un espace latent partagé pour comparer l'AS entre poissons à l'échelle du cerveau entier, via des RBMs entraînées d'abord dans un espace voxélisé partagé, puis selon un paradigme enseignant–élève (bi-training).
(3) la discrétisation de cet espace en un "vocabulaires" d'états cérébraux, dont les occupations et transitions sont étudiées avec un modèle de Markov.

Ce plan répond à deux défis clés de l'étude de l'AS : (i) l'absence de repère temporel commun entre individus (pas de stimuli), (ii) l'impossibilité d'aligner naïvement des neurones un-à-un d'un cerveau à l'autre. Les contributions ci-dessous contournent ces obstacles en alignant d'abord l'organisation fonctionnelle du cerveau entre individus, puis en segmentant la dynamique dans cet espace déjà commun.

Chapitre 2 : Lier les états cérébraux et comportementaux dans la locomotion des larves de poisson zèbre à l'aide de modèles de Markov cachés

Ce chapitre propose une ré-analyse unifiée par HMM de deux types de données. Côté comportement, la navigation des larve pendant la nage libre à différentes températures (18, 22, 26, 30, 33 °C). Cette navigation est décrite comme une succession de bouts (bouffées motrices) séparés par des intervalles d'environ 1-2 secondes, chaque bout étant associé à un angle de réorientation $\delta\theta$. Côté neuronal, l'activité du circuit ARTR enregistrée in vivo chez des larves immobilisées, aux mêmes températures. Dans les deux cas, on entraîne séparément un HMM à 3 états reflétant : côté comportement des bouts avant/gauche/droite, et côté ARTR l'asymétrie gauche/droite de l'activité neuronale.

Nous montrons que l'inférence jointe des émissions (distribution des angles par état) et des transitions produit un étiquetage non biaisé qui capte des régularités temporelles qu'un simple seuillage ignore. En particulier, la persistance intra-état (répéter plusieurs fois de suite le même type de bout) est mieux estimée, rendant la dynamique plus markovienne et rapprochant les diffusivités simulées de celles observées. À l'inverse, un HMM entraîné après seuillage ne reproduit pas correctement la diffusivité orientée. Cette supériorité du HMM justifie l'usage d'un cadre probabiliste dès la segmentation, autant pour le comportement que pour la neurodynamique de l'ARTR. Le cadre HMM rend aussi possible l'identification des individus à partir des matrices de transition, qui servent alors d'empreintes discriminantes (phénotypes de navigation).

Du côté de l'ARTR, nous trouvons que les trois états identifiés à partir du HMM correspondent à deux états latéralisés gauche/droite et à un état équilibré. La similitude entre les taux de transition observés à partir de l'ARTR et à partir du comportement suggèrent que l'ARTR pourrait contrôler tous les types de réorientation spontanées, y compris les bouts

avants.

Nous exploitons ensuite le caractère génératif des HMM pour convertir l'activité neuronal de l'ARTR en trajectoires de nages synthétiques via:
(i) décodage Viterbi des états de l'ARTR,
(ii) recalage temporel entre dynamique calcique et cadence des bouts,
(iii) échantillonnage des bouts à partir des distributions d'émission comportementales.
Les trajectoires générées reproduisent les statistiques de diffusivité des données réelles, confirmant que le circuit ARTR, modélisé par un HMM à trois états, pourrait contrôler directement la dynamique de réorientation.

En synthèse, ce chapitre montre qu'un HMM tri-états fournit un langage commun pour décrire, comparer et relier les dynamiques de réorientation et de l'ARTR, capte des effets de persistance temporelle, et révèle des signatures individuelles exploitables pour la comparaison inter-poissons.

Chapitre 3 : Construire une représentation partagée et interprétable de l'activité cérébrale spontanée de plusieurs larves de poisson zèbre

Dans ce chapitre, nous construisons un espace latent de l'AS à l'échelle du cerveau entier, partagé par plusieurs poissons. Nous présentons deux approches complémentaires.

Dans la première approche, les enregistrements d'activité spontanée de plusieurs poissons sont morphologiquement alignés dans un espace commun spatialement voxélisé, puis concaténés. Un RBM unique est entraînée sur ces données agrégées. Cet RBM reproduit les statistiques de l'activité cérébrale et révèle que, même si l'activité voxélisée brute varie entre individus, la représentation latente est au contraire stéréotypée. Nous montrons par exemple que l'activité moyenne de la couche cachée est comparable entre individus, et que la probabilité des états cérébraux d'après le modèle est également comparable. Ces résultats indiquent que l'espace latent décrit un domaine commun d'états accessibles.

La second approche repose sur un paradigme "enseignant–élève", où un RBM "enseignant" entraîné sur un seul poisson définit l'espace latent sur lequel les RBMs "élève" seront contraints. Concrètement, nous interpolons spatialement les poids de l'enseignant sur les positions neuronales des élèves, copions les potentiels de la couche cachée, et initialisons les champs visibles à partir de l'activité neuronale de chaque élève. Nous présentons ensuite un nouveau protocole d'entraînement des RBMs qui impose que les couches cachées des élèves maintiennent la même représentation latente que l'enseignant, tout en apprenant leurs propre statistique d'activité neuronale.

Malgré l'absence de contrainte spatiale explicite pendant l'apprentissage des élèves, les unités cachées (assemblées) retiennent des distributions spatiales proches de celles de enseignant, et leurs corrélations sont partiellement conservées. Au-delà de cet alignement,

l'espace partagé permet une traduction d'activité : une configuration neuronale d'un poisson A est projetée dans l'espace latent puis reconstruite dans un poisson B. Les configurations traduites ont une énergie libre comparable à celle des reconstructions intra-poisson, et elles préservent des corrélations spatiales élevées avec l'activité d'origine. Ceci démontre que l'espace latent encode un répertoire partagé de motifs probables et spatialisés, apte à transférer des configurations entre individus.

Comparée à d'autres méthode basées sur l'alignement anatomique ou l'alignement fonctionnelle, notre approche est générative (on peut échantillonner/évaluer), compositive (l'acrivité est décrite par une composition d'assemblées neuronales), et capable de passer naturellement de la représentation latente aux neurones sans stimuli ni labels ni correspondance neurone-par-neurone. Elle définit un a priori probabiliste partagé par plusieurs individus à une méso-échelle. Parmi les limites de cette méthode on trouve : le choix de l'enseignant (risque de biais), la prépondérance des motifs très structurés, et une transférabilité possiblement réduite dans des systèmes moins stéréotypés (p. ex. cortex de mammifère, ou mutants). Des extensions multi-enseignant ou des entraînements joints pourraient atténuer ces biais et accroître la sensibilité.

Chapitre 4 : Stéréotypie de la dynamique du cerveau entier pendant l'activité cérébrale spontanée chez les larves de poisson zèbre

Dans le 4ème chapitre, nous discrétisons l'espace latent commun construit dans le chapitre précédent, dans le but de comparer la dynamique de l'AS entre poissons.

Pour ce faire, nous superposons au RBMs cérébraux un second RBM qui prend pour entrées les configurations d'assemblées décrient précédemment, et les projettes vers une couche binaire de B unités. Chaque configuration b est interprétée comme un entier définissant un état neuronal s. Ces états décrivent des séquences temporelles de motifs latents. Crucialement, ce second RBM préserve l'espace latent appris précédemment, assurant un aller-retour entre l'espace des assemblées et l'espace des états.

L'analyse markovienne de ces séquences d'états permet de comparer directement les occupations stationnaires et les matrices de transition entre individus, et nous montrons ainsi que la dynamique de l'AS est surprenamment stéréotypée, particulièrement à l'intérieur de deux groupes d'animaux suggérant des phénotypes ou conditions expérimentales distinctes.

Nous montrons que le nombre d'états dominants est relativement indépendant du nombre total d'états présent dans le modèle, suggérant un noyau compact de 5–6 états majeurs, le reste constituant des raffinements plus rares.

Ces observations s'accordent avec l'idée de métastabilité : le cerveau alterne entre quelques configurations récurrentes longues, ponctuées par des passages plus brefs vers des sous-états.

Une analyse plus détaillée montre que certains états jouent le rôle de puits (recevant des transitions depuis la plupart des autres états), alors que d'autres agissent en sources. Le cartes neuronales associées aux différents états sont stéréotypée entre individues et présentent des patterns d'activité symétriques ou latéralisés cohérents avec l'implication de régions du hindbrain dont l'ARTR. Une analyse des transitions moyennées par groupe visualise cette hiérarchie, confirme la conservation des routes dominantes et clarifie en quoi les différences observées forment des profils stables entre groupes.

Méthodologiquement, l'approche complète d'autres méthodes plus classiques car, au lieu d'inférer des états directement depuis l'activité brute, on discrétise un espace latent déjà partagé et interprétable, ce qui aligne naturellement les sujets avant la segmentation et facilite la comparaison inter-individuelle.

Chapitre 5 : Discussion générale et perspectives

Pris ensemble, ces résultats proposent un cadre unifié pour décrire et comparer l'activité cérébrale spontanée entre individus :
(i) un lien circuit–comportement via des HMM tri-états reliant ARTR et réorientation.
(ii) un espace latent commun génératif et interprétable via RBM permettant traductions et comparaisons à résolution cellulaire.
(iii) une vocabulaire d'états cérébraux dont les occupations et transitions markoviennes sont stéréotypées entre individus.
Ces résultats confortent l'idée selon laquelle l'activité spontanée reflète des contraintes d'organisation robustes, partagées entre individus, tout en autorisant des modes d'exploration individuels au sein d'un sous-espace latent commun.

Au-delà du périmètre étudié, plusieurs perspectives pourraient être suivies dans le futur :
(a) utiliser les RBMs pour étudier les états spontanés et évoqués (ex. tâches/stimuli/perturbations) dans un même espace latent pour tester comment le vocabulaire spontané est recruté ou réorganisé durant des tâches sensori-moteurs.
(b) suivre des individus au cours du développement ou selon différent phénotypes afin de quantifier la maturation et la variabilité de la structure neuronale spontanée
(c) explorer des modèles multi-niveaux (arbres d'états, échelles multiples) afin d'articuler assemblées, réseaux mésoscopiques et dynamiques globales.

Ces directions prolongent l'idée d'un langage compositionnel des dynamiques cérébrales – neurones $\rightarrow$ assemblées $\rightarrow$ états $\rightarrow$ transitions, apte à faire le pont entre modèles mécanistes et la structure statistique de l'activité cérébrale spontanée.

When a physiologist speaks now of the life of a plant or of an animal, he sees rather an agglomeration, a colony of millions of separate individuals than a personality one and indivisible. He speaks of a federation of digestive, sensual, nervous organs, all very intimately connected with one another, each feeling the consequence of the well-being or indisposition of each, but each living its own life. Each organ, each part of an organ in its turn is composed of independent cellules which associate to struggle against conditions unfavorable to their existence. The individual is quite a world of federations, a whole universe in himself.

Peter Kropotkin - 1896 - published by San Francisco Free Society Press (1898)

In comparing social with cerebral organisations one important feature of the brain should be kept in mind; we find no boss in the brain, no oligarchic ganglion or glandular Big Brother. Within our heads our very lives depend on equality of opportunity, on specialisation with versatility, on free communication and just restraint, a freedom without interference. Here too local minorities can and do control their own means of production and expression in free and equal intercourse with their neighbours. If we must identify biological and political systems our own brains would seem to illustrate the capacity and limitations of an anarcho-syndicalist community.

W. G. Walter - Anarchy (1963)

Remerciements

There are a lot of people to thank... so brace yourselves!

First, I want to thank Volker and Georges for being incredible supervisors. You've helped me become a better scientist, as much through your guidance and advice as through your support. In particular, thank you for your help, patience, and protection when things got tough. There was a time when I probably would have quit if you hadn't been there. You make a great team!

I'm also grateful to my collaborators: Simona, Rémi, and especially Jorge, for being fantastic partners and for shaping a large part of the work presented in this thesis.

Then, a big thank you to the rest of the fish team. Thanks to Sharbat, Monica, Leonardo, and Alexandre for sharing this PhD adventure with me. Thanks to Natalia, Geoffrey, Julie, Guillaume, and Antoine for teaching me so much. And of course, thanks to Marcus, Elias, Ghislaine, Clémence, Guillaume, Sophia, Manon, and Hyppolyte who, along with those already mentioned, made my time in this team absolutely amazing.

I also want to thank all the members of the lab, too many to name, whom I had the chance to interact with. You're the reason I consider the LJP a home. Special mentions to Darka, Guillaume, Nidia, and Quentin.

Thanks to my TAC members, Maxime, Laurent, and Claire, for helping me realize that, at first, I didn't actually know what I wanted to work on, and then for your guidance and support along the way.

Merci à Malika, Laura et Anissa pour leur aide administrative, leur patience et pour avoir toujours été à l'écoute.

Merci à Karim, Marco, Jérémy, Édouard et à tous les autres membres de l'animalerie poisson pour avoir pris soin des poissons.

Merci à Thomas, Carou et à toute l'équipe de l'atelier mécanique pour leur aide technique.

Et bien sûr, merci à Skodou ... pour tout.

So long, so long, so long, and sorry to all the fish

Contents

List of Figures

List of Abreviations

ARTR Anterior Rhombencephalic Turning Region. i, ii, iv, 38, 39, 42–44, 72, 114, 126, 128, 137, 140

BM Boltzmann Machine. 23, 25–28

BMs Boltzmann Machines. 23, 25–27, 29, 30

bRBM brain RBM. 116, 117, 119–122, 127, 129

bRBMs brain RBMs. 116, 119

cRBMs compositional Restricted Boltzmann Machines. 31, 32

dReLU double-Rectified Linear Unit. 27, 31

EA Evoked Brain Activity. 3, 4, 7, 8, 11, 128

fMRI functional Magnetic Resonance Imaging. 4, 7, 15, 32

HMM Hidden Markov Model. i, ii, 32, 33, 35–39, 44, 72, 137, 139, 140

MC Markov Chain. 32, 33, 35, 36

RBM Restricted Boltzmann Machine. i, ii, 7, 11, 27–32, 38, 77–81, 83–85, 88, 92–94, 96–98, 100, 104–107, 114–116, 122, 124, 137, 139–141

RBMs Restricted Boltzmann Machines. 19, 25, 28, 29, 31, 32, 38, 73–75, 77, 78, 81, 83–85, 87, 92, 94–97, 100, 102, 106, 108, 109, 113, 119, 121, 124, 126–129, 133, 137, 139–141

rs-fMRI resting-state fMRI. 8

SA Spontaneous Brain Activity. xi, 3–5, 7–12, 19, 38, 113, 114, 119, 122, 126, 128, 137, 138

sRBM state RBM. 116, 117, 119–122, 124–129, 133, 134, 140

sRBMs state RBMs. 117, 123, 133

Chapter 1

Introduction

This is a film about a man and a fish
This is a film a about dramatic relationship between man and fish
The man stands between life and death
The man thinks
The horse thinks
The sheep thinks
The cow thinks
The dog thinks
The fish doesn't think
The fish is mute, expressionless
The fish doesn't think because the fish knows everything
The fish knows everything

Goran Bregović, This Is a Film

1.1 What are Brains Used For

At the broadest level, brains are biological machines that transform sensory inputs into adaptive behavior. They do so by building internal models that integrate perception with memory, emotion, motivation and expectations, ultimately giving rise to thought and conscious experience [1]. Although this functional description is simple, the underlying mechanisms span multiple spatial and temporal scales, from ion channels to whole-brain dynamics. A central challenge for neuroscience is to connect these levels in a way that also explains behavior and its underlying neuronal computation. Large-scale network structure and the constraints it imposes on neural dynamics are key to building those bridges [2, 3, 4].

Brain structure and connectivity. Brains are not amorphous: they are organized into regions and pathways linked by long-range projections. This structural connectivity provides a scaffold that shapes and constrains neural activity and, by extension, function [4, 3]. Many studies have mapped both global and regional structure, revealing densely interconnected hubs and modular organization [5, 6, 7, 8, 9]. Computational work further shows that some aspects of functional connectivity can be predicted from anatomy, underscoring how structure channels ongoing activity [10, 11].

A functionalist view of the brain. A useful way to organize our understanding of the brain is to ask what the system computes, how that computation is implemented, and on what physical substrate; an approach popularized by Marr's three levels [12]. In this functionalist view, modeling is not optional: quantitative theories make testable predictions about representation, dynamics, and behavior, from single-neuron spiking to whole-brain network models [9, 13, 14, 15]. These models help relate structural constraints to the computations that support all brain functions, from perception to learning and action.

Why behavior matters. Studying behavior is essential for interpreting neural activity. Modern computational ethology shows that actions can be decomposed into discrete, stereotyped modules that recur across time and individuals [16, 17, 18, 19, 20]. Such structure in behavior suggests corresponding structure in the neural circuits that generate it, and motivates analyses that seek common motifs in neural data across individuals.

From stereotyped behavior to stereotyped activity. Evolution and development further shape brains toward common solutions. Many neural cell types, circuit motifs and large-scale divisions are conserved across individuals, and some nervous systems exhibit near-invariant wiring diagrams [21, 22, 23]. These facts imply that at least some parts of brain organization are stereotypical, promoting species-typical behavior and potentially shared patterns of brain activity.

Stereotyped behaviors (for example, grooming sequences in flies [17] or modular locomotor bouts in mice [19]) suggest stereotyped neural programs. In sensory systems,

stimulus-evoked responses often reveal striking regularities across trials and individuals. At the same time, trial-to-trial variability in evoked responses is strongly influenced by the brain's ongoing state, indicating that the so called spontaneous activity conditions what gets expressed when stimuli arrive [24, 25].

Spontaneous activity as organized dynamics. Far from being noise, spontaneous activity exhibits coherent spatio–temporal structure (see Sec. 1.2). At the microscopic scale, spontaneous population events outline the *vocabulary* of possible evoked patterns, and learning can reshape this vocabulary to reflect environmental statistics [25, 26]. These observations suggest that ongoing dynamics embody internal models and constraints that are shared within a species to some degree.

Shared organization versus individual idiosyncrasy. How much of this organization is common across individuals? At coarse scales, resting-state networks are robustly similar, and group-averaged maps capture major features shared across individuals [27]. Yet precision studies reveal stable, individual-specific features of network topology and connectivity (*functional fingerprints*), alongside heritable variation in network topology [28, 29, 30, 31, 32, 33]. Thus, spontaneous activity seems to contain both shared structure and idiosyncratic detail.

The balance likely depends on scale, system and analysis method. In particular, most cross-individual comparisons focus on connectomics rather than neuronal dynamics, raising a key methodological issue: *how to compare spontaneous activity across individuals.*

The central question. And thus we arrive at the central question that motivates this thesis: *is spontaneous brain activity comparable between individuals ?*

Answering this will require operational definitions of *comparability*, cross-individual alignment methods, and modeling of neuronal activity. The remainder of this introduction sets up the conceptual and methodological tools we will use to address it.

1.2 Spontaneous Brain Activity

From stimulus–response paradigms to intrinsic dynamics. Systems neuroscience has traditionally embraced a reflexive, input–output framework: present a stimulus, record the neural response, average across trials, and treat the remaining variability as noise. This approach, well suited to sensory–motor questions, focuses on Evoked Brain Activity (EA) and permits straightforward trial averaging within individuals and response-map or tuning-curve comparisons between animals. Yet such averaging isolates only the fraction of brain activity that is time-locked to external events, effectively discarding the vast repertoire of Spontaneous Brain Activity (SA) [34].

Metabolic data underscore the brain's intrinsic activity. In adults, the brain already expends roughly 20% of the body's energy budget at rest, while goal-directed tasks raise

this baseline by merely ∼5% [35, 34, 36]. Pioneering work by Raichle and colleagues in the late 1990s identified a constellation of regions that are active during quiet wakefulness and attenuated during certain externally focused tasks : the *default-mode network* (DMN) [37, 34]. Subsequent studies have shown that the DMN is not an "anti-task" circuit but is linked to internally generated narrative, memory, and self-referential processing [38]. The DMN's discovery thus highlighted the need to interrogate brain function outside classic stimulus-driven paradigms.

Together, these observations imply that the brain is intrinsically busy and that EA represents only the visible tip of a far richer dynamical iceberg. The remainder of this section reviews what is known about SA, why it matters, and what its structure can reveal about fundamental brain function.

1.2.1 What is Spontaneous Brain Activity ?

Terminology across species. Spontaneous Brain Activity has been described in most major animal models, yet its naming and operational definition vary by discipline.
In humans, SA is often called *resting state activity*, and is commonly measured using functional Magnetic Resonance Imaging. This state is defined behaviorally as *quiet wakefulness*, often with eyes closed or fixated, and is associated with stimulus-independent thoughts like day-dreaming and mind-wondering.
In rodents, SA is sometimes referred to as *idle state activity*, again defined behaviorally as periods of immobility and/or non-engagement with tasks.
However, it is important to distinguish resting state (in the behavioral sense) from the brain's physiological state, which is never truly at rest [34].

Beyond the resting state. Although resting paradigms offer experimental convenience, SA is not confined to quiet wakefulness. Structured, intrinsic, activity is observed across sleep stages, under anesthesia, and even during ongoing sensory stimulation and motor output [36]. For this reason, many authors prefer the broader label *intrinsic brain activity*.

A working definition. While there seems to be a general intuition shared by most neuro-scientists, there is a lack of consensus in the literature on an exact definition of SA. Yet most descriptions share three hallmarks:

(i) **Ongoing:** present in all behavioral and consciousness states;

(ii) **Structured:** characterized by stable spatiotemporal patterns;

(iii) **Task-independent:** not directly time-locked to explicit stimuli or tasks.

We thus retain the following definition:

Spontaneous Brain Activity is ongoing, structured neural activity that is unrelated to current tasks or explicit sensory inputs.

The remainder of this section uses this working definition to survey empirical findings and theoretical frameworks that illuminate the origin, organization, and functional relevance of spontaneous activity.

1.2.2 Why is Spontaneous Brain Activity important ?

Spontaneous Brain Activity (SA) is implicated in a broad spectrum of fundamental brain functions. Below, we present a concise, though non-exhaustive, survey of evidence that connects spontaneous activity to neural development, memory consolidation, learning, the formation of internal models, and the shaping of ongoing behavior.

1.2.2.1 Early development

Many mammalian brains enter the world in an immature state, yet their circuits are already partially "pre-tuned" for future sensorimotor demands. A large body of work indicates that this early organization is driven by SA. In rodent neocortex, activity begins as largely uncorrelated single-neuron firing and then progressively organizes into weakly coordinated ensembles, often expressed as sparse, periodic waves [42, 43, 36]. Such early patterns guide circuit refinement, including synaptic pruning, plasticity, myelination, and axonal targeting.

Sensory systems. Spontaneous activity arises before overt sensory experience: in mice it is present in somatosensory cortex prior to exploratory behavior, in auditory pathways before ear opening, and in the visual system before eye opening [43]. In sensory regions, this activity helps establish coarse topographic maps [44]. A striking example comes from ferret visual cortex, where retinal-independent SA predicts the orientation maps measured after eye opening (Fig. 1.1 A), demonstrating a causal role in emergent functional architecture [39]. Remarkably, orientation-selective population structure still develops when ferrets are reared without exposure to oriented edges, supporting the idea that SA can shape functional organization on its own [45].

Precocial species. In contrast to rodents and humans, many species must function autonomously shortly after birth. Larval zebrafish, for instance, can hunt and evade predators by five days post-fertilization [46, 47, 48]. Accordingly, their visual circuitry matures rapidly. Optic-tectum studies show that visual experience sharpens this development, yet SA alone is sufficient to organize tectal circuitry even in enucleated fish [49, 50].

Taken together, these findings portray Spontaneous Brain Activity as a core driver of early circuit formation, an intrinsic "pre-training" phase that primes neural networks for later sensory experiences and behaviors.

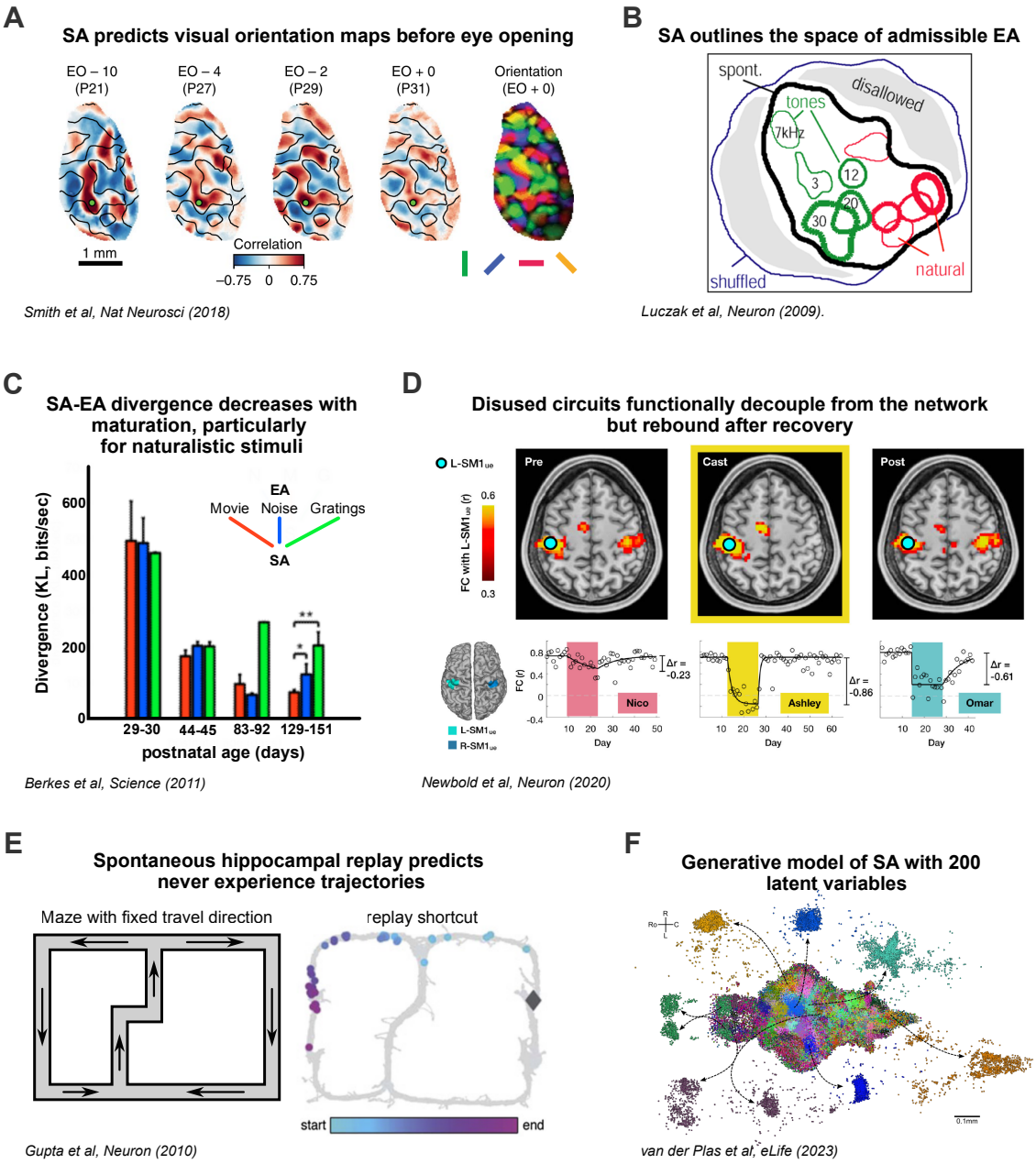


Figure 1.1: *(Caption on next page)*

Figure 1.1: **Spontaneous Brain Activity (SA) is integral to brain function**. **A**: Correlation maps of awake ferret visual cortex before eye opening (EO, correlation from green seed point). Contour lines indicate the domain as measured after EO. Right: orientation preference map after EO. Reproduced from Smith et al. [39]. **B**: Contour plot of neuronal population firing rate vector, measured in awake rat auditory cortex (projected to 2D space with multidimensional scaling). Sensory evoked activity lies within the subspace outlined by SA. Reproduced from Luczak, Barthó, and Harris [25]. **C**: Kullback–Leibler (KL) divergence between SA patterns and activity evoked by naturalistic movies (red), noise (blue), and sinusoidal gratings (green), in awake ferret visual cortex across development (29 to 151 days old). Adapted from Berkes et al. [26]. **D**: Human functional Magnetic Resonance Imaging functional connectivity (FC) measurement in somatomotor cortex, before, during, and after dominant arm casting. Top: FC from a seed voxel in the left primary somatomotor cortex (L-SM1). Bottom, FC between left and right SM1 over the whole experiment, shaded areas represent the casting period. Reproduced from Newbold et al. [40]. **E**: Hippocampal replay of rat trajectories in a T maze. Left: the maze can be traveled along 2 loops but always in the same direction. Right: Example replay trajectory never experienced by the animal (shortcut). Grey diamond indicates the rat's location at the time of replay. Dots represent locations encoded by place-cell activation, with color indicating the temporal sequence. Adapted from Gupta et al. [41]. **F**: Example neuronal assemblies identified with an Restricted Boltzmann Machine trained on zebrafish larvae whole-brain spontaneous neuronal activity. Reproduced from van der Plas et al. [11].

1.2.2.2 Similarity between Spontaneous and Evoked Activity

A key aspect of SA is its similarity with Evoked Brain Activity (EA).

Luczak, Barthó, and Harris [25] recorded ensembles of $40-100$ neurons in rat auditory and somatosensory cortex under diverse conditions (awake versus anesthetized, naturalistic versus artificial stimuli) and found that the sequential structure of population activity is largely conserved between EA and SA. Specifically, population firing-rate vectors elicited by stimuli occupied stimulus-specific subspaces *nested* within a broader region outlined by spontaneous events, suggesting a constrained neural "vocabulary" of admissible patterns (Fig. 1.1 B).

Importantly, this similarity between EA and SA is, in part, learned through experience. Using voltage-sensitive dye imaging in anesthetized rat V1, Han, Caporale, and Dan [51] showed that repeated presentation of a visual pattern increased the likelihood that subsequent spontaneous traveling waves matched the evoked template. The effect scaled with repetition number, persisted for minutes, and was specific to the trained pattern, indicating rapid experience-dependent reshaping of SA consistent with short-term memory formation.

Development further refines this correspondence. In awake, freely viewing ferrets, Berkes et al. [26] tracked the distribution of population activity in primary visual cortex

during natural-movie viewing (EA) and in darkness (SA). The divergence between the two distributions declined steadily with age and was minimal in mature animals for natural-scene responses, and less so for gratings or binary noise (Fig. 1.1 C). Importantly, this convergence was not explained by firing-rate statistics alone but required higher-order correlations, highlighting the role of structured population dynamics.

Collectively, these studies demonstrate that SA embodies priors over evoked patterns, priors which are learned through sensory experience and tuned to environmental statistics.

1.2.2.3 Plasticity, learning and memory

If SA encodes the brain's prior expectations, it should itself be plastic. Indeed, experience reshapes spontaneous patterns, yet SA is not merely a passive echo of EA.

Rapid circuit-level reorganization. Newbold et al. [40] immobilized the dominant arm of three adults for two weeks and recorded daily 30-min resting-state fMRI scans. Within 24–48h, cortical regions controlling the disused arm functionally disengaged from the broader somatomotor network (Fig. 1.1 D), while connectivity within the isolated circuit remained intact. During immobilization, large spontaneous "disuse pulses" propagated through the disconnected circuit. These pulses disappeared and functional connectivity rebounded soon after cast removal. These findings suggest that spontaneous bursts help preserve local integrity when a circuit is transiently uncoupled from the wider network.

Constructive replay. Hippocampal replay further illustrates that SA is not a simple recapitulation of past experiences. In an awake two-choice T-maze task, Gupta et al. [41] examined sharp-wave–ripple events and found forward and backward replays of trajectories that had not been recently experienced, and even of routes rarely or never experienced in that direction. They also observed novel "shortcut" sequences that stitched together unexperienced paths across the maze (Fig. 1.1 E). These results imply that replay actively maintains a balanced cognitive map and explores novel possibilities, rather than simply consolidating the most recent or frequent experiences.

Together, such evidence reveals that Spontaneous Brain Activity is a dynamic substrate for learning and memory, capable of both preserving circuit function during perturbation and generating prospective scenarios that guide future behavior.

1.2.2.4 Behavioral embedding and orthogonal coding

The sections above focused on SA recorded during quiescence, sleep, passive perception, or anesthesia. But what becomes of spontaneous activity when animals are actively behaving?

Orthogonal subspaces in mammalian cortex. Recording $>10{,}000$ neurons in awake mouse V1, Stringer et al. [52] combined darkness or visual-stimulus protocols with video

tracking of spontaneous facial movements. Ongoing behavior was robustly encoded in a high-dimensional latent state whose components predicted neuronal spontaneous activity. Sensory and behavioral signals resided in largely *orthogonal* subspaces, overlapping mainly along a single dimension that modulated the population's mean firing rate. Thus, incoming stimuli add to, rather than overwrite, the internal representation of behavior. This framework explains trial-to-trial variability not as noise but as the superposition of task-evoked responses onto a constantly present internal representation of ongoing endogenous behavior.

Behavioral embedding in small nervous systems. Whole-brain imaging in smaller organisms similarly shows tight links between SA and motor actions. In *Drosophila*, brain-wide activity ramps precede spontaneous walking, and the same movement-related manifold is revisited during immobilization [53]. In zebrafish, tectal assemblies predict tail flips [54]. In *C. elegans*, population activity traces a low-dimensional manifold that encodes motor sequences and is traversed even when the worm is paralyzed, indicating that internal dynamics can drive putative motor commands [55].

Dimakou et al. [36] concluded from these findings that Spontaneous Brain Activity continuously expresses *behavioral priors*: statistical regularities of the animal's own actions that can bias imminent sensory processing and decision making.

1.2.2.5 Allostasis

Beyond perception and action, a fundamental function of the brain is to sustain the body's physiological needs, a process termed *allostasis*. Doing so requires an internal model of bodily state that can bias other brain functions toward survival-promoting behaviors.

Animal studies. In mice, distributed networks encode motivational axes such as thirst versus satiety [56] and hunger versus fullness [57]. These internal-state signals modulate both behavior and the baseline pattern of spontaneous activity across widespread brain areas.

Human evidence. In humans, large-scale circuits involved in interoception and allostasis have been delineated with neuroimaging [58, 59]. Theoretical accounts propose that these networks generate predictive codes of present and future bodily states, which are compared with afferent visceral signals to guide perception and action [60].

Together, these findings imply that Spontaneous Brain Activity carries *interoceptive priors* : statistical regularities of internal physiology that continuously tune neural processing toward energy balance and bodily needs.

1.2.2.6 Spontaneous activity as an informative, adaptive signal

The examples presented above disprove the view of SA as mere noise. Instead, spontaneous patterns form an active substrate for development, circuit maintenance, internal modeling, and adaptive behavior. During early life, SA "pre-trains" developing circuits, shaping coarse sensory maps and preparing networks for later experience. In adults, it outlines the repertoire of admissible evoked configurations, yet does not simply replay past experiences. Rather, spontaneous bursts preserve local connectivity during disuse, construct novel trajectories, and update cognitive maps. Moreover, SA carries sensory, behavioral, and interoceptive priors that bias perception and action toward environmental regularities and bodily needs. Far from being neural noise, spontaneous activity is a dynamic, information-rich signal integral to brain function across the lifespan.

1.2.3 The brain as a generative model

A growing consensus views the brain as an inferential system that continuously predicts both the external world and the body within it. Because neuronal activity is intrinsically stochastic, this view is naturally formalized in probabilistic terms, most prominently by Bayesian inference [26, 61, 62, 36].

Bayesian inference in neuroscience. Let $\mathcal{C}$ denote a cause in the world (or the body), and $\mathcal{R}$ be neural responses (*i.e.* the internal representation). Bayes' rule expresses how the brain could combine prior expectations with incoming evidence under an internal model :

$$P(\mathcal{R} \mid \mathcal{C}) \;=\; \frac{P(\mathcal{C} \mid \mathcal{R})\, P(\mathcal{R})}{P(\mathcal{C})}. \tag{1.1}$$

Here $P(\mathcal{R} \mid \mathcal{C})$ is the *posterior* describing the internal representation of a given cause, $P(\mathcal{C} \mid \mathcal{R})$ the *likelihood* mapping that cause to an internal representation, $P(\mathcal{C})$ are the *prior* beliefs about the cause (*i.e.* what causes could be encountered), and $P(\mathcal{R})$ are the *prior* internal representations. This last distribution can be seen as the expected internal representations in the absence of any cause (*i.e.* the posterior when there is no explicit sensory input) [26].

Generative Models. Models which learn the conditional probability $P(\mathcal{C} \mid \mathcal{R})$ are usually called discriminative models as they can be used to "discriminate" the cause $\mathcal{C}$ given an observed brain response $\mathcal{R}$. In contrast, generative models learn the joint probability $P(\mathcal{R}, \mathcal{C})$, providing both *top-down* ($P(\mathcal{R} \mid \mathcal{C})$) and *bottom-up* ($P(\mathcal{C} \mid \mathcal{R})$) pathways to move between data and representation. Hence, these models can generate new data $\mathcal{R}$, contrary to discriminative models which are often used only as classifiers. [63]

Spontaneous activity as model priors. Pezzulo, Zorzi, and Corbetta [62] propose that SA is the signature of a generative model's top-down dynamics when decoupled from immediate action and perception. The brain *samples* its internal model, generating hypothet-

ical causes $\mathcal{C}$ and their neuronal responses. Dimakou et al. [36] distilled three predictions from this idea:

(i) SA should resemble EA, especially for familiar, ecologically relevant stimuli, reflecting experience-dependent priors.

(ii) SA should encode behavioral and interoceptive priors, biasing the system toward frequently encountered motor programs and internal states.

(iii) The metabolic cost of SA should reflect anticipatory readiness rather than direct task execution.

Predictions (i) and (ii) are supported by the evidence reviewed in the previous sections. Prediction (iii), while compelling, lies outside the scope of this thesis.

Link to predictive coding and active inference. Within hierarchical predictive-coding and active-inference frameworks, higher cortical areas generate top-down predictions that are compared with bottom-up sensory signals. Mismatches (*i.e.* prediction errors) drive both belief updates and action selection. SA may thus correspond to ongoing top-down "pre-play" that maintains model fidelity, explores counterfactual scenarios, and primes circuits for rapid adaptation, a view consistent with replay phenomena, disuse pulses, and internally driven motor manifolds discussed above.

Summary. With this perspective, SA represents the *prior* term $P(\mathcal{R})$ in Eq. 1.1, whereas EA approximates the *posterior* $P(\mathcal{R}|\mathcal{C})$ once sensory evidence arrives. Both are shaped by the same generative model, which is continually tuned to environmental statistics, bodily needs, and behavioral states.

1.2.4 Methodological frontiers in studying Spontaneous Brain Activity

Beyond the theoretical and experimental characterization of SA we have discussed until now, many methodologies leveraging generative approaches have been developed in recent years to analyze SA [64, 65].

Triplett et al. [66] developed a generative latent-variable model that simultaneously fits stimulus filters and low-dimensional spontaneous factors directly to fluorescence traces. They applied this method to recordings of the zebrafish tectum and mouse V1 and showed that they could decouple the spontaneous and single-trial evoked components of brain activity without trial averaging, showing that low-dimensional SA persists throughout stimulus presentation.

van der Plas et al. [11] fitted a Restricted Boltzmann Machine (see Sec. 1.5.4), a class of generative model, to the whole-brain neuronal activity of zebrafish larvae, and showed that

100-200 latent variables are sufficient to reproduce the first- and second-order statistics of SA. In particular, they showed that those latent variables represented spatially compact neuronal populations resembling known functional ensembles (Fig. 1.1 F), providing a model of SA which is interpretable at the mesoscale level.

Open challenge: cross-individual alignment of Spontaneous Activity. Most existing methods interrogate SA *within* a single animal. In human neuroimaging, cross-subject techniques such as hyperalignment and the shared-response model align representational spaces across brains [67, 68, 69]. Analogous tools are scarce for cellular-resolution data, where anatomical variability and the sheer dimensionality prevent direct comparison. Consequently, we still lack a principled framework for asking whether two animals "share" a spontaneous repertoire, how that repertoire varies with genotype or experience, and which neuronal ensembles are homologous across animals.

1.3 Zebrafish as a Model Organism

To investigate how spontaneous whole-brain neuronal dynamics are structured and conserved across individuals, we require an experimental system that allows brain-wide recordings at cellular resolution, ideally in a vertebrate. Larval zebrafish uniquely provide this combination of accessibility and scale, making them an ideal model for the questions addressed in this thesis.

Animals have been central to experimental biology since its earliest days. In modern life sciences a small set of *model organisms* has become disproportionately influential because they are tractable, standardized, and embedded in rich toolkits and communities (e.g., Drosophila, C. elegans, mouse, zebrafish) [70, 71]. A general discussion on the ethics of animal research can be found in Annex 1.

Zebrafish. Zebrafish (*Danio rerio*, Fig. 1.2 A) rose to prominence as a vertebrate model in the late 20th century through a sequence of enabling steps: adoption for genetic work by George Streisinger and colleagues, and then large-scale genetic screening that demonstrated its power for vertebrate development and disease genetics [72]. The subsequent consolidation of husbandry protocols, staging standards, genomic resources, and community repositories cemented its place across developmental biology, genetics, neuroscience, behavior, and biomedicine [73, 74].

Husbandry. Zebrafish are comparatively easy to maintain at scale. They breed easily, with external fertilization yielding large clutches and rapid embryogenesis. Developmental stages are standardized to hours or days post-fertilization (hpf/dpf), enabling tight age-matching across cohorts and facilitating statistical power thanks to large numbers of individuals [74, 75]. Practical guidelines for care and breeding are mature and widely available [73, 75].

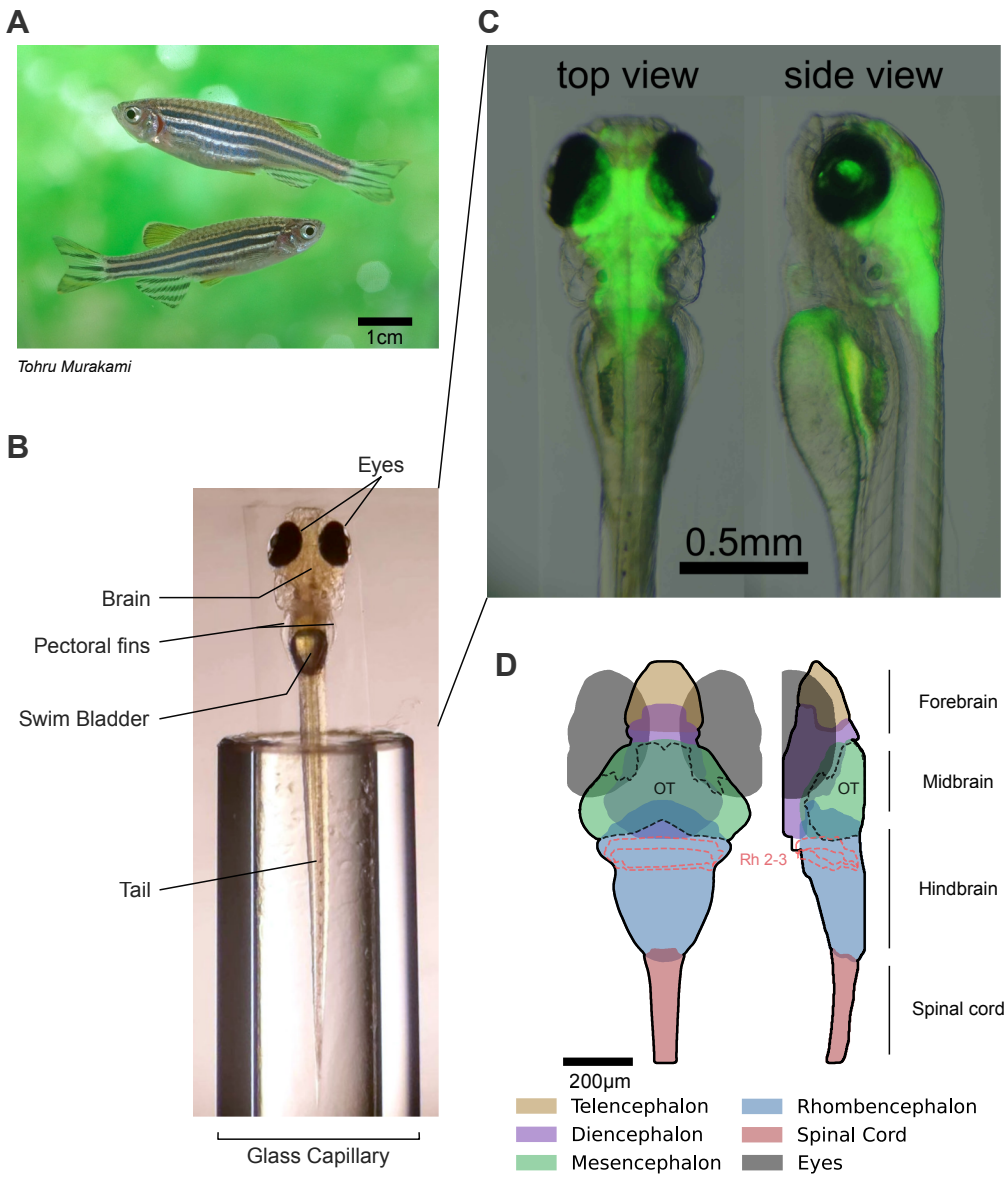


Figure 1.2: **Zebrafish larvae as model organisms**. **A**: Adult zebrafish (AB strain). Top: female, bottom: male. **B**: Zebrafish larva (nacre, 5dpf) embedded in 2% agarose and held by a glass capillary. **C**: Closeup view of a zebrafish larva head, top and side view. Two images are superposed : a bright-field of the anatomy, and a fluorescence image of pan-neural GCaMP6f expression. **D**: Brain anatomy of the zebrafish larvae, as referenced by the ZBrain atlas [6]. Large scale brain regions and anatomical divisions are shown. Optic Tectum (OT) and Rhombomere 2 and 3 (Rh 2-3) are shown for future reference.

Larval stage. By 4–7 dpf (Fig. 1.2 B), larvae exhibit a rich behavioral repertoire including routine and burst swimming, startle/escape, prey capture, and visually driven reflexes such as the optomotor (OMR) and optokinetic (OKR) responses [46, 76, 77, 78]. The combination of small size, transparency, and stereotyped behaviors has made larvae ideal for linking circuit activity to sensorimotor output under head-fixed or freely swimming conditions [79, 78, 80, 81]. Genetic backgrounds that reduce pigmentation (*e.g.* nacre, casper) further increase transparency and are now standard in systems neuroscience [82, 83]. Routine transgenesis (*e.g.* Tol2/Gal4-UAS) enables pan-neuronal (Fig. 1.2 C) or cell-type-specific expression of calcium indicators [84, 85].

1.3.1 The larval Zebrafish brain

Brain anatomy. Zebrafish and humans are both vertebrates and share many conserved brain divisions (forebrain, midbrain, hindbrain). Nevertheless, important differences exist, notably the pallium which is much smaller and has debated homologies to mammalian brain areas [86, 87].

Comprehensive atlases provide anatomical ground truth for mapping imaging data to structures [6, 88, 89, 90] (Fig. 1.2 D).

Spontaneous activity. Spontaneous activity is present from the earliest stages, with embryos showing spinal-cord–generated coiling and evolving locomotor rhythms before robust sensory-evoked behaviors emerge [91, 92, 93, 94].

In the larval brain, spontaneous population activity is structured from mesoscale to the whole-brain level [79], revealing network motifs whose spontaneous activation predicts aspects of behavior.

In the optic tectum, spontaneous assemblies align with retinotopy and behaviorally relevant stimulus features and can develop even without retinal drive [50, 49].

Across the brain, spontaneous activity exhibits rich, state-dependence, collective dynamics (e.g., neuronal *avalanches*) [95]. Recent work linked this spontaneous dynamics to structural and genetic determinants [96, 97], supporting the idea that it should be stereotypical across fish at mesoscales.

In the context of this thesis, larval zebrafish are uniquely positioned for four reasons :

1. their transparency and small brain size allow for whole-brain cellular-resolution imaging of brain activity.
2. transgenic lines are readily available for functional imaging.
3. the anatomical organization of their brain is well characterized.
4. developmental stages are highly stereotyped at ~ 5 dpf.

5. findings can be, to some extent, generalized to other vertebrates.

Together, this makes zebrafish larvae an optimal vertebrate system to quantify and compare spontaneous brain dynamics across individuals.

1.4 Recording Neuronal Activity

To investigate the organization of spontaneous whole-brain dynamics, we must measure neuronal activity at single-cell resolution across the entire brain, providing direct access to the activity patterns underlying conserved ensembles and states. In this section, we first outline the principles of neuronal excitability and population organization, then review available recording methods, and finally introduce light-sheet calcium imaging as the technique applied throughout this thesis.

1.4.1 From Neurons to Brains

The nervous system is made up of excitable cells (neurons) supported by diverse glia, embedded in a connectivity circuit that comprises synapses, microcircuits, and large-scale networks. Neurons convert synaptic input into action potentials via voltage-dependent membrane conductance, a biophysics phenomena first quantified in the squid giant axon by Hodgkin and Huxley [98]. At the systems level, cognition and behavior arise from coordinated activity across many regions. Information is not localized to single neurons but is distributed over many interacting populations and pathways.

The computational organization of the brain. Two ideas help organize this complexity. First, *cell assemblies* are transiently co-active groups of neurons proposed by Hebb [99], and later interpreted as the building blocks of population codes [100, 101, 102, 103].
Second, network architecture is characterized by neural graphs showing *small-world* features : high clustering with short path lengths. This is thought to support efficient communication and flexible computation [104, 105, 106].
Together these perspectives motivate measurements that capture not only single cells but also the mesoscale and whole-brain structure in which they are embedded.

1.4.2 Imaging Neuronal Activation

Early access to brain signals came from EEG which measures population-scale electromagnetic field fluctuations at the scalp. Electrophysiology then opened the cellular resolution: single-channel and whole-cell patch clamp provided direct, high-fidelity recordings of ion channels and spikes [108, 109]. Functional Magnetic Resonance Imaging (fMRI) added whole-brain coverage via the blood-oxygenation-level dependent (BOLD) contrast [110]. Each modality trades off spatial scale, temporal resolution, invasiveness, and throughput.

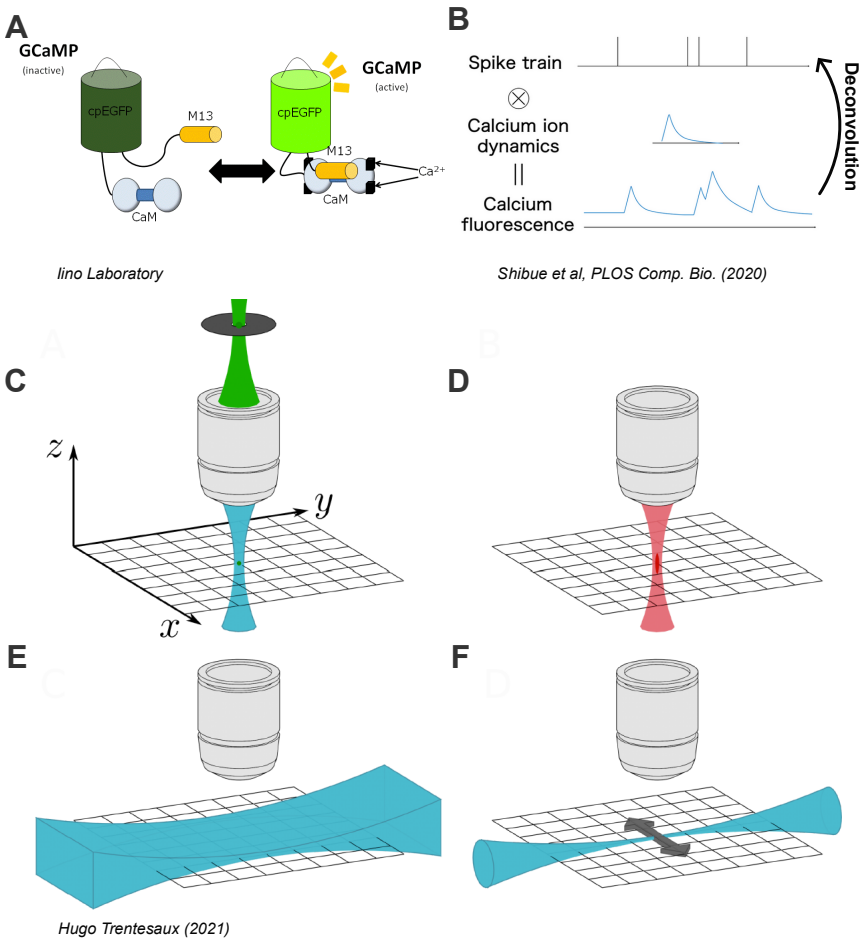


Figure 1.3: **Fluorescence microscopy and optical sectioning**. **A**: Diagram of the GCaMP genetically encoded calcium indicator (GECI), before and after the conformation change induced by the presence of Ca^{2+}. **B**: Convolution of a neuron spike train by the calcium kernel of a GECI. Reproduced from [107]. **C-F**: Methods of optical sectioning in fluorescence microscopy. Reproduced from the thesis of Hugo Trentesaux (2021). **C**: Confocal Microscopy: a pinhole is placed in the image plane to reject out-of-focus fluorescence. **D**: Two-photon Microscopy: non-linear absorption excites only a small volume at the focal point of the LASER. **E**: Static light-sheet : a cylindrical lens is used to excite a single plane of the sample. **F**: Scanning light-sheet : a galvanometric mirror scans a gaussian beam to excite a single plane of the sample.

Optical methods sit between electrophysiology and fMRI, offering cellular specificity over large fields of view.

Genetically encoded calcium indicators (GECIs). Optical readout of activity relies on indicators whose fluorescence reports intracellular signals. The GCaMP family is a canonical example which originated with single-GFP G-CaMP [111], and was progressively engineered to improve signal-to-noise, kinetics and dynamic range [112, 113, 114, 115].

GCaMPs fuse three modules into one protein: a circularly permuted GFP (cpGFP), flanked by calmodulin (CaM), and an CaM-binding peptide. When intracellular Ca^{2+} binds CaM, it triggers a series of conformation changes leading to a deprotonated form of cpGFP. The net effect is a Ca^{2+}-dependent increase in fluorescence (Fig. 1.3 A).

The fluorescence time course after a spike reflects (i) Ca^{2+} entry and buffering in the neuron, (ii) Ca^{2+} binding/unbinding to the indicator, and (iii) the rate of re-conformation. The result is a dynamic fluorescence signal much slower than that of neuron spiking itself, which can be approximated by a convolution of the spike train with an exponential kernel (Fig. 1.3 B). In the case of GCaMP6f in zebrafish, the typical time-scale of the kernel is $\sim 1.6s$ [116]. Consequently, spikes blur together when they occur within one "kernel time". An approximation of the original spike train can be obtained by deconvolution of the fluorescence signal [117], however it remains a proxy of true neuronal activity.

The importance of optical sectioning. In widefield fluorescence imaging, out-of-focus emission degrades contrast in thick tissue. Optical sectioning removes this background by focusing either the excitation or detection of fluorescence to a thin plane. Confocal microscopy rejects out-of-plane light with a pinhole [118, 119] (Fig. 1.3 C), whilst two-photon confines excitation to the focal volume through non-linear absorption, improving depth penetration in scattering brain tissue [120, 121] (Fig. 1.3 D). Both are laser-scanning techniques, meaning that images are high contrast, but they are typically acquired on a limited field-of-view and at limited rate due to serial scanning.

1.4.3 Light-sheet functional microscopy

Light-sheet (selective-plane) illumination decouples illumination and detection paths [123]. A thin sheet of light excites only a single plane orthogonal to the detection objective, and the camera images all neurons in this plane simultaneously (Fig. 1.4 A). The result is optical sectioning with dramatically reduced photobleaching and phototoxicity, and orders-of-magnitude faster volumetric rates when compared to point-scanning approaches [123, 124].

Principles. Practically, this can be implemented either by forming a static sheet with a cylindrical lens (Fig. 1.3 E)[123] or by rapidly scanning a focused Gaussian beam with a galvanometric mirror (Fig. 1.3 F)[124]. In both cases the axial confinement is set by the illumination numerical aperture (NA).

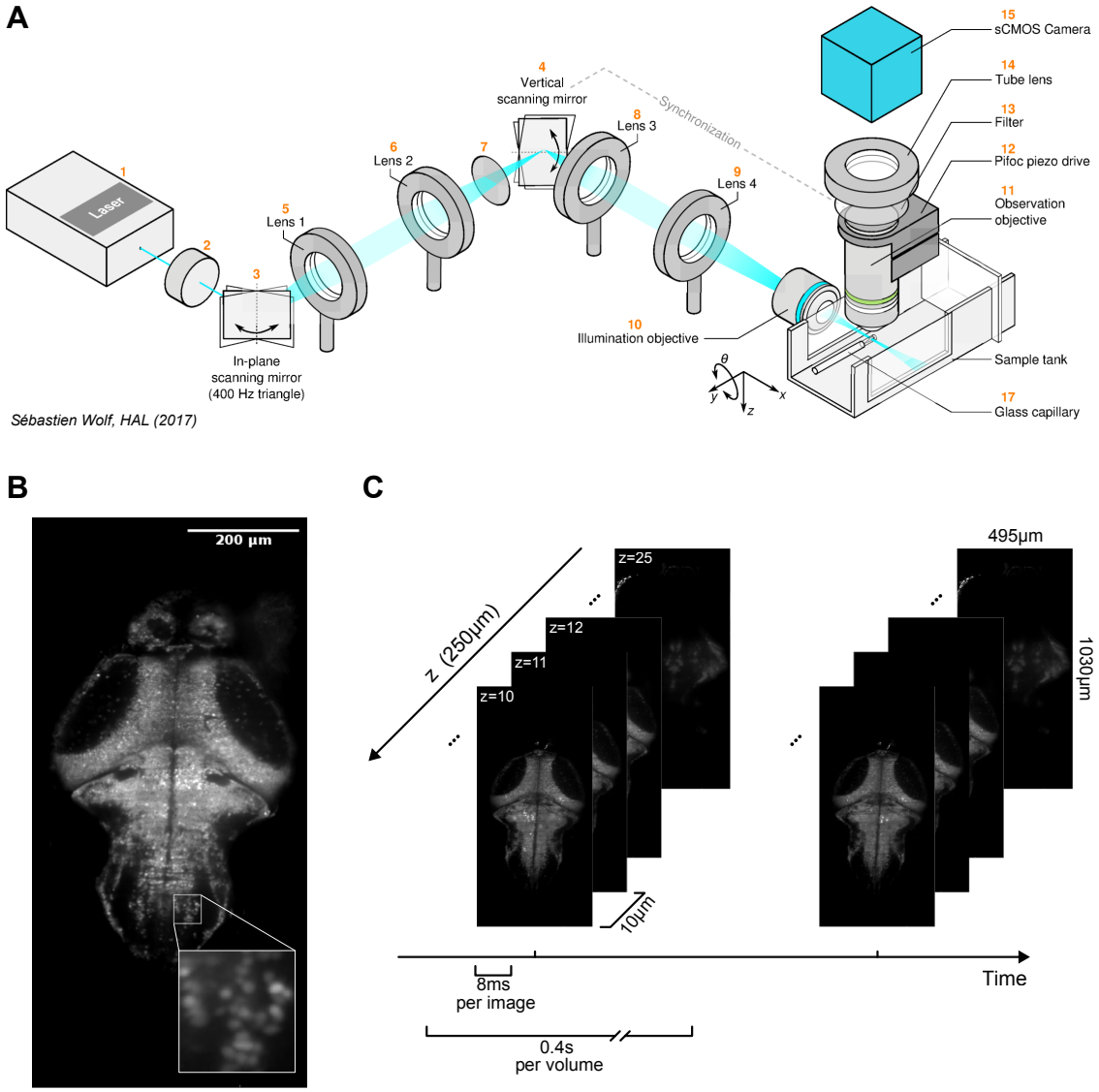


Figure 1.4: **Scanning Light Sheet functional imaging of zebrafish larvae**. **A**: Diagram of a scanning light-sheet microscope. Adapted from Wolf et al. [122]. **B**: Single layer high-resolution image of the zebrafish brain, as imaged with light-sheet microscopy . **C**: Functional imaging protocol. 25 z-layers separated by $10\mu m$ are imaged repeatedly at a volumetric scanning rate of $2.5Hz$.

Fluorescence is collected widefield along an axis orthogonal to the sheet, so every point in the illuminated plane is imaged simultaneously onto a camera. Axial sectioning is therefore set by the illumination (sheet thickness), while lateral resolution follows the detection NA and camera sampling.

Volumetric acquisition is obtained by rapidly stepping both the light-sheet and detection objective along the z direction, with the camera asynchronously triggered after each step to avoid motion blur (Fig. 1.4 C).

In neuroscience. In zebrafish larvae, light-sheet functional imaging records calcium signals from a large fraction of neurons across the whole brain, repeatedly and at high throughput. This was first done by Ahrens et al. [125] and Panier et al. [126] using GCaMP to image up to $\sim 80\%$ of neurons in the intact larval zebrafish brain with high volumetric scanning rates, establishing whole-brain single-cell functional mapping in vivo. Since then, light-sheet functional recording in zebrafish larvae as become a central tool to investigate mesoscale–to–whole-brain computations at cellular resolution.

1.4.4 On the importance of Whole-Brain Single-Cell recordings

Recent population studies reveal that neural representations are high-dimensional and distributed [127], and that inter-regions interactions occupy low-dimensional communication subspaces [128]. Brain-wide electrophysiology in mice similarly shows task variables encoded across many structures [129]. Light-sheet imaging complements these insights by delivering dense, brain-spanning, single-cell measurements suited to quantifying the shared structure of Spontaneous Brain Activity.

1.5 Modeling Brain Activity and Behavior

To make sense of whole-brain recordings of spontaneous activity, we need models that reduce high-dimensional data while preserving structure, generate realistic patterns, and remain interpretable across animals. In this thesis, Restricted Boltzmann Machines (RBMs) form the backbone of both our analysis and methodological developments, providing a generative framework to uncover latent neural assemblies and states. To capture how these states evolve over time, we complement RBMs with Markov Models, which provide a probabilistic description of temporal structure. In this section, we progressively introduce RBMs, beginning with simple network models, how they are trained and sampled to generate data, their applications in neuroscience, and we then outline how Markov models are used to analyze temporal dynamics.

1.5.1 A physicist's view of the brain

Spherical cows in a vacuum. Physicists are fond of stripping complex systems down to their essential components : the proverbial "spherical cow". The goal is not to deny biological richness but to gain traction by first modeling what matters most, and worrying about details later. In its barest form, a brain can be viewed as a collection of neurons that influence one another through synapses. Although real neurons exhibit intricate biophysical dynamics, we begin by approximating each neuron as a binary unit.

Binary network formulation. In this caricature, the brain is a network of nodes s_i that can occupy one of two states : *on* ($s_i = +1$) or *off* ($s_i = 0$ or, equivalently, $s_i = -1$) (Fig. 1.5 A). The network's wiring is encoded by connection strengths w_{ij}, and the instantaneous state of the system is the configuration vector $\mathbf{s} = (s_1, s_2, \ldots, s_N)$.

Structural vs. functional connectivity. There are however, a few subtleties we need to keep in mind when interpreting this simplified model, particularly regarding the connections w_{ij}.

Neurons are inter-connected anatomically through axons, dendrites, and synapses. These are the physical connections, also known as the *structural connectivity*. Acquiring such wiring diagrams at whole-brain scale is a formidable experimental challenge which has only recently started to be overcome [22, 8, 130, 131].

Instead, we often refer to connectivity as the effective connections between neurons which we can infer from observations of the network's activity [132]. This is known as the *functional connectivity*. A metric often used to measure this functional connectivity is the pairwise Pearson correlation

$$\rho_{ij} \;=\; \frac{\mathrm{cov}\big(s_i(t),\, s_j(t)\big)}{\sigma\big(s_i(t)\big)\,\sigma\big(s_j(t)\big)}$$

Structural connectivity shapes the activity, and therefore the pairwise correlation. Yet two neurons can be strongly correlated without a direct synapse, for example, when they share a common upstream partner.

In the rest of this thesis, we will not attempt to study neuronal activity through their structural connectivity. We will instead focus on models where the functional connectivity is naturally inferred from the data.

1.5.2 Configuration-based models

Ising. The simplest model one could use to model a network of interacting nodes is the Ising model. It was originally used to model ferromagnetic materials, where the nodes are

spins and the connections are the electromagnetic interactions between them. The energy of a configuration of this system is defined as :

$$E(\mathbf{s}) = -\sum_{i,j} J_{ij} s_i s_j - \mu \sum_i h_i s_i \tag{1.2}$$

where J_{ij} is the interaction between nodes i and j, h_i is an external field acting on node i, and μ is a constant describing the magnetic moment of each spin (*i.e.* how much the spin "feels" the magnetic field).

Energy-based models. It might be surprising to see an energy-based model in a neuroscience context. It is important to understand that when we talk about energy here, we are not talking about calories, but rather an abstract quantity describing the stability of the system. Using the Ising model, we can associate an energy to each configuration of the system $\mathbf{s}$. Some configurations will then have high energies, and others lower energies, defining an energy landscape in the space of all configurations (Fig. 1.5 B). As the system evolves from configuration to configuration, it tends to settle into valleys (local minima) but can also hop between them, especially at non-zero temperature.

Spin glass. In its simplest form, the Ising model is defined on a regular lattice, where each node is connected to its nearest neighbors only. Such a model is not very helpful in neuroscience as neurons tend to have long-distance connections. However, it can be generalized to any J_{ij}, a model known as Spin Glass, whose rich phase structure has proven useful for studying disordered systems and, by analogy, neural circuits [133, 134]. It is often used by first engineering the J_{ij}, and then studying the behavior of the system.

Hopfield networks. Hopfield [135] repurposed the spin-glass formalism to model associative memory. Adopting binary neurons $s_i \in \{+1, -1\}$ and symmetric weights w_{ij}, he defined the energy

$$E(\mathbf{s}) \;=\; -\sum_{i,j} w_{ij}\, s_i s_j \;-\; \sum_i \theta_i s_i, \tag{1.3}$$

where θ_i is the activation threshold of node i.

The idea behind Hopfield networks is to sculpt the energy landscape by manufacturing the w_{ij} in such a way that the system has a set of stable configurations which are energy minima. Each stable state encodes for a specific information, and as long as the system starts from a configuration within its basin of attraction, the system will evolve down-slope of the energy landscape towards the stable state. Each memory can therefore be recalled from an initial guess, similar to how we can remember a song from only a few notes.

This model became one of the foundations upon which the field of computational neuroscience was built.

Limitations. While Hopfield's formulation captures how discrete attractors can encode memories, it lacks a mechanism for *escaping* deep energy wells and therefore cannot explain the continual, metastable switching observed in real brains. To model such stochastic

exploration we require frameworks that explicitly embrace randomness, one step on the path toward Boltzmann machines and, ultimately, Restricted Boltzmann Machines (RBMs).

1.5.3 Boltzmann Machines

A major step beyond Hopfield networks is to *learn* an energy landscape *from data* rather than hand-crafting it. The aim is to approximate the probability distribution that generated the observations, not merely to store a fixed set of patterns.

This data-driven view took shape in the mid-1980s, when Hinton, Ackley, and Sejnowski introduced stochastic, probabilistic networks they called *Boltzmann machines* (BMs, Fig. 1.5 C). The framework fused statistical-physics ideas with machine learning and has influenced both fields ever since.

From energy to probability: the Boltzmann distribution. But how to include probabilities into an energy model ? Once again, we will call upon statistical physics, and in particular Ludwig Boltzmann. In the 1860s, Boltzmann was studying the behavior of gases, and was trying to understand the distribution of particles velocities in a gas depending on the temperature of the system. Whereas the energy of the whole system is fixed, each particle can have a range of energies, and Boltzmann was attempting to determine the probability distribution $P(E)$ that a given particle would have an energy E. Boltzmann's insight was to link the energy of the particle with a state s_E through the relation: $P(s_E) \propto e^{-E/T}$, where T is the temperature of the system.

We can build a good intuition of why this is true by discretizing the problem into infinitesimal states separated by an energy ϵ. Jumping up one state takes an energy ϵ, and happens with a probability p. Hence, jumping up n states takes a total energy $\Delta E = n\epsilon$, and happens with a probability $P(\Delta E) = p \times ... \times p = p^n$. Rearranging terms we get $P(\Delta E) = p^{\Delta E/\epsilon}$, which we can then re-parametrize into:

$$P(\Delta E) = e^{-\frac{\Delta E}{k_B T}} \tag{1.4}$$

where k_B is the Boltzmann constant linking thermal energy and temperature, which we will fix at $k_B = 1$.

This relation gives the relative probability of a particle transitioning between two states. In order to get the absolute probability $P(E)$ of an energy state E, we can remember that the probability distribution needs to be normalized ($\sum_E P(E) = 1$), and we can define an absolute reference energy E_R so that $P(E) = P(E - E_R) \propto e^{-E/T}$. The final form of the Boltzmann distribution become :

$$P(E) = \frac{1}{Z} e^{-E/T} \tag{1.5}$$

$$Z = \sum_E e^{-E/T} \tag{1.6}$$

with Z the partition function : the total probability of all possible states of the system.

This is a very powerful tool for our models. Indeed, given a neuronal configuration $\mathbf{s}$ and an energy landscape $E(\mathbf{s})$, we can now compute the probability of this configuration as :

$$P(\mathbf{s}) = \frac{1}{Z} e^{-E(\mathbf{s})/T} \tag{1.7}$$

The key point here is that low-energy configurations are exponentially more probable than high-energy ones.

Monte Carlo sampling. Computing Z is infeasible because Z sums over 2^N configurations (ex. with $N = 50$ neurons, we would already have more than 10^{15} possible configurations). Instead, we use Markov-chain Monte-Carlo (MCMC) methods to generate samples $\mathbf{s}$ that *asymptotically* follow $P(\mathbf{s})$ (Fig. 1.5 B). For example with the Metropolis–Hastings algorithm we:

1. flip a random unit in the current state $\mathbf{s}_0$ to propose a new state $\mathbf{s}_1$;
2. compute $\Delta E = E(\mathbf{s}_1) - E(\mathbf{s}_0)$;
3. accept the move with probability $p = \min\big[1,\, e^{-\Delta E/T}\big]$.

If the proposed state has lower energy ($\Delta E < 0$), it is always accepted, otherwise it is accepted with probability $e^{-\Delta E/T}$, allowing the chain to escape local energy wells.

Stochastic update rule. Unlike deterministic models like Ising or Hopfield, a Boltzmann Machines updates each unit probabilistically. This is done via Gibbs sampling, a MCMC method useful when sampling from the conditional distribution $P(s_i \mid \cdots, s_{i-1}, s_{i+1}, \cdots)$ is simpler than sampling from the marginal distribution $P(\mathbf{s})$.

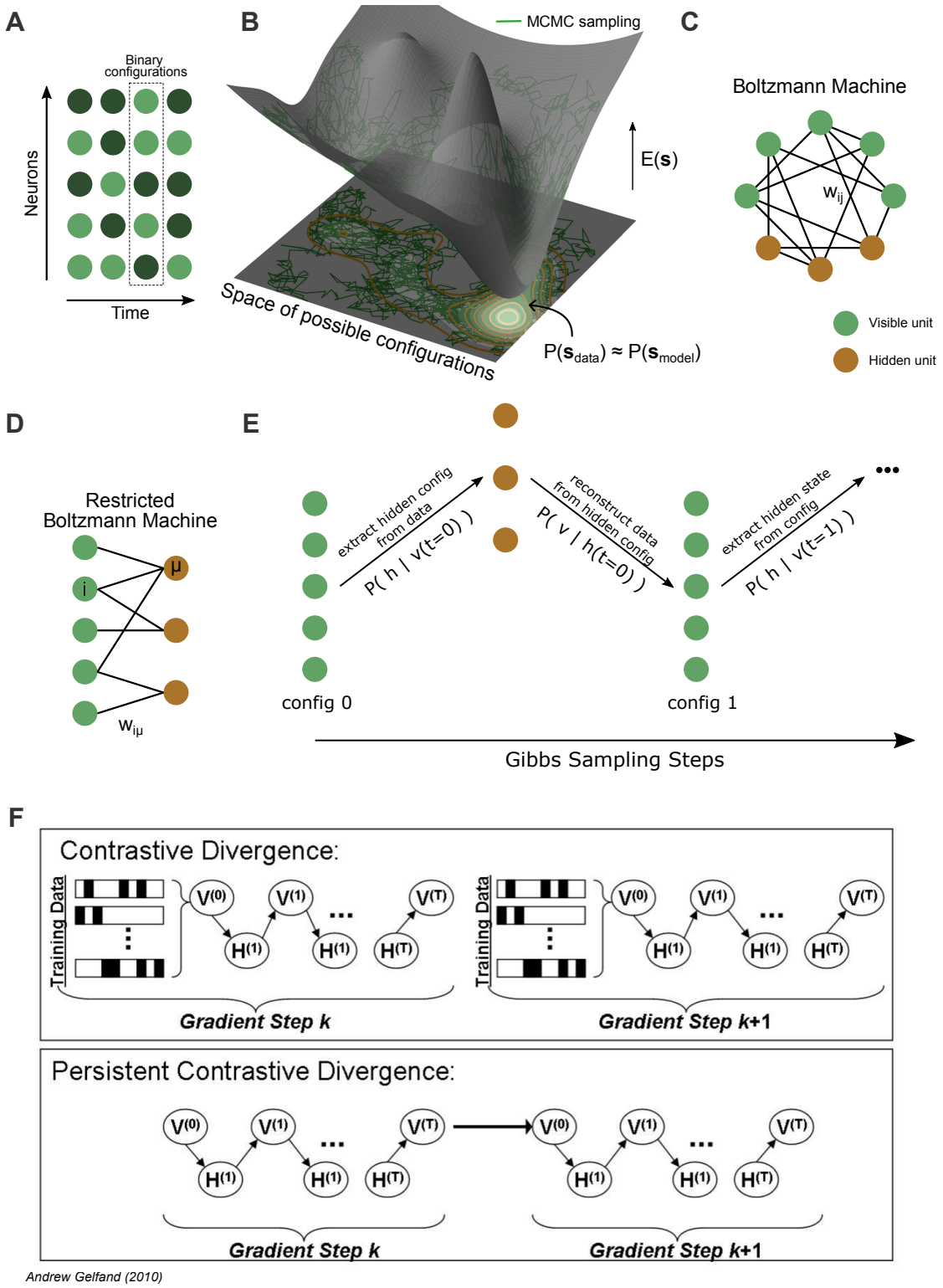
The energy of the system when a unit i is constrained ($s_i = 0$ or $s_j = 1$) but not the others can be written using Eq. 1.3 as:

$$\begin{aligned} E(\cdots, s_i, \cdots) &= -\sum_j w_{ij} s_i s_j - \theta_i s_i - \sum_{k \neq i,j} w_{kj} s_k s_j - \sum_{k \neq i} \theta_k s_k \\ &= E(s_i) + E(\text{others}) \end{aligned} \tag{1.8}$$

where $E(s_i)$ is the energy associated with unit i and $E(\text{others})$ is the energy associated with the rest of the network. Note that in BM units cannot connect to themselves ($w_{ii} = 0$).

As $E(\text{others})$ does not depend on the state of unit i, the probability that this unit is *on* is :

$$\begin{aligned} P_{i=\text{on}} &= \frac{1}{Z} e^{-E(\cdots, s_i=1, \cdots)} \\ &= \frac{e^{-E(\cdots, s_i=1, \cdots)}}{e^{-E(\cdots, s_i=0, \cdots)} + e^{-E(\cdots, s_i=1, \cdots)}} \\ &= \frac{1}{1 + e^{E(s_i=1) - E(s_i=0)}} = \frac{1}{1 + e^{-\sum_j w_{ij} s_j}} \end{aligned} \tag{1.9}$$



Andrew Gelfand (2010)

Figure 1.5: *(Caption on next page)*

Figure 1.5: **Energy-based models of brain activity**. **A**: Diagram of binary data configurations $\mathbf{s} = (s_1, s_2, \ldots, s_N) \in \{0, 1\}^N$. **B**: Illustration of a 2D energy landscape $E(\mathbf{s})$ (surface plot) and it's corresponding probability distribution $P(\mathbf{s})$ (contour plot). When trained, a generative model while match the distribution of data configurations ($P(\mathbf{s}_{\text{data}}) \approx P(\mathbf{s}_{\text{model}})$). Markov Chain Monte Carlo (MCMC) methods are used to sample $P(\mathbf{s}_{\text{model}})$ (green trajectories). **C**: Boltzmann Machines (BMs) are fully connected generative models containing hidden units to capture higher-order interaction between neurons. **D**: Restricted Boltzmann Machines (RBMs) are bipartite generative models with a layer describing neurons (visibles), and another layer captures latent features of the data (hiddens). **E**: Gibbs Sampling is a MCMC method for iteratively sampling an RBM's configurations from its conditional distributions. **F**: Contrastive Divergence (CD) and and Persistent Contrastive Divergence (PCD) are two methods used to obtain model statistics and train RBMs. In CD, at each gradient update step k during the training procedure, multiple Gibbs Sampling chains are initialized from data configurations, and run for a limited T steps. In PCD, a single set of chains is used throughout training, from gradient update to gradient update.

Therefore, the probability that unit i will be *on* is the logistic function of its input $I_i = \sum_j w_{ij} s_j$. Note that in the previous calculation we have omitted the temperature T for readability. However T controls the non-linearity of the logistic function. In the case of $T = 0$ the update rule becomes perfectly deterministic and is equivalent to the Hopfield network. For $T > 0$, the update rule permits the system to climb up the energy landscape, allowing the it to escape local energy minimas.

Training We want the BM to learn the empirical probability distribution $P(\mathbf{s})$ from the data. This means increasing the probability of regularly observed configurations and lowering the probability of configurations which are never observed. In other words, maximizing the likelihood :

$$\langle P(\mathbf{s} \mid w_{ij}, \theta_i) \rangle_{\text{data}} \tag{1.10}$$

With a set of observed configurations $\mathbf{s}_{\text{data}} = [\mathbf{s}_1, \cdots, \mathbf{s}_t, \cdots, \mathbf{s}_T]$

$$\begin{aligned} \log P(\mathbf{s}_{\text{data}}) &= \sum_t \log P(\mathbf{s}_t) \\ &= -\sum_t E(\mathbf{s}_t) - T \log Z \end{aligned} \tag{1.11}$$

Thus, maximizing $P(\mathbf{s}_{\text{data}})$ is equivalent to minimizing the energy $E(\mathbf{s}_t)$ of the observed configurations, and minimizing the partition function Z. As the partition function sums over all possibles configuration of the system, and because most configurations are random and not comparable to the data, this is equivalent to increasing the energy of non-observed states. In other words, we are trying to raise the energy landscape in unexplored regions and lower it in "touristic" regions.

Maximizing the likelihood can be done iteratively by gradient ascent of $\mathcal{L} = \langle \log P(\mathbf{s} \mid w_{ij}, \theta_i) \rangle_{\text{data}}$ with respect to the model parameters w_{ij} and θ_i [136, 137]:

$$\nabla \cdot \mathcal{L} = -\langle \nabla \cdot E(\mathbf{s}) \rangle_{\text{data}} + \langle \nabla \cdot E(\mathbf{s}) \rangle_{\text{model}}$$

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \theta_i} &= \langle s_i \rangle_{\text{data}} - \langle s_i \rangle_{\text{model}} \\ \frac{\partial \mathcal{L}}{\partial w_{ij}} &= \langle s_i s_j \rangle_{\text{data}} - \langle s_i s_j \rangle_{\text{model}} \end{aligned} \tag{1.12}$$

where $\langle \cdot \rangle_{\text{model}}$ is the expectation over the model distribution $P(\mathbf{s} \mid w_{ij}, \theta_i)$.

Therefore, maximizing the likelihood of the data given an BM amount to minimizing difference between data and model statistics, in particular the expected value $\langle s_i \rangle$ and pairwise interactions $\langle s_i s_j \rangle$.

Computing these statistics for the data is straightforward, however, obtaining the model statistics require samples from the current BM. Learning therefore has a *positive phase* (Hebbian: match data statistics) and a *negative phase* (anti-Hebbian: suppress configurations the model favors but the data does not). Contrastive divergence accelerates this process by running only a few Gibbs steps in the negative phase [138]. We will discuss it in details in Sec. 1.5.4.

A note on the Maximum Entropy principle. When constructing a probabilistic model $P(\mathbf{s})$ that replicates a particular data moment f_n, we want the model to make as few assumptions as possible. In other words we want the model to be "as random as possible" beyond the constraints we impose (Jaynes' maximum-entropy principle).

This is done by building a model with maximum informational entropy $H = -\sum_{\mathbf{s}} P(\mathbf{s}) \log P(\mathbf{s})$ [103]. It can be shown that this is the case when $P(\mathbf{s})$ can be expressed as a Boltzmann distribution over an energy function with the form [139]

$$P(\mathbf{s}) = \frac{1}{Z} \exp\left(-\log 2 \sum_n \Lambda_n f_n(\mathbf{s}) \right) \tag{1.13}$$

where Λ_n are Lagrange multipliers.

Boltzmann Machines respect this principle ensuring that the model will learn to reproduce the moments $\langle s_i \rangle$ and $\langle s_i s_j \rangle$ but will otherwise be agnostic.

Hidden units Classical Ising and Hopfield models include only *visible* units (in our case : neurons). BMs augment them with *hidden* units that capture higher-order structure and allow the model to represent multi-modal distributions (Fig. 1.5 C). During training hidden and visible units are updated by the same stochastic rule, but, during evaluation of model statistic, hidden units can be marginalized or sampled to infer latent causes. The next section introduces *Restricted* Boltzmann machines, whose bipartite architecture makes this inference and learning far more efficient.

1.5.4 Restricted Boltzmann Machines

Not so restricted. Boltzmann Machines (BMs) are extremely powerful and versatile generative models. However they suffer from the fact that they are fully connected, in the sense that all units (visible and hidden) are connected to each other. This means that to update the state of the model (and thus for training) one needs to update each unit iteratively, which scales exponentially with the number of units.

This problem was solved in 2002 by Hinton [140] when he introduced the Contrastive Divergence algorithm, a new and faster training paradigm for a variant of the BM called the Restricted Boltzmann Machine (RBM).

1.5.4.1 Definition

A Restricted Boltzmann Machine (RBM) is an extension of the Boltzmann Machine in which the N *visible* variables $\mathbf{v} = (v_1, \dots, v_N)$ are coupled to a set of M *hidden* variables $\mathbf{h} = (h_1, \dots, h_M)$ through a bipartite weight matrix $w_{i\mu}$ (Fig. 1.5 D). The absence of intra-layer connections makes both inference and learning tractable.

Throughout the rest of this introduction we will focus on the case of binary visible units $\mathbf{v} \in \{0,1\}^N$ (representing neurons) and continuous hidden units $\mathbf{h} \in \mathbb{R}^M$.

The joint probability distribution of visible $\mathbf{v}$ and hidden $\mathbf{h}$ configuration is given by :

$$P(\mathbf{v},\mathbf{h}) = \frac{1}{Z}\, e^{-E(\mathbf{v},\mathbf{h})} \qquad Z = \sum_{\mathbf{v}} \int d\mathbf{h}\, e^{-E(\mathbf{v},\mathbf{h})}.$$
$$E(\mathbf{v},\mathbf{h}) = -\sum_{i=1}^{N} g_i v_i + \sum_{\mu=1}^{M} \mathcal{U}_\mu(h_\mu) - \sum_{i,\mu} w_{i\mu} v_i h_\mu \tag{1.14}$$

with g_i the fields and $\mathcal{U}_\mu$ the potentials controlling respectively the activity levels of visible units and the marginal distributions of hidden units.

Hidden Potentials. The choice of potential $\mathcal{U}_\mu$ determines the statistics of hidden units h_μ, and has a strong impact on both the training and the interpretation of hidden units. Many different potential can be chosen, including Bernoulli and quadratic which are canonical in the RBM literature.

Following the work from Jerome Tubiana [11, 141, 142, 137], in this thesis we will consistently use the double-Rectified Linear Unit (dReLU) potential defined as :

$$\mathcal{U}(x) = \frac{1}{2}\gamma_+ x_+^2 + \frac{1}{2}\gamma_- x_-^2 + \theta_+ x_+ + \theta_- x_- \qquad x \in \mathbb{R} \quad \begin{cases} x_+ = max(0,x) \\ x_- = min(0,x) \end{cases} \tag{1.15}$$

Controlled by the parameters $\Theta_\mu = \{\gamma_+, \gamma_-, \theta_+, \theta_-, \}_\mu$, this potential can take many shapes, including standard ReLU ($\gamma_- \to \infty$), quadratic ($\gamma_+ = \gamma_-$, $\theta_+ = \theta_-$), and double-well [142]. We discuss in Sec. 1.5.4.4 the importance of using the dReLU potential in our use case.

From data to features and back. A key aspect of RBM is their ability to probabilistically *translate* a visible configuration to an hidden configuration ($\mathbf{v} \rightarrow \mathbf{h}$), and back ($\mathbf{h} \rightarrow \mathbf{v}$). This essentially permits to move from a empirical description of a configuration in data space, to a abstract representation in a latent space.

This is done by sampling the conditional distributions :

$$\begin{aligned} P(\mathbf{h} \mid \mathbf{v}) &= \prod_{\mu} P(h_\mu \mid \mathbf{v}) \propto \prod_{\mu} e^{-\mathcal{U}_\mu(h_\mu)+h_\mu I_\mu(\mathbf{v})} \\ P(\mathbf{v} \mid \mathbf{h}) &= \prod_{i} P(v_i \mid \mathbf{h}) \propto \prod_{i=1}^{N} e^{v_i(g_i+I_i(\mathbf{h}))} \end{aligned} \tag{1.16}$$

with $I_\mu(\mathbf{v}) = \sum_i w_{i\mu} v_i$ the input received by the hidden unit μ, and $I_i(\mathbf{h}) = \sum_\mu w_{i\mu} h_\mu$ the input received by the visible unit i.

The expected configurations from those conditional distribution are :

$$\begin{aligned} \mathbb{E}[\mathbf{h} \mid \mathbf{v}] &= \frac{\partial \Gamma_\mu}{\partial I}(I_\mu(\mathbf{v})) \\ \mathbb{E}[\mathbf{v} \mid \mathbf{h}] &= \frac{1}{1+e^{-I_i(\mathbf{v})+g_i}} \end{aligned} \tag{1.17}$$

with $\Gamma_\mu(I) = log\left[\int dh e^{-\mathcal{U}_\mu(h)+hI}\right]$ the cumulant generative function associated with he potential $\mathcal{U}_\mu$ (see Tubiana [137] p.51 for the full derivation).

In chapters 3 and 4 will often refer to $\mathbf{h}(t) = \mathbb{E}[\mathbf{h} \mid \mathbf{v}_t]$ as the *hidden activity*, as a way of converting a trajectory from neuronal space to latent space.

Marginal distribution. As before with the Boltzmann Machine, the goal is to model the data distribution $P(\mathbf{v})$, where $\mathbf{v}$ are binarized neuronal configurations.

In the case of RBMs, this can be obtained by marginalizing over the hidden configurations:

$$\begin{aligned} P(\mathbf{v}) &= \int \prod_{\mu} dh_\mu P(\mathbf{v}, \mathbf{h}) = \frac{1}{Z} e^{-E_{eff}(\mathbf{v})} \\ E_{eff}(\mathbf{v}) &= -\sum_i g_i v_i + \sum_\mu \Gamma_\mu(I_\mu(\mathbf{v})) \end{aligned} \tag{1.18}$$

where $E_{eff}(\mathbf{v})$ is an effective energy, sometimes called the *free energy*. We can think of it as the energy that a single visible configuration $\mathbf{v}$ would need in order to have the same probability as all the RBM configurations which contain $\mathbf{v}$ [143] :

$$e^{E_{eff}(\mathbf{v})} = \sum_{\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})} \tag{1.19}$$

This free energy is a very powerful tool as it provides a proxy for comparing the likelihood of different visible configurations without having to compute the partition function Z.

Effective Coupling. In the special case of a quadratic potential $\mathcal{U}_\mu$ (*i.e.*$\gamma_+ = \gamma_-$, $\theta_+ = \theta_-$), the free energy takes the form an Hopfield network with a coupling matrix between visible units $J_{ij} = \sum_\mu \frac{w_{i\mu}w_{j\mu}}{\gamma_\mu}$ [144, 11]. This describes a model of pairwise interaction where the couplings between visible units are mediated by the hidden units.

van der Plas et al. [11] introduced an approximation of the effective coupling between visible units for the general case of non-quadratic $\mathcal{U}_\mu$. The coupling is defined as the impact of the state of neuron j on neuron i in the context of an activity pattern $\mathbf{v}$:

$$J_{ij}(\mathbf{v}) = log\left(\frac{P(v_i = 1|v_j = 1)}{P(v_i = 1|v_j = 0)}\right) - log\left(\frac{P(v_i = 0|v_j = 1)}{P(v_i = 0|v_j = 0)}\right) \tag{1.20}$$

where all units $k \neq i, j$ are unconstrained. The effective coupling is then the average coupling observed across the whole data $J_{ij} = \langle J_{ij}(\mathbf{v})\rangle_{\text{data}}$. The authors show that J_{ij} can be approximated as :

$$J_{ij} \approx \sum_\mu w_{i\mu}w_{j\mu}\langle Var(h_\mu|\mathbf{v})\rangle \tag{1.21}$$

This provides a good intuition of how the RBM captures the interactions between neurons. The hidden units act as assemblies of neurons through which they interact.

1.5.4.2 Gibbs Sampling

As we have seen before in the case of Boltzmann Machines, direct sampling of $P(\mathbf{v})$ is impossible due to the large dimensionality of the system. In the case of RBMs we use Gibbs sampling to sample the joint distribution $P(\mathbf{v}, \mathbf{h})$ via the conditional distributions $P(\mathbf{v} \mid \mathbf{h})$ and $P(\mathbf{h} \mid \mathbf{v})$. This is particularly efficient as there are no connections within each layer, and therefore the hidden units are independent of each other when conditioned on the visible state, and conversely for the visible units.

The Gibbs sampling process is has follows (Fig. 1.5 E) :

1. choose a random starting configuration $\mathbf{v}$.
2. compute hidden inputs $I_\mu = \sum_i w_{ij}v_i$.
3. sample hidden units independently $h_\mu \sim P(h_\mu \mid I_\mu)$.
4. compute visible inputs $I_i = \sum_\mu w_{ij}h_\mu$.
5. sample visible units independently $v_i \sim P(v_i \mid I_i)$.
6. repeat steps 2-5 iteratively.

For a large number of repetitions the sampled configurations $(\mathbf{v}, \mathbf{h})$ will converges towards $P(\mathbf{v}, \mathbf{h})$.

Intuitively, steps 2-3 can be seen a stochastic feature extraction from $\mathbf{v}$, while steps 4-5 can be seen as stochastic reconstruction of $\mathbf{v}$ from features $\mathbf{h}$ [137].

1.5.4.3 Training : Persistent Contrastive Divergence

As with Boltzmann Machines, RBM training is performed by maximizing the likelihood of the data $\mathcal{L} = \langle \log P(\mathbf{v}) \rangle_{\text{data}}$.

Stochastic gradient ascent is then performed with :

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial g_i} &= \langle v_i \rangle_{data} - \langle v_i \rangle_{model} \\ \frac{\partial \mathcal{L}}{\partial \theta_\mu} &= \left\langle \frac{\partial \Gamma_\mu(I_\mu(\mathbf{v}))}{\partial \theta_\mu} \right\rangle_{data} - \left\langle \frac{\partial \Gamma_\mu(I_\mu(\mathbf{v}))}{\partial \theta_\mu} \right\rangle_{model} \\ \frac{\partial \mathcal{L}}{\partial w_{i,\mu}} &= \langle v_i \langle h_\mu | \mathbf{v} \rangle \rangle_{data} - \langle v_i \langle h_\mu | \mathbf{v} \rangle \rangle_{model} \end{aligned} \tag{1.22}$$

with θ_μ the parameters of the potential $\mathcal{U}_\mu$.

Computing the gradient thus requires the comparison of data and model statistics $\langle \cdot \rangle_{data} - \langle \cdot \rangle_{model}$. The RBM then learns to reproduce the statistics : $\langle \mathbf{v} \rangle$, $\langle \mathbf{h} \rangle$, and $\langle \mathbf{vh} \rangle$. As interactions between visible units are captured by common couplings to hidden units, a well trained RBM will also learn to reproduce covariances $\langle \mathbf{vv} \rangle$, and conversely for hidden units $\langle \mathbf{hh} \rangle$.

While computing $\langle \cdot \rangle_{data}$ is relatively easy and can be done at each gradient update (with mini-batch for efficiency), $\langle \cdot \rangle_{model}$ is much harder. Indeed, as $P(\mathbf{v})$ cannot be computed directly we need to use Gibbs sampling to generate *model-configurations*, however this would require an enormous number of samples to converge towards $P(\mathbf{v})$ making it intractable for large N.

To combat this problem, Hinton [140] introduced the Contrastive Divergence (CD) algorithm, where $n \gg 1$ Gibbs chains are initiated from data configurations instead of randomly, and are sampled in parallel only for a few steps (not waiting for convergence) before computing the gradient (Fig. 1.5 F). In practice, it is a good idea to use the same small number of visible configurations $\mathbf{v}_{\text{data}}$ to compute the data statistics $\langle \cdot \rangle_{data}$ and initialize the Gibbs chains, effectively computing a local "contrast" between data and model [137].

Tieleman [145] later introduced the Persistent Contrastive Divergence (PCD) algorithm, where the n Gibbs chains are maintained from one gradient update to the next, again performing only a few Gibbs sampling steps each time (Fig. 1.5 F). The intuition is that, for small enough gradient steps (*i.e.* a small learning rate), the model distribution $P(\mathbf{v}, \mathbf{h})$ changes only slightly at each update. Hence, if the previous Gibbs step lies within the distribution, then after the model update it will probably still be within the distribution. In practice, it is often necessary to decrease the learning rate progressively during training to reduce the chance of divergence.

During the rest of this thesis we will use PCD to train RBMs, however it is important to note that better methods exist, such as parallel tempering, mean-field approximation, and many others [146].

1.5.4.4 A note on compositionality.

An important question when attempting to interpret the *meaning* of the hidden units is there co-activation, *i.e.* how a visible configuration $\mathbf{v}$ is described in the representational space.

Due to the maximum entropy principle, and contrary to other methods like PCA or ICA, the RBM makes no assumption as to the statistical properties of data representations. This has been studied in detail by Tubiana and Monasson [141] and Tubiana, Cocco, and Monasson [142] who found three distinct behaviors :

In the ferromagnetic phase, a visible configuration $\mathbf{v}$ is represented in the hidden layer by a single strongly active unit ($m(\mathbf{v}) \sim 1$, with $m(\mathbf{v})$ the number of activated hidden units). While this phase might be desirable for a classification model where we want every $\mathbf{v}$ to be assigned to a single *label*, it does not provide a good mechanistical model of the functional organization of the brain.

In the spin-glass phase, a visible configuration $\mathbf{v}$ is represented by many moderately active hidden unit ($m(\mathbf{v}) \sim M$), making hidden units very hard to interpret. Indeed the weight matrix will be very dense/complex, and will not reflect the typical organization of visible configurations.

In the compositional phase, a visible configuration $\mathbf{v}$ is represented by a small number of active hidden unit ($1 \ll m(\mathbf{v}) \ll M$). Here, the typical organization of visible configurations can be inferred from the combination of a small subset of weights, yet the hidden units can be combined in a large variety of ways and thus capture rich neuronal activity.

Tubiana and Monasson [141] termed RBMs which lie in this last phase as compositional Restricted Boltzmann Machines (cRBMs), and showed that three conditions where sufficient for the emergence of this phase :

- real and unbounded hidden units with a non linear potential $\mathcal{U}_\mu$ (for example dReLU).
- a sparse weight matrix w_{ij}.
- normalized hidden values : $\langle h_\mu \rangle \sim 0$ and $\mathrm{Var}(h_\mu) \sim 1 \; \forall \mu$.

compositional Restricted Boltzmann Machines were studied extensively for neuronal modeling in van der Plas et al. [11], as we will discuss in the next section.

1.5.4.5 Applications in Neuroscience

RBMs have been used in biology in a large variety of domains, from proteins structure [144, 147, 148, 149, 150, 151, 152] to immunology [153, 154, 155].

In neuroscience, RBMs have recently started to be used to analyze large brain recordings.

Human fMRI. Hjelm et al. [156] used RBMs to identify intrinsic networks in human fMRI. They trained the model from the voxel BOLD activity recorded during an auditory task. Scans had been spatially aligned between 28 subjects and were concatenated to infer a single model capturing the combined brain activity distribution of all subjects. They showed that the features extracted by the hidden units could be interpreted as functional networks, and that the RBM performed at least as well as ICA to separate those components.

Zebrafish whole-brain. van der Plas et al. [11] used cRBMs to analyze the spontaneous whole-brain single-neuron activity of zebrafish larvae. They showed that the model could capture the mean activity and pairwise neuronal correlations between neurons. They found that hidden units represented spatially compact neuronal ensembles which were comparable to known circuits and were active on a longer time-scale than individual neurons. From this they posit that hidden units represent cell assemblies, *i.e.* strongly coupled neuronal populations which tend to co-activate.

Multi-fish RBMs. Building on this last publication, we will present in chapter 3 a novel training method for RBMs which allows to represent the spontaneous whole-brain neuronal activity of multiple zebrafish into the same hidden space.

1.5.5 Makov Models

All the probabilistic models discussed so far ignore temporal structure: they assign energies or probabilities to instantaneous neuronal configurations without an explicit representation of dynamics. In contrast, Markov models introduce time by specifying probabilistic rules for how a system transitions between states. In this section we review discrete-time (finite state) Markov Chains and their extension, Hidden Markov Models (HMMs), introducing concepts and procedures which will be relevant in later chapters.

We focus on *discrete* time because (i) most experimental recordings are sampled at regular intervals (imposing a natural discretization) and (ii) discrete models provide an analytically and computationally tractable framework. Continuous–time generalizations (e.g. Markov jump processes, stochastic differential equations) will not be discussed here.

1.5.5.1 Markov Chains

To illustrate Markov processes, we will use a simple example. Let's imagine we have a camera taking a picture of a PhD student every hour during the final three months before thesis submission. We observe that the student is always doing one of three things : Writing their manuscript (*Wr*), Procrastinating (*Pr*), or Panicking (*Pa*) (see Fig. 1.6 A). An expert scientist

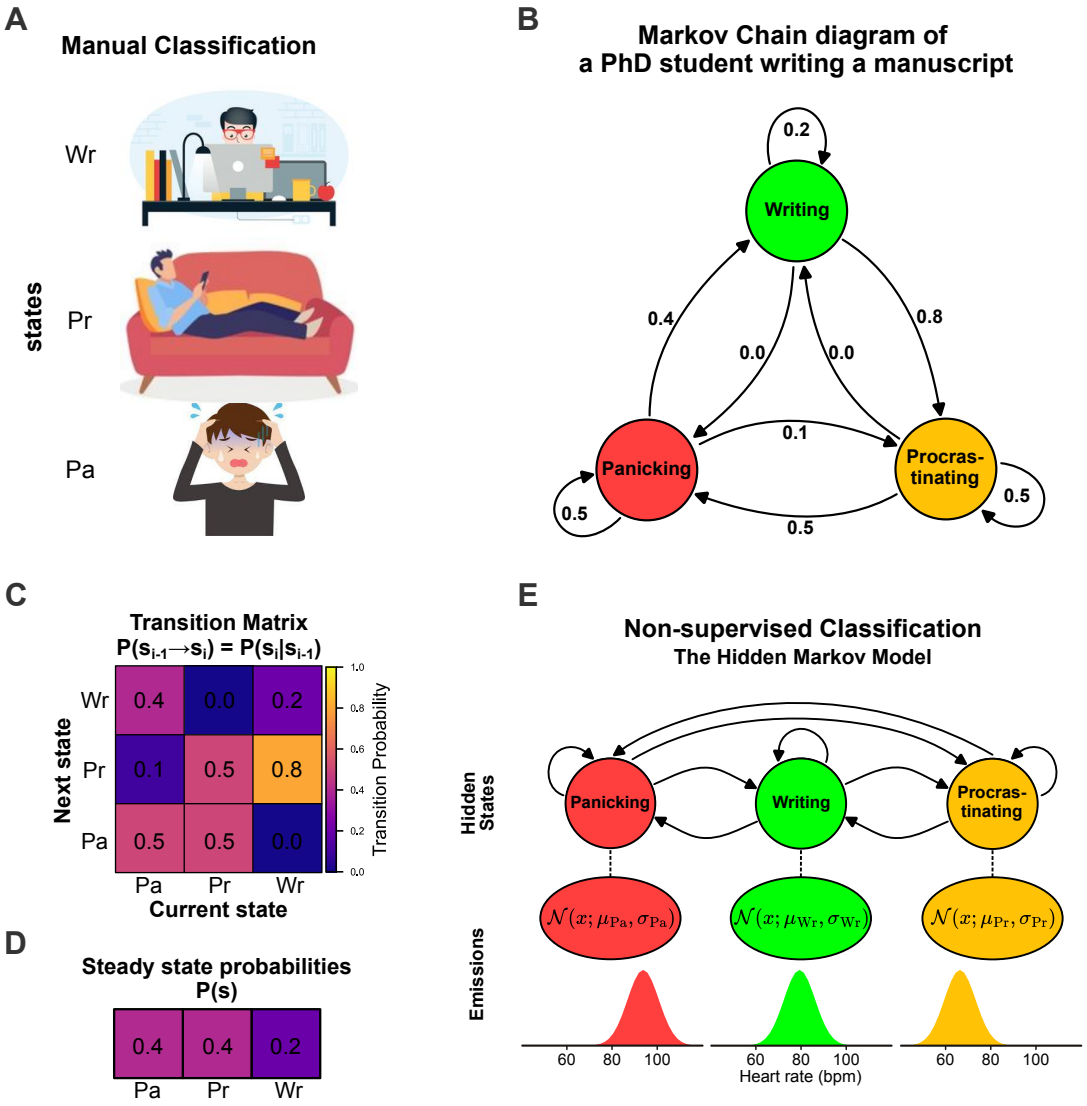


Figure 1.6: **Markov and Hidden Markov Models : an illustration**. **A**: Manual classification of the time spent by a student during the final three months of their PhD. Classified into 3 states : Writing their manuscript (*Wr*), Procrastinating (*Pr*), or Panicking (*Pa*) **B**: Markov Chain diagram of the 3 states described in panel A. **C**: Transition probability matrix of the Markov Chain described in panel B. **D**: Steady state probability extracted from the transition matrix in panel C. **E**: Hidden Markov Model with the same 3 hidden states as panel B, and emission distributions describing average heart rate during each state.

will classify each picture into one of those 3 states, yielding a sequence of observations $S = (\ldots, s_{t-1}, s_t, s_{t+1}, \ldots)$, where $s_t \in \{\text{Wr}, \text{Pr}, \text{Pa}\}\ \forall t$.

We might then ask : *what temporal structure underlies this labeled sequence ?*

Independent model. The simplest hypothesis posits independence across time: for all $t' \prec t$, $P(s_t \mid s_{t'}, t' \prec t) = P(s_t)$. Under this model the ordering of states carries no information; any apparent runs (e.g. prolonged procrastination) are attributed to chance. This independent baseline is essential as a null model when assessing temporal structure.

Transition matrix. Empirically, behavior clearly exhibits memory: what happens next depends (at least) on the current state. For example, Writing might lead to boredom and therefore Procrastination, while Panicking might prompt Writing. The *first–order Markov assumption* asserts that the only relevant history is s_{t-1}:

$$P(s_t \mid s_1, \ldots, s_{t-1}) = P(s_t \mid s_{t-1})$$

In Markov models, this temporal dependence is captured by the transition probability matrix $T_{ss'} = P(s_{t-1} = s \to s_t = s')$ (or written with conditional probabilities $P(s_t = s' \mid s_{t-1} = s)$). In practice, these probabilities can be estimated by counting their occurrences in the observed state sequence S, and normalizing such that $\sum_{s'} T_{ss'} = 1\ \forall s$.
While these models are very simple, they can identify key structures from data. For our student example (see Fig. 1.6 B-C), we find two vicious circles : Pr $\to$ Pr and Pa $\to$ Pa which both have a high transition probability of 0.5, meaning that once the student has entered the Procrastinating or Panicking state, we expect them to stay in this state for a long time, on average. These vicious circles interact via the transition Pr $\to$ Pa, and the student transitions to the Writing state only via Panicking. The Writing state itself is very unstable ($P(\text{Wr} \to \text{Wr}) = 0.2$) and has a high probability of transitioning to the Procrastinating state ($P(\text{Wr} \to \text{Pr}) = 0.8$).

Stationary distribution. A central question concerns the long–run fraction of time spent in each state. This is called the steady state probability or the stationary distribution $P(s)$ (often written π_s in the literature), defined such as :

$$\sum_{s \in \{\text{Wr,Pr,Pa}\}} P(s)P(s \to s') = P(s')$$

This distribution can be calculated via eigenvalue decomposition. However there exists an other method which provides a more intuitive definition of $P(s)$. Indeed, if the matrix $T = P(s_{t-1} \to s_t)$ describes the probability of transitioning between two states in one time step, the matrix power $T^n = \underbrace{T \times T \times \cdots \times T}_{n}$ describes the transition probabilities after n times steps (for example $P(\underbrace{\text{Wr} \to \cdots \to \text{Pa}}_{\text{n transitions}}) = T^n\,[\text{Wr}, \text{Pa}]$). In the limit of $n \to \infty$ where any autocorrelation in the sequence S becomes negligible, T^n converges to a matrix where each column is equal to $P(s)$. In our case, the steady state shows that over long

times scales, the student will spend only 20% of their time writing (see Fig. 1.6 D). Notice that the steady state $P(s)$ is not the same as the diagonal of the transition matrix $P(s \to s)$ which describes the state-persistence.

Sequence evaluation and generation. Given an observed state sequence of length N, the log–likelihood under a first–order chain is :

$$\log \mathcal{L} = \log P(s_1) + \sum_{t=2}^{N} \log P(s_{t-1} \to s_t)$$

From a transition matrix and stationary distribution, one can generate artificial sequences of states. This is done by first sampling the stationary distribution to obtain the initial state of the sequence, and then sampling the transition matrix to obtain sequentially the subsequent states :

$$\left.\begin{array}{l} s_1 \sim P(s) \\ s_2 \sim P(s_1 \to s) \\ \quad \cdots \\ s_t \sim P(s_{t-1} \to s) \end{array}\right\} [s_1, s_2, \ldots, s_{t-1}, s_t]$$

If the Markov assumption and labeling are adequate, simulated and empirical summary statistics (run lengths, auto-correlations, *etc*) should be undistinguishable from empirical sequences.

Higher order Markov Chains. However, real processes may depend on more than the last state. Indeed, we could imagine for example that the longer the student stays in the Procrastinating state, the more anxiety will accumulate, raising the probability of switching to the Panicking state. This can be modeled by a n-order Markov Chains, where instead of conditioning the state s_t only on the previous state s_{t-1}, it is conditioned on the n previous states $[s_{t-n}, \ldots, s_{t-1}]$:

$$P(s_{t-n} \to \cdots \to s_{t-1} \to s_t) = P(s_t \mid s_{t-n}, \ldots, s_{t-1})$$

Practically however, n-order Markov Chains require exponentially more parameters, making their use quite challenging.

1.5.5.2 Hidden Markov Models

A limitation of the basic chain is its reliance on *observed* discrete states. Manual labeling (e.g. *Wr*/*Pr*/*Pa*) can be subjective and error–prone. Often we record continuous or high–dimensional observations from which latent (unobserved) discrete states must be inferred. Hidden Markov Models address this by pairing a Markovian latent state process with an observation (emission) model.

Emissions. Let us imagine that instead of taking a picture of the student, we measure their average heart rate every hour, therefore recording a sequence $X = [\ldots, x_{t-1}, x_t, x_{t+1}, \ldots]$, where $x_t \in [40\text{Hz}, 120\text{Hz}]$. We want to simultaneously parse these observations into the 3 states *Wr*/*Pr*/*Pa* described above, and provide a description of their dynamics using a Markov transition matrix $P(s \to s')$. As these states are not observed directly, they are now called hidden states, and the model is thus called a Hidden Markov Model (HMM, see Fig. 1.6 E).

We start with the assumption that the heart rate x_t of our student will be different depending on the state s_t they are in, and that it can be modeled independently for each state as a normal distribution ($x \sim \mathcal{N}(\mu, \sigma)$). We therefore build so called emission distributions $P(x \mid s)$ as :

$$\begin{cases} P(x \mid \text{Pa}) = \mathcal{N}(\mu_{\text{Pa}}, \sigma_{\text{Pa}}) \\ P(x \mid \text{Wr}) = \mathcal{N}(\mu_{\text{Wr}}, \sigma_{\text{Wr}}) \\ P(x \mid \text{Pr}) = \mathcal{N}(\mu_{\text{Pr}}, \sigma_{\text{Pr}}) \end{cases}$$

Inference : the Baum-Welch Algorithm. Contrary to the Markov Chain where the transition matrix can be inferred by counting the transitions between annotated states, HMMs need to infer the transition matrix $P(s \to s')$ and the emission parameters $(\mu_s, \sigma_s)\ \forall s$ simultaneously. This is commonly done via expectation-maximization using the Baum-Welch algorithm [157].

We define $\theta = (P(s), P(s \to s'), P(x \mid s))$ the parameters of the model. These can be initialized from prior information or randomly, and the goal of the algorithm will be to find a set of parameters θ^* which is local maxima of the likelihood $P(X \mid \theta)$.

We start by defining $\alpha_s(t) = P([x_1, \ldots, x_t], s_t = s \mid \theta)$, the joint probability of observing the sequence $[x_1, \ldots, x_t]$ and being in state s at time t. This is often referred to as the forward probability, and is defined recursively by :

$$\begin{aligned} \alpha_s(1) &= P(s)P(x_1 \mid s) \\ \alpha_s(t+1) &= P(x_{t+1} \mid s) \sum_{s'} \alpha_{s'}(t) P(s' \to s) \end{aligned}$$

We also define $\beta_s(t) = P([x_{t+1}, \ldots, x_T] \mid s_t = s, \theta)$, the probability of observing a sequence $[x_{t+1}, \ldots, x_T]$ starting at time t and lasting until the end of the observed sequence, given an initial state s_t at time t. This is often referred to as the forward probability, and is defined recursively by :

$$\begin{aligned} \beta_s(1) &= 1 \\ \beta_s(t) &= \sum_{s'} \beta_{s'}(t+1) P(s \to s') P(x_{t+1} \mid s') \end{aligned}$$

Notably, the probability of the sequence $[x_1, \ldots, x_t]$ according to the model is $P([x_1, \ldots, x_t] \mid \theta) = \sum_s \alpha_s(t)\beta_s(t)$: the marginalization over all possible hidden states.

To update the model parameters θ, we first compute :

$$\begin{aligned}
\gamma_s(t) &= P(s_t = s \mid [x_{t+1}, \dots, x_T], \theta) \\
&= \frac{\alpha_s(t)\beta_s(t)}{P([x_1, \dots, x_t] \mid \theta)} \\
\xi_{ss'}(t) &= P(s_t = s, s_{t+1} = s' \mid [x_{t+1}, \dots, x_T], \theta) \\
&= \frac{\alpha_s(t)\beta_{s'}(t+1)P(s \to s')P(x_{t+1} \mid s')}{P([x_1, \dots, x_t] \mid \theta)}
\end{aligned}$$

and the new parameters can be updated as :

$$\begin{cases}
P^*(s) = \gamma_s(1) \\
P^*(s \to s') = \dfrac{\sum_t \xi_{ss'}(t)}{\sum_t \gamma_s(t)} \\
P^*(\mathrm{x} \mid s) = \dfrac{\sum_t \mathbb{1}_{x_t = \mathrm{x}} \gamma_s(t)}{\sum_t \gamma_s(t)}
\end{cases}$$

This process is repeated until θ has converged to a local maximum of log-likelihood θ^*.

Decoding : the Viterbi Algorithm. Once the parameters of a model has been inferred from the data, an important feature of HMMs is their ability to decode the observations into a sequence of hidden states. This is done using the Viterbi algorithm, which finds the most probable sequence of hidden states $\hat{S} = \arg\max_S P(S \mid X, \theta)$ which could have generated a sequence of observations X.
To do so, we calculate :

$$\begin{aligned}
P_s(t) &= \begin{cases} P(s)P(x_t \mid s) & \text{if } t = 1 \\ \max_{s'} P_{s'}(t-1)P(s' \to s)P(x_t \mid s) & \text{otherwise} \end{cases} \\
Q_s(t) &= \begin{cases} 0 & \text{if } t = 1 \\ \arg\max_{s'} P_{s'}(t-1)P(s' \to s)P(x_t \mid s) & \text{otherwise} \end{cases}
\end{aligned}$$

where $P_s(t)$ records the probability of the most likely sequence leading to the state s at observation t (out of all possible state sequences leading to s), and $Q_s(t)$ records the state preceding s in the most likely sequence described by $P_s(t)$.
The Viterbi state sequence is then constructed by selecting the state maximizing $P_s(T)$ at the final time T of the sequence, and following $Q_s(t)$ in reverse until $t = 1$.

Markov Models are foundational tools for the modeling of temporal sequences in terms of state sequences. We will use these models in chapters 2 and 4 to study the temporal structure of behavioral and neuronal recordings.

1.6 General question and Outline

This thesis asks a simple but demanding question:

Is Spontaneous Brain Activity (SA) comparable across individuals of the same species, and if so, how ?

This Introduction has argued that SA is structured and informative, that larval zebrafish enable brain-wide recordings at single-cell resolution, and that probabilistic, generative formalisms provide a common language for behavior and neural population dynamics. Building on this, the following three chapters pursue complementary routes from specific circuits to whole-brain organization and dynamics, each advancing the broader comparability question.

In Chap. 2, we restrict our analysis to the reorientation behavior of zebrafish larvae during free-swimming and the activity of the Anterior Rhombencephalic Turning Region (ARTR). If spontaneous behavior is organized into a small set of actions, and if a turning circuit like the ARTR expresses corresponding neural states, then we ask if a single state-space could describe both, providing an immediate bridge between behavior and brain and enabling direct comparison across animals and contexts. In particular, we assess whether larvae reorientation, expressed as a sequence of left/right/forward bouts, and neuronal activity of the ARTR can both be described by 3-states Hidden Markov Models with the same state-space.

In Chap. 3, we turn to the analysis of spontaneous, whole-brain, single-neuron calcium recordings of zebrafish larvae. Comparing such recordings across individuals is challenging as the activity cannot be aligned and there is no one-to-one correspondence between neurons of different brains. We hypothesize that, while single-neuron details vary, the structure of SA is a stable feature of the zebrafish brain, and therefore it should be alignable across individuals at coarse-grained scales. We therefore seek to build a shared latent space that captures conserved co-activation motifs comparable at the population level. We present two methods to do so based on Restricted Boltzmann Machines (RBMs), and show that such models allow us to *translate* neuronal activity from one fish to another.

Finally, in Chap. 3, we study the stereotypy of SA dynamics across zebrafish larvae. Leveraging the shared latent space built in Chap. 3, we ask whether SA explores a conserved repertoire of states with comparable usage and transitions across fish. To do so, we present an RBM-based method to segment the shared latent space into sequences of recurrent brain states, and study their dynamics in a Markovian framework.

Together, the methods and findings presented in this thesis present an operational framework in which to compare SA across individuals.

Chapter 2

Linking Brain and Behavior States in Zebrafish Larvae Locomotion using Hidden Markov Models

How does ongoing brain activity give rise to spontaneous behavior? This chapter tackles the problem in larval zebrafish by analyzing reorientation bouts during free-swimming, as well as the neural activity of the Anterior Rhombencephalic Turning Region (ARTR), a circuit known to be involved in the orientation of locomotion. We seek a unified statistical description that captures both the dynamics of locomotion and of the underlying neural activity, revealing how these organizational principles vary with environment and across individuals.

We use a Hidden Markov Model (HMM) to segment independently the behavior and neuronal activity into three states and model their dynamics in term of Markov chains. Behaviorally, hidden states represent forward, left-turn and right-turn bouts inferred directly from re-orientation angles. Neurally, the same architecture reveals states which map onto left-dominant, right-dominant and balanced ARTR activity.

This approach delivers three main insights. First, HMM labeling uncovers bout-type persistence which were previously underestimated by methods which used thresholds to segment reorientation behavior. This in turn creates a much more markovian description of this behavior. Second, the model is precise enough to identify individual fish from their trajectory, revealing inter-individual phenotypic variability. Third, we show that, by correcting for a temporal rescaling between the behavior and neuronal data, the transition rates between behavioral and neural states are comparable, confirming that left- and right-turn correspond to left- and right-dominant ARTR activity, but also that the forward state might be encoded by balanced ARTR activity. This implies that the ARTR alone could orchestrate all directional choices.

Together, these results show that a concise state-space model can simultaneously explain neuronal and behavioral dynamics, offering a tractable bridge from circuit to locomotion.

2.1 Introduction

2.1.1 On the segmentation of behavior

Why segment behavior? Behavior is expressed as a continuous stream of movements, yet many questions in ethology and neuroscience become tractable only once this stream is parsed into a finite set of recurrent *states* [158]. Segmenting action sequences enables us to (i) compare individuals through a shared symbolic vocabulary, (ii) quantify the temporal structure with which actions unfold, and (iii) link motor patterns to neuronal and genetic mechanisms.

The taxonomy problem: how many states, and how to find them? Choosing a behavioral vocabulary is far from trivial. Because biomechanical constraints merely bound, rather than enumerate, the space of possible actions, the theoretical repertoire is effectively infinite. In practice, however, animals use only a stable, stereotyped subset of that space, often conserved across individuals and even species [159, 19, 160]. Still, choosing the set, and particularly the number, of behavioral classes remains challenging [19].
Classical ethograms rely on expert-defined categories and can be heavily biased by experimental conditions and the research question at hand [158, 161]. This approach typically involves manual annotation of video or threshold-based rules applied to quantitative behavioral parameters [46, 162, 163, 164]. More complex methods like supervised machine-learning can inherits such biases by learning and reproducing expert segmentation.
Unsupervised clustering, by contrast, partitions behaviors without prior labels, thus limiting ad hoc thresholds, but often still requires the user to specify the number of clusters. Fully non-parametric methods can infer that number automatically, at the cost of interpretability when the resulting states lack obvious meaning.
The issue is well illustrated in zebrafish locomotion, where Marques et al. [162] extracted 13 basic swim patterns with a non-parametric clustering algorithm, whereas Johnson et al. [163] recovered 10 exploratory and 8 hunting states using a semi-unsupervised paradigm.

Variability across individuals and environments. Any segmentation must accommodate the fact that behavior is both individual- and context-dependent. Genotype, epigenetics and early development all affect the available repertoire [165], while environmental factors such as temperature [166], illumination [167] and social context can affect dynamics on short time scales. In larval zebrafish, strain differences are already evident during ontogeny [165], epigenetic manipulations of histone H4 acetylation compress or expand an individual's behavioral space [168], and ambient temperature biases the occupancy of exploratory states [166].

Temporal segmentation. Behavioral segmentation is done in terms of actions (the vocabulary), but also in time. Indeed, while classifying types of behavior allows to build a repertoire, discretizing in time allows the study of dynamics.
Larval zebrafish (5–7 dpf) do not swim continuously; instead, they execute stereotyped

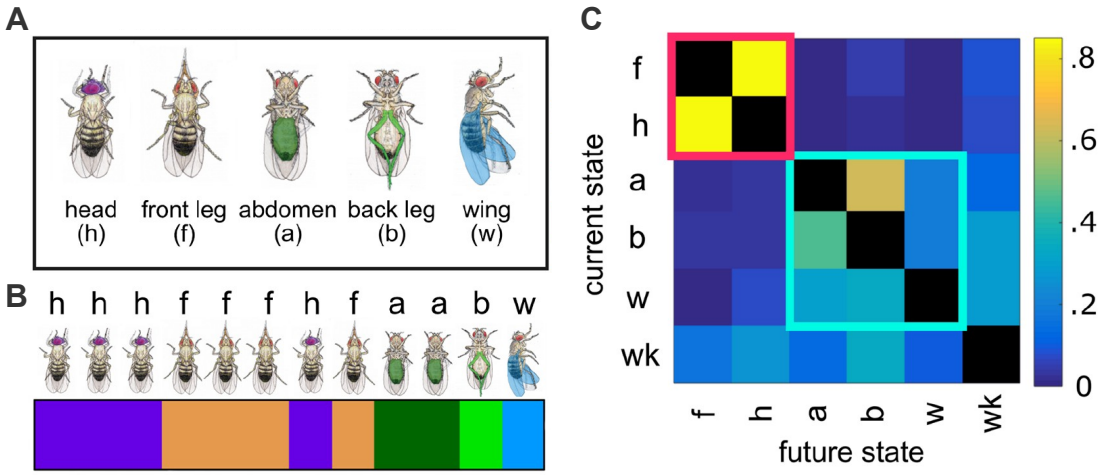


Figure 2.1: **Modeling behavior with Markov Chains : an example from *Drosophila melanogaster*.** **A**: Set of grooming states in *Drosophila melanogaster* after the animal is covered with dust. States are classified by the body-part being groomed (head, front legs, ...). **B**: Example ethogram of grooming events. **C**: Transition probability matrix between states inferred from discretized ethograms (*wk* corresponds to a walking state). Two grooming cycles are identified (in pink and cyan). Adapted from Mueller et al. [17].

bouts lasting $\sim 100ms$, separated by $\sim 1s$ pauses [159, 79]. Because bouts are discrete and short, they provide a built-in temporal segmentation that avoids many ambiguities inherent to continuous locomotion (for example in mammals).

From states to dynamics: discrete-time Markov chains. Once behavior is cast into a sequence of symbols, its temporal structure can be formalized as a discrete-time Markov chains. A great example of this is the study of *Drosophila melanogaster* grooming by Mueller et al. [17] (see Fig 2.1). In this article, researchers classified grooming behavior into 5 action-types, representing the grooming of 5 body-sections : the head, front legs, abdomen, back legs, and wings. They showed that the sequence of events was not random, but could be explained by a Markov Chain containing 2 main loops, grooming of the anterior body (head and front legs), and of the posterior body (abdomen, back legs and wings).

2.1.2 Zebrafish reorientation : Scientific context for the article.

Reorientation angles. Zebrafish navigation has been described as a sequence of swim events (*a.k.a.* bouts) where the animal can both displace and reorient its body. This type of locomotion has previously been characterized as containing two basic manoeuvres : turns that change the heading, and forward (or scoots) the propel the fish forwards [169]. In fact, bout-by-bout analyses shows that reorientation angles are not chosen uniformly but follow a three-modal distribution (Fig 2.2A-B) combining small corrective turns and rarer large-amplitude events [170, 162, 167, 166].

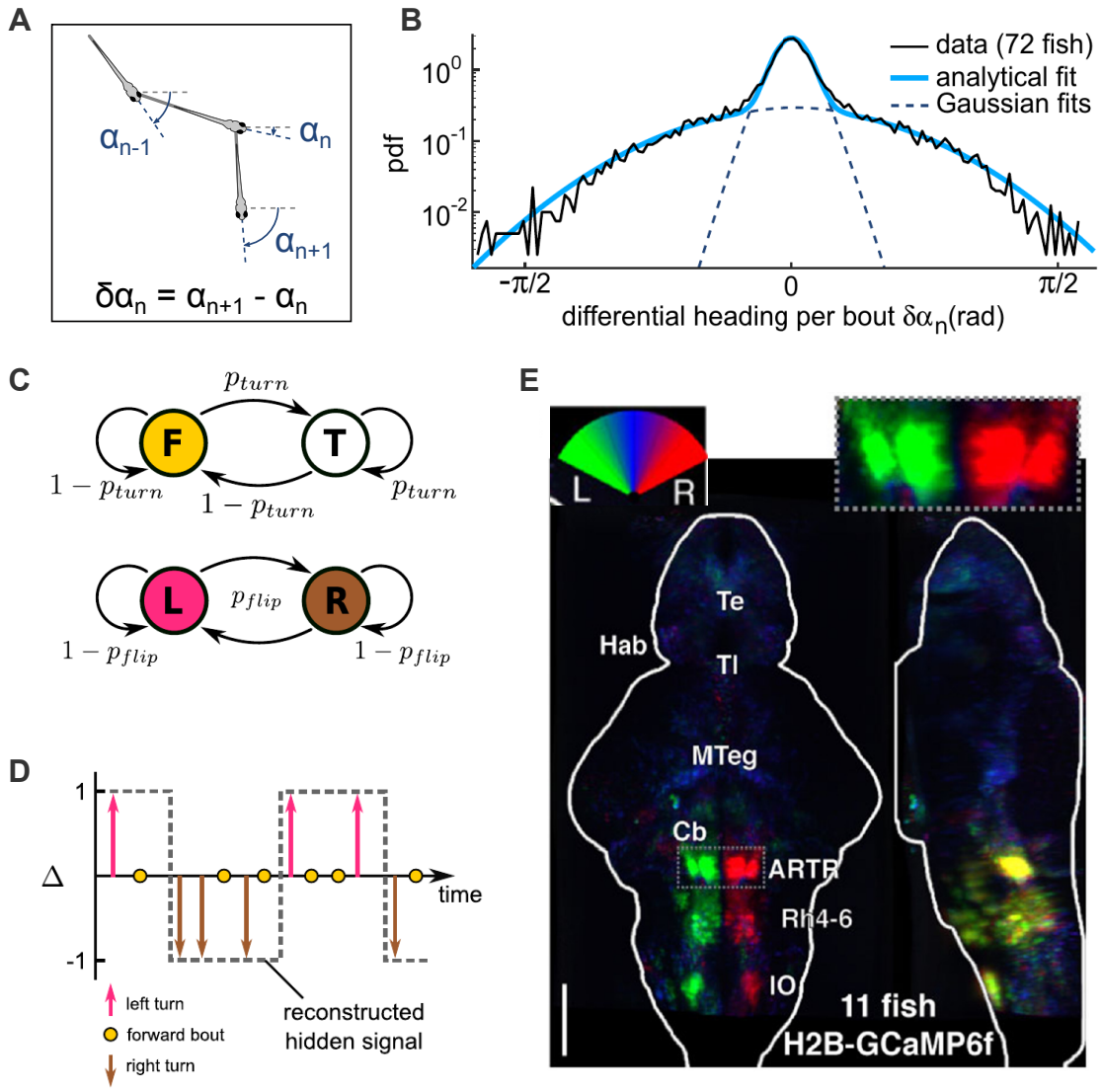


Figure 2.2: **Reorientation during swimming in Zebrafish larvae**. **A**: Definition of reorientation angle $\delta\alpha_n$ between swim bout n and $n+1$. **B**: Experimental (black) and analytical (blue) distribution of reorientation angle $\delta\alpha_n$. **A-B**: Reproduced from Karpenko [167]. **C**: 4-state Markov Chain model of reorientation [167, 166]. Two independent and parallel chains describe reorientation dynamics. The first selects the type of bout (Forward or Turn) with a switching probability p_{turn}, while the second determine if the fish in a Left or Right internal state with a switching probability p_{flip}. **D**: Schematic of a reorientation sequence as described by the model presented on panel C. Arrows and dots represent a sequence of observable turns and forward bouts. Dashed line represents the internal orientational state described by the second chain in panel C. This internal state is only exposed when the fish performs a turn. **C-D**: Reproduced from Le Goc et al. [166]. **E**: Functional stack of neuronal response to spontaneous swimming averaged across 11 fish. Regions active during left and right turns are shown respectively in green and red (see top left inset). The ARTR is predictive of future turn direction (inside dashed box and top right inset). (Reproduced from Dunn et al. [79])

Memory effects. These reorientations are not random but are structured by persistence in bout orientation. Dunn et al. [79] have shown that in the absence of stimuli larvae tend to turn repeatedly in the same direction before switching to the other direction. The authors simulate this behavior with a two-state (left-, right-turn) Markov Model and suggest that the low probability of switching direction favors foraging strategies. Their analysis focuses on turns and doesn't consider forward bouts as distinct events, thus the segmentation of behavior is quite straightforward and relies solely on the sign of the reorientation angle. This marks the first instance in the literature of a Markov Chain being used to model reorientation in zebrafish larvae.

Modeling reorientation with 4-state Markov Chains. Later studies of navigation have also used Markov Chains to model reorientation, with the goal of understanding how reorientation statistics are impacted by visual stimuli and environmental parameters.
Karpenko [167] used a parametric Markov model to understand how reorientation can be biased by stereo-visual stimulation. Here, reorientation is not segmented into left, right, and forward bouts. Instead, the distribution of reorientation angles is fitted by the sum of two normal distributions : one corresponding to forward bouts, and a one corresponding to turns (Fig 2.2B). The relative weights of the 2 distributions are then used to estimate the transition probabilities of the Markov model. Importantly, this model is constructed from two parallel Markov chains : one chain controls whether the animal is turning or going forward, and the other controls an internal directional state exposed when the animal performs a turns (Fig 2.2C-D). This model is parametrized by only two transition probabilities : p_{turn} which controls the switching between the turn and forward states, and p_{flip} which controls the switching between the left and right states.
The same 2-chains model was also used by Le Goc et al. [166] to investigate how bath temperature affects reorientation statistics. In this study the authors segmented the swimming into left/forward/right bouts by thresholding the distribution of reorientation angles (respectively $< -10°$, $-10° < \cdot < 10°$, and $> 10°$).
This 2-chain model effectively creates a 4-state Markov model : turning left, turning right, forward (with internal state left) and forward (with internal state right). It is coherent with a neuronal model of locomotion where one or more circuits trigger bouts as being either forwards or turns, while another circuit maintains a constant internal representation of directionality which affects turns. The best candidate for this second circuit is the Anterior Rhombencephalic Turning Region, as we will now describe in the next section.

ARTR as an internal representation of directionality. Situated in the anterior-most rhombomere of the larval zebrafish hindbrain, the Anterior Rhombencephalic Turning Region (ARTR, also sometimes referred to as the Hindbrain Oscillator or HBO) was first isolated by whole-brain calcium imaging during fictive locomotion (Fig 2.2E). Dunn et al. [79] revealed two bilaterally symmetric clusters whose antiphasic activity predicts whether the next fictive swim bout will be biased to the left or to the right. Each hemisphere inhibits the other, so that only one side is active at a time. The slow dynamics of the active cluster ($\approx 2s$ rise, $5-10s$ decay) mirrors the turn persistence observed behaviorally. Unilateral laser ablation of a subset of neurons in one side of the ARTR biases freely swimming larvae to turn toward the intact side. Collectively these findings establish the ARTR as a compact

pattern-selection hub that biases spontaneous locomotion, and map well onto the hidden left–right variable of the four-state Markov chain introduced earlier.
Beyond spontaneous exploration, the ARTR is linked to visual stimuli that guide reorientation. During phototaxis, unilateral or periodic light changes phase-lock the intrinsic oscillation of the ARTR, and bias the network toward the eye receiving brighter illumination, steering successive bouts toward the source [171]. Similar principles operate during the optomotor response, where binocular motion cues parsed in the Pretectum ultimately modulate hindbrain circuits, including the ARTR, to generate appropriate steering [172].
More recently, data-driven state-space analyses have described the ARTR as a low dimensional energy landscape. An Ising model built from large-scale recordings reproduces its spontaneous and visually driven dynamics, and accounts for temperature-dependent persistence [173]. This energy landscape reproduces the two stable states corresponding to left-dominant ad right-dominant ARTR activity, but interestingly, it also produces a stable state where both the left and right hemispheres have low activity. While this balance state is not accounted for by our current understanding of ARTR function, it hints to a more comprehensive view of this circuits as being more than a left-right modulator.

2.1.3 Using Hidden Markov Models to study reorientation.

In the summer of 2023, together with Leonardo Demarchi, Rémi Monasson, Simona Cocco and Georges Débregeas, I participated in the I-Bio Summer School "Advanced Computational Analysis for Behavioral and Neurophysiological Recordings" as a teaching assistant. The goal was to teach data analysis tools from physics to neuroscience PhD students.
Among those tools, we built a tutorial around Markov Chains and HMMs, and used the data from Le Goc et al. [166] as an illustration (the tutorial can be found here : `https://github.com/EmeEmu/IBIO-Banyuls2023-Python`). In order to make it simpler for the students to program and interpret, we used only 3 behavioral states (forward, left and right) instead of the 4 states described above. To our surprise, while a Markov chain inferred from thresholded data could not replicate the orientational diffusivity observed in the data, a trained HMM could.
This prompted a complete re-analysis of the data under the hypothesis that 3 states could be enough to describe reorientation in a markovian framework. In turn, this raised questions as to the role of the ARTR in this behavior, particularly whether the ARTR could control not only the left-right orientation, but also the forward bouts. This led to the article presented in the following section.

2.2 Article

The following article was written in collaboration with Jorge Fernandez-de-Cossio-Diaz, Monica Coraggioso, Volker Bormuth, Rémi Monasson, Georges Debrégeas, and Simona Cocco. It is available on bioRxiv since November 2024 under the doi : `https://doi.org/10.1101/2024.11.22.624881`, and is under review at PLOS Computational Biology since April 2025.

Linking Brain and Behavior States in Zebrafish Larvae Locomotion using Hidden Markov Models

Mattéo Dommanget-Kott,[1,2,*] Jorge Fernandez-de-Cossio-Diaz,[3,4,*] Monica Coraggioso,[1] Volker Bormuth,[1] Rémi Monasson,[3] Georges Debrégeas,[1,†] and Simona Cocco[3,‡]

[1] *Institut de Biologie Paris-Seine (IBPS), Laboratoire Jean Perrin, Sorbonne Université, CNRS, France*
[2] *Université Paris Cité, France*
[3] *Laboratory of Physics of the Ecole Normale Supérieure, CNRS UMR 8023 PSL Research, Sorbonne Université, Université Paris Cité, France*
[4] *Université Paris-Saclay, CNRS, CEA, Institut de Physique Théorique, 91191, Gif-sur-Yvette, France*

(Dated: August 27, 2025)

Understanding how collective neuronal activity in the brain orchestrates behavior is a central question in integrative neuroscience. Addressing this question requires models that can offer a unified interpretation of multimodal data. In this study, we jointly examine video-recordings of zebrafish larvae freely exploring their environment and calcium imaging of the Anterior Rhombencephalic Turning Region (ARTR) circuit, which is known to control swimming orientation, recorded *in vivo* under tethered conditions. We show that both behavioral and neural data can be accurately modeled using a Hidden Markov Model (HMM) with three hidden states. In the context of behavior, the hidden states correspond to leftward, rightward, and forward swimming. The HMM robustly captures the key statistical features of the swimming motion, including bout-type persistence and its dependence on bath temperature, while also revealing inter-individual phenotypic variability. For neural data, the three states are found to correspond to left- and right-lateral activation of the ARTR circuit, known to govern the selection of left vs. right reorientation, and a balanced state, which likely corresponds to the behavioral forward state. To further unify the two analyses, we exploit the generative nature of the HMM, using neural sequences to generate synthetic trajectories whose statistical properties are similar to the behavioral data. Overall, this work demonstrates how state-space models can be used to link neuronal and behavioral data, providing insights into the mechanisms of self-generated action.

Keywords: zebrafish; Hidden Markov Model; behavior; spontaneous neural activity

I. INTRODUCTION

Animal behavior unfolds as a structured sequence of stereotyped motor actions, much like language. Understanding behavior thus requires identifying the vocabulary, *i.e.* the elementary behavioral units, and characterizing the corresponding grammar, *i.e.* their relative organization in time [1]. Uncovering this underlying structure is non-trivial. Over the last decade, numerous approaches have been proposed, building on the rapid development of data-driven computational methods. State-space models, in particular, appear to be well adapted, as they offer an unsupervised approach to sparse high-dimensional data into discrete states, while simultaneously unveiling their temporal structure. These include various implementations of Hidden Markov Models (HMMs) [2–6] and other statistical models [7–9].

Since behavior is driven by the brain activity, one expects the behavioral structure to be reflected in the spontaneous brain dynamics in the form of a sequence of discrete "brain states" - defined as metastable patterns of activity [10]. Neural activity can, as behavioral data, be parsed to uncover neural states and their temporal sequences [11–14]. In general, however, behavioral or neuronal data are analyzed separately, as these experiments are typically conducted independently, limiting our ability to bridge the two processes. In contrast, a common modeling framework, when applied to both behavior and spontaneous neural activity, could help uncover a shared organizational structure linking self-generated neuronal dynamics and behavior.

Our model behavior is the spontaneous navigation of zebrafish larvae (see [9, 15–17]), which consists of discrete swimming bouts lasting $\sim$ 100 ms and triggered at $\sim 1-2$ Hz. In previous studies the categorization of bouts was carried out independently of the examination of their temporal organization. In Marques *et al.* [18], the authors used PCA-based automatic segmentation to distinguish 13 different bout types, a number that they found sufficient to encompass the entire behavioral repertoire of the animal, including hunting, escape, social behavior, etc. However, in more constrained conditions, when the fish merely explore its environment [19–25], a simple 3-state categorization is sufficient to describe their trajectories. In this case, the bouts are labeled as either forward, left-turn or right-turn based on the value of bout-induced body reorientation. The selection of these various bout types depends on sensory cues, resulting in the animal's capacity to ascend light [19, 22] or temperature [24, 26–28] gradients.

* These two authors contributed equally
† Correspondence: georges.debregeas@sorbonne-universite.fr
‡ Correspondence: simona.cocco@phys.ens.fr

Importantly, the neural circuit that controls the orientation of bouts has been identified as the anterior rhombencephalic turning region (ARTR), a bilaterally distributed circuit located in the anterior hindbrain. Using combined calcium imaging and motor nerve recordings, it was shown that the triggering of leftward and rightward bouts are correlated with increased activity on the corresponding side of the ARTR [20].

To characterize the behavioral and neural activities and their possible relationship, we hereafter re-analyze video recordings of freely swimming animals and ARTR recordings, performed at various water temperature, using Hidden Markov Models (HMM). First, we show that for the behavioral data, this approach provides an unbiased and therefore more consistent method of bout-type labeling compared to simple thresholding techniques as used in earlier studies. We further use the HMM inferred parameters to demonstrate and quantify inter-individual variability in exploratory kinematics. We then apply a 3-states HMM to the ARTR recordings performed in paralyzed tethered fish, leading to the comparison between the behavioral and neural HMMs. Finally we generate synthetic neuronal-based swimming sequences and compare the statistical structure of these synthetic trajectories with real ones to assess the consistency of the results across both behavioral and neural data.

II. RESULTS

A. Data

The behavioral data used in the present article comes from a publication that examined the kinematic of free exploration in zebrafish larvae [24]. The experimental design (Fig.1a) enables recording the trajectories of multiple freely swimming larvae aged 5-7 days at temperatures of 18^oC, 22^oC, 26^oC, 30^oC, and 33^oC. At each temperature, the trajectories of multiple fish are combined into a single dataset, and a set of kinematic parameters is extracted at each bout n, such as the angular change $\delta\theta_n$ in heading direction, the time elapsed since the previous bout and the traveled distance (see Material and Methods sec. IV A). Water temperature was found to systematically impact the statistics of navigation, leading to qualitatively different trajectories as illustrated in Figure 1b. As the temperature increases, trajectories tend to become more winding and erratic. We have also reanalyzed a second dataset of long-trajectories for 18 fish tracked individually for over two hours at 26^oC, in order to assess inter-individual variability (see Material and Methods sec. IV A).

The neural data comes from another publication in which the spontaneous activity of the *Anterior Rhombencephalic Turning Region* (ARTR) [29] (Fig.1e) was recorded from 5-7 days old immobilized larvae expressing the calcium indicator GCaMP6f, using light-sheet functional imaging. Several neural recordings (3-10) for each one of the five temperatures (from 18°C to 33°C (Fig.1b)) were analyzed. The fluorescence signal of each neuron was further deconvolved to estimate an approximate spike train (see Material and Methods sec. IV B).

B. Modeling of behavior

1. Markov Models

The distribution of reorientation angles after each bout, shown in Figure 1d, appears to be trimodal, suggesting a classification of the bouts in 3 types: forward (F), left-turn (L) and right-turn (R). In practice, this categorization is generally carried out by thresholding the distribution of re-orientation angles. Denoting the state of swim bout n by s_n we have:

$$s_n = \begin{cases} R, & \text{if } \delta\theta_n < -\delta\theta_0 \\ F, & \text{if } -\delta\theta_0 < \delta\theta_n < +\delta\theta_0 \\ L, & \text{if } \delta\theta_n > +\delta\theta_0 \end{cases} \quad (1)$$

The use of the same threshold (in absolute value) to detect left and right turns relies on the hypothesis that zebrafish larvae, as a group, have no preferred direction (*a.k.a.*. non-handedness). As the exact value of $\delta\theta_0$ has minimal qualitative impact on the results of the Markov Chains, we adopt the same value $\delta\theta_0 = 10°$ as in [24]; notice that $\delta\theta_0$ is the same across all temperatures to avoid introducing ad hoc, temperature-dependent biases. An example of the classification of states along a swimming trajectory is presented in Figure 2b.

Once the bout types are identified, we define a dynamical model for the trajectories $... \to s_{n-1} \to s_n \to s_{n+1} \to ...$ using a three-state Markov Chain (MC). Informally, the sequence of states (associated with the 3 different bout types) is described by the probabilistic automaton in Figure S3a. In this model, after each bout n, a new state s_{n+1} is drawn randomly, conditioned only on s_n (and not on previous states). The transition probabilities between states, $P(s = s_n \to s' = s_{n+1})$, are estimated by counting the numbers # of occurrences of the transitions $s \to s'$ along the trajectories:

$$P(s \to s') = \frac{\#(s \to s')}{\#(s \to F) + \#(s \to L) + \#(s \to R)} \quad (2)$$

with $s, s' \in \{F, L, R\}$.

The top right eigenvector of the 3×3 transition matrix gives access to the stationary probabilities $P(s)$ of the 3 states. As a self-consistency check, we confirmed that this stationary distribution matches the empirical fraction of threshold-labeled states measured in the same dataset, with a maximum absolute difference < 0.003 for every bout type and temperature.

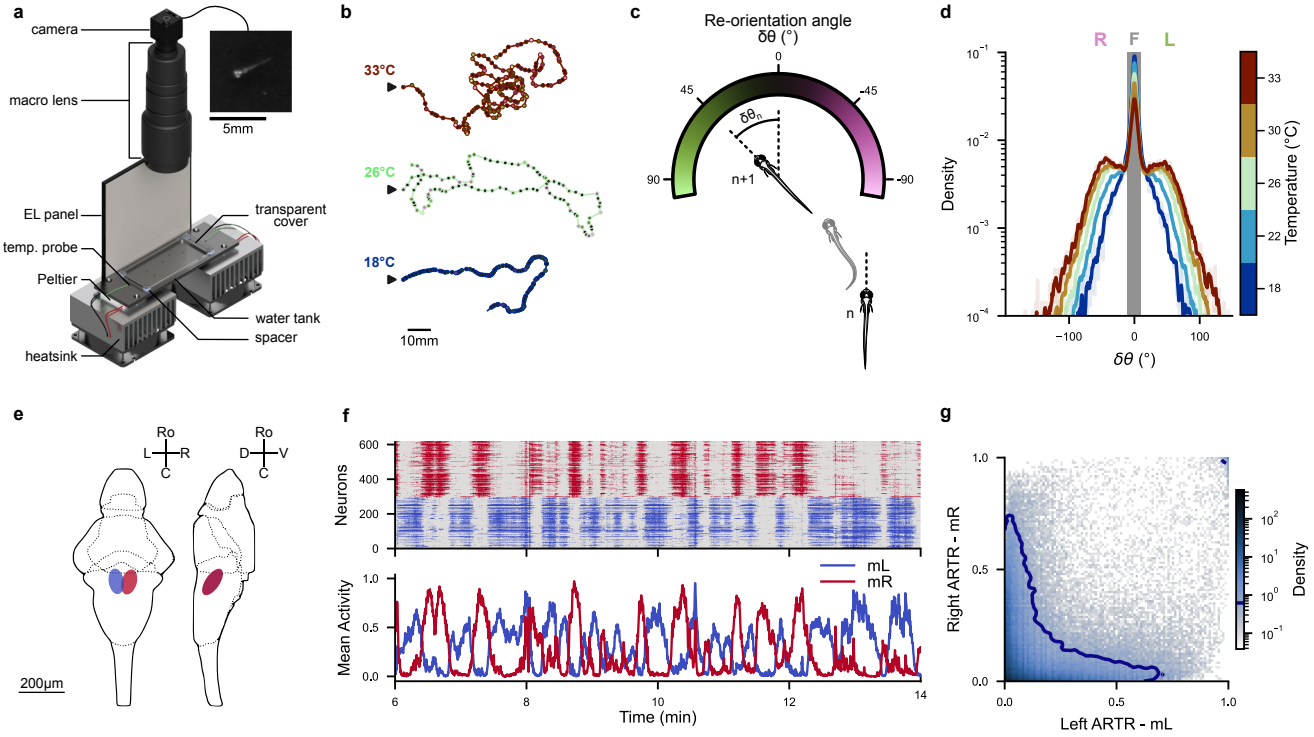


FIG. 1. **Behavioral and Neuronal Datasets**: **(a)** Overview of the experimental setup: Zebrafish larvae are free to move in a tank that is kept at a desired constant temperature by a Peltier module. An imaging system records images of the fish from above at a rate of 25 frames per second. The upper right panel provides a close-up view of a larva in a raw image. Adapted from Le Goc *et al.* [24]. **(b)** Example trajectories of zebrafish larvae in 2D space at various temperatures. Each point represents a swim bout, with the color indicating the corresponding re-orientation angle defined in panel c. The trajectories' starting points are denoted by black arrows. **(c)** Convention used for the reorientation angle ($\delta\theta_n$) between two consecutive swim bouts (n and $n+1$). **(d)** Distribution of re-orientation angles ($\delta\theta_n$) for each ambient temperature. The grayed-out area corresponds to the re-orientation angles classified as forward bouts by a thresholds at $\pm 10°$. **(e)** Diagram of the *Anterior Rhombencephalic Turning Region* (ARTR) in larval zebrafish. Adapted from Wolf *et al.* [29]. **(f)** Example ARTR activity at 22°C. Top : Raster plot of neurons located in the left and right ARTR (blue and red respectively) . Bottom : Mean activity m_L and m_R of neurons in the left and right ARTR. **(g)** Mean activities (m_L, m_R) of the ARTR for all recordings in the dataset. The blue contour line represents 90% of the joint distribution.

2. *Hidden Markov Model*

We then turn to an alternative categorization method, where states are inferred rather than *a priori* assigned. To do so, we consider a three-state Hidden Markov Model (HMM), see Fig.2a. Unlike MC, HMM makes a clear distinction between the observations (here the reorientation angles $\delta\theta_n$ treated as 'symbols') and the states of the system (here s_n, which are not directly accessible from the knowledge of $\delta\theta_n$, in contradistinction with the key assumption underlying MC). The HMM is defined by:

- The transition probabilities $P(s \to s')$ between the hidden states.
- The emission probabilities, $E(\delta\theta|s)$, relate the observations $\delta\theta$ to the hidden states s. For the forward state, we choose normally distributed reorientation angle emission distributions, centered in zero: $E(\delta\theta|F) = \mathcal{N}(\delta\theta; 0, \sigma)$. For turn states, we use Gamma distributed reorientation angles, with a positive or negative sign according to whether the state is Left or Right: $E(\delta\theta|L) = \Gamma(+\delta\theta; \alpha_L, \theta_L)$ and $E(\delta\theta|R) = \Gamma(-\delta\theta; \alpha_R, \theta_R)$, constraining $\theta_L, \theta_R > 0$ and $\alpha_L, \alpha_R > 1$. See Material and Methods sec. IV C for details about the validation of these emission distributions.
- A probability distribution for the initial state at the beginning of a trajectory.

We also consider symmetric HMM, assuming a priori the non-handedness of the model. This is done in practice by enforcing the following constraints over the transition probabilities,

$$
\begin{aligned}
P(F \to L) &= P(F \to R)\ , \\
P(L \to L) &= P(R \to R)\ , \\
P(L \to R) &= P(R \to L)\ , \\
P(L \to F) &= P(R \to F)\ ,
\end{aligned}
$$

and $\theta_L = \theta_R, \alpha_L = \alpha_R$ for the emission distributions. We have checked that relaxing these constraints on the transition matrix leads to equivalent results, see Figures S9 and S10, while maintaining them leads to faster training convergence.

We train both types of HMM models using Expectation Maximization (Baum-Welch algorithm) aggregating all trajectories at a given temperature in the first dataset, and for each individual fish in the long-trajectory data. We employ a customized Julia [30, 31] implementation (available at `https://github.com/ZebrafishHMM2023/ZebrafishHMM2023.jl/tree/bioRxiv`).

Results show that the inferred parameters for the unconstrained and symmetric models are very similar, see Fig. S9 and S10. This is expected as the datasets combine multiple individuals, and left-right asymmetry in bout orientation is very limited. In the following, we therefore employ symmetric HMM only, since imposing symmetry from the beginning results in faster and more robust training of the HMM. This in turn ensures that steady state bout probability is left-right symmetric ($P(L) = P(R)$).

C. State classification and behavioral persistence

1. Statistics of bout states

Since the Markov Chain inferred from thresholded data (MC, Fig.S3a) and the Hidden Markov Model (HMM, Fig.2a) share the same internal behavioral states, we can directly compare these two models and thus examine the impact of the labeling methods.

As illustrated with an example trajectory at 22°C in Figure 2b, MC and HMM labeling can differ significantly. MC-inferred sequences often exhibit multiple alternations between Forwards and Turns when the bouts reorientation angles are near the threshold, while for the same sequence, the HMM tends to consistently label these bouts as Turns. These differences result in a reclassification of approximately 60% of Forward bouts into Turning bouts at 22°C (Fig.S3e).

The HMM yields a relatively modest dependence of bout-type usage on temperature (see Fig.S3b). In contrast, the threshold classification method used in MC lead to a systematic and pronounced increase in the fraction of turning bouts with rising temperature. This strong temperature dependence, previously reported in Le Goc *et al.* [24], may have thus been overestimated, as it partly reflects the ad-hoc assumption of a fixed (temperature-independent) threshold $\delta\theta_0$. Conversely, the HMM approach infers a gradual widening of the forward bouts angular distribution with increasing temperature that corresponds to an increase in the effective angular threshold (see Fig.S2c-e).

2. Bout streaks and persistence

We further assessed how bout-type persistence, defined as the tendency to execute similar bouts in succession, depends on the chosen classification model. We start by describing trajectories as a series of streaks of similar bouts (forward, leftward or rightward), and then characterize the streak length distribution. For all bout types and models, the probability of observing a streak of ℓ consecutive bouts of the same type decays exponentially, $P(\ell) \propto e^{-\ell/\ell_1}$, with ℓ_1 defining the characteristic streak length (Fig.2c). For turning bouts, we found $\ell_1^{\mathrm{HMM}} \approx 1.4$ bouts while $\ell_1^{\mathrm{MC}} \approx 0.9$ bouts at $22°C$. Compared to MC, HMM-based labeling thus yield much longer turning streaks. In contrast, we find no significant difference in characteristic forward-streak length, with MC having slightly longer streaks than HMM. As temperature increases, we observe for both models that the characteristic streak length decreases (particularly for forward bouts, see Fig.1b).

Within the Markov or Hidden Markov Model frameworks, the average length $\ell_1(s)$ of a streak of bouts of type s is related to the probability $P(s \to s)$ of remaining in the same state through the simple relation $\ell_1(s) = -1/\ln P(s \to s)$. To distinguish the effects on bout-type persistence due to the presence of memory from the mere consequences of single-state frequencies, we introduce a null model, in which the transition probabilities are simply given by these frequencies, *i.e.* $P(s \to s') = P(s')$. In this null model without any memory, the average length of type-s bouts is simply $\ell_0(s) = -1/\ln P(s)$. The ratio $\ell_1(s)/\ell_0(s)$ is an estimator of the (relative) contribution of behavioral memory to bout-type persistence.

Results are shown in Figure 2d for the Markov (MC) and Hidden Markov (HMM) Models. The MC and HMM methods yield comparable outcomes for turning bouts at low temperature. However, HMM-based analysis further reveals a persistence for forward bouts at lower temperatures (Fig.2d), while this effect is absent in the MC model. This absence of forward persistence was previously reported in Karpenko *et al.* [22], and we hypothesize that it is due to the mis-labeling associated with the threshold method. Interestingly, such persistence effects vanish at higher temperatures, where the transition matrix becomes uniform (Fig.S3c,d), and all bouts become equiprobable ($P(F) \approx P(L) \approx P(R)$). One thus expect more erratic trajectories at higher temperatures, which is consistent with our observations (see Fig.1b).

3. Consistency of the MC and HMM descriptions of behavior

Taken together, the results above suggest that the Hidden Markov Model better captures persistence in reorien-

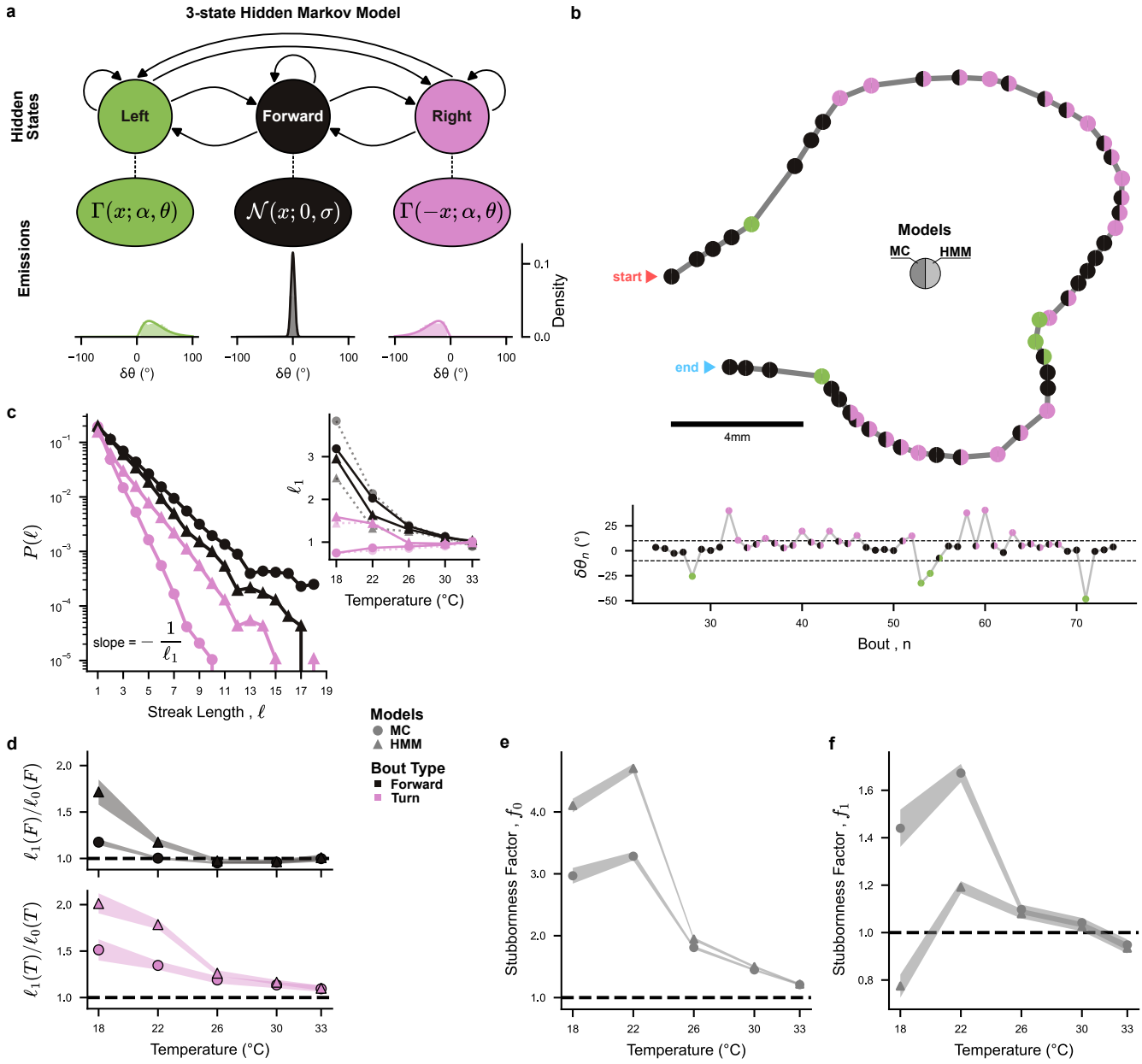


FIG. 2. **3-state Markov Chain and Hidden Markov Models - Stronger persistence emerges from better labeling**: **(a)** Diagram of the 3-state Hidden Markov Model (HMM) with normal emissions for Forward bouts, and gamma emissions for Turning bouts. Example emission distributions where taken at 26°C. **(b)** Example trajectory at 22°C. Each point represents a swim bout, with the left color for the threshold labeling (Markov Chain model), and the right color for the HMM labeling using the Viterbi algorithm. Top: 2D trajectory. Bottom: reorientation angle $\delta\theta_n$ for this trajectory, with the threshold $\delta\theta_0 = \pm 10°$ as a dashed line. **(c)** Probability $P(\ell)$ of observing a streak of ℓ consecutive forward bouts (black) or same-direction turning bouts (pink), for MC (circles) and HMM (triangles), at 22°C. Inset: Exponential decay characteristic length (ℓ_1, solid lines), and theoretical persistence length computed from the transition matrix ($\ell_1(s) = -1/\ln P(s \to s)$, dashed lines). **(d)** Ratio of persistence length ℓ_1/ℓ_0 (observed vs. no-memory null model) vs. temperature, for Forward ($s = F$, black) and turning ($s \in L, R$, pink) bouts. **(e)** *stubbornness* factor at $q = 0$ intermediary Forward bouts, $f_0 = \frac{P(L\to L)+P(R\to R)}{P(L\to R)+P(R\to L)}$. **(f)** *stubbornness* factor at $q = 1$ intermediary Forward bouts, $f_1 = \frac{P(L\to F\to L)+P(R\to F\to R)}{P(L\to F\to R)+P(R\to F\to L)}$. **(e-f)** Shaded bands represent the estimated errors from aggregated fish data (see Materials and Methods IV E).

tation by labeling bouts with small reorientation angles based on context. This leads to a more flexible and thus stable classification than the thresholding method. However, given the absence of a ground truth, it remains unclear whether the labeling produced by the Hidden Markov Models is more accurate than the one produced by the standard threshold-based approaches.

One way to address this question is to examine to what extent each of these methods are self-consistent, i.e. guarantees that the inferred labeled sequences are truly markovian such that the bout type at a given time only depends on the type of the preceding bout. It has been previously noted that the thresholding methods lead to significant non-markovianity. In particular, in a transition $T_1 \rightarrow F \rightarrow T_2$ with $T_1, T_2 \in \{L, R\}$, the two turning bouts tend to have the same orientation ($T_1 = T_2$). This means that the memory of orientation T_1 is maintained during the forward bout, in violation of the Markovian assumption. This observation led to propose a 4-state Markov system comprising two independent Markov chains, independently controlling the forward-turn bout transitions, and directional left-right bout transitions (see Fig.S4b for a diagram of this 4-state model) [22, 24].

Given that our 3-state Hidden Markov Model (HMM) re-labels numerous Forward bouts as Turn bouts, we ask whether this new classification might alleviate this non-Markovianity issue, such that the ad hoc 4-state model might no longer be needed. We thus propose a new test of Markovian violation specifically designed for our use case, that we apply to both the HMM and MC models.

We introduce the *stubbornness factor* f_q to empirically assess the tendency of larvae to retain their orientation after a sequence of q intermediary forward bouts (Fig.S4b, Materials and Methods sec. IV E):

$$f_q = \frac{\sum_{T_1=T_2} P(T_1 \rightarrow F^q \rightarrow T_2)}{\sum_{T_1 \neq T_2} P(T_1 \rightarrow F^q \rightarrow T_2)} \tag{3}$$

with $T_1, T_2 \in \{L, R\}$ and $F^q = \underbrace{F \rightarrow F \rightarrow \cdots \rightarrow F}_{q}$.

Owing to the loss of orientational memory after a forward bout, a non-handed 3-state Markovian model should have $f_q = 1$ for $q \geq 1$ (Materials and Methods sec. IV F). On the other hand, $f_{q=0}$ is a measurement of directional persistence during uninterrupted sequences of turning bouts.

We found that most of the memory effects captured by the HMM occur at $q = 0$, and that the *stubbornness* reaches $f_q \approx 1$ for $q \geq 1$, suggesting that the HMM-inferred bout sequences are quasi-Markovian. In comparison, and for lower temperatures, the thresholded MC classification displays lower persistence at $q = 0$ but higher *stubbornness* at $q = 1$ as seen on Figure 2e-f (and less significantly at $q = 2$, see Fig.S4d). This suggests that the thresholded labeling leads to Markov violation primarily due to the mislabeling of turn bouts as forward bouts during turning streaks, as anticipated in the previous section and illustrated on Figure 2b. As this *stubbornness* is mostly significant at $q = 1$, we expect that most mislabelings are one-off errors.

In summary, previous works using a ad hoc threshold to classify bouts had dismissed 3-state Markov models because the resulting sequences were non-markovian. We found that by using an unsupervised method to simultaneously label the data and infer a Markov Model, we could unveiled previously underestimated memory effects in zebrafish reorientation statistics. These results suggest that the HMM labeling is more markovian, making it a more coherent model than MC. This is probably due to a reclassification of forward bouts as turns during sequences of small reorientations.

D. Behavioral phenotyping from long individual fish trajectories

As HMM provides an unbiased quantification of the behavior, we now ask whether the approach is accurate enough to detect behavioral differences between specimen (inter-individual variability) and whether it can enable the unambiguous identification of each animal.

In the preceding sections, the dataset used to infer the models comprised trajectories from multiple fish, as the different individuals swimming together during a given assay could not be distinguished. To address the question of individuality, we used additional experiments reported in Le Goc *et al.* [24], in which individual fish were tracked at 26°C (see Materials and Methods IV A). A total of 18 fish were recorded for over 2 hours.

We first split the 2h-long recorded sequence of each individual fish into smaller periods (chunks) of $\approx$ 12 minutes each, and trained an HMM on each of these chunks (see diagram in Figure 3a-b). For each fish, the parameters of these HMMs exhibit significant variability (as shown by the vertical error bars in Figure 3c). This variability between the different chunks reflects both intra-individual (temporal) variability and, to a lesser extent, inference uncertainty due to the limited sampling of the HMM (see Fig.S5). We then also trained a single HMM on the entire dataset of a single fish (the "global" HMM). Figure 3c compares selected parameters of the global HMM for each fish, against the average parameters over several HMMs trained on the chunk trajectories (see Fig.S5 for all parameters). There is a clear trend between the global HMM and the average behavior of the chunk HMMs. Therefore, although a fish exhibits variability during a long sequence of bouts, the variability between distinct fish is larger.

These results suggest that the HMM models can be used to distinguish different fish from observations of their bout sequences. To test this hypothesis, we split the trajectories of each fish into a training and a withheld test set. After training the HMM on the train set for a particular fish, we computed the likelihood of all fish trajectories in the test set, and compared them. For 14 out

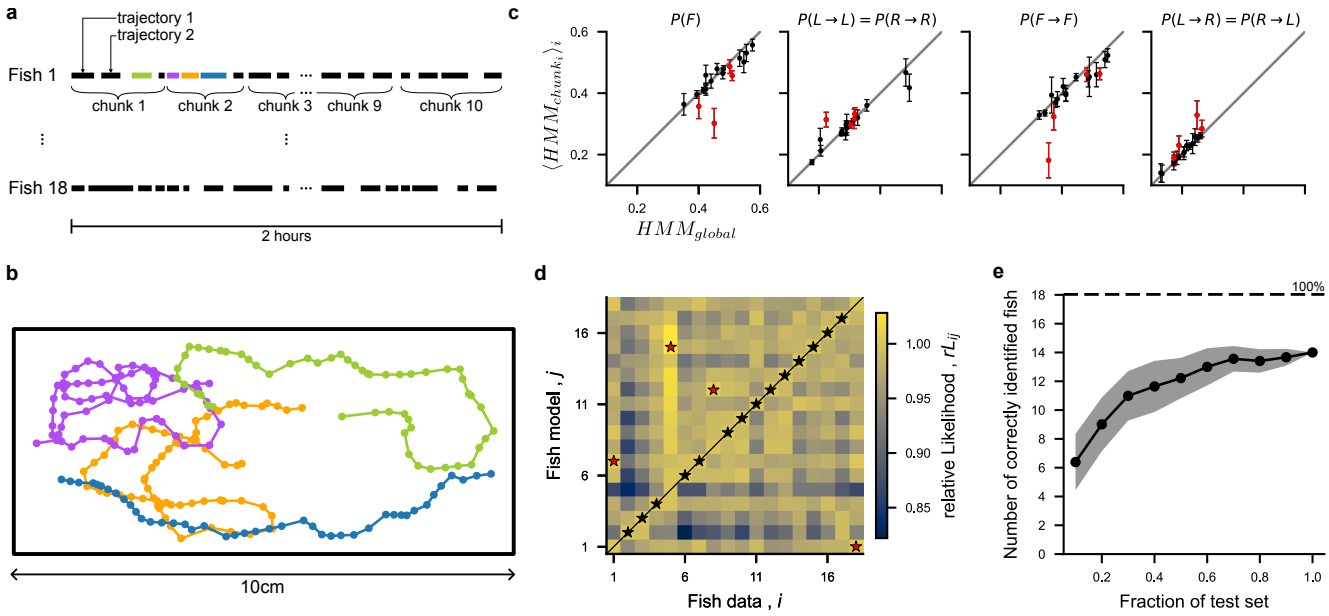


FIG. 3. **Fish identification from long trajectories**: **(a)** Dataset : 18 fish recorded individually for 2-hour sessions. Each session is split into 10 chunks (mean = 9.5 ± 0.5 trajectories per chunk) . **(b)** Example trajectories for fish 1. **(c)** HMM parameters inferred from all the trajectories of a fish (referred to as global) vs. inferred per-chunk. Only four HMM parameters are shown for clarity (see Fig.S5 for all parameters). Each dot represents a fish, and the error bars correspond to the standard error of the mean. Points labeled in red correspond to fish misidentified in panel d. **(d)** Confusion matrix of the relative likelihood $rL_{i,j} = \frac{L(data_i|model_j)}{L(data_i|model_i)}$ of data coming from fish i and HMM trained on fish j. The fish identity most likely according to each model is indicated with a star (black : correctly identified fish, red : misidentified). **(e)** Average number of correctly identified fish when a fraction f of the test data is used for identification. Shaded band : standard deviation across 100 trials. In each trial, the trajectories of each fish were randomly split into train and test sets (50%).

of the 18 fish, the test set that yield the maximum likelihood rightly identifies the fish used to train the HMM (Fig.3d). This finding suggests that the HMM captures behavioral parameters which are distinctive enough to discriminate between different fish. Given the large variability exhibited by a single fish, one expects this discriminative ability to increase with the duration of the training sequences. To quantify this, we further evaluated the likelihoods of subsets of the test fish trajectories, and recorded the number of times that the maximum likelihood HMM corresponded to the correct fish (Figure 3e). Even when withholding 80% of the sequence, we were able to correctly identify 10 out of the 18 fish. These results suggest that individual fish exhibit variable but distinctive behavior which can be captured by the HMM.

E. Modeling of neural data

The selection of turning bouts orientation in zebrafish is known to be controlled by a small bilaterally distributed circuit in the anterior hindbrain, called *Anterior Rhombencephalic Turning Region* (ARTR). This circuit displays self-sustained alternating activity between its left- and right-lateral sub-population, with a period of the order of tens of seconds (Fig.1e). The animal tends to execute left turns when the left ARTR is active while the right ARTR is inactive (and vice versa for right turns) [20].

In contrast, no specific circuit has yet been identified for the selection of turn vs forward bouts. The hypothesis that two distinct circuits are involved in bout-type selection is consistent with the 4 states Markovian model of navigation, in which two independent Markov chains drive the two selection processes. However, the 3-states Markovian model supported by the HMM analysis suggests that the same circuit (ARTR) could drive the selection of all 3 bout-types.

In order to test this hypothesis, we re-analyzed the ARTR recordings reported in Wolf *et al.* [29] using a 3-state HMM (Fig.4a). We posit an independent neural model for the activity of the N recorded neurons, yielding, for each state, the emission probability:

$$P(\sigma_1, \sigma_2, \dots \sigma_N|s) = \prod_{i=1}^{N} \frac{e^{h_i^s \sigma_i}}{(1 + e^{h_i^s})} \qquad (4)$$

where $(\sigma_1, \sigma_2, \dots \sigma_N)$ is a neuronal configuration, s is the hidden state, and h_i^s is the local field representing the effective excitability of neuron i in state s. The model thus includes $3 \times N$ parameters h_i^s, associated to each neuron and each hidden states. Notice that for

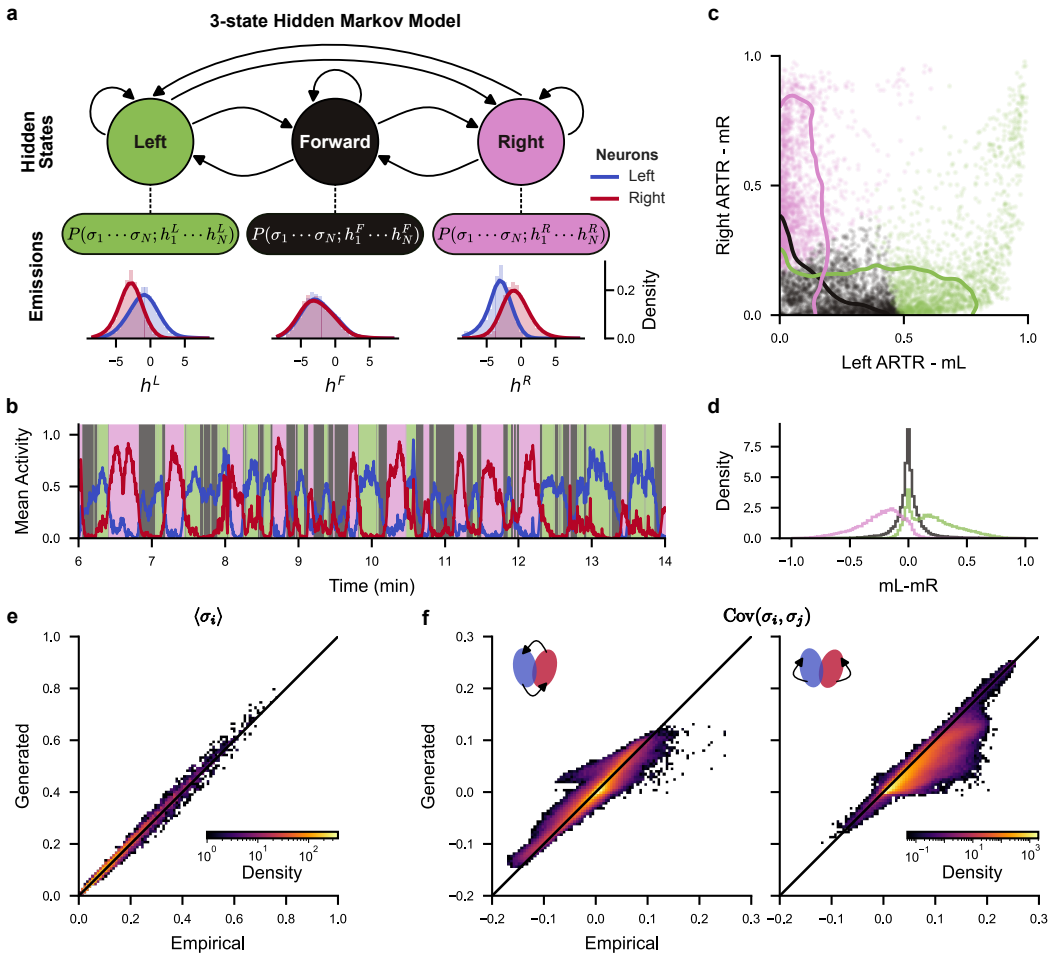


FIG. 4. **3-state Hidden Markov Model (HMM) describes ARTR neuronal statistics**: **(a)** Diagram of the 3-state Hidden Markov Model (HMM) with emissions described as independent models of the ARTR neuronal population, see Eq. (4). Distributions of fields h_i^s are shown for all fish for neurons in the left (blue) and right (red) ARTR. **(b)** Example ARTR activity (see Fig. 1f) classified by the 3-state HMM. Solid lines represent the mean activity of neurons in the left (m_L, blue) and right (m_R, red) ARTR. **(c)** HMM classification in (m_L, m_R) space. Dots : neuronal configurations from the example recording in panel b. Solid lines : 90% of the distributions for all recordings combined. **(d)** Distributions of $m_L - m_R$ per state (all recordings combined). **(e-f)** Empirical vs. HMM-generated neuronal statistics (all recordings combined). **(e)** Mean activity $\langle \sigma_i \rangle$ of neuron i. **(f)** Covariance $\mathrm{Cov}(\sigma_i, \sigma_j)$ of neurons i and j on opposite sides (left plot) and on the same sides (right plot) of the ARTR.

the neural HMM, the non-handedness of the behavioral HMM is not enforced. We also performed comparisons of neural HMM with different numbers of hidden states in a cross-validation test, and found that including more than 3 hidden states results in marginal improvement, see Fig. S8.

The distribution of fields h_i^s for the 3 hidden states, shown in Fig.4a, are used to assign labels to the three states (see Materials and Methods IV G). Consistent with our current understanding of the ARTR function for turn selection, the state with large values of the fields on the left and smaller values of the fields on the contralateral side is labeled "left" (and vice versa for "right"). The third state exhibits similar distributions of fields for neurons on the left and right side of the ARTR, and is labeled forward in analogy with behavior. The ARTR activity is thus modeled as a sequence of left-right-forward states.

With this classification, the forward state corresponds to a low mean neuronal activity of both the left (m_L) and right (m_R) sides of the ARTR, while turning states are associated with large activity on the ipsilateral side of the ARTR (left state : $m_L > m_R$, right state : $m_L < m_R$, see Fig.4b-d).

This model accurately captures the mean activity of each neuron (Fig.4e), as well as the pairwise correlations between contralateral neurons. However, ipsilateral pairwise correlations are not as well reproduced, showing lower covariance in the generated data (Fig.4f). This mismatch presumably comes from the fact that the

activities of neurons within a state are uncorrelated in our emission probabilities, while recurrent interactions in the ARTR circuit produce correlations. These would be better modeled with emission probabilities including effective interactions between neurons [29].

F. Comparison of Behavior and Neuronal HMMs

In the preceding sections, we demonstrated that both the reorientation behavior and the neuronal activity of the *Anterior Rhombencephalic Turning Region* (ARTR) can be effectively modeled using three-state Hidden Markov Models (HMMs). However, it remains unclear whether the three states identified in the Behavioral HMM (B-HMM) directly correspond to those inferred in the Neuronal HMM (N-HMM).

Unfortunately, there is currently no publicly available dataset offering simultaneous recordings of freely swimming larvae kinematics and neuronal activity, which would enable direct comparison of B-HMM and N-HMM states for individual bouts. Current research addressing this question largely relies on experimental paradigms where larvae are either paralyzed with electrophysiological recording of motor nerve signals (fictive swimming preparations)[20, 32, 33], or head-embedded with a free-moving tail (head tethered preparations)[34–37]. In fictively swimming preparations, whilst the classification of left-vs-right bouts is feasible based on the asymmetric nature of the motor command, such experiments are not, to the best of our knowledge, capable of discriminating forward-vs-turning bouts [20]. On the other hand, head tethered preparations allow forward-left-right bout classification [34, 36], but typically rely on visual stimuli to elicit behavior [34–37] as the spontaneous sequence of bouts is strongly disrupted in comparison with freely swimming contexts [38]. The disruption of behavioral and neuronal dynamics caused by animal immobilization is a general problem in behavioral neuroscience, as illustrated by studies in C. *elegans* where the direct comparison of neuronal dynamics between freely-moving and immobilized worms was quantified [39].

We hereafter propose to circumvent these experimental challenges by comparing the statistical structures of the reorientation sequences inferred from the two datasets presented in sections II A and II E. The transition probabilities $P(s_n \to s_{n+1})$ obtained from B-HMM and N-HMM at all recorded temperatures are shown in Fig.5b. Comparison of these transition rates require to first correct them for differences in sampling rates. Indeed, neural transition rates are computed from neuronal recordings performed at $\sim$ 6Hz (depending on the dataset, see Materials and Methods IV B), while for behavior, the sequences are divided into swim bouts triggered at an average rate of $\sim$ 1Hz, depending on the temperature.

To bridge the gap between neuronal and behavioral datasets, one needs to estimate how the behavior is subsampled from the neuronal activity. Previous observations have shown that internal dynamics, and precisely ARTR pseudo-period, tends to be slower in head-tethered assays [29], as well as experiments comparing free and fictive swimming which showed that the fictive swimming frequency is reduced in paralyzed animals [20]. To verify this, we computed the distribution of sojourn times Δt_s of all three states in both B-HMM and N-HMM, where $\Delta t_s = t_k - t_1$ is the duration of a sequence $(s_1, ..., s_k)$ of k consecutive states s observed at times $(t_1, ..., t_k)$. We found the neuronal sojourn times to be significantly longer than the behavioral sojourn times (Fig.5a). The optimal temporal scaling factor $f_{N/B}$ for which the distribution of neuronal sojourn times matches the distribution of behavioral sojourn times (see Materials and Methods IV H) was $f_{N/B} \approx 0.44$. Interestingly, this value appears to be consistent with findings from Dunn *et al.* [20], which reported the mean interbout interval for fictive swimming to be 0.41 times slower than for freely swimming.

Using this temporal re-scaling factor, we find that the transition probabilities $P(s_n \to s_{n+1})$ for behavior and ARTR models are similar (RMSE = 0.1, see Fig.5b), indicating that the behavioral and neuronal state sequences share similar underlying structures. This is remarkable as the number and meaning of the neuronal internal states were not *a priori* fixed, but entirely assigned by N-HMM after training.

This result supports our hypothesis that the ARTR not only governs the selection between rightward and leftward turning bouts, but also controls the bout-type selection, forward *vs* turn. To test this claim further, we analyzed in more detail the statistics of trajectories in the bout space inferred from the ARTR dynamics and from behavioral data. We specifically examined the bout sequences leading to a change in orientation, such as transitions from L to R and vice-versa. Such orientational switches can be either direct, e.g. $L \to R$, or may include an intermediate forward bout, $L \to F \to R$ (Fig.5c). Using the ARTR signal, we found that the second path is strongly favored as evidenced by the the fact that $\frac{P(L\to R)}{P(L\to F)} << 1$. A comparable value of this ratio is observed in the behavioral data (Fig.5d), indicating that fish indeed tend to execute a forward bout when changing orientation. This statistical bias would be difficult to understand under the standard model that posit the existence of independent neural circuits governing orientation and bout-type selection, respectively. In contrast, in our model, it emerges naturally from the the phase space structure of the ARTR dynamics as shown in Figure 4c and Figure 5c. The L-Shaped distribution of $\{m_l, m_R\}$ constrains the Left-to-right (or Right-to-Left) trajectories to pass through a symmetrical, low activity state, thus favoring intermediate forward bouts.

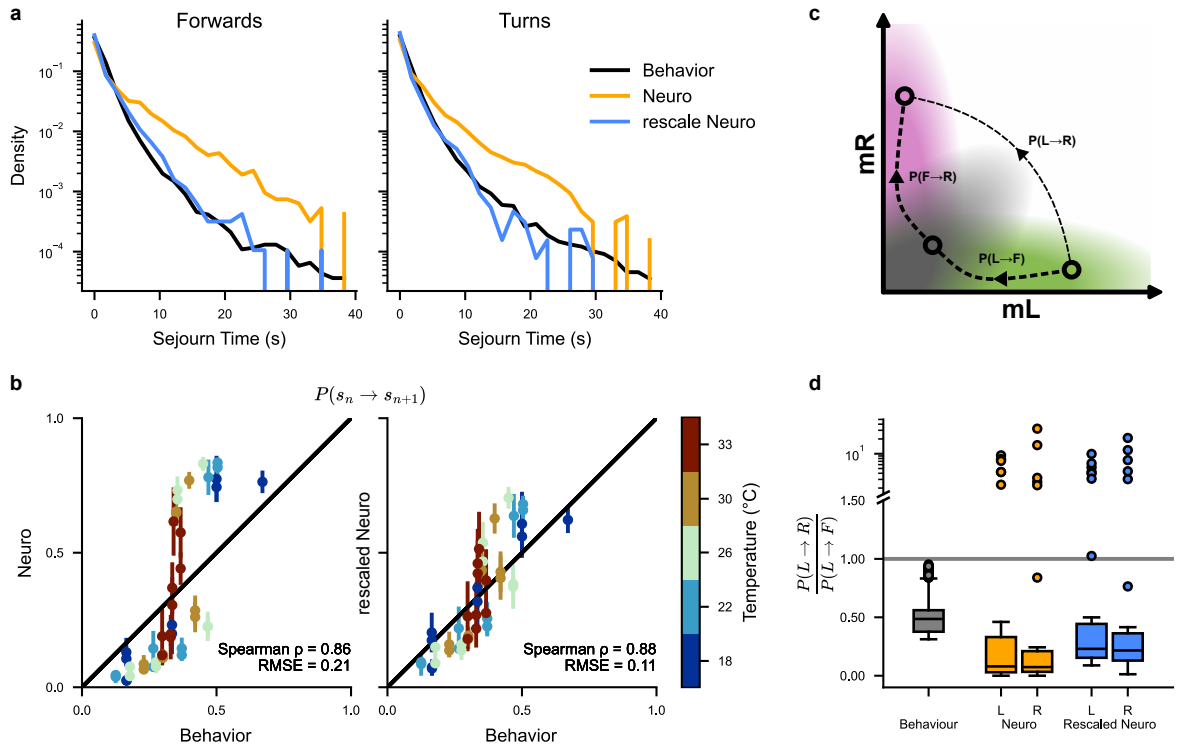


FIG. 5. **Behavior vs. Neuronal temporal structure**: **(a)** Sojourn-time distribution for forward (left) and turn states (right) : behavior (black), neuronal before (orange) and after (magenta) temporal re-scaling. A single re-scaling factor is used for forward and turning states, for all temperatures, and recordings. **(b)** Behavior vs. neuronal state-transition probabilities $P(s_n \to s_{n+1})$ (for all state pairs and all temperatures), before (left) and after (right) temporal re-scaling. Each dot represents a single transition probability at a given temperature. For neuronal state-transition, we show the mean and standard error over all recordings. **(c)** Diagram showing two possible transition trajectories between left and right states in ARTR (m_L, m_R) space. Transitions through the forward state are more probable (see panel d). **(d)** Distributions of $\frac{P(L\to R)}{P(L\to F)}$ for behavior (black) and neuronal data before (orange) and after (magenta) temporal re-scaling (all temperatures combined). These distributions are shown as standard box plots (median, quartiles, and outliers beyond 1.5× the inter-quartile range from the median).

G. Generation of synthetic behavior with the neural model

Until now, we compared neuronal and behavioral data by examining only the short-scale statistical structures of the HMM-inferred state sequences. We now wish to test whether it is possible to compare full trajectories by leveraging the generative nature of the HMM. Specifically, we use the N-HMM model to generate long synthetic trajectories and compare their statistics with those of freely swimming fish. This approach allows us to assess whether the N-HMM, when combined with appropriate scaling and behavioral parameters, can reproduce the complex statistical properties of exploration at various temperatures.

1. *Generation of synthetic neural and reorientation trajectories*

As stochastic processes, Hidden Markov Models (HMMs) can be sampled to generate new sequences of internal states. Following the previous section II F, we hypothesize that the internal states of a Neural HMM (N-HMM) match the behavioral internal states, after proper temporal rescaling. Therefore, we expect that it should be possible to generate artificial swim trajectories from the N-HMMs.

Using the N-HMMs associated to individual fish recordings, we started by generating synthetic temporal sequences of neural states $s_n^N \in \{F, L, R\}$. We then sampled the behavioral distribution of inter-bout intervals δt_n, rescaled by the scaling factor $f_{N/B} \approx 0.44$ obtained in the previous section II F. This simulates a stochastic bout-initiation process with the correct temporal characteristics, yielding synthetic sequences of bout internal states b_n for the behavior. For each state, we then sample the emission probability $E(\delta\theta_n|b_n)$ associated to the

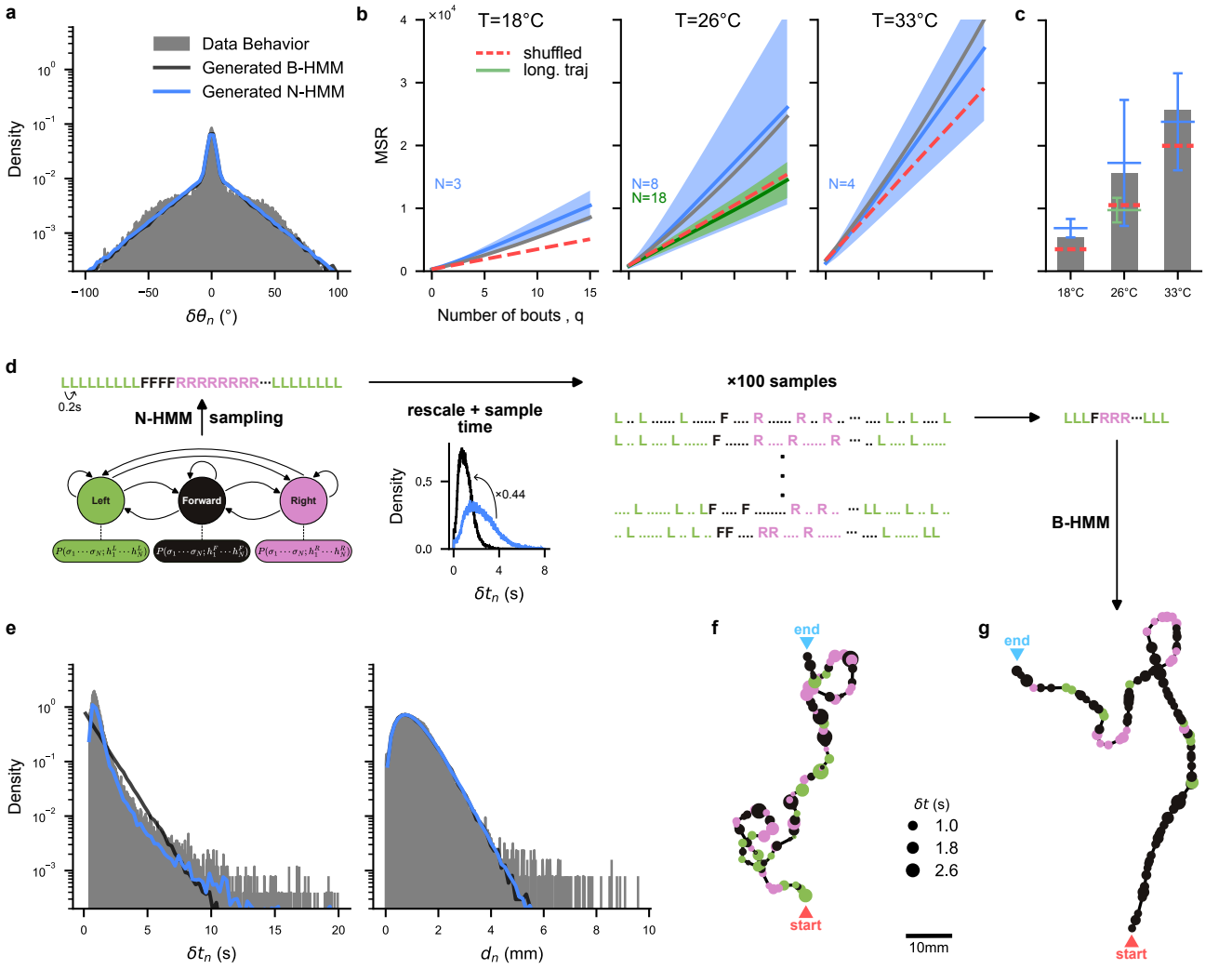


FIG. 6. **Generative ability of HMM models and trajectory reconstruction**: **(a)** Distribution of reorientation angles $\delta\theta_n$ for the aggregated multiple-fish trajectories (gray), trajectories generated from Behavioral Hidden Markov Models (B-HMM ; black), and generated from Neuronal Hidden Markov Models (N-HMM ; blue), at 22°C. **(b)** Mean Square Reorientation (MSR) accumulated after q bouts for aggregated multiple-fish trajectories (normal : grey, shuffled : red dashed), single-fish long trajectories (green), and trajectories generated from N-HMM (blue). For both long and N-HMM-generated trajectories, we show the mean and standard deviation over all individuals (solid line and band). (see Fig.S7 for individual trajectories and all temperatures) **(c)** MSR at $q = 10$ bouts, with mean (horizontal bars) and standard deviation (vertical bars). **(d)** Pipeline to convert N-HMM-generated state sequences to swim trajectories. The N-HMM is first sampled to generate a sequence of forward/left/right internal states. Time is then re-scaled as described in Fig 5, and bout sequences are sampled 100 times based on the interbout-interval distribution. A swim trajectory is constructed for each bout sequence by sampling the B-HMM emission distributions : bout distances d_n and inter-bout intervals δt_n. **(e)** Distribution of inter-bout intervals δt_n and bout distances d_n for the aggregated multiple-fish trajectories (gray), generated trajectories from B-HMM (black), and generated trajectories from N-HMM (blue), at 22^{o}C. **(f)** Example B-HMM-generated trajectory at 26^{o}C (color = bout type (left, right, forward), point size = inter-bout interval). **(g)** Same as panel f for a N-HMM-generated trajectory at 26^{o}C.

Behavioral HMM (B-HMM) inferred from all fish data to get a realization of the reorientation angle $\delta\theta_n$ (Fig. 6d). As expected, the distribution of these angles is in very good agreement with the ones observed in the behavioral data (Fig. 6a).

We further characterize the trajectories using the Mean Square Reorientation (MSR) after q bouts:

$$\mathrm{MSR}(q) = \left\langle \Big(\sum_{n=t+1}^{t+q} \delta\theta_n \Big)^2 \right\rangle_t \qquad (5)$$

where the average is taken across all times and all trajectories.

Figure 6b shows the values of $MSR(q)$ obtained from N-HMM-generated trajectories at different temperatures (see Fig.S7 for the remaining temperatures), as well as the MSR directly obtained from multiple-fish trajectories and long single-fish trajectories (only at 26°C).

We first notice that long individual fish trajectories at 26°C display large variability in MSR(q) values, compatible with the presence of fish-to-fish variability. This variability is washed out for the multiple-fish dataset (since individual trajectories are combined) providing a averaged MSR(q) for each temperature. Interestingly, the MSR for the long sequences of individual animals significantly differ from the MSR obtained from multiple-fish trajectories. This could be due to differences in experimental conditions, and in particular the effects of collective vs. isolated navigation [24].

N-HMM-generated trajectories have a MSR distribution compatible and encompassing the behavioral data in their variability. Such large variability is expected from the large fluctuations in neural brain states. Some trajectories generated with N-HMM show anomalously large angular persistence (see Fig.S7a), which may correspond to brain states where the *Anterior Rhombencephalic Turning Region* (ARTR) displays no left-right alternating behavior. This is expected, as the N-HMMs were established from spontaneous activity recordings of immobilized fish which were not constrained to swim-like behaviors. In Fig. 6c, we summarized and compared the results for the Mean Square Reorientation after 10 bouts, MSR($q = 10$), for all temperatures. We found, consistently across temperatures, that the MSR of behavioral data are comprised within the one-standard-deviation confidence interval of N-HMM-generated trajectories.

As we show in the Appendix 1, the MSR of HMM-generated trajectories can be decomposed as the sum of a purely diffusive contribution, associated to the variance of bout angles, and of terms arising from time-correlations in bout type selection along a trajectory (see Eq. (A.28)). The increase of bout-angle variance with temperature is sufficient to explain the increasing trend of the mean MSR with temperature observed in Figure 6c (see Fig.S7).

2. *Generation of synthetic 2D trajectories*

In the previous sub-section, we reconstructed reorientation sequences from synthetic neural state sequences. For the sake of completeness, we also used the N-HMM model to generate full synthetic 2D trajectories. To do so, for each bout state identified with the procedure reported above, we sampled an inter-bout interval duration δt_n and traveled distance d_n from their experimental distributions, ignoring the potential dependence of δt_n and d_n to bout type. We then reconstructed the coordinates of the virtual fish after k bouts, (x_k, y_k), through

$$x_k = \sum_{n=1}^{k} d_n \cos(\theta_n) \ , \quad y_k = \sum_{n=1}^{k} d_n \sin(\theta_n) \ , \qquad (6)$$

where $\theta_n = \sum_{i=1}^{n} \delta\theta_i$ is the orientation angle of the fish at bout n, constructed as the cumulative sum of re-orientation angles at previous bouts.

An example trajectory is shown in Fig. 6g.

For comparison, we show a synthetic trajectory generated from behavioral HMM (Fig. 6f) where the dependence of δt_n and d_n to bout type is learned from the data. In practice, we expanded on the B-HMM introduced in section II B 2 by adding two new emission distributions corresponding to δt_n and d_n. As done previously, the HMMs were first trained only on the re-orientation angles $\delta\theta_n$, before learning the emission distributions for δt_n and d_n. We then plot the corresponding trajectories using Eq. (6), which are qualitatively similar to the N-HMM-generated ones, as illustrated by an example in Fig.6f. The similarity is quantified by the comparison of the distributions of bout angles, inter-bout duration intervals and traveled distances, see Fig.6e.

We report in Figure S7e-h, the outcome of an intermediate generative model, in which 2D swim trajectories are generated directly from the experimental neural recordings. This is done by first identifying neural states from the recordings using the Viterbi algorithm, emitting inter-bout intervals using the same procedure as described in the previous subsection, and then feeding the resulting state sequences through the B-HMM.

III. DISCUSSION

With the advancement of video-tracking and brain recording methods, behavioral neuroscience has changed radically in the last decade. It is now possible to study in great details animal behavior in unconstrained naturalistic conditions [40–42], while new recording methods give access to extended circuit activity encompassing several brain regions. Such experiments produce vast amounts of high-dimensional data, requiring automated yet robust and interpretable analysis methods.

An essential task is the identification of behavioral or neural states from the segmentation of the recorded time series, in order to extract low-dimensional representations that are easier to interpret. However, no definitive procedure exists for selecting the optimal number of states or for defining valid labeling criteria. This choice typically depends on available observables and involves a compromise between interpretability and accuracy of representation.

In our case, the difficulty stems from three main factors: (i) the dependence of swim bout kinematics with bath temperature, (ii) the inter-individual variability, and (iii) the overlapping distributions of reorientation angles for distinct bout types, in particular

at low temperatures. Because they can accommodate such overlaps while taking into account the temporal regularities in the bout sequences, Hidden Markov Models (HMMs) are ideally suited for such a task. They are easily interpretable as the dynamics between the different internal states is described by a Markovian process. This makes HMM a powerful alternative to deep-learning-based methods, whose predictive power comes at the cost of interpretability.

In this study, we successfully applied a three-state HMM to parse behavioral and neural time series associated with exploratory dynamics. We showed that for behavioral data, HMM provided a less biased and more consistent method for bout-type labeling compared to standard threshold-based Markov Chain (MC) methods used in earlier studies.

This robustness proved essential as we investigated the effect of bath temperature on navigation. Zebrafish being cold-blooded animals, the water temperature is expected to directly affect muscle efficiency, leading to a systematic increase in the amplitude of reorientation elicited by bouts as temperature rises. When using threshold-based MC methods, this may lead to a systematic but artifactual increase in the fraction of bouts labeled as turns with temperature. With HMM, this physiological effect of temperature is naturally accounted for through an adaptive adjustment of the effective threshold angle between turn and forward bouts. With this unbiased labeling, we found that the fractions of forward and turn bouts were only weakly dependent on temperature, in contrast with previously published analysis [24]. The primary effect of temperature of rising temperature is to progressively decrease bout-type persistence, i.e. the tendency of the animal to chain similar bouts. Interestingly, we found that all three bout types, and not just turns as previously reported, exhibit comparable persistence.

HMMs also demonstrated remarkable sensitivity to individual phenotypic variability. Inter- and intra-individual variability are ubiquitous traits of animal behavior [43, 44] and are necessary to ensure a trade-off between flexibility and adaptability to changing environmental demands and robustness in neural development [45, 46]. In Le Goc *et al.* [24], inter-individual differences were demonstrated on the same dataset using multiple kinematic parameters (including inter-bout interval, forward travel distance or reorientation amplitude). In contrast, our study shows that HMM can identify individual fish solely based on the dynamics of bout-type sequences. Moreover, HMM provides explicit likelihood evaluation for bout sequences for various individual-specific models, providing a quantitative measure of phenotypic proximity between animals or across time.

Since our approach is based on gait phenotyping and is independent of image features, it is compatible with low-resolution videos (in which only the animal's position and orientation can be accessed) while still keeping versatility, reliability, and fast execution. This opens new opportunities for studying phenotypic variation in swimming behavior, potentially uncovering subtle effects on behavior of genetic, developmental, or environmental cues. This ability to precisely capture behavioral variability might also prove fruitful in order to explore the neural basis of individuality.

The fact that the fish directional dynamics can be described by a three-state Markovian sequence, suggests that bout-type selection is likely governed by a single circuit, with the ARTR being the most plausible candidate. Since its discovery in 2012 [32], the ARTR has been viewed as a direction-selection hub, controlling lateralized behaviors such as tail flick and ocular saccade orientation [20, 47]. It also responds to lateralized visual stimuli, including binocular contrast and whole-field lateral motion [22, 47].

In the present study, we showed that a three states HMM can accurately describe ARTR neuronal data, and that this model is structurally and temporally similar to behaviorally-trained HMMs. This result suggests that the ARTR may also govern forward bout selection, unifying the control of all directional bout types within a single circuit. This interpretation is reinforced by the generation of synthetic, neurally driven swimming sequences that closely matched the statistics of observed trajectories. However, we acknowledge that this hypothesis rests only on the comparison of a behavioral and a neuronal model obtained from data recorded separately. In order to confirm it, one would need to perform experiments with joint behavioral and neuronal recordings, where swim bouts can be segmented into left/right/forward bouts. However, such a dataset is not yet available. The closest which exist are of paralysed fish where fictive behavior is inferred from electrophysiological recordings of motor neurons in the spinal cord. While those recordings provide a good estimation of motor command asymmetry [20, 48], to the best of our knowledge they cannot be segmented into left/right/forward bouts.

Bout-type persistence, as observed in behavioral assays, is mirrored in the slow sequential exploration of the three hidden states identified in neural recordings of the ARTR. Although the HMM enables the identification of these neuronal states, they provide no interpretation of how they emerge from interactions among the ARTR neuronal population. In fact, our implementation of HMM assumes the activity of neurons to be independent of each other when conditioned to a state.

In a recent study, we trained data-driven graphical models (Ising model) on ARTR activity sequences [29]. The Ising model uses activity patterns to learn the interactions between neurons but, unlike HMM, it ignores any temporal information in the data. Interestingly, the inferred Ising models tended to display three metastable states, two with high activity on either side and one "equilibrated" state with intermediate, balanced activity on both sides, consistent with the

three hidden states found with HMM. This convergence underlines the complementary strengths of state-space and energy-based models in elucidating neural dynamics. While the former might enable capturing the temporal structure in collective neural activity, the latter offer insights into the underlying network interactions driving these states, and how metastability emerges within neural populations.

Current evidence suggests that the decision to initiate a swim bout and the choice of its direction are mediated by partially distinct circuits. Pre-motor nuclei such as the nucleus of the medial longitudinal fasciculus (nMLF) and the mesencephalic locomotor region (MLR) act as "swim-drive" centers. The first was shown to be activated by visual stimuli inducing forward bouts [49], and the second was shown to elicit forward bouts and scale swimming speed when stimulated [50].
Turning, by contrast, relies on lateralized descending inputs. A subset of ventral spinal projection neurons (vSPNs, including RoV3, MiV1 and MiV2) which fire asymmetrically during the first tail undulation, biasing the bout toward the turning side; subsequent undulations then resemble those of forward swims. These vSPN follow a sigmoid profile, with low activity during contraversive swims, high during ipsiversive swims, and intermediate during forward swims [33], which is coherent with the left-active, right-active, and balanced states uncovered by our HMM analysis.
While the exact role of the ARTR in controlling bout selection remains unclear, anatomical tracing shows that it projects onto the vSPN domain [20], suggesting that it might drive this bias. Our data are consistent with a model in which the left and right ARTR subpopulations inhibit contraversive turns: when both sides are equally active they suppress turning, leaving forward bouts as the default. Such a motor suppression mechanism is consistent with observations by Dunn *et al.* [20], who showed a monotonic relationship between the activity of ARTR neurons, and the lateralization of fictive bouts. Visually driven locomotion may bypass this mechanism via an ARTR-independent pathway, as suggested by Naumann *et al.* [51], whereas spontaneous swimming could rely more heavily on ARTR-mediated bout selection.
Overall, we have found no incompatibility in the literature with our proposed model of the ARTR as a modulator of both forward and left-right swim orientations. We suggests two tests of this hypothesis. First, a detailed analysis of ARTR dynamics during forward swims, for example during forward OMR stimulation, in order to test whether its intrinsic oscillation persists or becomes phase-locked in a balanced state. Second, a bilateral ablation of all ARTR neurons to assess their necessity for spontaneous versus visually evoked locomotion, by comparing bout frequency and orientation after lesions.

Last of all, to facilitate the accessibility and adoption of Hidden Markov Model (HMM) formalism for analyzing behavioral sequences, we provide a comprehensive and instructive Python tutorial (`https://github.com/EmeEmu/IBIO-Banyuls2023-Python`). This tutorial can be adapted for specific datasets or used as a resource for broader educational goals.

IV. MATERIALS AND METHODS

A. Behavioral Datasets

The behavioral dataset used in the present study is derived from Le Goc *et al.* [24], and can be accessed directly at `https://doi.org/10.5061/dryad.3r2280ggw`. This dataset comprises spontaneous swimming trajectories of 5 to 7 dpf zebrafish larvae, collected at controlled bath temperatures of 18°C, 22°C, 26°C, 30°C, and 33°C. A camera was used to continuously record the swimming behavior of the fish within an arena of $100\times45\times4.5\text{mm}^3$ for 30 minutes at 25 frames/second. To eliminate border effects, a Region of Interest (ROI) was defined at a distance of 5mm from the arena's walls. Fish that swam outside the defined tracking ROI were considered lost, and a new trajectory was initiated upon their re-entry into the ROI. The identity of the fish is thus lost each time it exits the ROI. Therefore, the dataset contains a varying number of fish trajectories, ranging from 532 to 1513 trajectories across the different temperatures (mean = 1148). Individual trajectories were tracked offline using the open-source FastTrack software [52], and were then discretized into sequences of swimming bouts.

Each trajectory consists of a sequence of swim bouts, spanning from 9 to 748 bouts per trajectory (mean=60, distributions shown in Fig.S1a). From this extensive dataset, we primarily utilized the re-orientation angles, defined as the difference between the heading direction at bout $n+1$ and the heading direction at bout n:

$$\delta\theta_n = \theta_{n+1} - \theta_n \tag{7}$$

(a graphical illustration of this definition can be found in Fig.1c). This parameter encapsulates the angular change between consecutive bouts, providing insight into the fish's ability to modify its orientation during swimming.

We also used the interbout interval $\delta t_n = t_{n+1} - t_n$ representing the elapsed time between 2 consecutive bouts, and the traveled distance $d_n = \|\vec{r}_{n+1} - \vec{r}_n\|$.

On top of these multi-fish trajectories, we used in sections II D and II G a second dataset from Le Goc *et al.* [24] consisting in single-fish recordings. For this dataset, each fish (N=18) is placed alone in the arena at 26°C, and is recorded for 2 hours. With this experimental paradigm, the identity of the fish is conserved across trajectories, even when the fish leaves and re-enters the ROI.

B. Neuronal Datasets

The neuronal dataset used in the present study is derived from Wolf *et al.* [29], and can be accessed directly at `https://gin.g-node.org/Debregeas/ZF_ARTR_thermo`. This dataset contains 32 one-photon Light-Sheet Microscopy recordings of spontaneous brain activity, for 13 zebrafish larvae (5 to 7 dpf) at 18°C, 22°C, 26°C, 30°C, and 33°C. It focuses on neurons from the *Anterior Rhombencephalic Turning Region* (ARTR), with $\sim$ 300 neurons (mean 307, std 119), recorded during $\sim$ 20min (mean 23, std 4 min) at $\sim$ 6Hz (mean 5.9, std 2.1 Hz). The approximate binarized spike trains of segmented neurons were inferred from fluorescence signals using a previously described deconvolution algorithm [53].

C. Emission of reorientation angles in the Hidden Markov Model

To validate the hypothesis that the re-orientation angles can be modeled using normal and gamma distributions, we compared the distribution of the data with a Gaussian Mixture Model (GMM) and a Gaussian & Gamma Mixture Model:

$$p(\delta\theta) = w_F\mathcal{N}(\delta\theta; 0, \sigma) + w_L\Gamma(\delta\theta; \alpha, \theta) + w_R\Gamma(-\delta\theta; \alpha, \theta)$$

where $w_F+w_L+w_R = 1$, and w_F, w_L, and w_R denote the weights for forward, left, and right states, respectively.

Using Quantile-Quantile (QQ) plots, we show that this last mixture model accurately reproduces the observed distribution of $\delta\theta_n$ in the data, and is much better than a GMM, especially in the tails of the distributions (Fig. S2a). We also confirmed that, once trained, the emission distributions do indeed match the observed reorientation distributions (Fig. S2b-c).

D. Hidden Markov model training

Given a data set of trajectories (neuronal or behavioral), the Hidden Markov models were trained using the standard Baum-Welch algorithm [54]. We train for a maximum of 500 expectation-maximization iterations, stopping earlier if the gain in the total log-likelihood of all the data in one iteration becomes less than a small threshold, that we fix at 10^{-6}. In practice we find that all HMMs trained in this study converge before reaching 500 iterations.

The transition matrix parameters are initialized at random, respecting left-right symmetry in the case of the behavioral models.

E. Stubbornness factor

The *stubbornness* factor f_q is a measurement of the animal's preference towards turning in the same direction over changing direction, after q intermediary forward bouts (Fig.S4c), as defined in (3).

It can be computed from a sequence of classified bouts b_n by first identifying and counting the q-plets $T_1 \rightarrow F^q \rightarrow T_2$ where $T_1 = T_2$ and where $T_1 \neq T_2$:

$$\begin{cases} N_{=} & = \#(T_1 \rightarrow F^q \rightarrow T_2, T_1 = T_2) \\ N_{\neq} & = \#(T_1 \rightarrow F^q \rightarrow T_2, T_1 \neq T_2) \end{cases} \tag{8}$$

and then computing their ratio:

$$f_q = \frac{N_{=}}{N_{\neq}} \tag{9}$$

In practice, this ratio has a physical interpretation only for long sequences of bouts where $N_{=} >> 1$ and $N_{\neq} >> 1$. As the trajectories in our dataset can be quite short (Fig. S1a), we compute f_q from all trajectories at a specific temperature, increasing the chance of observing a high number of stubborn ($N_{=}$) and non-stubborn ($N_{\neq}$) trajectories.

By considering that the probability of a given q-plet is stubborn follows a binomial distribution ($\mathbb{E}(N_{=}) = pN$ and $\mathbb{E}(N_{\neq}) = (1-p)N$ with $N = N_{=} + N_{\neq}$), we can evaluate the uncertainty in *stubbornness* as:

$$\Delta f_q = f_q \frac{1}{N_{=} + N_{\neq}} \left(\sqrt{\frac{N_{=}}{N_{\neq}}} + \sqrt{\frac{N_{\neq}}{N_{=}}} \right) \tag{10}$$

It is to be noted that these uncertainties are conservative estimates, as there exits a bias inherent to the dataset. Indeed, a very stubborn fish will tend to stay longer within the Region Of Interest (ROI) of the camera, leading to longer trajectories and therefore weighing more on the final result. Hence, it is unclear whether a *stubbornness* factor $f_q = 1 \pm 0.2$ is truly significant (as suggested by the estimated error bars on Fig.S4d).

Furthermore, as the stubbornness factor is computed from all trajectories (and thus all fish) at a particular temperature, it represents an average behavior rather than an individual fish.

F. Stubbornness factor and 3-state Markov Chain

The *stubbornness* factor can be defined directly from the transition matrix:

For $q = 0$, calculations are simple:

$$f_{q=0} = \frac{P(L \rightarrow L) + P(R \rightarrow R)}{P(L \rightarrow R) + P(R \rightarrow L)} \tag{11}$$

*For $q \geq 1$ *, the *stubbornness* factor is defined from the transition matrix as:

$$\begin{aligned} S_{L,q} &= P(L \to F^q \to L) \\ &= P(L)P(L \to F)P^q(F \to F)P(F \to L) \\ W_{L,q} &= P(L \to F^q \to R) \\ &= P(L)P(L \to F)P^q(F \to F)P(F \to R) \\ f_q &= \frac{S_{L,q} + S_{R,q}}{W_{L,q} + W_{R,q}} \end{aligned}$$

with $S_{L,q}$ the probability of a trajectory which starts and ends with a left bout, $W_{L,q}$ the probability of a trajectory which starts with a left bout and ends with a right bout, and $S_{R,q}$ $W_{R,q}$ their symmetrical opposites.

For a 3-state model, the forward-forward bout probability cancels out, giving:

$$f_q = \frac{P(L)P(L \to F)P(F \to L)}{P(L)P(L \to F)P(F \to R)} \frac{+P(R)P(R \to F)P(F \to R)}{+P(R)P(R \to F)P(F \to L)}$$

and with our non-handedness hypothesis: $P(L) = P(R)$, $P(L \to F) = P(R \to F)$, and $P(F \to L) = P(F \to R)$, yielding:

$$f_q = 1 \quad \forall q > 0 \tag{12}$$

G. Labeling of states in the neuronal Hidden Markov Model

The internal states of the Hidden Markov Models (HMMs) trained from neuronal activity are not *a priory* assigned to the Left, Right and Forward labels, and must therefore be re-ordered post-training.

We expect a certain symmetry in the system, where neurons in the left side of the ARTR will be more active during a Left state (and vice versa). Hence, we can use the excitability h_i^s of neuron i in each internal state s, as defined in the emission distribution of the HMM (see Eq. 4). We define the lateralized excitability:

$$\Delta h^s = \left\langle l(h_i^s) \right\rangle_{i \in \mathfrak{L}} - \left\langle l(h_i^s) \right\rangle_{i \in \mathfrak{R}} \tag{13}$$

where $l(x) = \frac{1}{1+e^{-x}}$ is the standard logistic function, and $\mathfrak{L}$ and $\mathfrak{R}$ are the sets of neurons located respectively in the left and right side of the ARTR. We thus label the HMM states such that

$$\Delta h^L > \Delta h^F > \Delta h^R \tag{14}$$

with F, L, and R the Forward, Left and Right internal states.

H. Temporal re-scaling

To find the temporal re-scaling factor $f_{N/B}$ between behavioral and neuronal models, we first compute the distributions of sojourn times Δt_s for all states $s \in \{F, L, R\}$ in both behavioral and neuronal Hidden Markov Models.

We then find the optimal re-scaling factor $f_{N/B}$ for which the combined distributions $\Delta t_b = [\Delta t_F^b, \Delta t_L^b, \Delta t_R^b]$ and $\Delta t_n = [\Delta t_F^n, \Delta t_L^n, \Delta t_R^n]$ are as close to each other as possible :

$$f_{N/B} = \min_{f \in [0,1]} \mathrm{RMSE}\Big(Q(\Delta t_b), f.Q(\Delta t_n)\Big) \tag{15}$$

where $Q(D)$ is the quantiles of a distribution D, and $\mathrm{RMSE}(\mathbf{x}, \mathbf{y})$ is the Root Mean Squared Error between vectors $\mathbf{x}$ and $\mathbf{y}$ (see Fig. S6a).

For Markov chains, the transition matrix $P = P(s_n = s \to s_{n+1} = s')$ represents the probability of transitioning in one step from state s to state s'. The transition probability $s \to s'$ in $k \in \mathbb{N}_1$ steps $P(s_n = s \to s_{n+k} = s')$ is then the matrix power P^k.

In order to apply the temporal re-scaling $f_{N/B}$ between behavioral and neuronal models, we can thus compute the re-scaled transition matrix :

$$P_n^* = P_n^{\lfloor \frac{\nu}{f_{N/B}} \rceil} \tag{16}$$

where P_n is the transition matrix inferred from neuronal data recorded at a frequency ν Hz.

I. Mean Square Reorientation

To characterize the orientational diffusivity of the trajectories, we use the Mean Square Reorientation (MSR) accumulated after q bouts, as defined in equation (5) [22].

For infinitely large datasets with no left-right bias, we expect a centered distribution of reorientation angles $\langle \delta\theta_n \rangle_n = 0$. However, this is not the case, particularly for the neuronal dataset where experimental limitations can induce strong biases. In particular, two of those limitations are due to the one-sided illumination in our Light Sheet Fluorescence Microscope [55]. First, due to scattering, the illumination beam is not uniform left-right across the brain, which can induce biases in the detection of neurons and their activity. Second, the non-uniform perception of light by the zebrafish larvae can elicit a phototaxis response, which is known to bias the activity of the *Anterior Rhombencephalic Turning Region* (ARTR) [47].

Since a non-zero bias can result in a distortion of the MSR (see Appendix 1), the MSR is computed from $\delta\hat{\theta}_n = \delta\theta_n - \langle \delta\theta_n \rangle_n$ instead of $\delta\theta_n$.

Acknowledgment. We acknowledge the following funding:

Author	Funder
M. DK.	École Doctorale Frontière de l'Innovation en Recherche et Education - Programme Bettencourt
J. FdCD.	Université PSL, AI Junior Fellow program
M. C.	European Union, Horizon 2020 Programme (H2020 MSCA ITN Project SmartNets GA-860949)
V. B.	European Research Council (ERC) under the European Union's Horizon 2020 research innovation program grant agreement number 715980
R. M., G. D., S. C.	Locomat ANR-21-CE16-0037

Data and Code availability. All the data and code used in the present article are available under GNU General Public License version 3 at `https://github.com/ZebrafishHMM2023/ZebrafishHMM2023_CodeAndData/tree/bioRxiv`.

Our Julia implementation of the Hidden Markov Models used is available under MIT License at `https://github.com/ZebrafishHMM2023/ZebrafishHMM2023.jl`.

We also provide two tutorials for the use of Hidden Markov Models for behavioral sequence analysis. The first one was created for the Cogmaster "Machine learning for cognitive science" course, at the École Normale Supérieure in Paris, and is available at:`https://github.com/CoccoMonassonLab/ZebrafishHMM`. The second one was created for the i-Bio Summer School "Advanced Computational Analysis for Behavioral and Neurophysiological Recordings" held in Banyuls-sur-Mer in the summer of 2023, and is available under GNU General Public License version 3 at `https://github.com/EmeEmu/IBIO-Banyuls2023-Python`.

[1] N. Tinbergen, *The study of instinct* (Pygmalion Press, an imprint of Plunkett Lake Press, 2020).

[2] A. B. Wiltschko, M. J. Johnson, G. Iurilli, R. E. Peterson, J. M. Katon, S. L. Pashkovski, V. E. Abraira, R. P. Adams, and S. R. Datta, Mapping sub-second structure in mouse behavior, Neuron **88**, 1121 (2015).

[3] J. M. Mueller, P. Ravbar, J. H. Simpson, and J. M. Carlson, Drosophila melanogaster grooming possesses syntax with distinct rules at different temporal scales, PLoS computational biology **15**, e1007105 (2019).

[4] T. Gallagher, T. Bjorness, R. Greene, Y.-J. You, and L. Avery, The geometry of locomotive behavioral states in c. elegans, PloS one **8**, e59865 (2013).

[5] L. Tao, S. Ozarkar, J. M. Beck, and V. Bhandawat, Statistical structure of locomotion and its modulation by odors, Elife **8**, e41235 (2019).

[6] D. S. Mearns, J. C. Donovan, A. M. Fernandes, J. L. Semmelhack, and H. Baier, enDeconstructing Hunting Behavior Reveals a Tightly Coupled Stimulus-Response Loop, Current Biology **30**, 54 (2020).

[7] S. Linderman, A. Nichols, D. Blei, M. Zimmer, and L. Paninski, Hierarchical recurrent state space models reveal discrete and continuous dynamics of neural activity in c. elegans, bioRxiv 10.1101/621540 (2019).

[8] G. J. Berman, W. Bialek, and J. W. Shaevitz, Predictability and hierarchy in drosophila behavior, Proceedings of the National Academy of Sciences **113**, 11943 (2016), https://www.pnas.org/doi/pdf/10.1073/pnas.1607601113.

[9] R. E. Johnson, S. Linderman, T. Panier, C. L. Wee, E. Song, K. J. Herrera, A. Miller, and F. Engert, Probabilistic models of larval zebrafish behavior reveal structure on many scales, Current Biology **30**, 70 (2020).

[10] M. Breakspear, enDynamic models of large-scale brain activity, Nat Neurosci **20**, 340 (2017).

[11] L. Mazzucato, A. Fontanini, and G. La Camera, Dynamics of multistable states during ongoing and evoked cortical activity, Journal of Neuroscience **35**, 8214 (2015), https://www.jneurosci.org/content/35/21/8214.full.pdf.

[12] A. J. Quinn, D. Vidaurre, R. Abeysuriya, R. Becker, A. C. Nobre, and M. W. Woolrich, Task-evoked dynamic network analysis through hidden markov modeling, Frontiers in neuroscience **12**, 603 (2018).

[13] Y. Zhang and S. Saxena, Inference of neural dynamics using switching recurrent neural networks, in *The Thirty-eighth Annual Conference on Neural Information Processing Systems* (2024).

[14] T. Li and G. La Camera, A sticky poisson hidden markov model for spike data, bioRxiv 10.1101/2024.08.07.606969 (2024), https://www.biorxiv.org/content/early/2024/08/08/2024.08.07.606969.

[15] M. B. Orger and G. G. de Polavieja, Zebrafish behavior: opportunities and challenges, Annual review of neuroscience **40**, 125 (2017).

[16] J. R. Meyers, Zebrafish: development of a vertebrate model organism, Current Protocols Essential Laboratory Techniques **16**, e19 (2018).

[17] J. H. Bollmann, The zebrafish visual system: from circuits to behavior, Annual review of vision science **5**, 269 (2019).

[18] J. C. Marques, S. Lackner, R. Félix, and M. B. Orger, Structure of the zebrafish locomotor repertoire revealed with unsupervised behavioral clustering, Current Biology **28**, 181 (2018).

[19] X. Chen and F. Engert, Navigational strategies underlying phototaxis in larval zebrafish, Frontiers in Systems Neuroscience **8**, 10.3389/fnsys.2014.00039 (2014).

[20] T. W. Dunn, Y. Mu, S. Narayan, O. Randlett, E. A. Naumann, C.-T. Yang, A. F. Schier, J. Freeman, F. Engert, and M. B. Ahrens, Brain-wide mapping of neural activity controlling zebrafish exploratory locomotion, Elife **5**, e12741 (2016).

[21] E. J. Horstick, Y. Bayleyen, J. L. Sinclair, and H. A. Burgess, Search strategy is regulated by somatostatin signaling and deep brain photoreceptors in zebrafish, BMC biology **15**, 1 (2017).

[22] S. Karpenko, S. Wolf, J. Lafaye, G. Le Goc, T. Panier, V. Bormuth, R. Candelier, and G. Debrégeas, From be-

havior to circuit modeling of light-seeking navigation in zebrafish larvae, eLife **9**, e52882 (2020), publisher: eLife Sciences Publications, Ltd.

[23] E. J. Horstick, Y. Bayleyen, and H. A. Burgess, Molecular and cellular determinants of motor asymmetry in zebrafish, Nature Communications **11**, 1170 (2020).

[24] G. Le Goc, J. Lafaye, S. Karpenko, V. Bormuth, R. Candelier, and G. Debrégeas, Thermal modulation of Zebrafish exploratory statistics reveals constraints on individual behavioral variability, BMC Biology **19**, 208 (2021).

[25] D. L. Barabási, G. F. Schuhknecht, and F. Engert, Functional neuronal circuits emerge in the absence of developmental activity, Nature Communications **15**, 364 (2024).

[26] M. Haesemeyer, D. N. Robson, J. M. Li, A. F. Schier, and F. Engert, A brain-wide circuit model of heat-evoked swimming behavior in larval zebrafish, Neuron **98**, 817 (2018).

[27] M. Haesemeyer, D. N. Robson, J. M. Li, A. F. Schier, and F. Engert, A brain-wide circuit model of heat-evoked swimming behavior in larval zebrafish, Neuron **98**, 817 (2018).

[28] V. Palieri, E. Paoli, Y. Wu, M. Haesemeyer, I. Grunwald Kadow, and R. Portugues, The preoptic area and dorsal habenula jointly support homeostatic navigation in larval zebrafish, Curr Biol **3**, 34 (2024).

[29] S. Wolf, G. Le Goc, G. Debrégeas, S. Cocco, and R. Monasson, Emergence of time persistence in a data-driven neural network model, eLife **12**, e79541 (2023).

[30] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah, Julia: A fresh approach to numerical computing, SIAM review **59**, 65 (2017).

[31] G. Dalle, Hiddenmarkovmodels.jl: generic, fast and reliable state space modeling, Journal of Open Source Software **9**, 6436 (2024).

[32] M. B. Ahrens, J. M. Li, M. B. Orger, D. N. Robson, A. F. Schier, F. Engert, and R. Portugues, Brain-wide neuronal dynamics during motor adaptation in zebrafish, Nature **485**, 471 (2012), 22622571.

[33] K.-H. Huang, M. B. Ahrens, T. W. Dunn, and F. Engert, Spinal Projection Neurons Control Turning Behaviors in Zebrafish, Current Biology **23**, 1566 (2013).

[34] L. Petrucco, H. Lavian, Y. K. Wu, F. Svara, V. Štih, and R. Portugues, Neural dynamics and architecture of the heading direction circuit in zebrafish, Nature Neuroscience **26**, 765 (2023).

[35] R. Portugues, C. E. Feierstein, F. Engert, and M. B. Orger, Whole-Brain Activity Maps Reveal Stereotyped, Distributed Networks for Visuomotor Behavior, Neuron **81**, 1328 (2014), 24656252.

[36] E. I. Dragomir, V. Štih, and R. Portugues, Evidence accumulation during a sensorimotor decision task revealed by whole-brain imaging, Nature Neuroscience **23**, 85 (2020).

[37] K. E. Severi, R. Portugues, J. C. Marques, D. M. O'Malley, M. B. Orger, and F. Engert, Neural control and modulation of swimming speed in the larval zebrafish, Neuron **83**, 692 (2014), 25066084.

[38] S. Karpenko, *Naviguer Avec La Lumière : Du Comportement Aux Circuits Neuronaux Chez La Larve de Poisson Zèbre*, These de doctorat, Université Paris sciences et lettres (2020).

[39] K. M. Hallinen, R. Dempsey, M. Scholz, X. Yu, A. Linder, F. Randi, A. K. Sharma, J. W. Shaevitz, and A. M. Leifer, Decoding locomotion from population neural activity in moving *C. elegans*, eLife **10**, e66135 (2021).

[40] A. E. Brown and B. De Bivort, Ethology as a physical science, Nature Physics **14**, 653 (2018).

[41] T. D. Pereira, J. W. Shaevitz, and M. Murthy, Quantifying behavior to understand the brain, Nature neuroscience **23**, 1537 (2020).

[42] A. Kennedy, The what, how, and why of naturalistic behavior, Current opinion in neurobiology **74**, 102549 (2022).

[43] K. Honegger and B. de Bivort, Stochasticity, individuality and behavior, Current Biology **28**, R8 (2018).

[44] A. K. Shaw, Causes and consequences of individual variation in animal movement, Movement ecology **8**, 12 (2020).

[45] P. R. Hiesinger and B. A. Hassan, The evolution of variability and robustness in neural development, Trends in Neurosciences **41**, 577 (2018).

[46] G. Sridhar, M. Vergassola, J. C. Marques, M. B. Orger, A. C. Costa, and C. Wyart, Uncovering multiscale structure in the variability of larval zebrafish navigation, Proceedings of the National Academy of Sciences **121**, e2410254121 (2024).

[47] S. Wolf, A. M. Dubreuil, T. Bertoni, U. L. Böhm, V. Bormuth, R. Candelier, S. Karpenko, D. G. Hildebrand, I. H. Bianco, R. Monasson, *et al.*, Sensorimotor computation underlying phototaxis in zebrafish, Nature communications **8**, 651 (2017).

[48] X. Chen, Y. Mu, Y. Hu, A. T. Kuan, M. Nikitchenko, O. Randlett, A. B. Chen, J. P. Gavornik, H. Sompolinsky, F. Engert, and M. B. Ahrens, Brain-wide Organization of Neuronal Activity and Convergent Sensorimotor Transformations in Larval Zebrafish, **100**, 876, 30473013.

[49] M. B. Orger, A. R. Kampff, K. E. Severi, J. H. Bollmann, and F. Engert, Control of visually guided behavior by distinct populations of spinal projection neurons, **11**, 327, 18264094.

[50] M. Carbo-Tano, M. Lapoix, X. Jia, O. Thouvenin, M. Pascucci, F. Auclair, F. B. Quan, S. Albadri, V. Aguda, Y. Farouj, E. M. C. Hillman, R. Portugues, F. Del Bene, T. R. Thiele, R. Dubuc, and C. Wyart, The mesencephalic locomotor region recruits V2a reticulospinal neurons to drive forward locomotion in larval zebrafish, **26**, 1775.

[51] E. A. Naumann, J. E. Fitzgerald, T. W. Dunn, J. Rihel, H. Sompolinsky, and F. Engert, From Whole-Brain Data to Functional Circuit Models: The Zebrafish Optomotor Response, **167**, 947, 27814522.

[52] B. Gallois and R. Candelier, Fasttrack: an open-source software for tracking varying numbers of deformable objects, PLoS computational biology **17**, e1008697 (2021).

[53] J. Tubiana, S. Wolf, T. Panier, and G. Debregeas, Blind deconvolution for spike inference from fluorescence recordings, Journal of Neuroscience Methods **342**, 108763 (2020).

[54] L. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, Proceedings of the IEEE **77**, 257 (1989), conference Name: Proceedings of the IEEE.

[55] T. Panier, S. Romano, R. Olive, T. Pietri, G. Sumbre, R. Candelier, and G. Debrégeas, Fast functional imaging of multiple brain regions in intact zebrafish larvae using Selective Plane Illumination Microscopy, Frontiers in Neural Circuits **7** (2013).

APPENDICES

1. Mean squared reorientation

The mean-square reorientation (MSR) of a trajectory at lag q is defined as [22]:

$$M_q = \mathrm{E}\left[\left(\sum_{i=1}^{q} \delta\theta_{t+i-1}\right)^2\right] = \sum_{i,j=1}^{q} \mathrm{E}\left[\delta\theta_{t+i-1}\delta\theta_{t+j-1}\right] \tag{A.17}$$

where it is assumed that $\mathrm{E}[\delta\theta] = 0$. The average is taken over time t. Assuming stationarity, this is independent of t, and should thus only depend on the separation $|i-j|$,

$$\mathrm{E}\left[\delta\theta_{t+i-1}\delta\theta_{t+j-1}\right] = \mathrm{E}\left[\delta\theta_i\delta\theta_j\right] = A_{|i-j|} \tag{A.18}$$

where $A_{|i-j|}$ stands for the time equilibrated autocorrelation function:

$$A_t = \lim_{t_0\to\infty} \mathrm{E}[\delta\theta_{t_0}\delta\theta_{t_0+t}] \tag{A.19}$$

In particular $A_0 = \mathrm{E}[\delta\theta^2]$ is just the variance of $\delta\theta$. It follows that,

$$\begin{aligned} M_q &= \sum_{i,j=1}^{q} A_{|i-j|} = \sum_{t=0}^{q-1}\left(\sum_{i,j=1}^{q}\delta_{|i-j|,t}\right)A_t \\ &= qA_0 + 2\sum_{t=1}^{q-1}\left(\sum_{i<j}^{q}\delta_{j-i,t}\right)A_t \\ &= qA_0 + 2\sum_{t=1}^{q-1}(q-t)A_t \end{aligned} \tag{A.20}$$

Note that for a random walk without any correlations across time, $A_t = 0$ for $t > 0$. In this case, $M_q = qA_0$ grows linearly with q.

On the other hand, it is expected that $A_t \to 0$ as $t \to \infty$, and usually this decay is exponentially fast in time. Therefore, for large q, we get the following asymptotic expression for M_q:

$$M_q \sim \left(A_0 + 2\sum_{t=1}^{\infty} A_t\right) q - 2\sum_{t=1}^{\infty} tA_t \tag{A.21}$$

Notice that this is affine in q, with the coefficient $A_0 + 2\sum_{t=1}^{\infty} A_t$. Therefore, M_q is initially linear in q with slope A_0 for small q, then has an elbow and eventually approaches the asymptotic slope $A_0 + 2\sum_{t=1}^{\infty} A_t$ as $q \to \infty$. This asymptotic slope is different from A_0 only if the process exhibits non-trivial autocorrelations in time.

a. MSR for the HMM

As an illustration, we can compute all these quantities exactly for the HMM as follows. For the autocorrelation, we have:

$$\begin{aligned} A_t &= \operatorname*{tr}_{h_0,\dots,h_t} P(h_t|h_{t-1})\dots P(h_2|h_1)P(h_1|h_0)P(h_0) \\ &\quad\times\left[\int P(\delta\theta|h_0)d\delta\theta\right]\left[\int P(\delta\theta|h_t)d\delta\theta\right] \\ &= \operatorname*{tr}_{h,h'} [\Omega^t]_{h',h} P(h)\langle\delta\theta|h\rangle\langle\delta\theta|h'\rangle \end{aligned} \tag{A.22}$$

where $[\Omega]_{h',h} = P(h'|h)$ is the transition matrix of the HMM. We will assume here that the initial state is sampled from $P(h) = p_{\mathrm{eq}}(h)$, the equilibrium distribution of hidden states of the HMM, which satisfies the stationarity equation:

$$\operatorname*{tr}_{h} \Omega_{h',h} p_{\mathrm{eq}}(h) = p_{\mathrm{eq}}(h') \tag{A.23}$$

Note also that $\mathrm{E}[\delta\theta] = 0$ implies that $\sum_h p_{\mathrm{eq}}(h)\langle\delta\theta|h\rangle = 0$. Now let $p_1(h), \dots, p_L(h)$ denote the remaining eigenvectors of Ω, with the associated eigenvalues $\lambda_1, \dots, \lambda_L$. By the Perron-Frobenious theorem, these remaining eigenvalues are all smaller than one in absolute value. The vector $P(h)\langle\delta\theta|h\rangle$ can be writen in the basis of this eigenvectors,

$$P(h)\langle\delta\theta|h\rangle = \alpha_{\mathrm{eq}} p_{\mathrm{eq}}(h) + \sum_{i=1}^{L} \alpha_i p_i(h) \tag{A.24}$$

for some coefficients $\alpha_{\mathrm{eq}}, \alpha_1, \dots, \alpha_L$. Then it follows that,

$$\begin{aligned} A_t &= \operatorname*{tr}_{h'}\left[\alpha_{\mathrm{eq}} p_{\mathrm{eq}}(h') + \sum_i \lambda_i^t \alpha_i p_i(h')\right]\langle\delta\theta|h'\rangle \\ &= \sum_i \alpha_i \lambda_i^t \operatorname*{tr}_{h'} p_i(h')\langle\delta\theta|h'\rangle \end{aligned} \tag{A.25}$$

Since the $|\lambda_i| < 1$ it follows that $A_t \to 0$ exponentially fast as $t \to \infty$. Moreover we can compute,

$$\sum_{t=0}^{\infty} A_t = \sum_i \frac{\alpha_i}{1-\lambda_i} T_i, \quad \sum_{t=0}^{\infty} tA_t = \sum_i \frac{\alpha_i\lambda_i}{(1-\lambda_i)^2} T_i \tag{A.26}$$

where

$$T_i = \operatorname*{tr}_{h} p_i(h)\langle\delta\theta|h\rangle \tag{A.27}$$

These expressions then give a complete and exact characterization of the MSR for the HMM.

b. Standardized MSR

The MSR as defined in Eq. A.17 includes both the diffusive contribution from the initial term A_0 and contributions arising from non-trivial time correlations in the process coming from the terms A_t for $t > 0$. As already pointed out, this initial term $A_0 = \mathrm{E}[\delta\theta^2]$ is just the variance of the distribution of bout angles and is insensitive

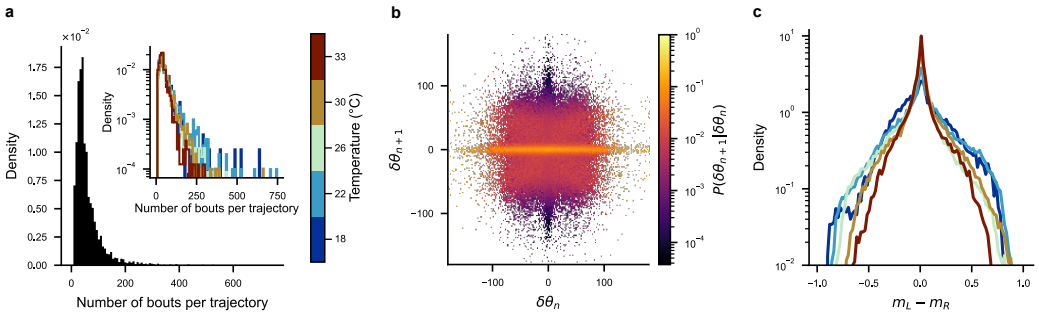


FIG. S1. **Supplementary panels to Fig.1**: **(a)** Number of bouts per trajectory for the entire behavioral dataset (black), and per temperature (inset, colored). **(b)** Observed transition probabilities between reorientation angles for the entire behavioral dataset. **(c)** Difference between mean activities in the left (m_L) and right (m_R) *Anterior Rhombencephalic Turning Region* for all fish at each temperature.

to time correlations. To emphasize the time correlations we may normalize the trajectories by defining:

$$\hat{M}_q = \frac{M_q}{A_0} \tag{A.28}$$

By comparing with Eq. A.21, we see that $\hat{M}_q$ has initially a slope ≈ 1 for small q, then has an elbow and eventually approaches the asymptotic slope $1 + 2\sum_{t=1}^{\infty} A_t/A_0$ for large q.

In contrast to M_q, the quantity $\hat{M}_q$ is better suited to compare the time correlations of very diverse trajectories because it is insensitive to variations of $\mathrm{E}[\delta\theta^2]$. Figure S7c-d plots the normalized MSR from Eq. (A.28) for the various trajectories and temperatures considered before in Figure S7c-d. We observe that the standardized MSR exhibits comparable behavior across various temperatures, suggesting that the trend of the unnormalized MSR observed in Figures 6b-c and S7a-b is just due to an increase in the bout angle amplitudes $\mathrm{E}[\delta\theta^2]$ with temperature, but not due to changes in the structure of their time correlations.

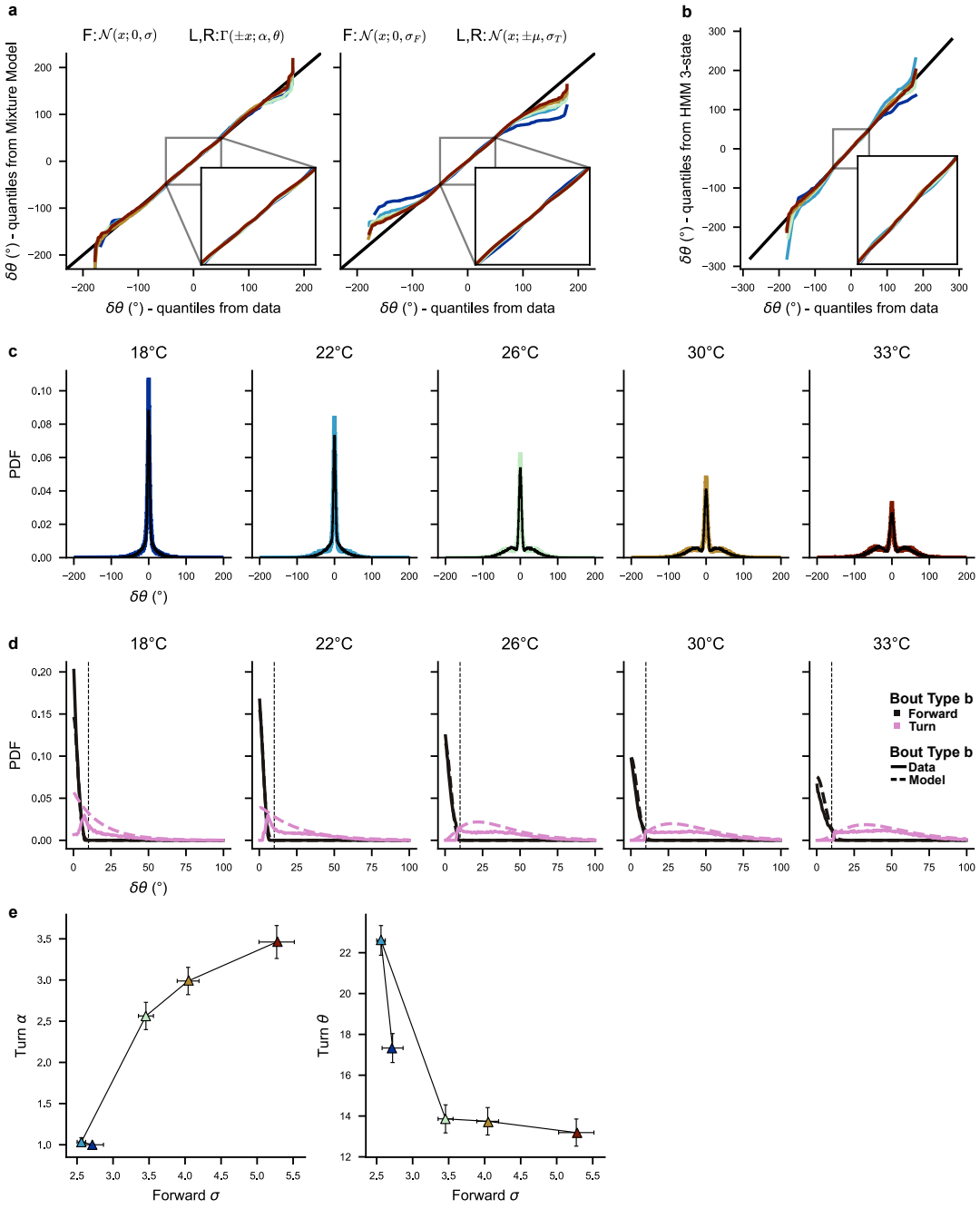


FIG. S2. **Supplementary panels to Fig.2 - Emission distributions**: **(a)** Quantile-Quantile plot between the empirical distributions of reorientation angles and two Mixture Models. *Left:* Mixture Model defined from a central Normal distribution (forward bouts) and two Gamma distributions (left and right turning bouts), corresponding to the model of HMM emissions. *Right:* Gaussian Mixture Model. *Insets:* Zoom on $\pm 50°$. **(b)** Quantile-Quantile plot between the empirical distributions of reorientation angles and the distributions of reorientation angles generated by HMM. *Insets:* Zoom on $\pm 50°$. **(c)** Empirical distributions of reorientation angles (colored) vs. distributions generated by the 3-state Hidden Markov Model (HMM; black), at each temperature. **(d)** Distributions of absolute reorientation angles labeled as forward bouts (solid black) and turning bouts (left or right; solid pink) by the Hidden Markov Model (HMM). Dashed lines show the HMM emission distribution for forward and turning bouts (black and pink respectively). The vertical black line shows the threshold $\delta\theta_0 = 10°$ used in the Markov Chain model. **(e)** HMM emission parameters : σ the standard deviation of the central Normal distribution (forward state), α and θ the shape and scale of the Gamma distribution (turning states). Each dot corresponds to one temperature, and error bars were computed from the minimum-maximum of 100 cross-validations (trained on randomly selected 50% of the datasets).

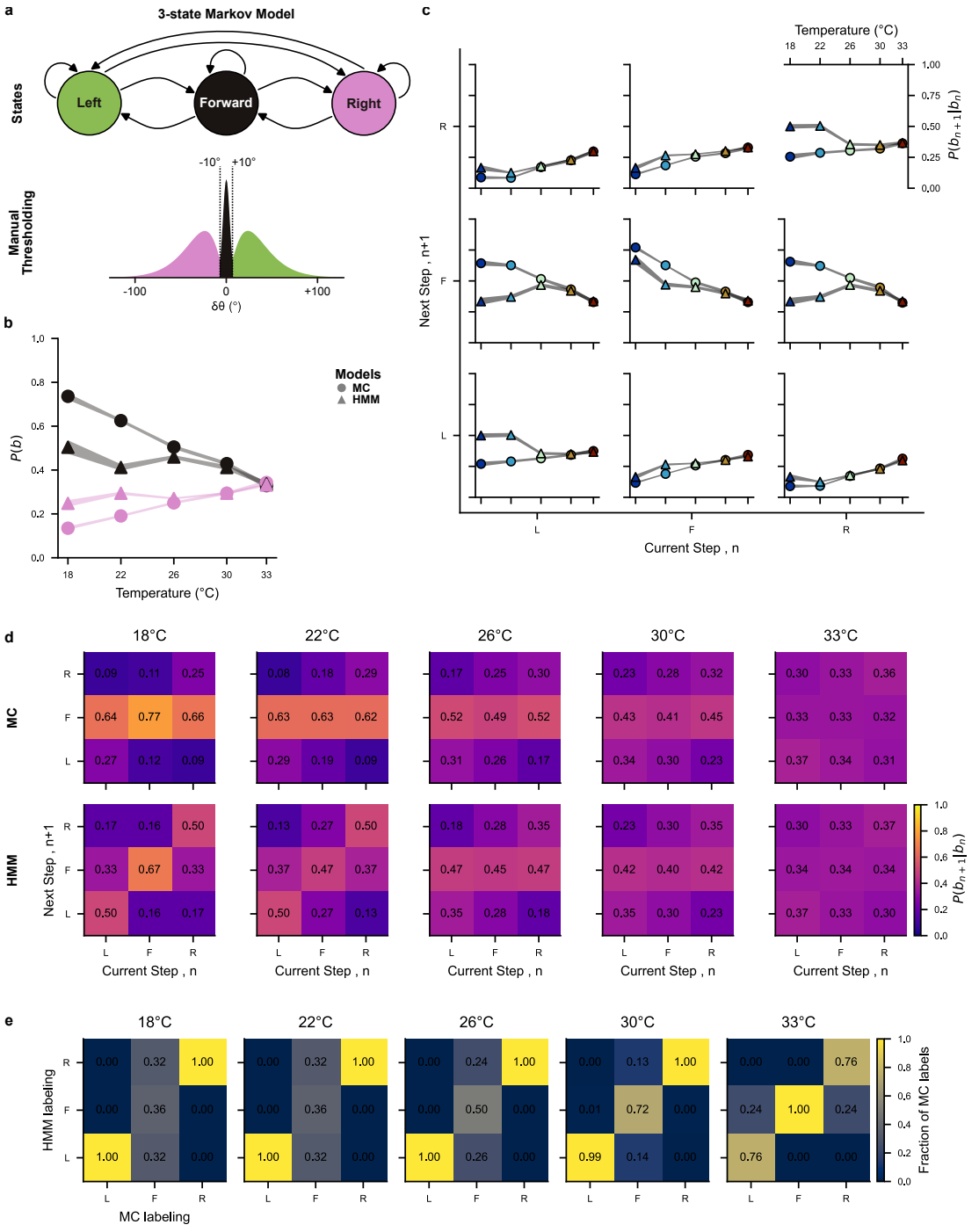


FIG. S3. **Supplementary panels to Fig.2 - Comparison between Markov Chain and Hidden Markov Model**: **(a)** 3-state Markov Model (HMM) with a threshold at $\pm\delta\theta_0 = 10°$ to classify reorientation angles $\delta\theta$ into forward/left/right states. **(b)** Steady state bout probabilities $P(s)$ vs. temperature, for forward ($s = F$, black) and turning bouts ($s \in L, R$, pink), and for both Markov Chain inferred from threshold (MC, circles) and Hidden Markov Models (HMM, triangles). **(c)** Transition probabilities $P(s_n \to s_{n+1})$ vs. temperature , for both Markov Chain (MC, circles) and Hidden Markov Model (HMM, triangles). **(b,c)** Shaded curves represent the minimum-maximum of 100 cross-validations of both models inferred from randomly selecting 50% of the data. **(d)** Transition matrices between forward (F), left (L) and right (R) states, for both the Markov chains inferred from thresholded data (MC, top) and Hidden Markov Model (HMM, bottom), at each temperature. **(e)** Confusion matrices between labeling of MC and HMM at each temperatures (normalized with respect to MC labeling).

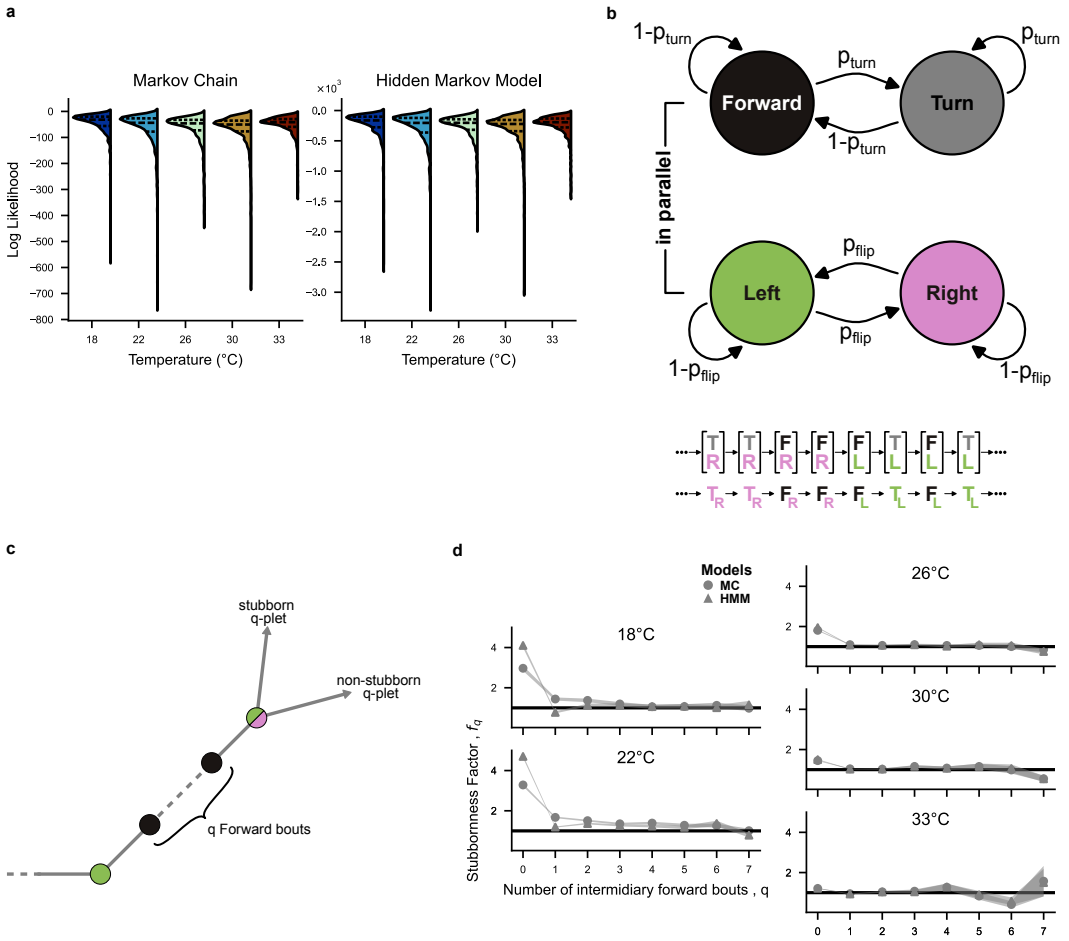


FIG. S4. **Supplementary panels to Fig.2** - Markovianity: **(a)** Distribution of log Likelihoods (LLHs) for both the Markov chains inferred from thresholded data (left) and Hidden Markov Model (right). For each model and each temperature, LLHs were computed for 100 models inferred from 50% of the trajectories (randomly constructed training set) and on the remaining 50% of the trajectories (testing set). Dashed lines show the quartiles of each distribution. **(b)** 4-state Markov chain used in previous publications [22, 24]. Two Markov Chains run in parallel, with the first chain controlling bout type (forward or turn) and the second controlling direction (left or right). The system can be in one of four states: $[T, L], [T, R], [F, L], [F, R]$, thus left and right states represent internal directional states (not just observed behavioral orientations). **(c)** Graphical definition of the stubbornness. For a q-plet of bouts $T_1 \rightarrow F \rightarrow \cdots \rightarrow F \rightarrow T_2$ with q intermediary forward bouts, a stubborn sequence is defined as one where directionality is conserved (i.e. $T_1 = T_2$), whilst a non-stubborn sequence will lose the memory of the initial turn (i.e. $T_1 \neq T_2$). **(d)** Stubbornness factor f_q (see Eq. 3) vs. number of intermediary forward bouts q, for both Markov Chain inferred from thresholded trajectories (MC, dots) and the Hidden Markov Model (HMM, triangles) trained directly from reorientation angles, at each temperature. The width of the shaded curves represent the estimated error in *stubbornness* factor (see Materials and Methods IV E).

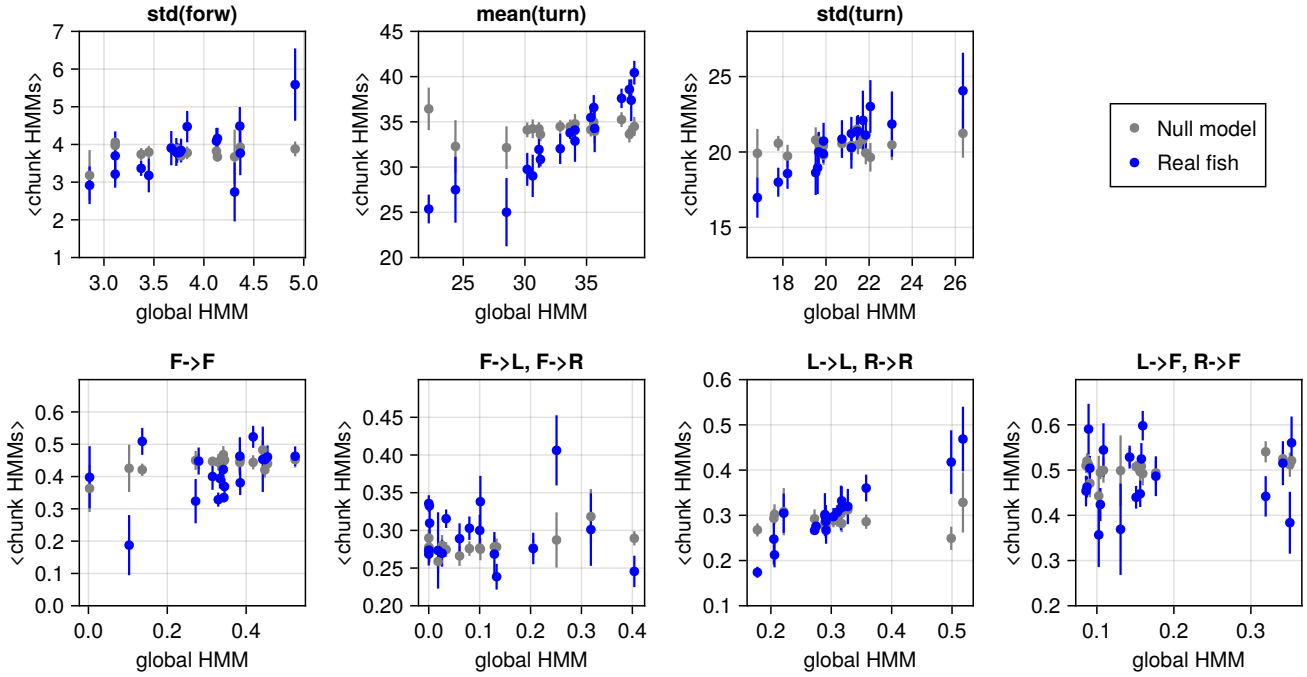


FIG. S5. **Supplementary panels to Fig.3** Hidden Markov Model parameters inferred from all trajectories from an individual fish, compared with the average parameters inferred from chunks of that fish's trajectories. All HMM parameters are shown. Each dot represents a fish, with error bars corresponding to standard error of the mean. Blue color corresponds to real individual fish data. Gray points are obtained by sampling long trajectories from a single HMM trained on all fish bundled together, thus representing a null model for the fish individuality.

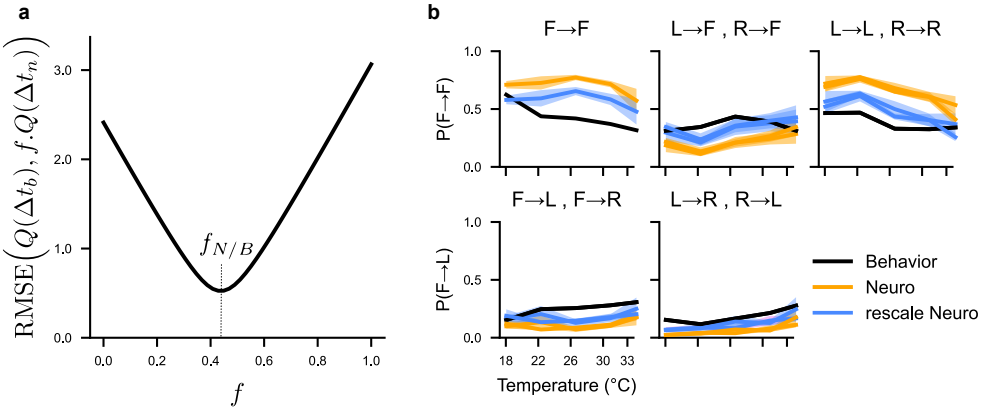


FIG. S6. **Supplementary panels to Fig.5 (a)** Root Mean Squared Error (RMSE) between quantiles of the behavior and neuronal sojourn distributions presented in Figure 5a at different values of the rescaling factor f. The optimal rescaling factor corresponds the minimal RMSE at $f = f_{N/B} \approx 0.44$. **(b)** Transition probabilities $P(s \to s')$ between hidden states F, L, and R, for the behavioral HMM (black), and neuronal HMMs before (orange) and after rescaling by $f_{N/B} = 0.44$ (magenta), at each temperatures. Shaded curves represent the standard error of the mean over all fish.

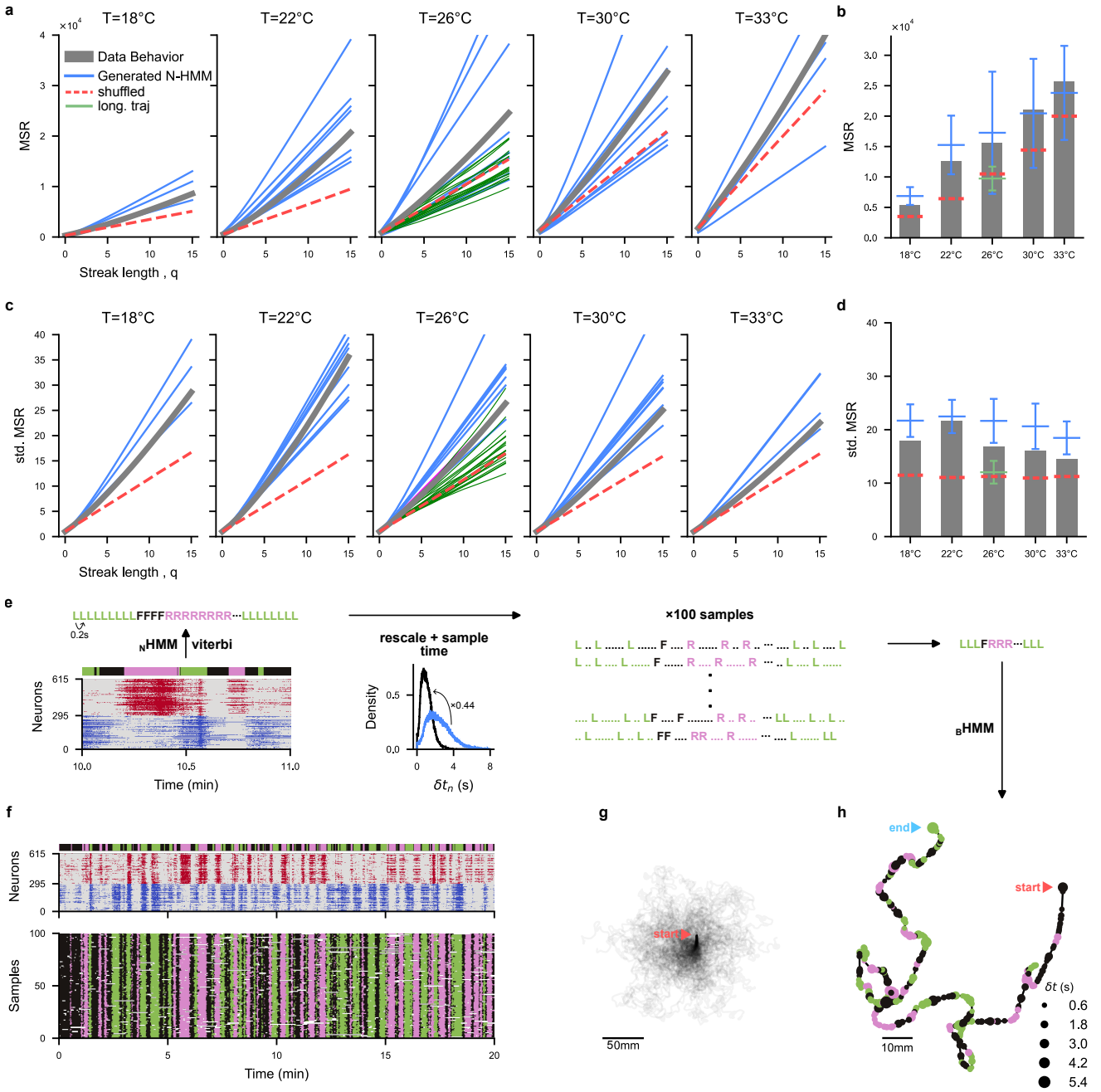


FIG. S7. **Supplementary panels to Fig.6 (a)** Mean Square Reorientation (MSR) after q bouts from aggregated multiple-fish trajectories at 18-33 ^{o}C (grey), long-individual trajectories at 26^o C (green) and trajectories generated from Neural HMM (N-HMM, blue). Red dashed lines are MSR obtained from shuffled aggregated multiple-fish trajectories. **(b)** MSR(q=10) for data and N-HMM-generated trajectories, with mean (horizontal bars) and standard deviations (vertical bars). **(c-d)** Same as panels a-b but for the standardized MSR where trajectories are normalized such that the bout angles have unit variance. See Eq. (A.28). **(e)** Pipeline to convert neuronal activity to swim trajectory. ARTR activity is first converted to the most likely sequence of forward/left/right hidden states using the Viterbi algorithm on the N-HMM. Time is then re-scaled using the scaling factor identified in Fig 5, and bout sequences are sampled based on the interbout interval distribution. A swim trajectory is constructed for each bout sequence by sampling the bout distances d_n and inter-bout intervals δt_n emission distributions from the behavioral HMM. **(f)** Example empirical ARTR activity at 26^{o}C (top) and corresponding state sequences after temporal re-scaling and bout sampling (bottom). **(g)** Reconstructed trajectories from the empirical neuronal activity presented in panel f (for each sampled state sequence). **(h)** Example reconstructed trajectory from the recorded ARTR activity presented in panel f.

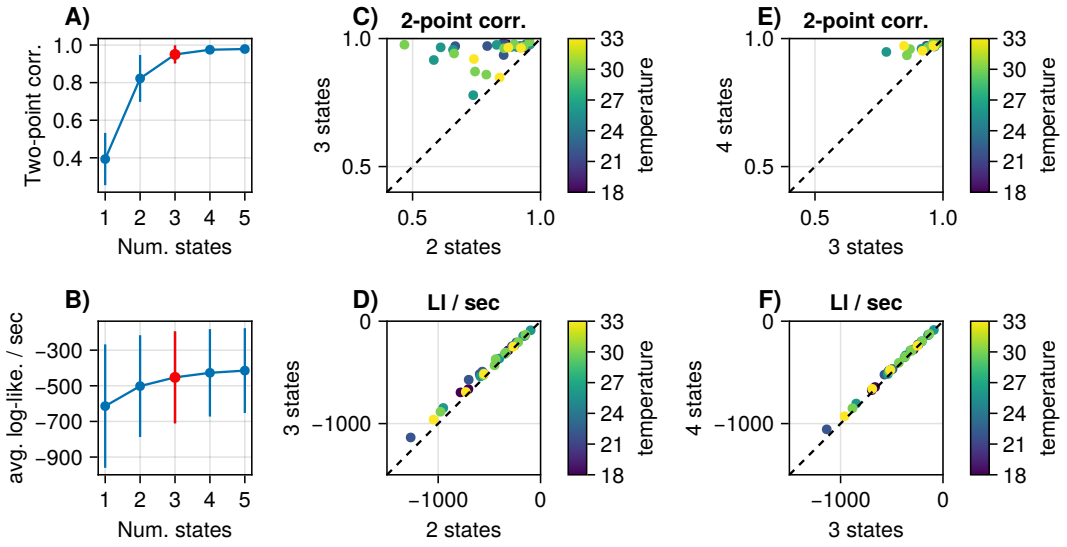


FIG. S8. **Cross-validation of 3-state HMM for Neural ARTR data A)** Correlation between two-point averages $\langle \sigma_i \sigma_j \rangle$ estimated by the HMM and their empirical counterparts, as a function of the number of hidden states in the HMM. This correlation is computed for each temperature, and the plot shows only the average and the standard deviation. **B)** Average log-likelihood per unit time of withheld data as a function of the number of hidden states in the HMM. This correlation is computed for each temperature, and the plot shows only the average and the standard deviation. **C)** Comparison of correlation of the HMM two-point averages to their empirical counterparts for each temperature, in the 2-state (x-axis) vs. the 3-state models (y-axis). **D)** Same as C), but comparing the log-likelihood per unit time. **E)** Same as C), but comparing 3-state HMM to 4-state HMM. **F)** Same as D), but comparing 3-state HMM to 4-state HMM.

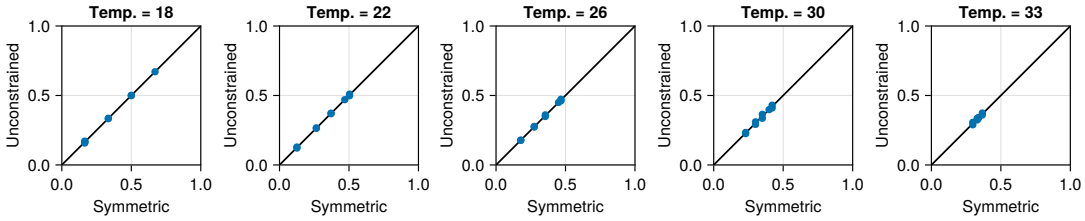


FIG. S9. **Comparion of transition probabilities inferred with unconstrained and symmetric behavioral HMMs.** We trained behavioral HMMs as defined in the main text on the first dataset. The figures compares the nine inferred transition probabilities $P(s \to s')$ (y-axis) defining the unconstrained HMM with their counterparts in the symmetric HMMs, in which Left-Right symmetry is imposed in the transition matrix. The excellent agreement confirms that the models spontaneously learn symmetric transitions. In the main text, we impose Left-Right symmetry to speed up training and improve accuracy. See also Fig. S10. In both cases the emission distributions for left and right bouts are the same.

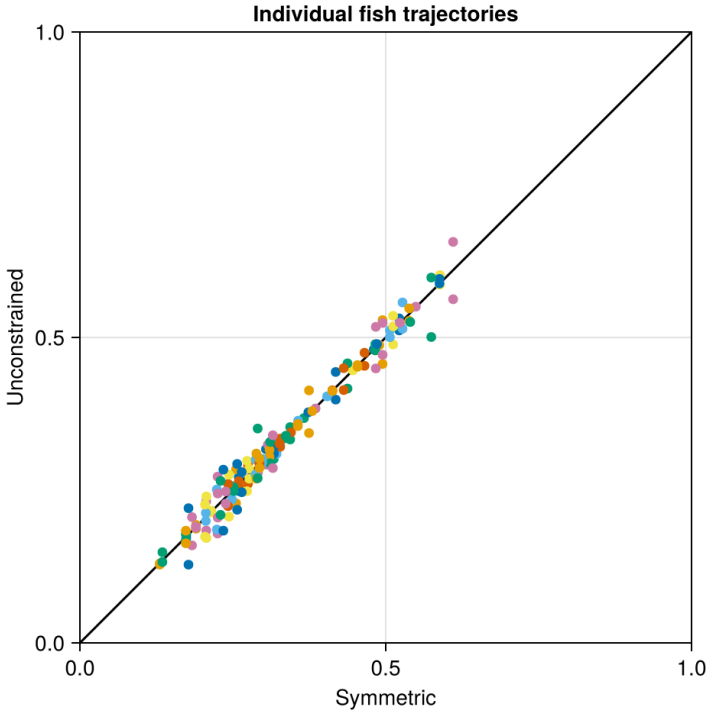


FIG. S10. **Comparison of transition probabilities inferred with unconstrained and symmetric behavioral HMMs on long trajectories.** Same as Fig. S9, but for long trajectories of individual fish. Different colors correspond to different fish.

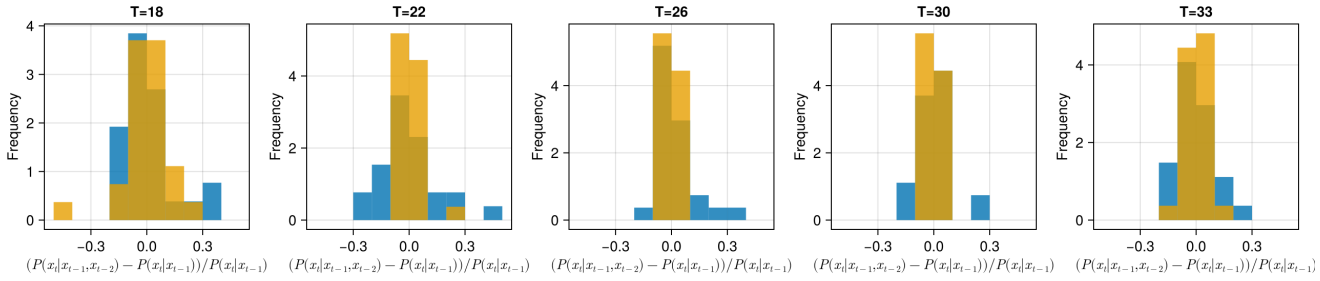


FIG. S11. Histogram of empirical $(P(x_t|x_{t-1}, x_{t-2}) - P(x_t|x_{t-1}))/P(x_t|x_{t-1})$ across experimental trajectories, where x_t are states (F, L, R) determined by thresholding bout angles at 10°. Actual data is shown in blue. For comparison, we generated fictitious trajectories from a Markov model with transition probabilities $P(x_t|x_{t-1})$ calculated from the empirical transition frequencies. The corresponding histogram is shown in yellow.

2.3 Discussion

Beyond the results already mentioned in the discussion of the article, two points are of particular importance to the work presented in this manuscript : inter-individual variabilities in behavioral dynamics, and the building of a neuronal model from the brain recordings of multiple fish.

Individuality and Dynamics. An important methodological contribution of this article to this manuscript is the use of Markov models to study behavioral dynamics, particularly across individuals. We discussed how the transition probabilities between behavioral states are affected by environmental variables such as temperature, and how those probabilities vary across fish allowing for a kind of phenotyping. This was possible because, while the behavior itself might be complex, it can be segmented into a set of states comparable between individuals and across environmental conditions. While we did not discuss it in the article, this can be extended to brain recordings as well. For example with the ARTR, the variability in transition probabilities across brain recordings was small enough that we were able to compare them with behavior. We could therefore imagine using the transition probabilities between neuronal states to compare the brain recordings of multiple animals. We will discuss this point further in chapter 4 where we will use a markovian framework to compare the dynamics of spontaneous neuronal activity across individuals.

Combined neuronal model of multiple fish. Another important contribution of this article to this manuscript is the use of a state model to map the neuronal activity of multiple individuals to a single latent space. Indeed, when trying to compare the activity of neurons in the ARTR between fish, we encounter two problems.
First, recordings of different larvae contain different number of neurons. This is in part due to variabilities in imaging conditions, but, more fundamentally : there are no biological constraints enforcing a circuit like the ARTR to contain the exact same number of neurons across individuals. This means that, for a state model like the HMM, the cardinality of the observable space (*i.e.* the emissions) needs to be adapted to each fish.
Second, the function of a circuit like the ARTR arises from the collective activation modes of its neuronal population. This means that, although the macro-scale activity of the circuit might be comparable between individuals, we can't map one-to-one the activity of individual neurons across brains.
Both of these problems make it challenging to create a single neuronal model which can be applied to all neuronal recordings. In the article, we solved this problem by training a HMM for each fish, with the same state space (3 states), but with emissions tailored to each brain (parametrized by fields h_i^s, see Equation 4 of the paper). The *meaning* of each state was not enforced, but emerged systematically as a left-dominant, a right-dominant, and a balanced state ofARTR activity. This means that while the emissions are individual-specific, the state space is common to all individuals, and thus the neuronal activity of different fish can be compared through this shared latent space. This point will be discussed further in chapter 3, where the same concept is used with another probabilistic model, with the goal of investigating how whole-brain spontaneous activity generalises across individuals.

Chapter 3

Building a shared and interpretable representation of spontaneous brain activity across multiple zebrafish larvae

Spontaneous brain activity exhibits structured dynamics that vary across individuals, making it challenging to identify conserved principles of neural organization.

Here, we develop a framework to uncover shared latent representations of spontaneous whole-brain activity in larval zebrafish, that does not require neuron-level correspondence. Using Restricted Boltzmann Machines (RBMs), we identify latent co-activation motifs, or cell assemblies, that generalize across animals and provide interpretable building blocks of population-wide activity. A novel bi-training approach constrains RBMs from different individuals to share a common latent space, enabling alignment of functional cell assemblies and translation of activity patterns between brains. We show that these shared representations are more stereotyped across fish than raw activity or correlation patterns, and retain spatial and statistical structure.

Our results reveal that the functional organization of spontaneous activity is conserved at the level of latent population codes, offering a new approach for comparing brain-wide dynamics across individuals.

This chapter is an article soon to be submitted for publication, written with :
Jorge Fernandez-de-Cossio-Diaz, Guillaume Faye-Bédrin,
Georges Debrégeas, and Volker Bormuth.

3.1 Introduction

Spontaneous brain activity is a fundamental feature of neural systems, shaping sensory processing, behavior, and internal state dynamics even in the absence of external stimuli [36, 174, 175]. In humans, population-level analyses of resting-state fMRI have revealed large-scale networks that are spatially and functionally conserved across individuals [176]. These so-called resting-state networks (RSNs) have been interpreted as a manifestation of intrinsic brain organization and are increasingly used to study brain development [177], aging [178], and psychiatric disorders [179]. However, their interpretability is limited by low spatial and temporal resolution.

At the other end of the spectrum, exhaustive recordings in small organisms, such as *C. elegans*, have enabled dynamic models that directly link activity to connectome structure. These models can predict transitions between behavioral states and reveal low-dimensional manifolds underlying brain-wide activity [180, 181, 182]. However, they are largely restricted to small nervous systems with stereotyped anatomy and limited neuron count.

In between these two extremes lies a major challenge: how to model spontaneous activity at single-cell resolution in vertebrates with non-stereotyped, brain-wide architectures at the single-cell level, while retaining interpretability and enabling cross-individual comparisons. Zebrafish larvae offer a unique opportunity to fill this gap. Their optical transparency and compact brains enable functional imaging of nearly all neurons at single-cell resolution over extended periods [126, 125]. Recent studies have leveraged this animal model to study sensory-motor transformations [171, 116, 172], behavioral state dynamics [183, 81], spontaneous brain activity [11, 54, 95], or internal models [184].

Approaches to inter-individual comparison in neuroscience have mainly relied on anatomical registration combined with stimulus- or behavior-guided voxel-wise averaging or clustering. Although useful, these methods assume voxel-level correspondence and overlook the rich variabilities at neuronal scales. More recent frameworks such as Shared Response Models (SRMs) [69], hyperalignment-attempt [67, 185, 186, 187], spatial autoencoders [188] and contrastive frameworks for cross-subject alignment [189] aim to learn shared embeddings across subjects. However, most analyses remain limited to individual subjects or rely on anatomical alignment combined with stimulus- or behavior-guided averaging, making them unsuitable for extracting shared structure from spontaneous activity at the neuron population level.

Probabilistic generative energy-based models such as Restricted Boltzmann Machines (RBMs) offer a powerful alternative [190]. They model the joint distribution of neural activity via latent variables (hidden units) that can be interpreted as co-activating neuronal assemblies. RBMs have been used to model retinal activity, hippocampal replay [191, 192], and patterns of spontaneous activity [11, 193]. However, these studies have generally focused on single datasets, and it remains unclear whether RBMs can be used to identify conserved structure across individuals, especially when applied to full-brain recordings at cellular resolution.

Here, we introduce a framework for discovering shared, interpretable latent structure in spontaneous brain-wide activity across individuals. Using whole-brain calcium imaging

in larval zebrafish, we train RBMs with a shared latent space, a strategy we term bi-training, in which the hidden units are constrained to maintain their prior functional organization and activation distribution across individuals. This enables us to extract a common set of latent motifs, corresponding to spatially organized assemblies, that are reused across different brains. Crucially, we go beyond statistical comparison by projecting activity from one fish into the shared latent space and decoding it into the neural space of another fish. We show that spontaneous patterns are translatable between individuals, while preserving both spatial correlations and probabilistic plausibility.

Our approach fills a key gap between connectome-informed modeling in *C. elegans* and population-level RSN analysis in humans. It enables comparative, interpretable, and probabilistically grounded modeling of spontaneous brain-wide activity across individuals. This provides a concrete route for identifying shared circuit motifs that shape spontaneous dynamics at the mesoscale, bringing us closer to a common representational framework for a vertebrate brain.

3.2 Results

3.2.1 Latent representations of spontaneous brain-wide activity are variable and non-alignable across individual RBMs

We showed previously in van der Plas et al. [11] that spontaneous whole-brain neuronal activity of zebrafish larvae could be modeled by Restricted Boltzmann Machines (RBMs). In this framework, the binarized spiking activity of individuals neurons is mapped onto a latent space comprised of hidden units which we showed correspond to functional ensembles of co-activating neurons (Fig. 3.1 A). They are characterized by spatial distribution of weights $w_{i\mu}$ between neuron i and hidden unit μ which are sparse, spatially compact, and correspond to functional circuits and anatomical structures. This suggests that hidden units represent cell-assemblies within the model. Furthermore, we showed that an effective couplings J_{ij} between neurons i and j could be inferred from the model (see Methods 3.4.4.5). Importantly, J_{ij} is not a measurement of the correlation between neurons, but reflects a model of connectivity which takes into account indirect network interactions (see Fig. 3.1 B). We showed previously that this measure of functional connectivity reflects, at least in part, structural connectivity measured between brain regions.

Because both the latent ensembles and the couplings are biologically interpretable, one might hope that RBMs trained on the *same* data (or even on data from different fish) would converge on similar representations, which could then be used to compare multiple trainings and individuals.

RBMs produce stable functional connectivity model only at coarse-grained scales
To test this hypothesis, we trained multiple RBMs on the same neuronal dataset (fish 1

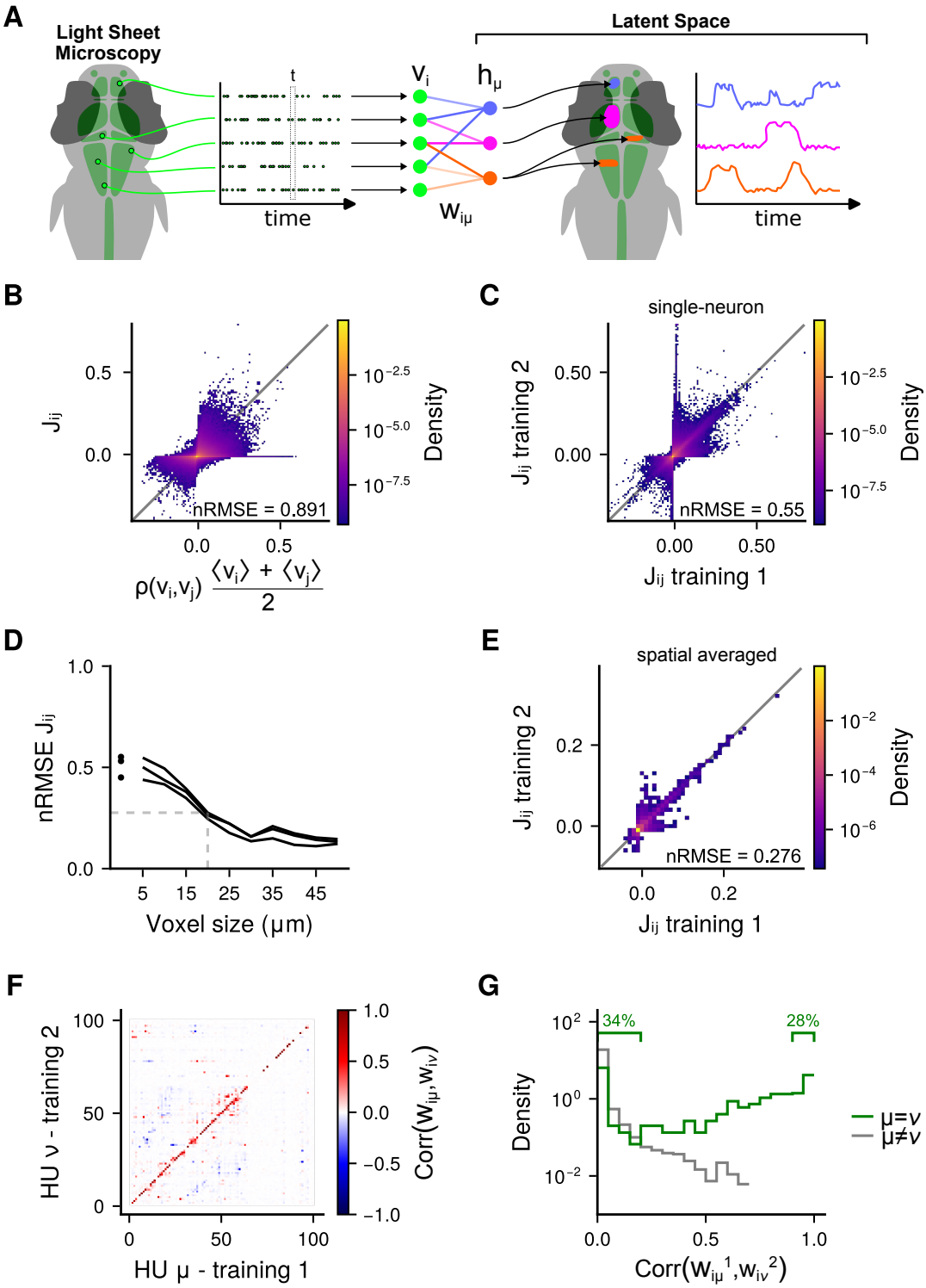


Figure 3.1: *(Caption on next page)*

Figure 3.1: **Restricted Boltzmann Machines (RBMs) produce degenerate representations of neuronal functional organization.** **A**: Zebrafish larvae neuronal activity is recorded using Light Sheet Fluorescence Microscopy and deconvolved into binarized spike trains. RBMs are trained with binary visible units v_i corresponding to neurons and connected to dReLU hidden units h_μ via a weight matrix $w_{i\mu}$ (see Methods 3.4.4). The hidden layer describes a latent space capturing the activation modes of the whole neuronal population as groups of co-activating neurons [11]. **B**: Rescaled pairwise Pearson correlation $\rho(v_i, v_j)\frac{\langle v_i\rangle+\langle v_j\rangle}{2}$ between neurons i and j vs. coupling matrix J_{ij} (see Eq. 3.14) inferred from an example RBM. **C**: Comparison of two example coupling matrices inferred from two RBMs trained on the same neuronal recording. **D**: nRMSE (see Methods 3.4.7) between coupling matrices inferred from three different training on the same neuronal recording. Dots correspond to J_{ij} matrices as presented in panel C. Lines correspond to coupling matrices spatially averaged on a cubic-voxel grid : $J_{mn} = \frac{1}{|\mathcal{V}_m|\,|\mathcal{V}_n|}\sum_{i\in\mathcal{V}_m}\sum_{j\in\mathcal{V}_n} J_{ij}$ with $\mathcal{V}_m$ and $\mathcal{V}_n$ the sets of neurons in voxels m and n respectively. **E**: Comparison of two example coupling matrices J_{mn} spatially-averaged at $20\mu m$ voxel size. **F**: Pairwise Pearson correlation between the weight matrices $w_{i\mu}$ and $w_{i\nu}$ of two example RBMs trained on the same neuronal recording. The matrix rows have been reordered to obtain the best alignment between the two trainings. **G**: Distribution of correlations from panel F, for best alignment pairs ($\nu = \mu$, green) and other pairs ($\nu \neq \mu$, grey).

in Tab. 3.1), with the same training parameters and performance (see Methods 3.4.6.1 and 3.4.4.4). We compared the effective couplings J_{ij} inferred from the different RBMs. We found that, while they were mostly comparable between training, coupling were highly variable, with a significant portion of neuron pairs being strongly coupled in one model and completely decoupled in another (Fig. 3.1 C). However, when couplings are spatially averaged between cubic-voxels of side $\gtrapprox 20\mu m$, we found that J_{ij} was much more stereotypical between trainings (Fig. 3.1 D-E). This suggests that, while the inferred couplings between neurons is degenerate between RBM trainings, it is not at the scale of cell assemblies or brain regions.

RBMs represent neuronal statistics from a degenerate set of latent features Next we investigated whether different trainings produce the same latent space description of neuronal activity. To do so we compared the hidden units inferred from multiple training as described above. We started by computing the Pearson correlation $\rho_{\mu\nu} = \rho(w_{i\mu}, w_{i\nu})$ between the weight matrices $w_{i\mu}$ and $w_{i\nu}$ of two trainings, and found that, for the best alignment, only $\approx 28\%$ of hidden units could be unequivocally paired between trainings ($\rho_{\mu\nu} > 0.9$), while $\approx 34\%$ had no matching pair ($\rho_{\mu\nu} < 0.2$) (see Fig. 3.1 F-G). This suggests that a core set of hidden units can be systematically identified across trainings, but a large number cannot.
We also performed the same analysis on the expected hidden activity $h_\mu(t) = \mathbb{E}[h_\mu \mid \mathbf{v}(t)]$ by computing the correlation $\rho_{\mu\nu} = \rho(h_\mu(t), h_\nu(t))$, and found that $\approx 54\%$ of hidden units had correlations $\rho_{\mu\nu} > 0.9$ and $\approx 18\%$ $\rho_{\mu\nu} < 0.2$ (see Supp. 3.5). This result reflects the fact

that, despite being connected to different neuronal populations, many hidden units share similar temporal signals, reflecting dominant dynamics of brain activity spanning large brain regions.

In summary, we found that RBMs provide degenerate representations of neuronal activity, even for the same dataset, and under the same architecture and training parameters. This reflects the probabilistic nature of the model and its training algorithm, as well as the fact that the RBMs try to model causal relationships between neurons even though multiple explanations can fit the data. Inferred couplings between neurons are degenerate, which can be explained by a different modeling of indirect connections between neurons. However, this degeneracy collapses at coarse grained scales, suggesting that the model provides stable description of the population-level functional connectivity. Lastly, we found that multiple RBMs could represent neuronal activity with the same statistical accuracy but through a different set of latent feature. These findings caution against naively comparing RBMs trained on different individuals and motivate multi-subject training paradigms that explicitly align latent spaces.

3.2.2 A global voxel-level RBM uncovers shared structure in spontaneous brain activity across individuals

We first asked whether spontaneous whole-brain activity across individual zebrafish larvae could be described by a common latent representation, despite inter-individual anatomical and functional variability.

To enable cross-individual comparisons, we spatially registered six fish to a common brain coordinate system using affine and non-linear transformations (see Methods: Morphological Registration). Each neuron was then assigned to a voxel in a 3D grid ($20\ \mu m$ cube size), and voxel activity was computed as the mean of the normalized fluorescence ($\Delta F/F$) of all neurons within it (Fig. 3.2 A and Methods 3.4.5.1). Only voxels that contained at least two neurons in each fish were retained, ensuring that anatomical coverage was consistent across individuals. This filtering step yielded a total of 2,995 shared voxels, with an average of 11 ± 5 neurons per voxel (Supp. 3.6 D, mean and standard deviation). Importantly, this voxel size was the smallest possible to conserve $> 75\%$ of all neurons in all fish, thus balancing spatial resolution with inter-individual consistency (Supp. 3.6 B).

We then trained a Restricted Boltzmann Machine on the concatenated voxelized activity from all fish (Fig. 3.2 A-B). This global RBM consisted of 2995 Gaussian visible units (corresponding to the number of shared voxels) and 40 hidden units. Training was performed using L_1^2-regularization on the weights with a regularization parameter of $\lambda_{21} = 0.1$. These hyperparameters were found by optimization via cross-validation (see Methods 3.4.5.2 and Supp. 3.7). Before training, we z-scored each voxel's activity within each fish to account for differences in mean and variance across individuals [156]. Without this normalization, voxelized training resulted in the RBM capturing fish identity rather than shared features.

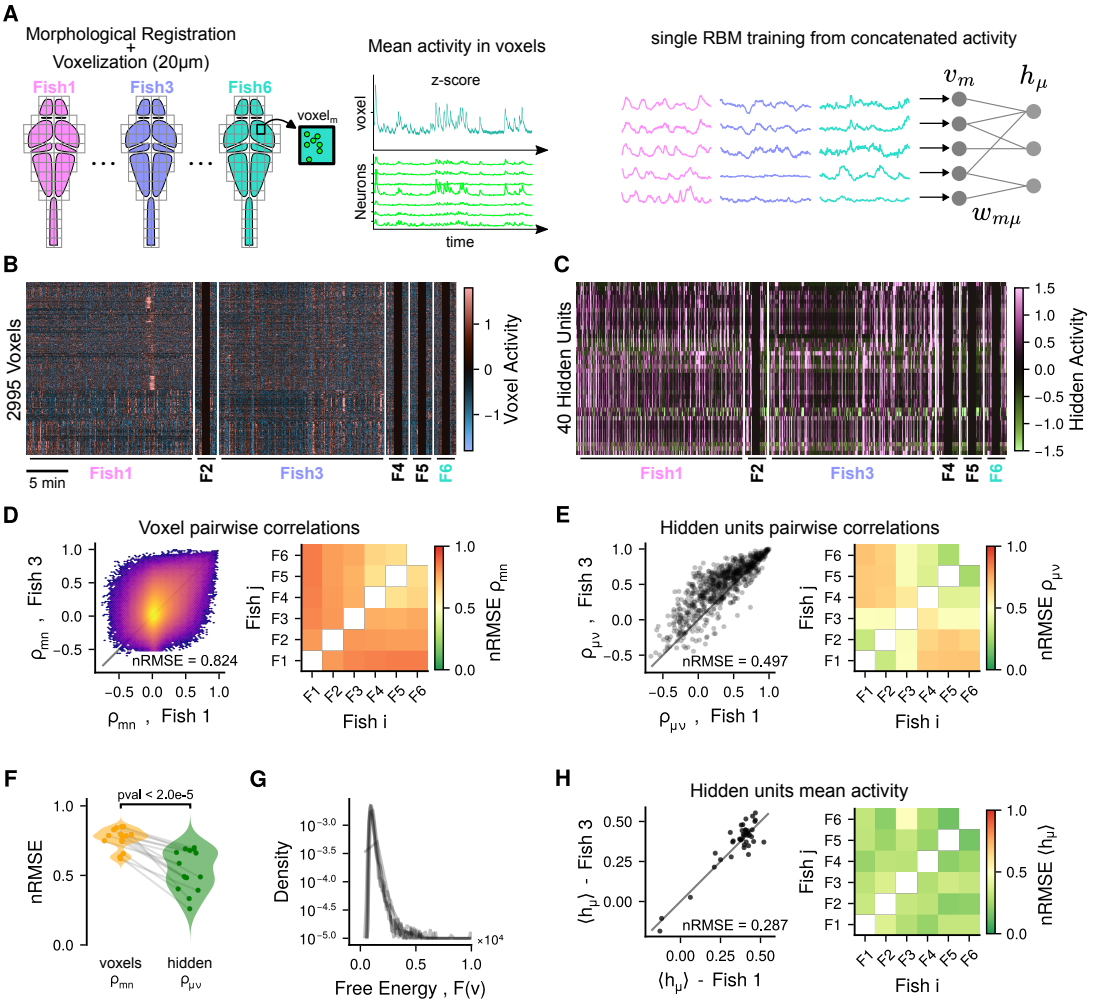


Figure 3.2: **Concatenation and voxelization allow for common RBM representation.** **A**: Pipeline of voxelized RBM training. Averaged brain scans are first morphologically registered to a common reference brain. A grid of cubic voxels with side $20\mu m$ is then used to group neurons spatially. The neuronal activity within each voxel is averaged and z-scored. The voxel activity for all 6 fish is concatenated to construct one dataset which is then used to train an RBM (see Methods 3.4.5) **B**: Concatenated voxelized activity of all 6 fish. Fish 1 and 3 are shown in full, while the activity of fish 2, 4, 5 and 6 are truncated for visualization purposes. **C**: The activity of panel B *translated* into the hidden space of the RBM (see Methods 3.4.6.4). **D**: Stereotypy of the pairwise Person correlation coefficient ρ_{mn} between voxels m and n. Left: example fish pair. Right: nRMSE (see Methods 3.4.7) of ρ_{mn} between all pairs of fish. **E**: Stereotypy of the pairwise Person correlation coefficient $\rho_{\mu\nu}$ between hidden units μ and ν. Left: example fish pair. Right: nRMSE of $\rho_{\mu\nu}$ between all pairs of fish. **F**: Distributions of fish-to-fish nRMSE for pairwise Pearson correlations between voxels (left orange, same as panel D right) and between hidden units (right green, same as panel E right). Each dot represents a pair of fish. The p-value was computed with a one-tailed Mann-Whitney U test. **G**: Distribution of free energy $F(\mathbf{v})$ (see Methods 3.4.4) of voxel activity configurations $\mathbf{v}$. One line per fish. **H**: Stereotypy of the mean activity $\langle h_\mu \rangle$ of hidden units μ. Left: example fish pair. Right: nRMSE of $\langle h_\mu \rangle$ between all pairs of fish.

The global RBM is a good model for data from individual fish. Validation of this global RBM model on each fish separately demonstrated that the model accurately captured pairwise voxel correlations in 5 out of 6 fish, and more generally, provided a good statistical model of voxelized activity across individuals (see Supp. 3.8). This indicates that, despite voxel-level variability, the RBM can extract a common statistical structure. This is particularly notable because no fish-specific information was provided during training.

Latent representations are more stereotyped than raw activity. Projecting each fish's activity into the RBM's hidden space (Fig. 3.2 B-C) revealed that correlations between hidden units were more conserved across individuals than correlations between voxels (Fig. 3.2 D-F). This was quantified using the normalized Root Mean Squared Error (nRMSE). Pairwise voxel correlations had high fish-to-fish variability (nRMSE up to 0.8) with weak voxel size dependence (Supp. 3.6 G). In contrast, hidden-unit correlations were significantly more stereotyped (nRMSE 0.3–0.7, Mann–Whitney U test, $p < 0.01$). Interestingly, we observed that fish from the same clutch (F1–F3 vs. F4–F6) grouped together based on their latent representations (see Fig. 3.2 G), despite no explicit labeling. This suggests that the RBM captures subtle commonalities in spontaneous activity linked to biological or experimental context.

RBM captures shared structure without encoding individual idiosyncrasies. To verify that the model did not simply encode for fish-specific features, we considered evaluating the probability $P(\mathbf{v})$ of neuronal configurations. Evaluating $P(\mathbf{v})$ is computationally impractical and we used the free energy as an alternative :

$$F(\mathbf{v}) = -\log Z - \log P(\mathbf{v}) \tag{3.1}$$

where Z is the partition function (see Methods 3.4.4.1). We found that free energy distributions for all fish overlapped closely (Fig. 3.2 G), indicating that each individual's activity was equally well-explained by the model.

We further compared the mean hidden activity $\langle h_\mu \rangle$ across fish, and found it to be comparable between individuals (Fig. 3.2 H), suggesting that the RBM did not merely classify individuals.

Overall, these findings confirmed that the latent space was shared and not specialized to any specific fish. The dynamics of activity in this space might be different for different fish, but the overall distribution is similar. As such, it describes the common realm of possible hidden configurations $\mathbf{h}$.

Despite anatomical variability, spontaneous brain activity across zebrafish larvae can be represented in a shared latent space. A single RBM trained on voxelized and concatenated data successfully captures second-order activity statistics across individuals. Importantly, latent representations show greater inter-fish stereotypy than voxelized activity, revealing higher-order regularities in collective dynamics. Crucially, this increased alignment is not trivially explained by dimensionality reduction but rather reflects the ability of the RBM to

extract conserved co-activation motifs that generalize across animals. These findings establish the feasibility of constructing unified, interpretable representations of spontaneous neural activity across individuals.

3.2.3 Bi-trained single-cell resolved RBMs reveal conserved spatial cell assemblies via a shared latent space

While voxel-based registration enabled us to construct a common latent space at a coarse spatial resolution, this approach sacrifices single-neuron interpretability. Next, we asked whether we could preserve interpretability while enforcing a shared representation of spontaneous whole-brain activity across individual fish.

We developed a training framework in which a reference RBM, the *teacher*, defines the latent space, and subsequent *student* RBMs are trained on different fish while being constrained to remain in the same latent space (Fig. 3.3 A). Each RBM models the binarized spiking activity of 43480 ± 2755 neurons (mean and std, see Materials 3.1) recorded in individual fish at single-cell resolution.

In practice, we first trained the teacher RBM on one fish, with binary visible units, 100 dReLU hidden units to define the latent space, and with L_1 normalization on the weights with $\lambda_1 = 0.02$ determined by cross-validation [11] (Methods 3.4.6.1). Each hidden unit in this trained teacher model represents a distributed functional assembly, a spatial pattern of coactive neurons, as defined by its weight vector $w_{i\mu}$ (see examples in Fig. 3.3 B and Supp. 3.11 B). Student RBM have the same number of hidden units, but can have a different number of visible units to account for variability in neuron-count per fish. To initialize the student RBM parameters, we interpolated these spatial weight patterns from the trained teacher onto the neuron positions in the student brains, we copied the hidden unit potentials $\mathcal{U}_\mu(h)$ of the teacher, and we estimated the fields of the visible units from the corresponding neuron's baseline activity in the student (see Methods 3.4.6.2). These pre-initialized student RBMs could not at this stage capture the statistical structure of their respective fish, as they had not yet been trained on the new data. If trained without additional constraints, the student representations would adapt solely to their local data and gradually drift away from the teacher's latent structure.

We therefore introduced a new paradigm, which we call bi-training, in which the student RBM is trained not only to maximize the likelihood of its own observed neuronal activity statistics, but also to capture the distribution of hidden-unit activations sampled from the teacher RBM:

$$\mathcal{L} = (1-\lambda)\,\langle \log P(\mathbf{v}^S_{\text{data}})\,\rangle + \lambda\,\langle \log P(\mathbf{h}^T_{\text{sampled}})\,\rangle \tag{3.2}$$

where $\mathbf{v}^S_{\text{data}}$ is the neuronal activity of the student, $\mathbf{h}^T_{\text{sampled}}$ are hidden configurations sampled from the teacher RBM, and $\lambda \in [0, 1]$ is an hyperparameter controlling the relative

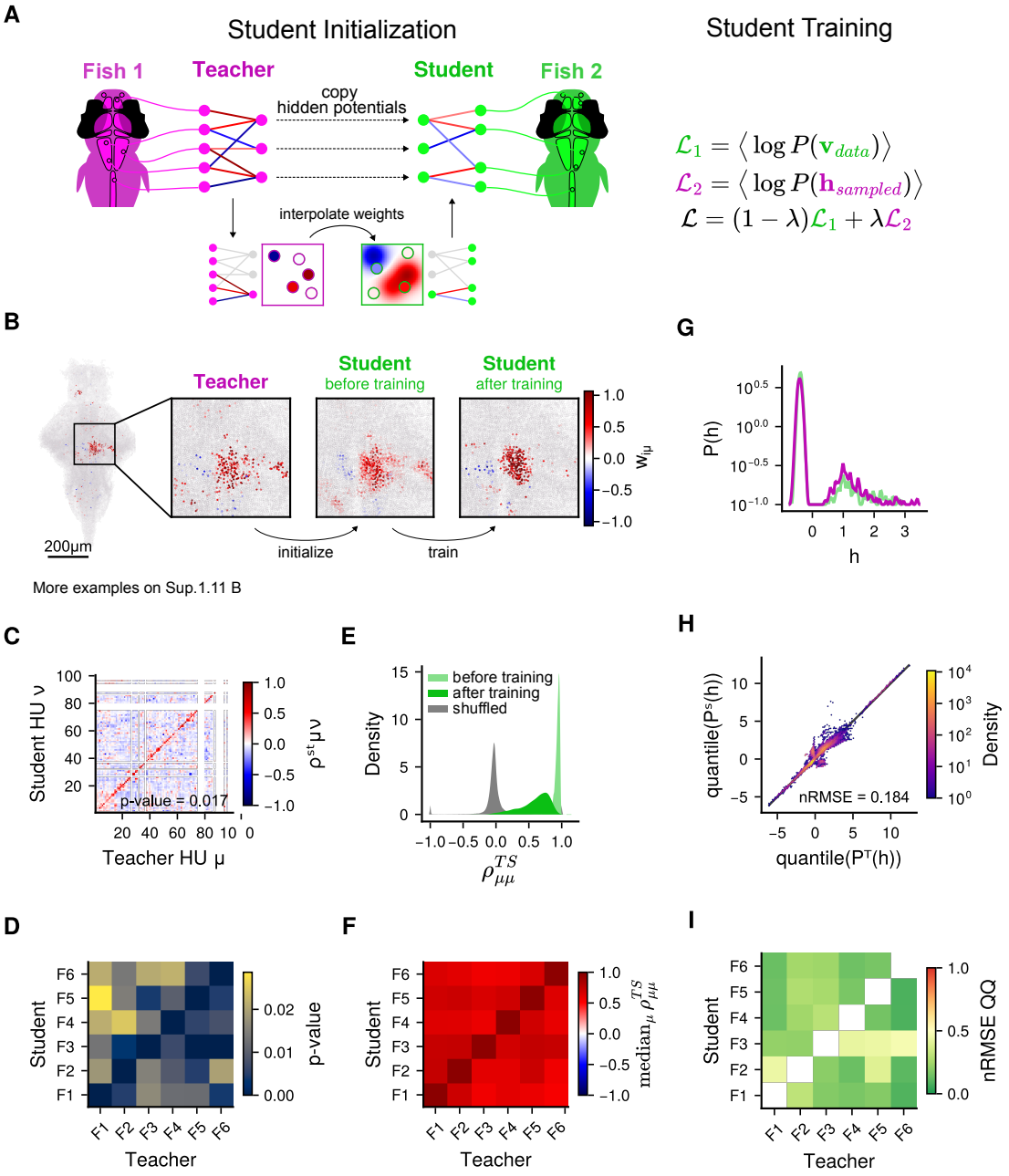


Figure 3.3: *(Caption on next page)*

Figure 3.3: **RBMs can be trained from priors and constrained to enforce a similar hidden space.** **A**: Teacher/Student paradigm. A *student* RBM is initialized from a pretrained *teacher* RBM by copying its hidden layer and spatial interpolation of its weights. The student RBM is then trained on both student neuronal data and hidden configurations sampled from the teacher RBM. **B**: Example spatial interpolation of weights. Left: z-projection map and inset of teacher weights $w^T_{i\mu}$ for each neuron i and a single hidden unit μ. Middle: interpolated weights in student. Right: weights after student training. **C**: Example spatial correlation matrix $\rho^{TS}_{\mu\nu}$ (see Methods 3.4.8) between weights maps of hidden unit μ of teacher RBM and hidden unit ν of student RBM. **D**: P-value of diagonal dominance of the $\rho^{TS}_{\mu\nu}$ matrix (see Methods 3.4.9) for every pair of teacher and student. **E**: Distribution of spatial correlations $\rho^{TS}_{\mu\mu}$ between corresponding hidden units in all teacher-student pairs for students before training (light green), students after training (green), and shuffled pairs of hidden units (grey). **F**: Median of spatial correlation $\mathrm{median}_\mu\, \rho^{TS}_{\mu\mu}$ for every teacher-student pair. **G**: Prior distribution of hidden value $P(h)$ for the example hidden unit presented in panel B, in both teacher (magenta) and trained student (green) RBMs. **H**: Quantile-quantile (Q-Q) plot of $P(h)$ in the example teacher and student pair for all hidden units combined. **I**: nRMSE (see Methods 3.4.7) of Q-Q plot in panel H for every teacher-student pair.

impact of each term on training (see Methods 3.4.6 for the detailed method and Fig. 3.3 A for a diagram). This approach ensures that all student RBMs can access the same subspace of hidden configurations while retaining sufficient flexibility to adapt to individual neuronal datasets.

Bi-trained students converge faster and more reliably. The bi-trained student RBMs converged rapidly, requiring only one-tenth the number of training steps compared to unconstrained and randomly initialized models. This suggests that the initialization step acts as kind of pretraining, and student training only finetuned the model to match student-statistics. Indeed we found that, contrary to teachers where multiple RBMs needed to be trained to obtain a good model, student RBMs were consistently able to reproduce the first- and second-order statistics of neuronal activity, sometimes even exceeding that of the teachers (see Supp. 3.9 A-B and Supp. 3.10). Despite being trained with a latent-space constraint, student RBMs remained faithful to the distribution of neuron firing rates and pairwise covariances in their respective datasets. Notably, even though the hidden-unit correlations were not directly constrained during training, we found that their pairwise covariances were partially preserved between the teacher and student (see Supp. 3.9 C-D). This consistency across animals supports the hypothesis that the RBM's hidden units reflect stereotyped functional connectivity motifs.

Hidden units retain their spatially localized patterns. Each hidden unit in the teacher RBM represents a spatially localized pattern of co-activating neurons. After interpolating the teacher weights into a student RBM and training the model, we observed that these spatial patterns were largely retained (examples in Fig. 3.3 B and Supp. 3.11 B). The spatial

correlation between corresponding hidden units in the teacher and student RBMs was on average above 0.6 and significantly higher than between non-matching pairs (see Fig. 3.3 C-E, bootstrap p-value ≈ 0.02, see Methods 3.4.8 and 3.4.9), indicating that most hidden units can be reliably matched between models based on their spatial weight patterns. We observe that student weight maps were generally of lower amplitude and more compact, i.e., involving fewer strongly connected neurons $N_\mu^S = |w_{i\mu} > 10^{-5}|_i$ with $N_\mu^S \approx \frac{1}{2} N_\mu^T$ (see Supp. 3.11 A-B), this is consistent with the weight regularization applied during training. Crucially, the core spatial identity of the assemblies was maintained.

Student RBMs preserve the teacher's hidden-unit priors. Finally, we examined whether student RBMs accessed the same regions of the latent space as the teacher. We compared the prior distributions $P(h_\mu)$ of each hidden unit and found close agreement between the teacher and student RBMs (Fig. 3.3 G-I). Quantile–quantile (Q–Q) plots showed strong linear correspondence (Fig. 3.3 H), and the identity error (nRMSE) across all teacher-student pair was low (Fig. 3.3 I). This indicates that, despite anatomical and statistical variability, spontaneous activity across individuals can be mapped to the same latent sub-space.

Together, these results demonstrate that RBMs trained on different individuals can share a common latent space in which functional assemblies are reliably matched based on spatial structure and prior activation. However, this alone does not guarantee that individual brain states (*i.e.* specific whole-brain spontaneous activity patterns) are encoded similarly across fish

3.2.4 Cross-individual decoding confirms conserved encoding of spontaneous activity in the shared latent space

To assess whether spontaneous activity patterns generalize across individuals, we tested whether they are similarly encoded in the shared latent space. Specifically, we asked whether a neuronal configuration $\mathbf{v}^T$ from a teacher fish, encoded as the expected hidden activity $\mathbb{E}[h^T \mid \mathbf{v}^T]$, could be decoded by the student RBM into an activity pattern $\mathbb{E}[\mathbf{v}^S \mid h^T]$ (see Methods 3.4.6.4) that is both highly probable under the student model and has the same spatial structure as the original teacher pattern. Because the mapping is bidirectional, it allowed us to benchmark both teacher-to-student ($T \to S$) and student-to-teacher ($S \to T$) translations against within-fish reconstructions and randomized controls (Fig. 3.4 A). We show in Movie 3 and Movie 4 two example movies of $\mathbf{v}^{T\to S}$ and $\mathbf{v}^{S\to T}$ respectively to illustrate this activity translation. Qualitatively, these movies show that this procedure is indeed capable of generating activity which resembles real brain activity, with translated patterns recapitulating salient features of spontaneous brain-wide activity, including spatially organized assemblies and bilateral coordination.

Statistical features are preserved during mapping. We first compared the mean activity $\langle v \rangle$ and pairwise covariances $\langle vv \rangle$ of translated versus empirical neuronal configurations. In both directions and for all pairs ($T \to S$ and $S \to T$), translated configurations

preserve mean activity with high fidelity, and captured substantial pairwise covariances, except for a few outliers discussed bellow (see Fig. 3.4 B-C). These values were comparable to, and in some cases better than, within-animal reconstructions, indicating that translation via the latent space preserved meaningful structure. A detailed look at these outliers reveals that we fail to reproduce the mean activity $\langle v_j \rangle$ of 2% to 18% of their neurons (see Supp. 3.13 A-B). These neurons tend to be spatially structured like functional ensembles (see Supp. 3.13 C), and we show in Supp. 3.13 D that this is consistently due to < 5 hidden units which apparently failed to maintain the teacher's spatial distribution. It is possible that by re-training those outlier student RBMs, the failing hidden units would retain teacher characteristics and therefore remove this problem. However, these examples illustrate that just a few badly learned hidden units can have a large impact, confirming the robustness of the method the rest of the time.

Translated patterns are probable under the receiving model. To test whether translated configurations were plausible under the receiving RBM, we evaluated their free energy $F(\mathbf{v})$, which measures the log-probability of the configurations up to an additive constant. Translated patterns ($T \to S$ and $S \to T$) had free energies comparable to those of within-fish reconstructions ($T \to T$, $S \to S$), and well within the expected distribution for each fish (Fig. 3.4 D). In contrast, randomized activity patterns had significantly higher free energy and thus were unlikely under the receiving model, confirming that this is not caused by the model, but by an accurate representation of naturalistic neuronal configurations in hidden space. These results held across nearly all teacher–student pairs, with the exception of the outliers mentioned above (Fig. 3.4 E).

Spatial structure is retained across fish. We finally asked whether transferred activity patterns retained the spatial features of the source configuration. We computed the spatial correlation between the original and translated configurations at all time points. With the exception of the outliers described above (see Supp. 3.14), both $T \to S$ and $S \to T$ translations achieved significantly higher spatial correlations ($\mathrm{median} \approx 0.55$) than time-shuffled controls ($\mathrm{median} \approx 0.3$; Fig. 3.4 F–G), indicating that large-scale spatial motifs are preserved during transfer.

These results demonstrate that spontaneous neuronal activity patterns can be robustly translated across individuals through the shared RBM latent space. These translated configurations preserve key statistical and spatial features, are assigned high probability by the recipient model, and closely resemble the native activity observed in each fish. These findings indicate that the learned latent space encodes a shared and stereotyped repertoire of spontaneous brain-wide activity motifs. This framework offers a powerful new approach for comparing high-dimensional neuronal dynamics across animals, even in the absence of cell-by-cell correspondence.

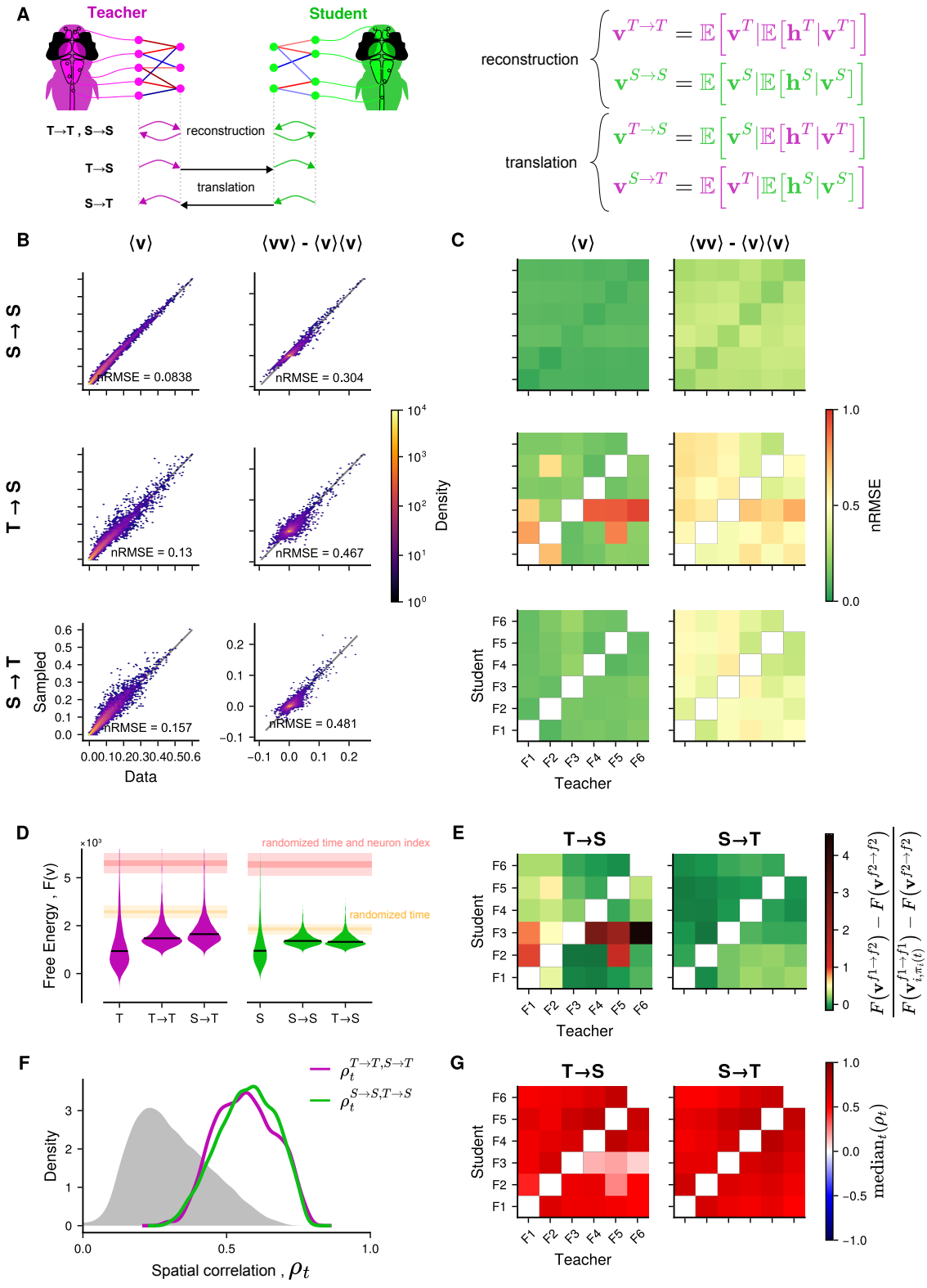


Figure 3.4: *(Caption on next page)*

Figure 3.4: **Spontaneous neuronal activity can be translated through the latent space to another fish. A**: Neuronal configurations can be represented in hidden space and then either reconstructed in the original neuronal space, or translated to another neuronal space. (see Methods 3.4.6.4) **B**: Neuronal statistics for the example teacher-student pair. Mean neuronal activity $\langle v \rangle$ (left) and pairwise covariance $\langle vv \rangle - \langle v \rangle \langle v \rangle$ (right) between empirical data and reconstructed student configurations ($S \to S$, top) and translated configurations in both teacher-to-student ($T \to S$, middle) and student-to-teacher ($S \to T$, bottom). **C**: nRMSE (see Methods 3.4.7) for each panel of B and every teacher-student pair. **D**: Distributions of free energy for empirical (T or S), reconstructed ($T \to S$ or $S \to T$), and translated ($S \to T$ or $T \to S$) configurations in an example teacher (left) and student (right) RBMs. Black lines indicate median. Horizontal bands indicate 50% and 99% of free energy distribution for randomized data with shuffled time frames (orange), and shuffled time frames and neurons (red). **E**: Median free energy contrast $\frac{F(\mathbf{v}^{f_1 \to f_2}) - F(\mathbf{v}^{f_2 \to f_2})}{F(\mathbf{v}^{f_1 \to f_1}_{i,\pi_i(t)}) - F(\mathbf{v}^{f_2 \to f_2})}$ of translated configurations ($\mathbf{v}^{f_1 \to f_2}$) with respect to reconstructed configurations ($\mathbf{v}^{f_2 \to f_2}$) and time-randomized configurations ($\mathbf{v}^{f_1 \to f_1}_{i,\pi_i(t)}$). **F**: Spatial correlation $\rho_t^{T \to T, S \to T}$ (magenta) and $\rho_t^{S \to S, T \to S}$ (green) between reconstructed and translated neuronal configuration for each time frame of an example teacher-student pair. In gray, the distribution of spatial correlations between shuffled pairs of time frames. **G**: Median of the distributions of spatial correlations from panel F for every teacher-student pair.

3.3 Discussion

In this study, we uncovered a shared latent structure underlying spontaneous brain-wide activity in larval zebrafish. By bi-training Restricted Boltzmann Machines (RBMs) across individuals, we identified a latent space composed of functional assemblies: hidden units that represent recurrent co-activation patterns. The activity of one animal could be projected into this space and transferred into the neuronal space of another, preserving spatial structure and probabilistic consistency. These results reveal that spontaneous activity is structured by a shared repertoire of latent cell assemblies, suggesting conserved circuit-level constraints across brains.

Even in the absence of constraints that enforce a shared latent space, many co-activation motifs were spontaneously conserved across individuals. When RBMs were trained independently on individual fish, approximately 30% of the hidden units could be quantitatively matched across all animals based on spatial similarity. Visual inspection revealed that up to 50% displayed a conserved spatial density core, likely reflecting functional assemblies shaped by developmental or anatomical regularities. Although bi-training does not uncover additional motifs beyond those of the teacher, it demonstrates that the teacher's latent space acts as a shared non-linear functional basis, capable of representing spontaneous activity across diverse individuals. The success of this transfer supports the presence of stereotyped latent population codes that generalize beyond individual variability.

The RBMs' shared hidden units exhibited spatial and functional consistency across indi-

viduals. This is remarkable because although student models were initialized using the spatial weight distribution of the teacher's RBM, they were trained solely on functional activity without any spatial priors. However, hidden units retained consistent spatial patterns and, in particular, pairwise correlations between hidden units were partially preserved across teacher and student. This correlation structure was not explicitly enforced during training, suggesting that the RBM uncovers latent activity motifs that recur across individuals. These motifs are functionally consistent, as reflected in preserved hidden-unit correlations, but also spatially conserved, despite the absence of explicit spatial constraints.

The ability to transfer activity patterns across individuals suggests that similar brain states are encoded similarly across fish. Importantly, this is not a trivial outcome: the model could have allowed degenerate mappings, where divergent hidden configurations represent equivalent activity patterns. Instead, we observe consistent population codes, suggesting that the shared latent space supports interpretable and conserved representations. Moreover, the fact that the transferred activity patterns preserve low free energy in the recipient model indicates that the latent space captures not only structural regularities but also statistical plausibility, an essential property for generative models. Analogously to Helmholtz free energy in statistical physics, the free energy of an activity pattern measures the cost of maintaining a particular neural configuration (macrostate) by marginalizing over all compatible hidden configurations (microstates). Low free energy implies high probability and interpretability under the model. This framework enables cross-individual comparison of brain-wide activity based on statistically grounded latent structure, allowing for quantitative evaluation of functional similarity grounded in principles from statistical physics.

Compared to other approaches for inter-subject alignment, such as anatomical registration, shared response modeling (SRM), hyperalignment, or contrastive autoencoders, our method offers several advantages. It is generative, allowing brain-wide activity patterns to be sampled, evaluated, and transferred probabilistically. It defines a probabilistic prior over spontaneous neural configurations, and after bi-training, this prior is shared across individuals, reflecting common structure in brain-wide activity. It is compositional, combining latent assemblies non-linearly to capture a richer repertoire of motifs than possible in linear subspaces. And it retains single-cell interpretability, unlike deep autoencoder methods. In addition, our approach is fully unsupervised and does not require stimuli, behavioral labels, or neuron correspondence.

Despite these strengths, the current implementation of our approach has limitations. It assumes, on the mesocal, a spatial organization of brain function. Although this assumption is appropriate for the larval zebrafish brains at 20–30 μm resolution, it may not generalize to a more variable or less stereotyped system, such as the mammalian cortex. Furthermore, the need to select a reference (teacher) fish introduces a source of bias. Although we observed robust generalization across multiple teacher choices, future extensions could mitigate this issue by implementing joint training or multi-teacher constraints. A symmetric, group-constrained approach, where all models are trained jointly with shared constraints, could enable the discovery of a richer and more balanced repertoire of conserved motifs. This would avoid teacher-specific bias and improve the statistical power to detect rare or subtle shared features. The resulting latent space could serve as a functional atlas onto which new data could be projected and compared across developmental stages, genotypes, or exper-

imental paradigms. This framework could be used for functional phenotyping, where an individual's deviation from the typical use of latent space may indicate neurodevelopmental or pathological alterations.

The emergence of spatially and statistically conserved latent motifs suggests that spontaneous dynamics are not arbitrary, but shaped by shared developmental programs and circuit architecture. The ability to identify and compare these motifs across individuals, without relying on anatomical correspondence or external stimuli, provides a new lens on how vertebrate brains are organized and how they differ. Together, our findings lay the foundations for a comparative neuroscience framework based on shared latent structure in spontaneous activity, enabling quantitative, model-based comparisons of brain-wide dynamics across individuals, developmental stages, or experimental conditions.

3.4 Materials and Methods

3.4.1 Data and Code Availability

All code, models, and post-processed data used in the present article are available in a centralized repository at : `https://github.com/EmeEmu/MultiFishRBM/tree/v.PhD.1`. Raw data is available upon request (approximately 400GB).

Calcium imaging pre-processing was performed using previously published protocols and software implemented in MATLAB (Mathworks) [126, 171, 116, 117].

Morphological Registrations was performed using the Advanced Normalization Tools (ANTs) [194].

The Restricted Boltzmann Machine modeling is based on the general-purpose Julia package `https://github.com/cossio/RestrictedBoltzmannMachines.jl` [195]. An extension to this package implementing the *Bi-Trained* procedure is available at `https://github.com/BiRBMs2024/BiTrainedRBMs.jl`.

3.4.2 Fish Husbrandry

All experiments were performed on nacre Tg(elavl3:H2B-GCaMP6f) Danio *rerio* larvae aged 5 to 6 days post-fertilization (dpf). Larvae were reared in groups of 20-30, in Petri dishes in embryo medium (E3) on a 14/10 hour light/dark cycle at $28°C$, and were fed $1mL$ of rotifers at $10^6/\ell$ daily from 5 dpf.
Larvae were screened at 4 and 5dpf for GCaMP6f expression, checking for fluorescence intensity and uniform expression patterns. At 5dpf, larvae were also screened for correct inflation of the swim bladder.

The experimental protocols were approved by Le Comitée d'Ethique pour l'Expérimentation Animale Charles Darwin C2EA-05 (02601.01 and #32423-202107121527185 v3).

3.4.3 Experimental Protocol

3.4.3.1 Fish Preparation

Larvae were paralyzed by immersion in a $1mg/mL$ solution of alpha-bungarotoxin (Invitrogen™ B1601) for one to two minutes in groups of 5 fish. After treatment, the larvae were washed and placed in E3 solution for 20 minutes. Startle response and blood flow were then checked to assess correct paralysis.
A single larvae is then mounted tail first in a capillary (Wiretrol® II $25 - 50\mu L$) in 2% low melting point agarose (Invitrogen™ 16520-100), and placed in the imaging tank, in E3 solution, under the microscope for an habituation period of 1 hour.

3.4.3.2 Imaging

Spontaneous neural activity was recorded using Light-Sheet Microscopy, at 2.5 volumes per second during 25 minutes, with 25 z-planes separated by $10\mu m$ [126].

A total of 17 fish were recorded. Out of those, 11 were discarded : 3 due to failing paralysis, 2 died during recording, 4 presented uncorrectable drift in z, and 2 presented very low fluorescence signal.
This leaves 6 fish which were pre-processed and used in the present study. Fish 1 and 3 were randomly chosen as examples in all main figures, with Fish 3 as teacher and Fish 1 as student in all relevant figures (see Tab. 3.1).

Fish #	Clutch	Session	Age (dpf)	Time of day	Neurons
1	A	1	6	14:30	44378
2	A	1	6	17:30	48179
3	A	1	6	13:00	43379
4	B	2	5	16:30	41349
5	B	2	5	17:45	40242
6	B	3	6	12:00	43354

Table 3.1: List of fish

3.4.3.3 Data pre-processing

Image pre-processing was performed offline using MATLAB according to previously published protocols [126, 171, 116].

Two dimensional drift correction was applied to each imaging layer to correct for movements of the sample during recording. Watershed was then used for cell segmentation, and a mean fluorescence value was computed from each cell ROI i and frame t : F_{it}. This was then normalized to $\Delta F/F_{it} = (F_i - \langle F_i \rangle_t)/(\langle F_i \rangle_t - F_0)$, where $\langle F \rangle_t$ is the baseline fluorescence of neuron i and F_0 is the image-wide background fluorescence.
The $\Delta F/F_{it}$ was then deconvolved into binarized spike trains $s_{it} \in [0, 1]$ using Blind Sparse Deconvolution (BSD) [117], using the calcium kernel time constants for GCaMP6f previously inferred using BSD [116, 11] : rise time 0.15s and decay time 1.6s. This method was used over other published methods as it not only deconvolves the fluorescence into spike trains, but also provides a rationale for the binarization of this train.

The first 4 minutes (600 frames) of each recording were excluded. The fish were accustomed to the imaging chamber for 1 hour with the laser turned off to avoid photobleaching. When the laser is turned on at the start of recording, the mean calcium signal rises sharply and then decays to a stable level within about 1 min. By discarding the initial 4 min, we ensure that all analyzed data come from the steady-state phase of neuronal activity under continuous illumination.

Samples of the binarized whole-brain activity for each fish can be seen in Movie 1.

3.4.3.4 Morphological Registration

For each recording, a high-resolution stack was acquired with 250 z planes separated by $1\mu m$ and a camera exposure time of $200ms$ per layer. A low-resolution stack was then created by averaging all frames of the functional recording, and then rigidly registered to the high-resolution stack using the Advanced Normalization Tools (ANTs) [194].
The high-resolution stacks off all fish were then registered together using both affine and warp transformations to create a single reference space in which all segmented neurons could be projected. The ANTs command and parameters were adapted from Vladimirov et al. [196], and are given bellow.

```
antsRegistration
    --float 1 --dimensionality 3 --interpolation WelchWindowedSinc
    --use-histogram-matching 0
    --output [outputdir/LowRes_TO_HighRes_ ,outputdir/LowRes_TO_HighRes.nrrd]
    --initial-moving-transform [HighRes.nrrd,LowRes.nrrd,1]
    --transform Rigid[0.1] --metric MI[HighRes.nrrd,LowRes.nrrd,1,32,Regular,0.25]
    --convergence [200x200x200x0,1e-8,10] --shrink-factors 12x8x4x2
    --smoothing-sigmas 4x3x2x1vox --transform Affine[0.1]
    --metric MI[HighRes.nrrd,LowRes.nrrd,1,32,Regular,0.25]
    --convergence [200x200x200x0,1e-8,10] --shrink-factors 12x8x4x2
    --smoothing-sigmas 4x3x2x1vox



antsRegistration
    --float 1 --dimensionality 3 --interpolation WelchWindowedSinc
    --use-histogram-matching 0
    --output [HighRes_TO_Reference_ ,HighRes_TO_Reference.nrrd]
    --initial-moving-transform [Reference.nrrd,HighRes.nrrd,1]
    --transform Rigid[0.1] --metric MI[Reference.nrrd,HighRes.nrrd,1,32,Regular,0.25]
```

```
    --convergence [200x200x200x0,1e-8,10] --shrink-factors 12x8x4x2
    --smoothing-sigmas 4x3x2x1vox --transform Affine[0.1]
    --metric MI[Reference.nrrd,HighRes.nrrd,1,32,Regular,0.25]
    --convergence [200x200x200x0,1e-8,10] --shrink-factors 12x8x4x2
    --smoothing-sigmas 4x3x2x1vox --transform SyN[0.05,6,0.5]
    --metric CC[Reference.nrrd,HighRes.nrrd,1,2]
    --convergence [200x200x200x200x10,1e-7,10] --shrink-factors 12x8x4x2x1
    --smoothing-sigmas 4x3x2x1x0vox



antsApplyTransforms
    --dimensionality 3 --float --input LowRes.nrrd
    --reference-image Reference.nrrd
    --output outputdir/LowRes_TO_Reference.nrrd
    --interpolation WelchWindowedSinc --default-value 0
```

3.4.4 Restricted Boltzmann Machines

3.4.4.1 Definition

Restricted Boltzmann Machines (RBMs) [143] are two-layer energy-based models, over N visible units $\mathbf{v} = (v_1, \ldots, v_N)$ representing the neural activity with binary values $v_i \in \{0, 1\}$, and M real-valued hidden units $\mathbf{h} = (h_1, \ldots, h_M)$.

The RBM defines a probability distribution over configurations of all the units given by:

$$P(\mathbf{v}, \mathbf{h}) = \frac{1}{Z} e^{-E(\mathbf{v}, \mathbf{h})} \tag{3.3}$$

where Z is a normalization constant known as the partition function, and $E(\mathbf{v}, \mathbf{h})$, the energy, is given by:

$$E(\mathbf{v}, \mathbf{h}) = -\sum_{i=1}^{N} \mathcal{V}_i(v_i) + \sum_{\mu=1}^{M} \mathcal{U}_\mu(h_\mu) - \sum_{i=1}^{N} \sum_{\mu=1}^{M} w_{i\mu} v_i h_\mu \tag{3.4}$$

The functions $\mathcal{V}_i(v_i)$ and $\mathcal{U}_\mu(h_\mu)$ are potentials biasing the activities of single units, while the weights $w_{i\mu}$ account for interactions between visible and hidden units. In the case of neuronal data, we take $\mathcal{V}_i(v_i) = g_i v_i$ with g_i called the field. Following van der Plas et al. [11], we use dReLU potentials for the hidden units,

$$\mathcal{U}_\mu(h_\mu) = \frac{1}{2} \frac{\gamma_\mu}{1 + \eta_\mu} h_{\mu,+}^2 + \frac{1}{2} \frac{\gamma_\mu}{1 - \eta_\mu} h_{\mu,-}^2 + \theta_\mu h_\mu + \frac{\Delta_\mu}{1 + \eta_\mu} h_{\mu,+} - \frac{\Delta_\mu}{1 - \eta_\mu} h_{\mu,-} \tag{3.5}$$

with real parameters $\gamma_\mu, \eta_\mu, \Delta_\mu, \theta_\mu$, satisfying $\gamma_\mu > 0$. Here $h_{\mu,+} = \max(h_\mu, 0)$ and $h_{\mu,-} = \min(h_\mu, 0)$.

To obtain the probability of neural activity, the hidden units are marginalized. In this way, one defines a marginal likelihood over the visible units:

$$P(\mathbf{v}) = \int P(\mathbf{v}, \mathbf{h}) \mathrm{d}\mathbf{h} = \frac{1}{Z} e^{-E_{\text{eff}}(\mathbf{v})} \tag{3.6}$$

where $E_{\text{eff}}(\mathbf{v})$ is a free energy function that incorporates the effective interactions induced by the marginalized hidden variables,

$$E_{\text{eff}}(\mathbf{v}) = \sum_{i=1}^{N} \mathcal{V}_i(v_i) - \sum_{\mu=1}^{M} \ln \int e^{\sum_i w_{i\mu} v_i h_\mu - \mathcal{U}_\mu(h_\mu)} dh_\mu \tag{3.7}$$

3.4.4.2 Standardized RBMs

Following [197], we employ a generalized *centering trick* [198], where the activity of the hidden units is not only centered, but also standardized.

The Standardized Restricted Boltzmann machine (StdRBM) is a re-parameterization of the RBM, where the energy function is defined as follows:

$$E(\mathbf{v}, \mathbf{h}) = \cdots - \sum_i g_i v_i - \sum_\mu \theta_\mu h_\mu - \sum_{i\mu} w_{i\mu} \frac{v_i - \lambda_i}{\sigma_i} \frac{h_\mu - \lambda_\mu}{\sigma_\mu} + \sum_{i\mu} \frac{w_{i\mu}}{\sigma_i \sigma_\mu} \lambda_i \lambda_\mu \tag{3.8}$$

where we only show the fields for the unit potentials for simplicity. During training, the parameters λ_i, λ_μ and σ_i, σ_μ track a moving average of the mean and standard deviations of the corresponding unit activities. After training, the StdRBM can be converted into an equivalent RBM by the transformations:

$$g_i \to g_i - \sum_\mu \frac{w_{i\mu}}{\sigma_i \sigma_\mu} \lambda_\mu, \quad \theta_\mu \to \theta_\mu - \sum_i \frac{w_{i\mu}}{\sigma_i \sigma_\mu} \lambda_i, \quad w_{i\mu} \to \frac{w_{i\mu}}{\sigma_i \sigma_\mu} \tag{3.9}$$

As has been observed empirically in previous works [198, 197, 199], such centering and standardization of unit activities leads to improved training convergence and stability.

3.4.4.3 Training

All RBM parameters ($g_i, \gamma, \eta, \Delta, \theta, w_{i\mu}$) are trained to maximize the likelihood of a neural activity dataset. If a data set of neural activity recordings is given as $\mathcal{D} = \{\mathbf{v}^t\}$, where t is the index of the sample (or time), then the RBM is trained by maximizing:

$$\frac{1}{|\mathcal{D}|} \sum_{t \in \mathcal{D}} \ln P(\mathbf{v}_t) - \mathcal{R}(W) \tag{3.10}$$

Here, $\mathcal{R}(W)$ is a regularization term applied over the RBM weights. Following van der Plas et al. [11], we employ an L1 regularization,

$$\mathcal{R}(W) = \lambda_1 \sum_{i\mu} |w_{i\mu}| \tag{3.11}$$

where λ_1 controls the regularization strength. As in van der Plas et al. [11], we set $\lambda_1 = 0.02$, and $M = 100$ hidden units. The optimization is carried out by gradient ascent of Eq. (3.10). We use the ADAM algorithm [200].

3.4.4.4 Evaluation

Trained RBMs are evaluated on their ability to reproduce data statistics. As they are trained to maximize the log-likelihood of the data, a well trained RBM is expected to generate data with statistics $\langle v_i \rangle$, $\langle h_\mu \rangle$, and $\langle v_i h_\mu \rangle$ matching the corresponding empirical data statistics. We also evaluate the RBMs' ability to reproduce the second-order statistics $\langle v_i v_j \rangle - \langle v_i \rangle \langle v_j \rangle$ and $\langle h_\mu h_\nu \rangle - \langle h_\mu \rangle \langle h_\nu \rangle$ to ensure that first order interactions between neurons are captured by the model [11]. An RBM has to perform well all five statistics to be considered usable in our use case.

Measurement of statistics matching between model-generated and empirical data is done using the normalized Root Mean Squared Error (nRMSE) as introduced by van der Plas et al. [11] (see 3.4.7). This measurement is standardized such that a value of 1 corresponds to shuffled data statistics, and a value of 0 corresponds to the optimal matching expected from the difference between train and test sets. The evaluation of an RBM is thus represented by a vector of length 5 containing the nRMSE for all five statistics, and multiple RBMs can be compared using the L_∞ norm.

3.4.4.5 Inferred neuronal couplings

van der Plas et al. [11] showed that an effective coupling matrix J_{ij} could be inferred from a trained RBM model. This is done by manually perturbing the activity of each neuron and quantifying the impact on other neurons by evaluating the marginal distribution $P(\mathbf{v})$. Given a neuronal configuration $\mathbf{v}$, the coupling is then defined as

$$\begin{aligned} J_{ij}(\mathbf{v}) = &\log\left(\frac{P\left(v_i = 1 \mid v_1, \cdots, v_j = 1, \cdots v_N\right)}{P\left(v_i = 1 \mid v_1, \cdots, v_j = 0, \cdots v_N\right)}\right) \\ &- \log\left(\frac{P\left(v_i = 0 \mid v_1, \cdots, v_j = 1, \cdots v_N\right)}{P\left(v_i = 0 \mid v_1, \cdots, v_j = 0, \cdots v_N\right)}\right) \end{aligned} \tag{3.12}$$

The overall coupling matrix is then obtained from all empirical neuronal configurations as:

$$J_{ij} = \langle J_{ij}(\mathbf{v}) \rangle_{\text{data}} \tag{3.13}$$

However, calculating $J_{ij}(\mathbf{v})$ is computationally impractical, and we use the following approximation introduced by van der Plas et al. [11]:

$$J_{ij} \approx \sum_{\mu=1}^{M} w_{i\mu} w_{j\mu} \langle \mathrm{Var}(h_\mu \mid \mathbf{v}) \rangle_{\text{data}} \tag{3.14}$$

3.4.5 Voxelised RBMs

3.4.5.1 Voxelisation

To obtain a coarse-grained description of the datasets, we voxelize the brain into cubes of side v (typically $v = 20\mu m$). This grid is defined in the common registered brain space.

Neurons are assigned to the voxels after segmentation, according to their center of mass. The activity of each voxel for each fish is then defined as :

$$v_m(t) = \frac{1}{|\mathcal{V}_m|} \sum_{i=1}^{|\mathcal{V}_m|} \frac{\Delta F}{F}_i(t) \tag{3.15}$$

where v_m is the activity of voxel m containing the set of neurons $\mathcal{V}_m$ and $\frac{\Delta F}{F}_i(t)$ is the normalized fluorescence signal of neuron i at time t.
This activity is then normalized for each fish f and each voxel using z-score :

$$\mathrm{v}_m^f(t) = \frac{v_m^f(t) - \langle v_m^f \rangle_t}{\sigma_t(v_m^f)} \tag{3.16}$$

This process yields a series of voxelized activities $\mathrm{v}_m^f(t)$ where each voxel m corresponds to the same anatomical position across all fish.
To ensure that the z-scored activity of each voxel is well defined, we discard voxels for which $|\mathcal{V}_m^f| < 2$ for at least one fish.

3.4.5.2 Training

Contrary to the neuronal RBMs described above (3.4.4), in the case of voxelized data, we used Standardized RBMs with gaussian visible units, as the voxelized activity approximately follows a normal distribution (see Supp. 3.7 C).

$$\mathcal{U}_i(v) = \frac{1}{2}\gamma_i v^2 + \theta_i v \tag{3.17}$$

We found empirically that for voxelized RBMs, L_1 regularization produced weight matrices where the majority of hidden units were decoupled from the visible layer. Hence we use the L_1^2 regularization introduce by Tubiana, Cocco, and Monasson [142] :

$$\mathcal{R}(W) = \frac{\lambda_{21}}{N} \sum_{\mu} \left(\sum_i |w_{i\mu}| \right)^2 \tag{3.18}$$

where N is the number of voxels and λ_{21} is the regularization factor.

We performed a cross-validation over the number of hidden units M and regularization factor λ_{21} on Fish 1 (Supp. 3.7 A). We identified 3 sets of hyperparameters which we then tested on the concatenated voxelized data of all 6 fish in the dataset (Supp. 3.7 B), and chose $M = 40$ and $\lambda_{21} = 0.1$.

The rest of the training parameters are listed here :

- 20000 gradient updates.
- 50 Markov Chain steps between gradient updates.

- 100 time points per batch.
- $5{\cdot}10^{-4}$ learning rate for the first $1/4$ of the training. The learning rate is then annealed with a geometrical decay to reach 10^{-5} at the end of training.

3.4.6 Bi-trained RBMs

3.4.6.1 Teacher RBMs

Teacher standardized RBMs were constructed and trained from whole-brain binarized neuronal activity with the following parameters :

- $M = 100$ hidden units (instead of either 200 or 100 in van der Plas et al. [11] as most of our fish have weak activity in the optic tectum).
- $\lambda_1 = 0.02$ for weight L1 regularization (as obtained by van der Plas et al. [11] after cross-validation).
- 200000 gradient updates.
- 15 Markov Chain steps between gradient updates.
- 256 time points per batch.
- $5{\cdot}10^{-4}$ learning rate for the first $1/4$ of the training. The learning rate is then annealed with a geometrical decay to reach 10^{-5} at the end of training.

Before training, individual fish datasets were split into a 70% training and 30% validation sets using the method described by van der Plas et al. [11]. The validation set was used to evaluate RBM-generated vs empirical data statistics as described in Methods 3.4.4.4.

For every fish, we trained 10 RBMs and kept the one which performed best as described in Methods 3.4.4.4 (see Supp. 3.9 for all training statistics). Indeed, we found that convergence can be very variable, with some models converging satisfactorily, and other either converging poorly or not converging at all.

3.4.6.2 Student Initialization

Student RBMs are initialized from the teacher with the following steps.

First, the fields g_j of the visible layer are estimated from logit of the average neuron activity $\langle v_j \rangle$: $g_j = -\log\left(\frac{1}{\langle v_j \rangle} - 1\right)$. The visible scaling ($\sigma_j$) and bias ($\lambda_j$) parameters used to standardize the RBM are also estimated from neuronal activity as : $\sigma_j = \frac{1}{2}\operatorname{var}(v_j) + 1$ and $\lambda_j = \langle v_j \rangle$.

Second, hidden potentials $\mathcal{U}_\mu(h_\mu)$ (i.e. all parameters of the dReLU $\gamma_\mu, \eta_\mu, \Delta_\mu, \theta_\mu$) are copied directly from the teacher RBM.

Third, the teacher weight matrix $w_{i\mu}$ is spatially interpolated at the locations of student neurons. To do so, we first build a 3D map of weights for each hidden unit μ :

$$W_\mu(\vec{x}) = \frac{\sum_i w_{i\mu}\delta(\vec{x}, \vec{x}_i)}{\sum_i \delta(\vec{x}, \vec{x}_i)} \tag{3.19}$$

$$\delta(\vec{x}, \vec{y}) = \begin{cases} 1, & \text{if } ||\vec{x} - \vec{y}|| < r \\ 0, & \text{otherwise} \end{cases} \tag{3.20}$$

where $\vec{x}$ is any 3D location in shared brain space, $\vec{x}_i$ is the 3D location of teacher neuron i, and $\delta(\vec{x}, \vec{x}_i)$ is a boolean ball of radius $r = 4\mu m$, the average size of neurons. These maps are then convolved with a gaussian kernel $G(\vec{x}, \sigma)$ of width $\sigma = 4\mu m$ to smooth out local discontinuities : $\mathcal{W}_\mu(\vec{x}) = W_\mu(\vec{x}) * G(\vec{x}, \sigma)$. Because weight matrices are very sparse, this smoothing step has a tendency to attenuate the weight map. We therefore add a bias :

$$\beta_\mu = \min_\beta \sqrt{\frac{1}{N}\sum_{i=1}^{N} \left(w_{i\mu} - \beta \mathcal{W}_\mu(\vec{x_i})\right)^2} \tag{3.21}$$

Finally, a weight matrix for the student can be created by evaluating $\mathcal{W}_\mu(\vec{x})$ at the locations $\vec{x}_j$ of neurons from the student fish :

$$w_{j,\mu} = \beta_\mu \mathcal{W}_\mu(\vec{x}_j) \tag{3.22}$$

3.4.6.3 Student training

Once initialized, student RBMs are trained with the same set of parameters as the teacher, except for the number of gradient updates for which we found 1/10th (20000) to be sufficient for convergence.

In absence of any constraints, the space of representations learned by the student can drift arbitrarily far from the one of the teacher. This is undesirable because we cannot compare the hidden unit activities from the teacher and the student. In order to avoid this, we introduce a new training paradigm to maintain an alignment between the hidden units of the student and the teacher. Mathematically, we would like the marginal distribution of hidden unit activities, $P(\mathbf{h}) = \sum_\mathbf{v} P(\mathbf{v}, \mathbf{h})$, to be similar in the teacher and the student. To achieve this, we first sample hidden unit activity from the teacher RBM to create a dataset of hidden unit activities, $\mathcal{T} = \{\mathbf{h}_t\}$ from the teacher. The student is trained following the modified objective:

$$\frac{1}{|\mathcal{D}|}\sum_s \ln P(\mathbf{v}_s) + \frac{\kappa}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} \ln P(\mathbf{h}_t) - \mathcal{R}(W) \tag{3.23}$$

The first and last terms are the same appearing in the original training objective, Eq. (3.10). The second term is the average log-likelihood of the teacher hidden unit activity evaluated

under the student. By maximizing this term, we force the student to align its marginal hidden unit distribution to match that of the teacher. The coefficient κ controls the relative importance of this term: for $\kappa = 0$, we recover the original training objective of Eq. (3.10), while for κ large, the student RBM is forced to match the teacher hidden unit statistics more closely.

3.4.6.4 Mapping from one fish to another

In principle, translating activity from the visible layer of one RBM, through the hidden layer, to the visible layer of another RBM should be done via sampling the hidden layer given the first fish visible layer, and for each sample, sampling the second fish visible layer given the hidden layer, and then averaging over all samples, which yields $\mathbb{E}_{\mathbf{h}^T|\mathbf{v}^T}\mathbb{E}_{\mathbf{v}^S|\mathbf{h}^T}\left[\mathbf{v}^S\right]$. However, in practice, this is computationally intensive and thus impractical in our use case. Instead, we use $\mathbb{E}_{\mathbf{v}^S|\mathbb{E}_{\mathbf{h}^T|\mathbf{v}^T}}\left[\mathbf{v}^S\right]$ as we have found empirically that for a high number (~ 1000) of samples the two method give similar results (see Supp. 3.12 A-D).

This similarity can be explained with a Taylor expansion on $f_i : \mathbf{h}^T \mapsto \mathbb{E}_{\mathbf{v}^S|\mathbf{h}^T}\left[\mathbf{v}_i^S\right]$ at $\mathbf{h}^{T*} = \mathbb{E}_{\mathbf{h}^T|\mathbf{v}^T}\left[\mathbf{h}^T\right]$:

$$\begin{aligned} f_i\left(\mathbf{h}^T\right) \underset{\mathbf{h}^T \to \mathbb{E}_{\mathbf{h}^T|\mathbf{v}^T}[\mathbf{h}^T]}{=} & f_i\left(\mathbb{E}_{\mathbf{h}^T|\mathbf{v}^T}\left[\mathbf{h}^T\right]\right) \\ & + \sum_\mu \left(\mathbf{h}_\mu^T - \mathbb{E}_{\mathbf{h}^T|\mathbf{v}^T}\left[\mathbf{h}_\mu^T\right]\right) \frac{\partial f_i}{\partial \mathbf{h}_\mu^T}\left(\mathbb{E}_{\mathbf{h}^T|\mathbf{v}^T}\left[\mathbf{h}^T\right]\right) \\ & + \frac{1}{2}\sum_{\mu,\nu} \left(\mathbf{h}_\mu^T - \mathbb{E}_{\mathbf{h}^T|\mathbf{v}^T}\left[\mathbf{h}_\mu^T\right]\right)\left(\mathbf{h}_\nu^T - \mathbb{E}_{\mathbf{h}^T|\mathbf{v}^T}\left[\mathbf{h}_\nu^T\right]\right) \frac{\partial^2 f_i}{\partial \mathbf{h}_\mu^T \partial \mathbf{h}_\nu^T}\left(\mathbb{E}_{\mathbf{h}^T|\mathbf{v}^T}\left[\mathbf{h}^T\right]\right) \\ & + \dots \end{aligned} \tag{3.24}$$

Because of the conditional independence of $\mathbf{h}^T$ given $\mathbf{v}^T$,

$$\begin{aligned} \mathbb{E}_{\mathbf{h}^T|\mathbf{v}^T}\left[f_i\left(\mathbf{h}^T\right)\right] \underset{\mathbf{h}^T \to \mathbf{h}^{T*}}{=} & f_i\left(\mathbb{E}_{\mathbf{h}^T|\mathbf{v}^T}\left[\mathbf{h}^T\right]\right) \\ & + \frac{1}{2}\sum_\mu \underset{\mathbf{h}^T|\mathbf{v}^T}{\mathrm{var}}\left[\mathbf{h}_\mu^T\right] \frac{\partial^2 f_i}{\partial \mathbf{h}_\mu^{T2}}\left(\mathbb{E}_{\mathbf{h}^T|\mathbf{v}^T}\left[\mathbf{h}^T\right]\right) \\ & + \dots \end{aligned} \tag{3.25}$$

Since $f_i\left(\mathbf{h}^T\right) = \mathbb{E}_{\mathbf{v}^S|\mathbf{h}^T}\left[\mathbf{v}_i^S\right] = \sigma\left(g_i + \sum_\mu w_{i\mu}^S \mathbf{h}_\mu^T\right)$, where $\sigma(z) = \frac{1}{1+\exp(-z)}$, we find

$$\mathbb{E}_{\mathbf{h}^T|\mathbf{v}^T}\mathbb{E}_{\mathbf{v}^S|\mathbf{h}^T}\left[\mathbf{v}^S\right] = \mathbb{E}_{\mathbf{v}^S|\mathbf{h}^{T*}}\left[\mathbf{v}^S\right] + \Delta \tag{3.26}$$

where

$$\Delta_i \approx \frac{1}{2}\sum_\mu \underset{\mathbf{h}^T|\mathbf{v}^T}{\mathrm{var}}\left[\mathbf{h}_\mu^T\right]\left(w_{i\mu}^S\right)^2 \sigma''\left(g_i + \sum_\mu w_{i\mu}^S \mathbf{h}_\mu^T\right) \tag{3.27}$$

We show on Supp. 3.12 E that Δ is indeed negligible compared to $\mathbb{E}_{\mathbf{v}^S|\mathbf{h}^{T*}}\left[\mathbf{v}^S\right]$.

3.4.7 Measuring identity

As introduced by van der Plas et al. [11], we quantify the goodness of fit to the identity using the normalized Root Mean Squared Error (nRMSE). It is defined as :

$$\mathrm{nRMSE}(X,Y) = 1 - \frac{\mathrm{RMSE}(X,Y) - \mathrm{RMSE}(\Pi(X),\Pi(Y))}{\mathrm{RMSE}(X^O,Y^O) - \mathrm{RMSE}(\Pi(X),\Pi(Y))} \tag{3.28}$$

where X and Y are two vectors of same length, $\mathrm{RMSE}(A,B) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(A_i - B_i\right)^2}$ the standard Root Mean Squared Error, (X^O, Y^O) is the pair of optimal vectors corresponding to the best expected fit of X and Y, and Π is a shuffling operator such that $\Pi(A)_i = A_{\pi(i)}$.

Unlike the standard RMSE, this measure is independent of the length of the vectors X and Y, and is easily interpretable. Indeed, a value of $\mathrm{nRMSE}(X,Y) = 1$ means that X and Y are as uncorrelated as shuffled vectors, while $\mathrm{nRMSE}(X,Y) = 0$ means that X and Y are as identical as can be expected from the optimal (X^O, Y^O).

In the case where no optimal (X^O, Y^O) can be defined other than $X^O = Y^O$, $\mathrm{RMSE}(X^O, Y^O) = 0$ and therefore :

$$\mathrm{nRMSE}(X,Y) = \frac{\mathrm{RMSE}(X,Y)}{\mathrm{RMSE}(\Pi(X),\Pi(Y))} \tag{3.29}$$

$\mathrm{nRMSE}(X,Y) = 0$ is then interpreted as $X_i = Y_i \; \forall i$.

3.4.8 Measuring Spatial Similarity

To compare how a spatially defined variable (such as neuron weight or firing rate) is distributed across two different fish, we define a spatial similarity measure between animals A and B.

Let ϕ^A and ϕ^B be two spatial fields that assign scalar values $y_i^A \in \mathbb{R}$ and $y_j^B \in \mathbb{R}$ to neurons at 3D positions $\mathbf{x}_i^A$ and $\mathbf{x}_j^B$ in fish A and B, respectively: $\phi^A(\mathbf{x}_i^A) = y_i^A$ and $\phi^B(\mathbf{x}_i^B) = y_i^B$. We wish to estimate how well ϕ^A aligns with ϕ^B.

Since neurons are not spatially matched between individuals, we interpolate the values y_i^B from fish B to the locations of neurons in fish A using Gaussian kernel smoothing :

$$\hat{y}_j^{A\to B} = \frac{\sum_i y_i^A e^{-\frac{\|\mathbf{x}_j^B - \mathbf{x}_i^A\|^2}{2\sigma^2}}}{\sum_i e^{-\frac{\|\mathbf{x}_j^B - \mathbf{x}_i^A\|^2}{2\sigma^2}}} \tag{3.30}$$

where $\|\mathbf{x}_j^B - \mathbf{x}_i^A\|$ is the euclidean distance between neuron i from fish A and neuron j from fish B, and σ is the standard width of a gaussian kernel, which we take to be a neuron size ($\sigma = \frac{8\mu m}{2} = 4\mu m$). This provides an estimate of what the spatial distribution of y from

fish A would look like in fish B.

We can then evaluate the similarity between the two spatial fields ϕ^A and ϕ^B by computing the Pearson correlation coefficient between y_j^B and $\hat{y}_j^{A\to B}$:

$$\rho^{A\to B} = \frac{\mathrm{cov}(y_j^B, \hat{y}_j^{A\to B})}{\sigma_{y_B^2}\sigma_{\hat{y}_j^{A\to B}}} \tag{3.31}$$

In practice, the Person correlation is used over other identity tests such as the nRMSE (see Methods 3.4.7) because our observables tend to be sparse and the gaussian interpolation will thus naturally lower $\hat{y}$ ($\hat{y}_i^{A\to A} \approx \alpha y_i^A$ with $\alpha < 1$).

In the case of the weight maps presented in Fig. 3.3 B-F, $\rho^{A\to B}$ is only computed for neurons whose weights are significant ($> 10^{-5}$), as we are only interested in whether student maps are compatible with teachers, and the regularization imposed on weights during RBM training will force a non-spatial linearity in weight distribution.

3.4.9 Measuring Diagonal Dominance

In Fig. 3.3 C-D, we ask whether the spatial distribution of hidden units' weights is maintained between teacher and student RBMs, in a way that allows for their unambiguous pairwise identification. In other words, we ask whether the spatial distribution of $w_{i\mu}^S$ is closer to $w_{i\mu}^T$ than to any other hidden unit $w_{i\nu}^T$ with $\nu \neq \mu$.

To do so we first compute the matrix $\rho_{\mu\nu}^{TS}$ ($\mu, \nu \in (1, ..., M)$) of spatial correlation (see Methods 3.4.8) between the weight map of hidden unit μ in the teacher RBM, and hidden unit ν in the student RBM.

Our hypothesis is confirmed not only if each teacher hidden unit μ is most similar to its corresponding student unit (i.e., if $\arg\max_\nu \rho_{\mu\nu}^{TS} = \mu$), but also if this match is significantly better than any other (i.e., $\rho_{\mu\mu}^{TS} \gg \rho_{\mu\nu}^{TS}$ for all $\nu \neq \mu$).

To test for this systematically, we perform a bootstrap analysis. We start by computing a *distance* between the diagonal and the next largest element of the matrix :

$$d_\mu = \rho_{\mu\mu}^{TS} - \max_{\nu\neq\mu} \rho_{\mu\nu}^{TS} \tag{3.32}$$

We next compute the same distance for $n = 5000$ random row permutations of the matrix $\rho_{\mu\mu}^{TS}$:

$$\tilde{d}_\mu = \rho_{\pi(\mu)\mu}^{TS} - \max_{\nu\neq\pi(\mu)} \rho_{\pi(\mu)\nu}^{TS} \tag{3.33}$$

where $\pi(\mu)$ is a random permutation of $1, ..., M$ (see Supp. 3.11 C for an example). We can then compute a p-value measuring the probability that $\rho_{\mu\mu}^{TS} \gg \rho_{\mu\nu}^{TS}$ could be explained by

random permutations of the matrix :

$$p = \frac{1}{Mn} \sum_{j=1}^{n} \sum_{\mu=1}^{M} \mathbb{1}\{\tilde{d}_\mu > d_\mu\} \tag{3.34}$$

It is to be noted that, as some hidden units are completely disconnected from the visible layer ($w_{i\mu} < 10^{-5}, \forall i$), some weight maps are undefined. These are therefore omitted from this measurement.

Acknowledgments

We would like to thank Simona Cocco, Rémi Monasson, and Jérôme Tubiana for helpful discussions and insights. Ans Imran for his help programming some useful function.
We would also like to thank Abdelkrim Mannioui, Marco Amaral, Edouard Manzoni and the rest of the team at the fish facility for taking care of the animals, as well as Malika Pierrat and Laura Bacerra-Zapata for taking care of the administration and contracts.

Supporting information

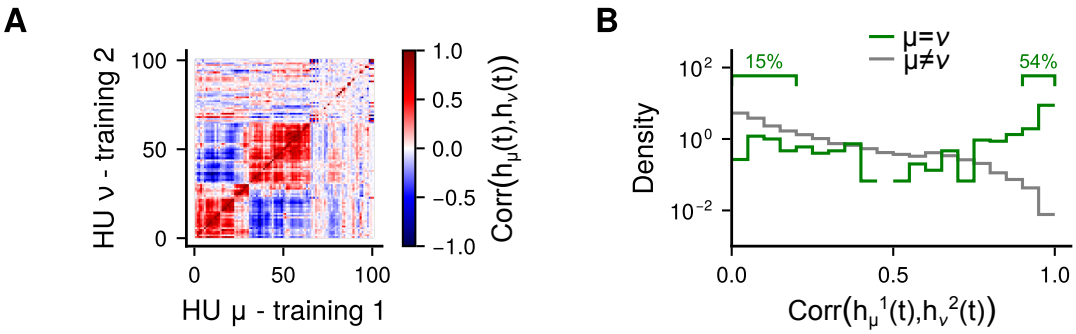


Figure 3.5: **Supplementary for Fig. 3.1. A**: Pairwise Pearson Correlation between hidden activity $h_\mu(t) = \mathbb{E}[\mathbf{h} \mid \mathbf{v}_t]$ and $h_\nu(t)$ of two example RBMs trained on the same neuronal recording. The matrix rows have been reordered to obtain the best alignment between the two trainings. **B**: Distribution of correlations from panel A, for best alignments pairs ($\nu = \mu$, green) and other pairs ($\nu \neq \mu$, grey)

Movie Supplementary Movies and captions are available at `https://doi.org/10.5281/zenodo.16886749`.

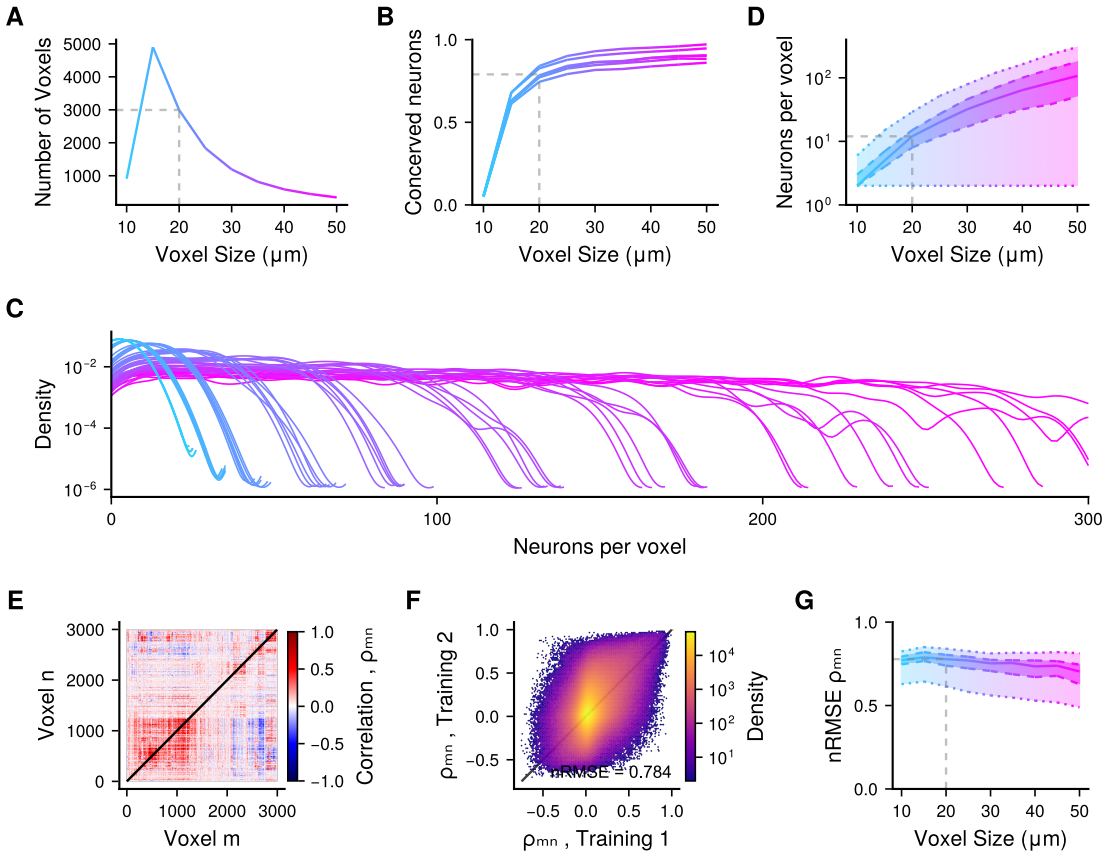


Figure 3.6: **Supplementary for Fig. 3.2 investigating the impact of voxel size. A**: Number of populated voxels (each voxel needs to contain $\geq$ 2 neurons for each fish, see Methods 3.4.5) as function of voxel size. **B**: Fraction of neurons conserved after voxelization. Each line is a fish. **C**: Distributions of the number of neurons per voxel, for each voxel-size tested (same colors as in panel A), and for each fish (one line per fish). **D**: Median (solid line), 25-75% range (dashed lines), and min-max range (dotted lines) of the number of neurons per voxel (all fish combined). **E**: Pairwise Person correlation coefficients ρ_{mn} between voxels m and n for an example fish at $20\mu m$ voxel size. **F**: Stereotypy in ρ_{mn} between 2 example fish at $20\mu m$ voxel size. **G**: nRMSE (see Methods 3.4.7) of ρ_{mn} between 2 example fish as a function of voxel size.

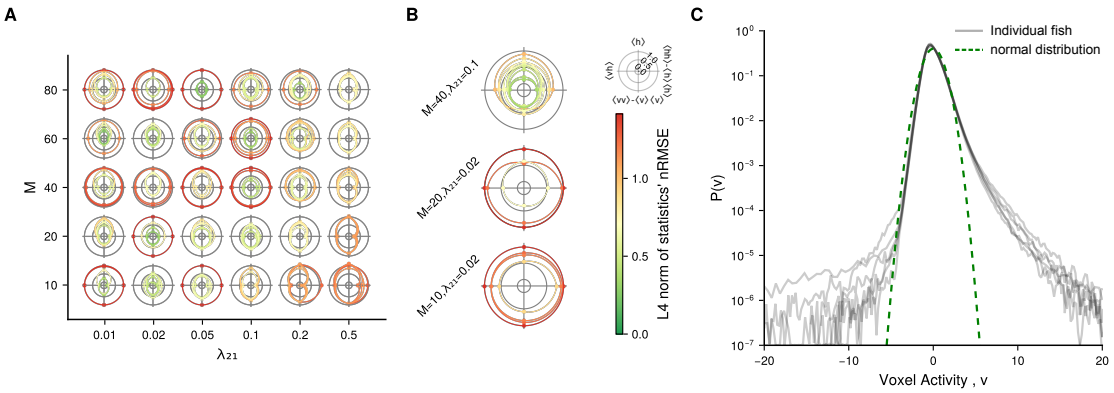


Figure 3.7: **Supplementary for Fig. 3.2 cross-validation of voxelized RBM hyperparameters.** **A**: nRMSE (see Methods 3.4.7) of all four moments (see Methods 3.4.4.4) used to evaluate RBM convergence ($\langle v \rangle$ is omitted as voxelized data is normalized by z-score). Each RBM is represented by a line in polar coordinates where every ray corresponds to a moment $m \in \left\{ \langle \mathbf{vh} \rangle, \langle \mathbf{h} \rangle, \langle \mathbf{vv} \rangle - \langle \mathbf{v} \rangle \langle \mathbf{v} \rangle, \langle \mathbf{hh} \rangle - \langle \mathbf{h} \rangle \langle \mathbf{h} \rangle \right\}$, and the ray length corresponds to the $\mathrm{nRMSE}_m(\text{data}, \text{generated})$ between empirical data and data generated by the RBM. Five training were performed and evaluated per hyperparameter pair (M, λ_{21}), where M is the number of hidden units, and λ_{21} is the regularization factor (see Methods 3.4.5.2). All trainings were done on voxelized data from Fish 1. **B**: Same panel A for 3 sets of hyperparameters, and for the concatenated voxelized data of all 6 fish in the dataset. **C**: Distributions of voxelized activity values, with one grey line per fish. Normal distribution in green (dashed).

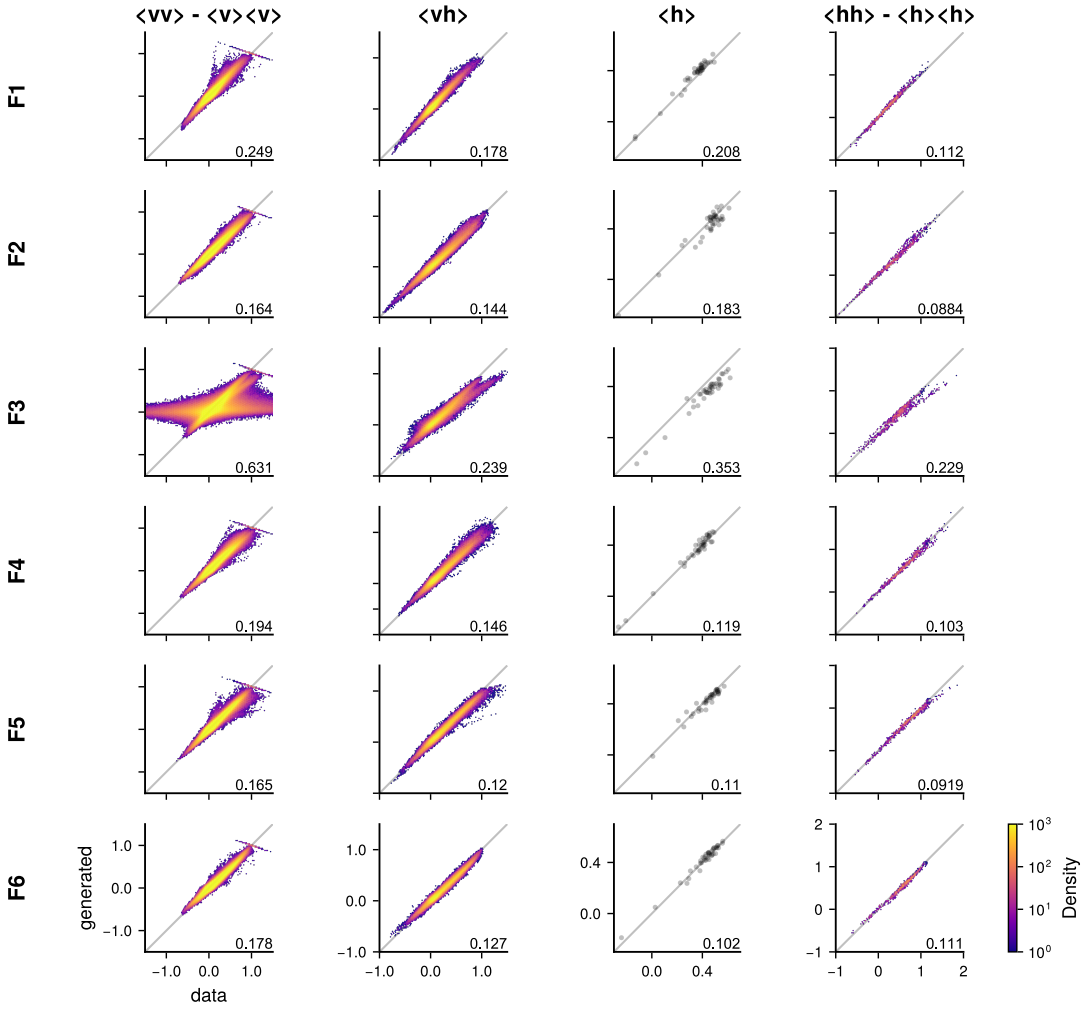


Figure 3.8: **Supplementary for Fig. 3.2 showing the voxelized training statistics for each fish.** Identity plots for all 4 moments (see Methods 3.4.4.4) used to evaluate voxelized RBM convergence (notice that moment $\langle v \rangle$ is not represented as voxelized activity is normalized by z-score, see Methods 3.4.5) for each fish.

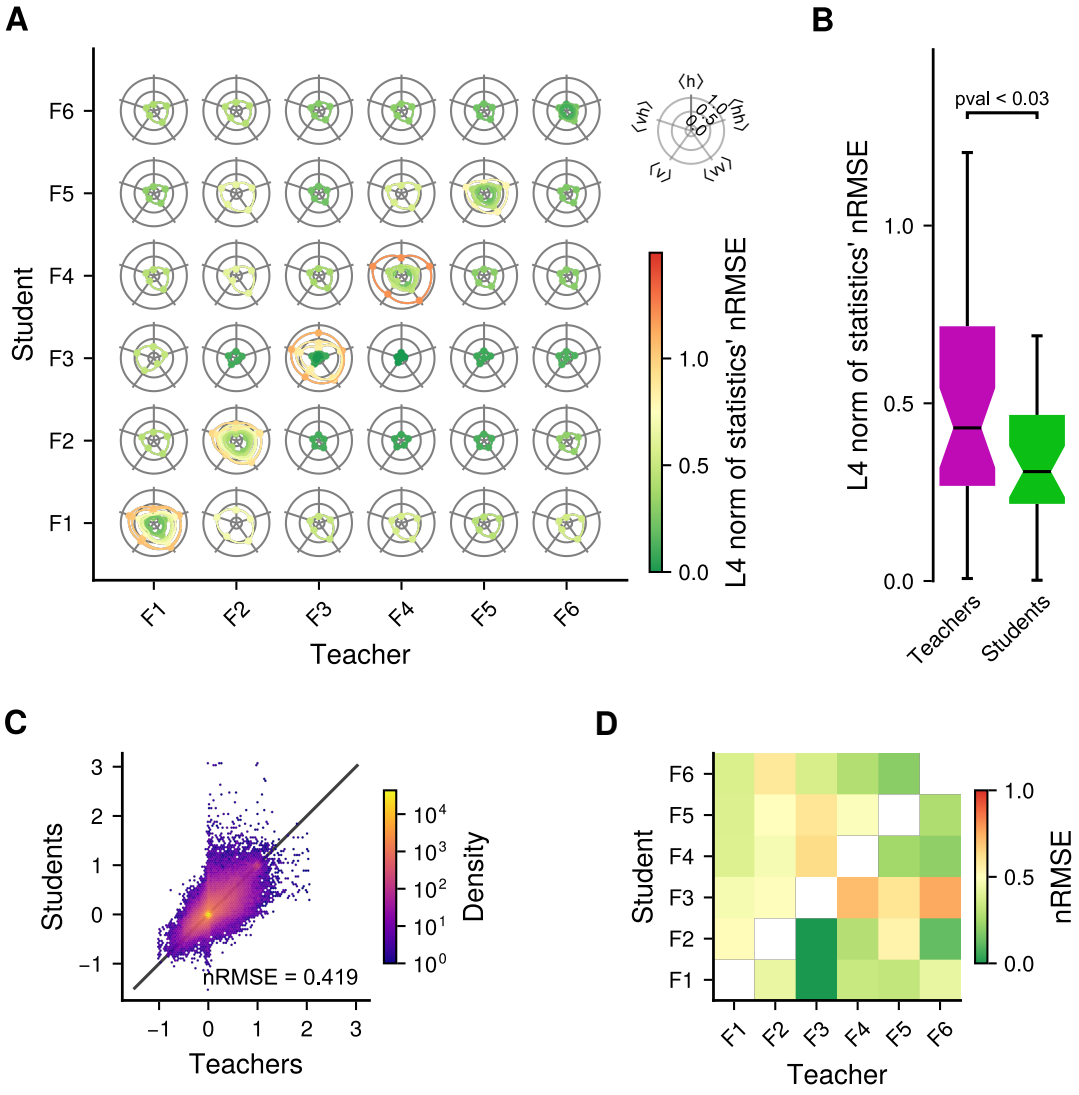


Figure 3.9: **Supplementary for Fig. 3.3 showing the training statistics of every teacher-student pairs** **A**: nRMSE (see Methods 3.4.7) of all five moments (see Methods 3.4.4.4) used to evaluate RBM convergence. Each RBM is represented by a line in polar coordinates where every ray corresponds to a moment $\langle \mathbf{v} \rangle$, $\langle \mathbf{vh} \rangle$, $\langle \mathbf{h} \rangle$, $\langle \mathbf{vv} \rangle - \langle \mathbf{v} \rangle \langle \mathbf{v} \rangle$, or $\langle \mathbf{hh} \rangle - \langle \mathbf{h} \rangle \langle \mathbf{h} \rangle$, and the ray length corresponds to the $\mathrm{nRMSE}_m(\text{data}, \text{generated})$ between empirical data and data generated by the RBM. The line color corresponds to the infinity norm of all five statistics. RBMs along the diagonal correspond to teacher RBMs trained classically (see Methods 3.4.4.2, 10 repetitions). Of-diagonal RBMs are students. (column 3 of this graph corresponds to Supp. 3.10) **B**: Comparison of training accuracy between teachers (left) and students right, measured as the L4 norm of the moments nRMSE in panel A. The p-value was computed with a one-tailed Mann-Whitney U test. **C**: Identity plot between teacher and student hidden unit pairwise covariance $\langle \mathbf{hh} \rangle - \langle \mathbf{h} \rangle \langle \mathbf{h} \rangle$, for all teacher-student pair combined. **D**: nRMSE of $\langle \mathbf{hh} \rangle - \langle \mathbf{h} \rangle \langle \mathbf{h} \rangle$ for every teacher-student pair.

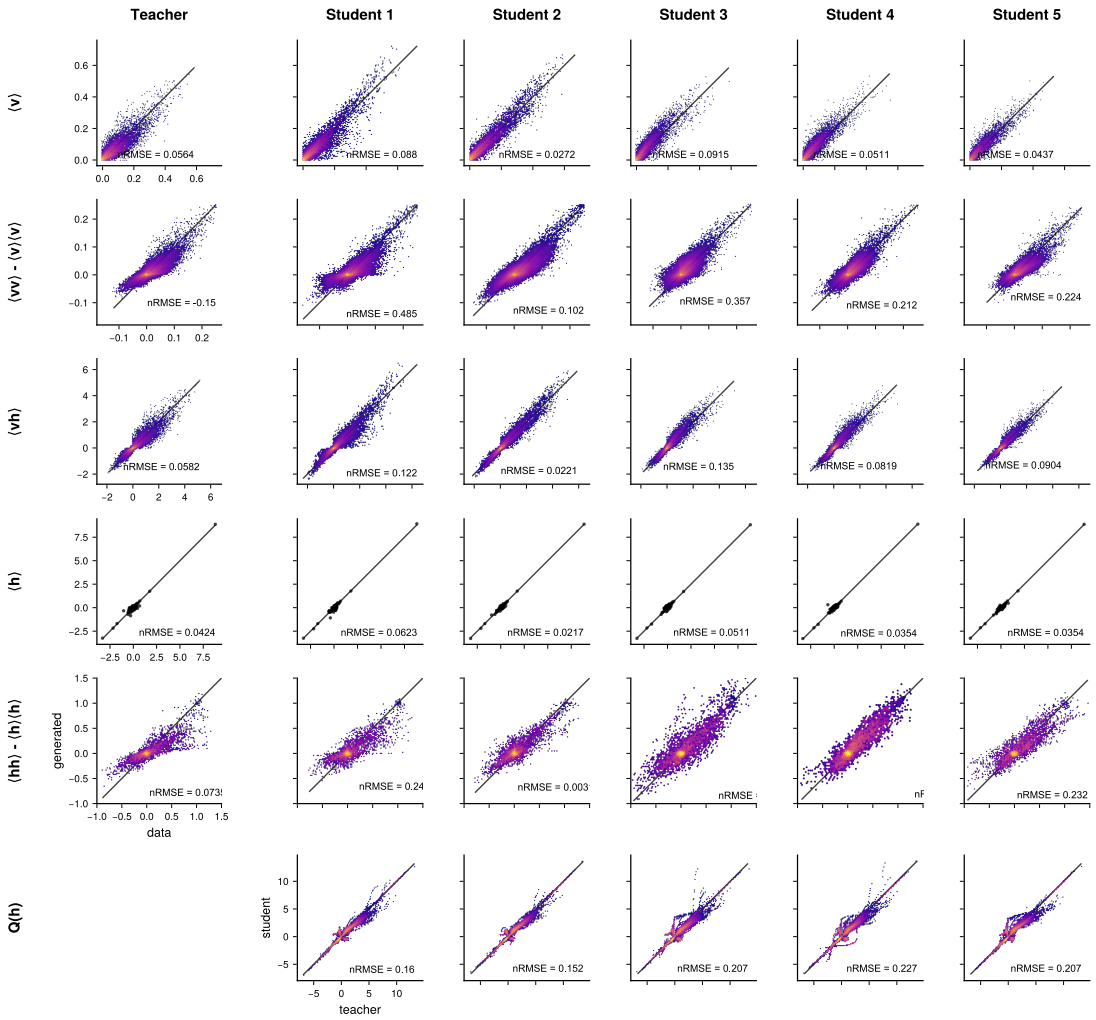


Figure 3.10: **Supplementary for Fig. 3.3 showing the training statistics and hidden distribution of all students of the example teacher fish.** Identity plots for all five moments (see Methods 3.4.4.4) used to evaluate RBM convergence and quantile-quantile plot of hidden unit prior distributions $P(h)$ between teacher and students, for the example teacher from Fig. 3.3 and its five students. (corresponds to column 3 of Supp. 3.9 A).

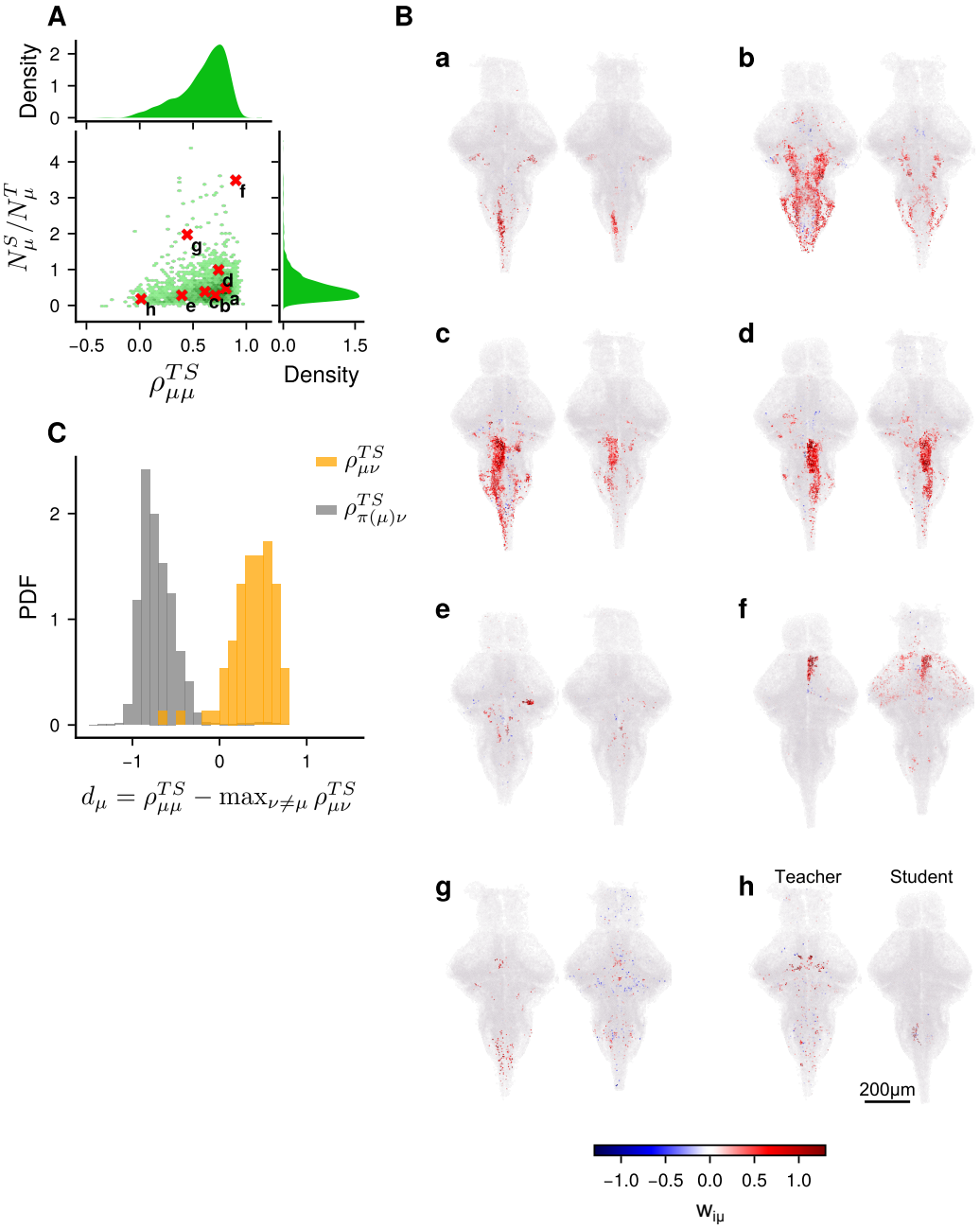


Figure 3.11: **Supplementary for Fig. 3.3 on the measurement of distances between weight maps.** **A**: Joint distribution for the spatial correlation $\rho_{\mu\mu}^{TS}$ (see Methods 3.4.8) of weight maps between teacher and students, and ratio N_μ^S/N_μ^T of the number of neurons $N_\mu^T = |w_{i\mu}^T > 10^{-5}|_i$ significantly connected with hidden unit μ in the teacher and in the student N_μ^S. This distribution was created for all hidden units in all teacher-student pairs. Labeled red crosses correspond to the maps presented in panel B. **B**: Example hidden unit weight maps for teacher (left) and student (right) RBMs. Each map pair corresponds to a red cross in panel A. **C**: Distribution of $d_\mu = \rho_{\mu\mu}^{TS} - \max_{\nu\neq\mu} \rho_{\mu\nu}^{TS}$ (see Methods 3.4.9) for all hidden units μ of the example teacher-student pair from Fig. 3.3. In orange the distribution for the observed matrix $\rho_{\mu\nu}^{TS}$, and in gray for the shuffled matrix $\rho_{\pi(\mu)\nu}^{TS}$.

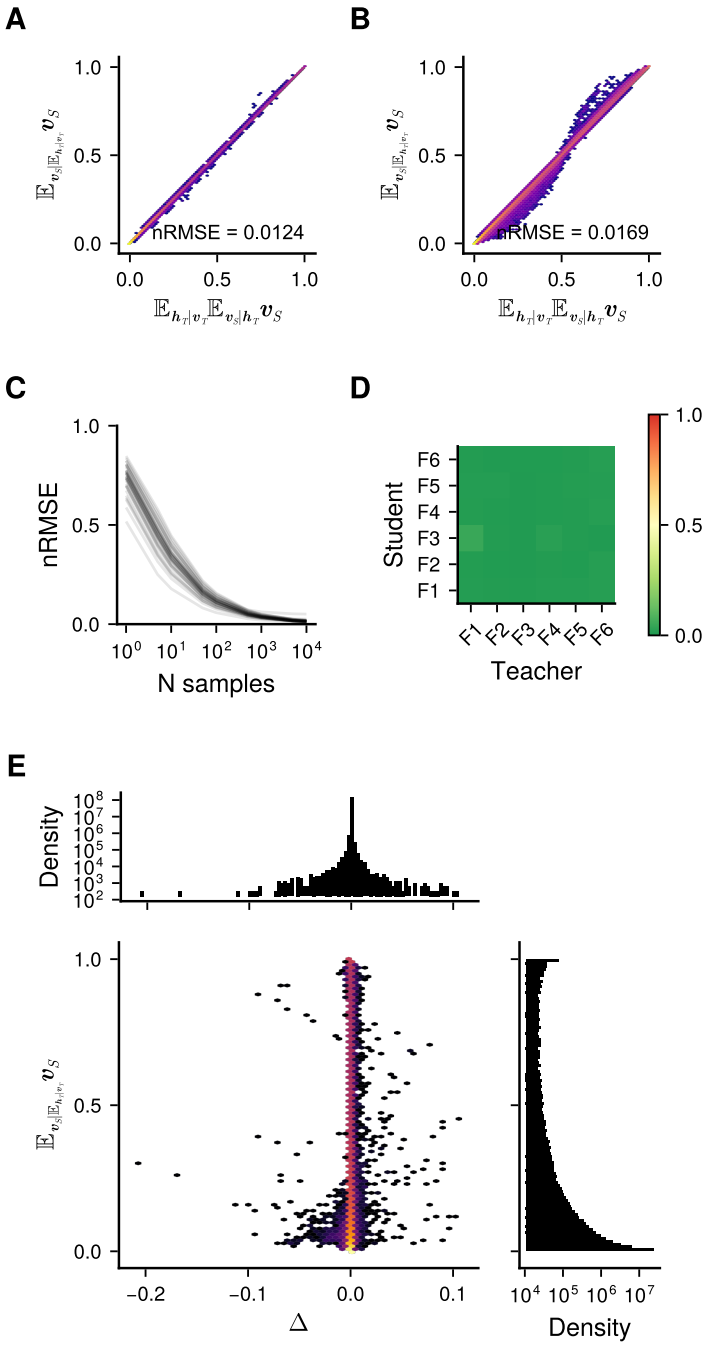


Figure 3.12: **Supplementary for Fig. 3.4 and Methods 3.4.6.4 on the method of translating neuronal configurations between fish.** **A**: Identity plot between the two methods (sampling and expectation) activity transfer presented in Methods 3.4.6.4 for an example teacher-student RBMs, and 10000 samples. **B**: Same as panel A for all pairs of teacher and student combined. **C**: nRMSE (see Methods 3.4.7) between the 2 methods (see panel A) for all teacher-student pair (one line per pair), as a function of the number of samples used. **D**: nRMSE for activity translation between every teacher-student pair (and activity reconstruction of teacher along the diagonal) for 10000 samples. **E**: Joint distribution of the sampling method and the corrective factor Δ.

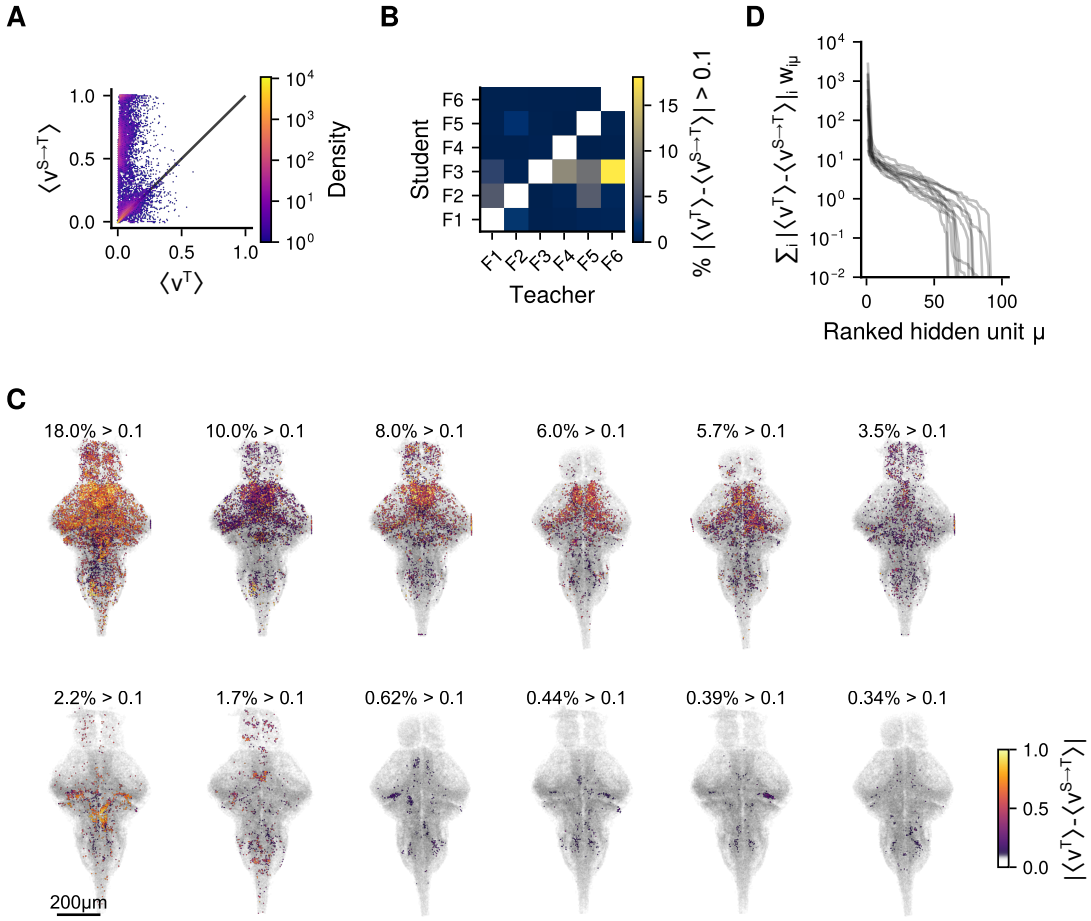


Figure 3.13: **Supplementary figures for Fig. 3.4 of the outliers presented in Fig. 3.4 C,E,G A**: Stereotypy of mean neuron activity between an example teacher $\langle \mathbf{v}^T \rangle$ and student translated in to the teacher $\langle \mathbf{v}^{S \to T} \rangle$. We chose specifically the worst outlier. **B**: Percentage of neurons in each student with a translation residual $|\langle \mathbf{v}^T \rangle - \langle \mathbf{v}^{S \to T} \rangle|$ greater that 0.1. **C**: Neuron map of translation residual for the 12 worst teacher-student pairs. **D**: Translation residuals projected onto hidden units $\sum_i |\langle \mathbf{v}^T \rangle - \langle \mathbf{v}^{S \to T} \rangle| w_{i\mu}$. One line per teacher-student pair presented in panel C.

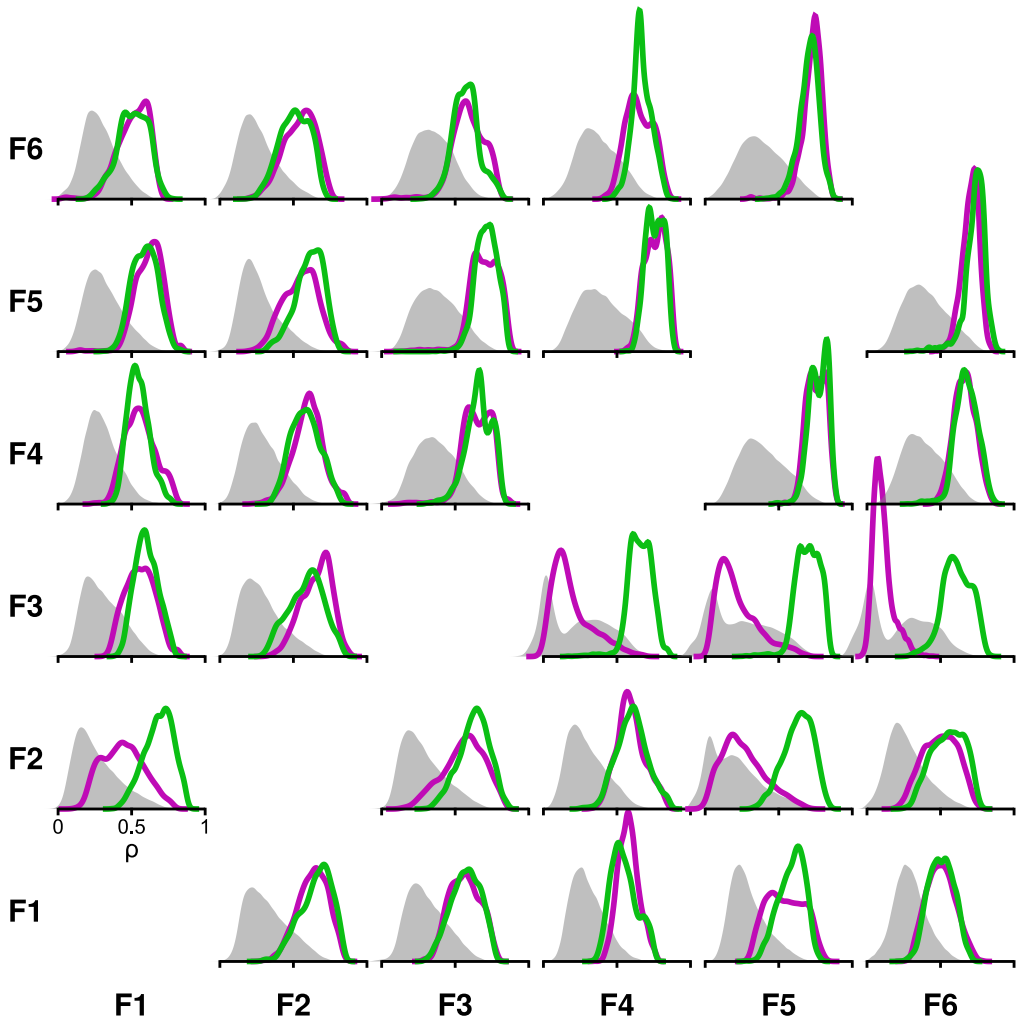


Figure 3.14: **Supplementary figures for Fig. 3.4 of the measurement of distances between weight maps.** Same as Fig. 3.4 F for every teacher-student pair. Spatial correlation $\rho_t^{T\to T,S\to T}$ (magenta) and $\rho_t^{S\to S,T\to S}$ (green) between reconstructed and translated neuronal configuration for each time frame of an example teacher-student pair. In gray, the distribution of spatial correlations between shuffled pairs of time frames.

Chapter 4

Conserved spontaneous whole-brain dynamics revealed by a compact neuronal vocabulary

4.1 Introduction

Animal behavior unfolds as a structured sequence of stereotyped motor actions [17, 20]. This suggests that brain dynamics itself should also display a common temporal structure across individuals. Evidence from *C. elegans* has shown that sequences of motor actions underlying exploratory behavior can be decoded from neuronal activity as trajectories on a low-dimensional manifold. This manifold is consistent across individuals and allows prediction of future motor commands up to 30 seconds in advance [180, 201]. Similarly, in mice and monkeys, Safaie et al. [202] demonstrated that motor cortex activity of animals performing the same task could be expressed as trajectories in a PCA latent space and aligned across individuals. Together, this exemplifies that neural population activity during behavior displays stereotypical dynamics. In contrast, extending such comparative dynamic analysis to Spontaneous Brain Activity (SA), particularly when no linked behavior is available for sequence alignment, remains considerably more challenging.

In the previous chapter, we introduced a method based on bi-trained Restricted Boltzmann Machines (RBMs) to map brain-wide spontaneous neuronal activity recorded from multiple zebrafish larvae into a single, interpretable latent space. This latent space comprises $M{=}100$ hidden units, each representing a co-activating, spatially-clustered neuronal population resembling known functional networks [11]. Such assemblies, commonly referred to as cell-assemblies, have been reported across species and brain regions, and are thought to represent fundamental units of neural computation in the brain [100, 101, 103, 102].

In this chapter, we investigate the temporal dynamics of SA across larval zebrafish by building on this shared latent representation. We extend this framework by discretizing trajectories in the latent space into finite temporal sequences of neuronal states, thereby

searching for regularities in how this space is explored by different individuals.

We demonstrate that stacking a second RBM on top of the first produces a discrete representation of brain dynamics, represented as sequences of states which capture dominant patterns of co-activation. These states are stereotypical across individuals, and we find that both state occupancies and Markovian state transitions are partially conserved. Notably, model granularity has little impact on state occupancy, with $\sim$5–6 states sufficient to account for most of the activity, suggesting the existence of a compact core repertoire with rarer refinements.

Altogether, this chapter introduces a methodology for constructing a compositional and generative model of spontaneous brain activity. This model provides an intrinsic path from neurons to assemblies to discrete states, naturally aligned across individuals, thereby enabling direct cross-fish comparisons.

4.2 Results

4.2.1 Dynamics in the shared latent space reveals conserved timescales and individual exploration patterns

To assess how spontaneous activity unfolds in the shared latent space, we first analyzed the temporal dynamics of hidden unit activity. While individual neurons fluctuate on the order of milliseconds, these assemblies evolve on substantially slower time scales [11] consistent with behavioral dynamics [20].

Hidden units evolve on conserved, second-long time scales. We find that the average hidden-unit autocorrelation $\langle \mathrm{ACF}(h_\mu) \rangle_\mu$ (see Fig. 4.1 C and Supp. 4.6 for individual hidden units) is largely consistent across fish, with a characteristic decay time $\tau \approx 2$–$3\,\mathrm{s}$. This finding is stable across teacher choices. One exception is fish 3 which displays longer dynamical persistence of $\sim 4\,\mathrm{s}$. Fish 1 and 2 exhibit weak oscillations in their autocorrelation with a period of $\sim 30\,\mathrm{s}$. This is most likely due to activity of the Anterior Rhombencephalic Turning Region which is continuously active during long periods in fish 1 and 2, but less so in the other fish (see Movie 1 in previous chapter).

The seconds-long persistence of hidden unit activity implies that latent trajectories are predictable over short time scales. This in turn suggests that zebrafish SA dynamics presents a natural temporal segmentation on the order of ≈ 2 seconds.

Fish-specific exploration within a shared latent subspace. As established previously, many hidden units have bimodal marginal distributions $P(h_\mu)$ and, during student training, these priors are enforced to be conserved across fish (see Fig. 3.3 G). Consequently, only a restricted region of the latent space is typically occupied, and this accessible subspace is shared across individuals (Fig. 4.1 D, left).

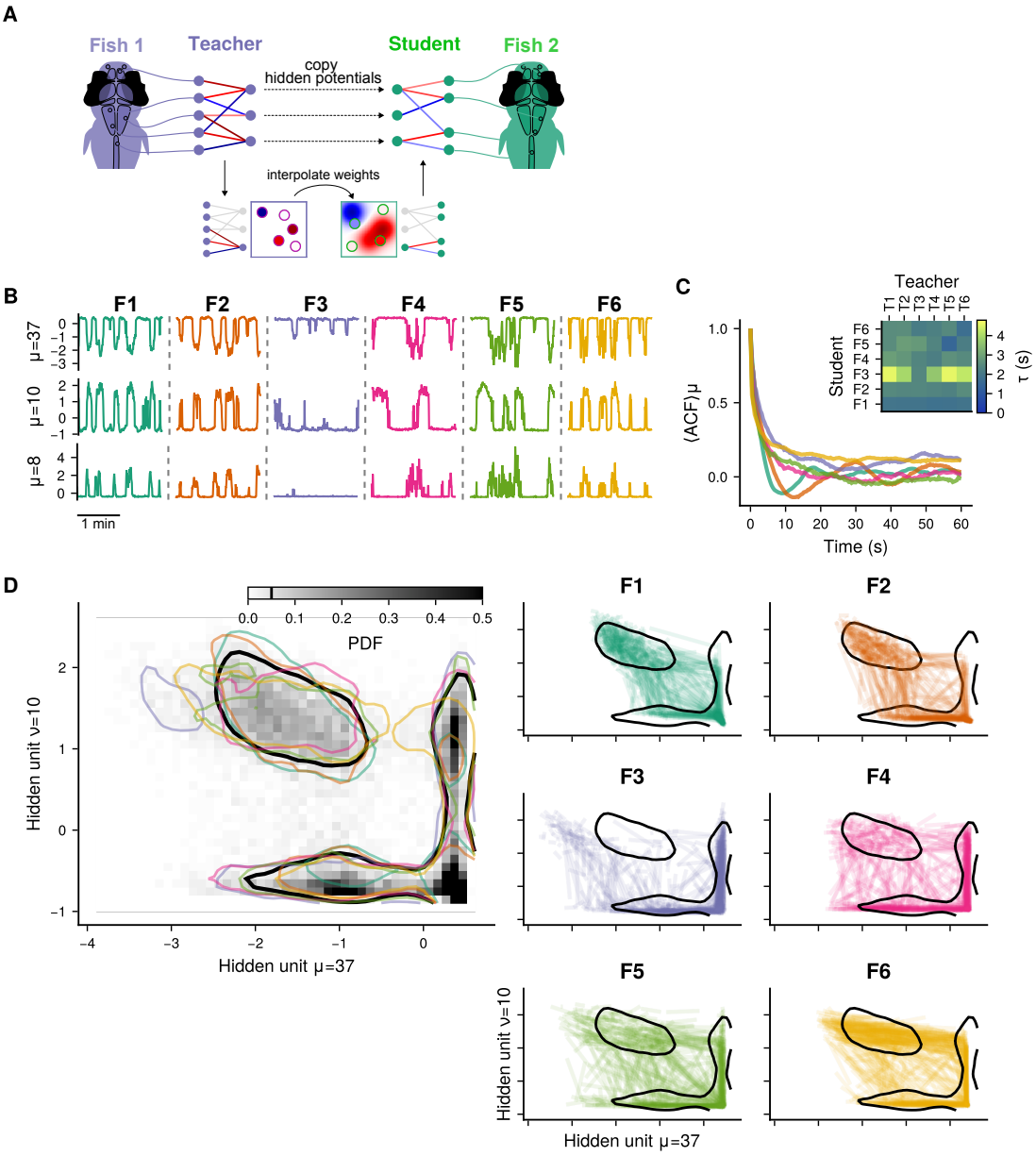


Figure 4.1: **A shared latent space with individual exploration signatures.** **A**: Schematic of the Restricted Boltzmann Machine (RBM) teacher–student (bi-training) paradigm used to construct a shared latent space across fish. **B**: Hidden activity time series $h_\mu(t) = \mathbb{E}[h_\mu \mid \mathbf{v}_t]$ for three example hidden units μ shown for all six fish. **C**: Mean autocorrelation $\left\langle \mathrm{ACF}\big(h_\mu(t)\big)\right\rangle_\mu$ across hidden units for each fish (colors match panel B). Inset: exponential fit $\langle \mathrm{ACF}\big(h_\mu(t)\big)\rangle_\mu \approx e^{-t/\tau}$ and corresponding decay time τ for every teacher–student pair. **D**: Example 2D slice of the 100D latent space (additional slices in Fig. 4.7). *Left*: joint distribution $P(h_{\mu=37},\, h_{\mu=10})$ pooled across fish (gray with black contour) with per-fish contours (colors). *Right*: empirical transitions $h_\mu(t) \to h_\mu(t{+}1)$ in the same 2D plane, shown separately for each fish.

However, by design, there is no constraint on *how* each fish's neuronal activity explores this subspace.

Examining one-step transitions $\mathbf{h}_t = \mathbb{E}[\mathbf{h} \mid \mathbf{v}_t] \to \mathbf{h}_{t+1} = \mathbb{E}[\mathbf{h} \mid \mathbf{v}_{t+1}]$ in a representative 2D slice (h_1, h_2) reveals a rich diversity of exploration patterns together with conserved motifs (Fig. 4.1 D, right). For example, fish 1 and 2 tend to access the $(h_1 = \text{high},\ h_2 = \text{high})$ quadrant symmetrically from both $(\text{low}, \text{high})$ and $(\text{high}, \text{low})$, whereas fish 6 transitions preferentially from $(\text{low}, \text{high})$. Thus, while the *accessible* latent region is shared, the *dynamics* within it are fish-specific.

In summary, the shared latent space is characterized by second-long persistence of hidden unit dynamics, a stereotyped accessible subspace, and inter-individual diversity in how this subspace is explored. These properties establish the shared RBM's hidden layer as a coherent coordinate system for comparing spontaneous neuronal dynamics across individuals.

4.2.2 Partitioning hidden activity into a neuronal state vocabulary

While hidden unit activity dynamics revealed conserved time scales and individual-specific exploration, analyzing trajectories in a 100-dimensional space can be unwieldy and difficult to interpret. To obtain a more compact and interpretable description, we extend the RBM framework by stacking a binary RBM on top of the bi-trained RBM. This model partitions the shared latent space into a finite *neuronal state vocabulary*. This yields a probabilistic mapping from spontaneous neuronal activity to a sequence of symbolic states, which we can then compare across individuals.

Implementation of the stacked RBM We will refer to the teacher and student models which map neuronal configurations to the shared latent space as brain RBMs (bRBMs). On top of a bRBM, we stack a new RBM termed the state RBM (sRBM), whose visible layer is identical in size and potential to the bRBM's hidden layer ($M{=}100$ units with $\mathcal{U}_\mu = \mathrm{dReLU}$), and whose hidden layer consists of B binary units (Fig. 4.2 A). The binary configuration $\mathbf{b} = [b_1, \ldots, b_B]$ with $b_k \in \{0, 1\}$ is interpreted as a binary number and converted to a decimal state symbol

$$s = \sum_{k=0}^{B-1} b_k 2^{B-k-1} \tag{4.1}$$

For convenience, we summarize the terminology used in the rest of this chapter:

1. **Visible layer $\mathbf{v}$**: N binary units corresponding to *neurons* (visible layer of the bRBM).
2. **Hidden layer $\mathbf{h}$**: M dReLU units corresponding to *cell-assembly*-like ensembles (hidden layer of the bRBM and visible layer of the sRBM).

3. **Binary layer** $\mathbf{b}$: B binary units representing states in binary format (hidden layer of the sRBM).

4. **State** s: one of $S = 2^B$ symbols (the decimal encoding of $\mathbf{b}$).

Importantly, the bRBM and sRBM do not share the hidden layer (*i.e.* parameters are not copied from one RBM to the other, see Fig 4.2A). The sRBM must be trained so that the prior over $\mathbf{h}$ matches, allowing hidden configurations from the bRBM to be transferred to the sRBM (see Methods 4.3).

We found that sRBMs converge reliably (Supp. 4.8 D) and accurately reproduce the prior distributions $P(h_\mu)$ across fish and teachers (Fig. 4.2 B–C), ensuring that hidden configurations can be transferred between bRBM and sRBM without informational degradation.
However, the sRBM only partially preserves pairwise covariances $\langle hh\rangle - \langle h\rangle\langle h\rangle$ from the bRBM (Fig. 4.2 D–E). This is expected given the low dimensionality of the binary layer. While it does not affect the forward mapping $\mathbf{h} \to \mathbf{b} \to s$, a reverse mapping will not, in general, reconstruct the exact pairwise interactions captured by the bRBM.

Inferring discrete state sequences from neuronal activity. Having established that the stacked RBM can reliably represent the distribution of hidden unit activities, we next use it to define the forward mapping from temporal sequences of neural configurations to discrete state sequences. This mapping is probabilistic: each neural pattern is associated with a distribution over states rather than a single deterministic label.

The forward transformation $\mathbf{h} \to \mathbf{b} \to s$ is performed by sampling $\mathbf{b} \sim P(\mathbf{b} \mid \mathbf{h})$ from the sRBM and then converting $\mathbf{b}$ to s via Eq. 4.1. Figure 4.2 F shows $\mathbb{E}[\mathbf{h} \mid s]$, which describe the "center of mass" of each state in hidden space and illustrates that states are composed of coordinated activations across multiple hidden units.

Mapping neuronal activity to states, $\mathbf{v} \to \mathbf{h} \to \mathbf{b} \to s$, is performed by first sampling $P(\mathbf{h} \mid \mathbf{v})$ under the bRBM and then $P(\mathbf{b} \mid \mathbf{h})$ under the sRBM. As this process is probabilistic, each neuronal configuration $\mathbf{v}_t$ can by mapped to every state s with the conditional probability :

$$P(s \mid \mathbf{v}_t) = \mathbb{E}\Big[P\big(s \mid P(\mathbf{h} \mid \mathbf{v}_t)\big)\Big] \tag{4.2}$$

An example time series of $P(s \mid \mathbf{v}_t)$ for $S = 32$ states is shown in Fig. 4.2 H. Two features stand out. First, many states remain probable over several consecutive frames, consistent with the slow hidden-unit dynamics mentioned previously (Fig. 4.1 C, Supp. 4.8 E). Second, at any given time, only a small subset of states concentrates most of the probability mass.

To quantify this, we track the most probable state $s_{\max}(t) = \arg\max_s P(s \mid \mathbf{v}_t)$ and its probability $P(s_{\max}) = \max_s P(s \mid \mathbf{v}_t)$ (Fig. 4.2 I). Because $\sum_s P(s \mid \mathbf{v}_t) = 1$, $P(s_{\max})$ provides a measure of certainty: values near 1 indicate that a single state dominates, while smaller values indicate ambiguity among several states.
The distributions in Fig. 4.2 G show that most neuronal configurations map to a single state with $P(s_{\max}) > 0.9$, although this fraction decreases as the number of states S increases (from 2^3 to 2^7). This pattern suggests a hierarchical organization in which larger sRBMs refine broader states into more specific sub-states.

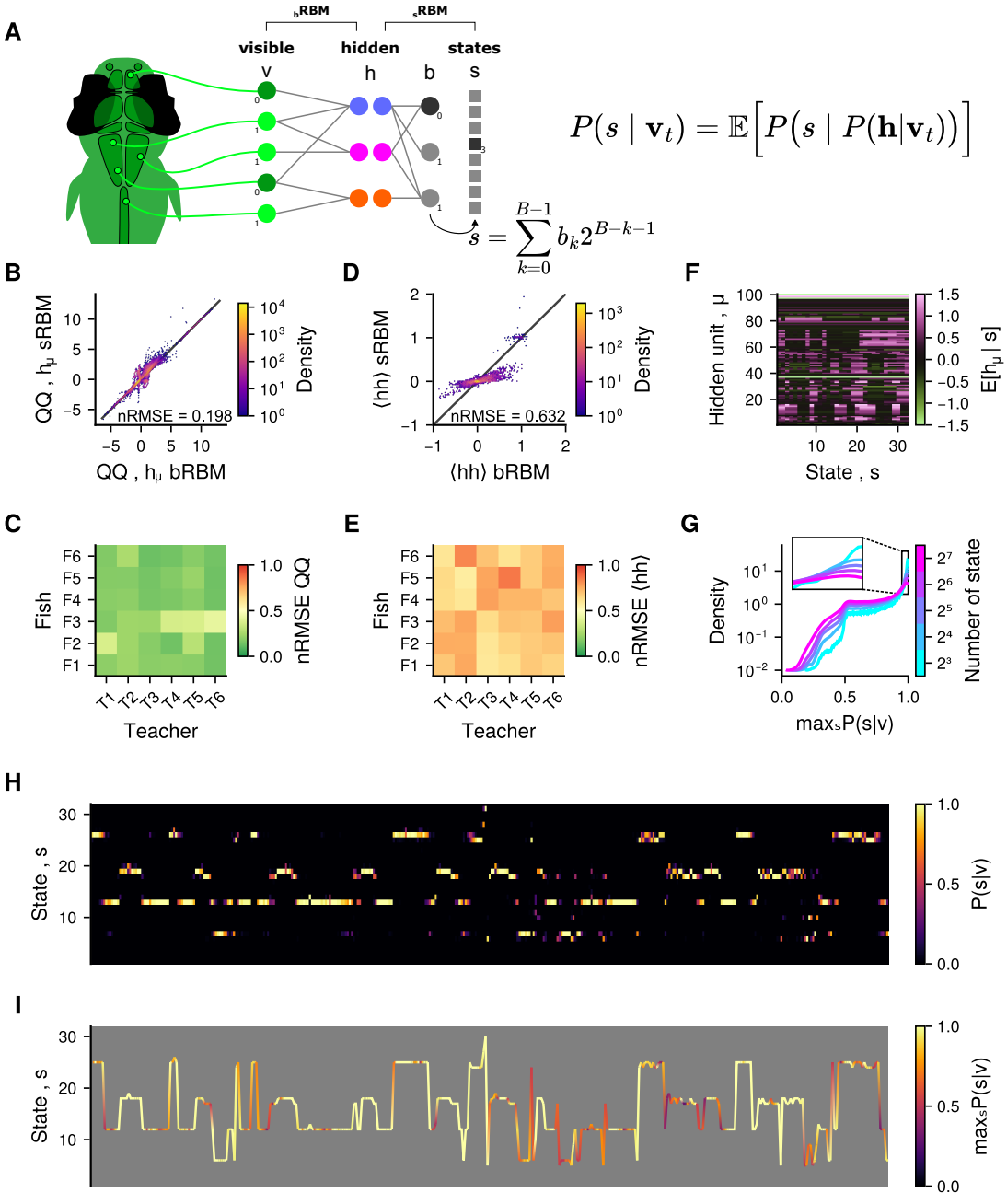


Figure 4.2: *(Caption on next page)*

Figure 4.2: **A staked RBM identifies neuronal states from regularities in cell assembly co-activations**. **A**: Diagram of the stacked architecture. The previously described brain RBM (bRBM) maps neuronal configurations $\mathbf{v}$ to hidden configurations $\mathbf{h}$ representing the collective modes of activation of the network. A second RBM, termed the state RBM (sRBM), takes $\mathbf{h}$ as its visible layer and maps it to a binary hidden layer with configuration $\mathbf{b}$. Each $\mathbf{b}$ is interpreted as a binary number and converted to a decimal state label s. **B**: Quantile–quantile plot comparing an example prior $P(h_\mu)$ under the bRBM and the sRBM. Alignment with the identity indicates that hidden configurations can be transferred between models without loss of information. **C**: nRMSE (see Methods 3.4.7) of the quantile–quantile fit for $P(h_\mu)$, aggregated over all hidden units μ, shown for each fish and teacher. **D**: Identity plot of pairwise hidden-unit covariance $\langle hh\rangle - \langle h\rangle\langle h\rangle$, computed from configurations generated by an example bRBM and sRBM. **E**: nRMSE of $\langle hh\rangle - \langle h\rangle\langle h\rangle$ between bRBM- and sRBM-generated hidden configurations, for each fish and teacher. **F**: Expected hidden configuration $\mathbb{E}[\mathbf{h} \mid s]$ (centers of mass of states in hidden space) for an example sRBM. **G**: Distribution of $P(s_{\max}) = \max_s P(s \mid \mathbf{v}_t)$, pooled across fish, for sRBMs with $S \in \{2^3, \ldots, 2^7\}$ states (one curve per S). **H**: Example time series of state probabilities $P(s \mid \mathbf{v}_t)$ (Eq. 4.2). **I**: Most-probable state sequence $s_{\max}(t) = \arg\max_s P(s \mid \mathbf{v}_t)$ for the series in panel H; line color encodes $P(s_{\max})$.

In summary, stacking a sRBM on top of bRBMs partitions the shared latent space into a common set of states. These states capture regularities in the coordinated activation of cell assemblies, offering a natural symbolic description of how functional networks are recruited during SA. Although the mapping is probabilistic, most neuronal configurations concentrate on a single, or a few, states at any given time, supporting the use of discrete state sequences for cross-fish comparisons.

4.2.3 From states to neurons: Composition and cross-individual consistency

Next, we ask which neuronal populations compose each state, and whether these compositions are conserved across individuals. To address this, we leverage the fact that RBMs model the joint distribution over visible and hidden variables, which makes conditional expectations such as $\mathbb{E}[\mathbf{v} \mid \mathbf{h}]$ directly accessible. This property underpins the interpretability of the learned representations, enabling the creation of state-conditioned neural maps which can be compared across individuals.

Binary units represent compositions of hidden assemblies. To understand how binary units relate to hidden activity, we first analyze the sRBM weight matrix $w_{\mu b}$ that connects bRBM hidden units to sRBM binary units. In the example presented in Fig. 4.3 A, most hidden units contribute (positively or negatively) to multiple binary units. This is consistent with the view that the sRBM aggregates activation patterns of the hidden layer,

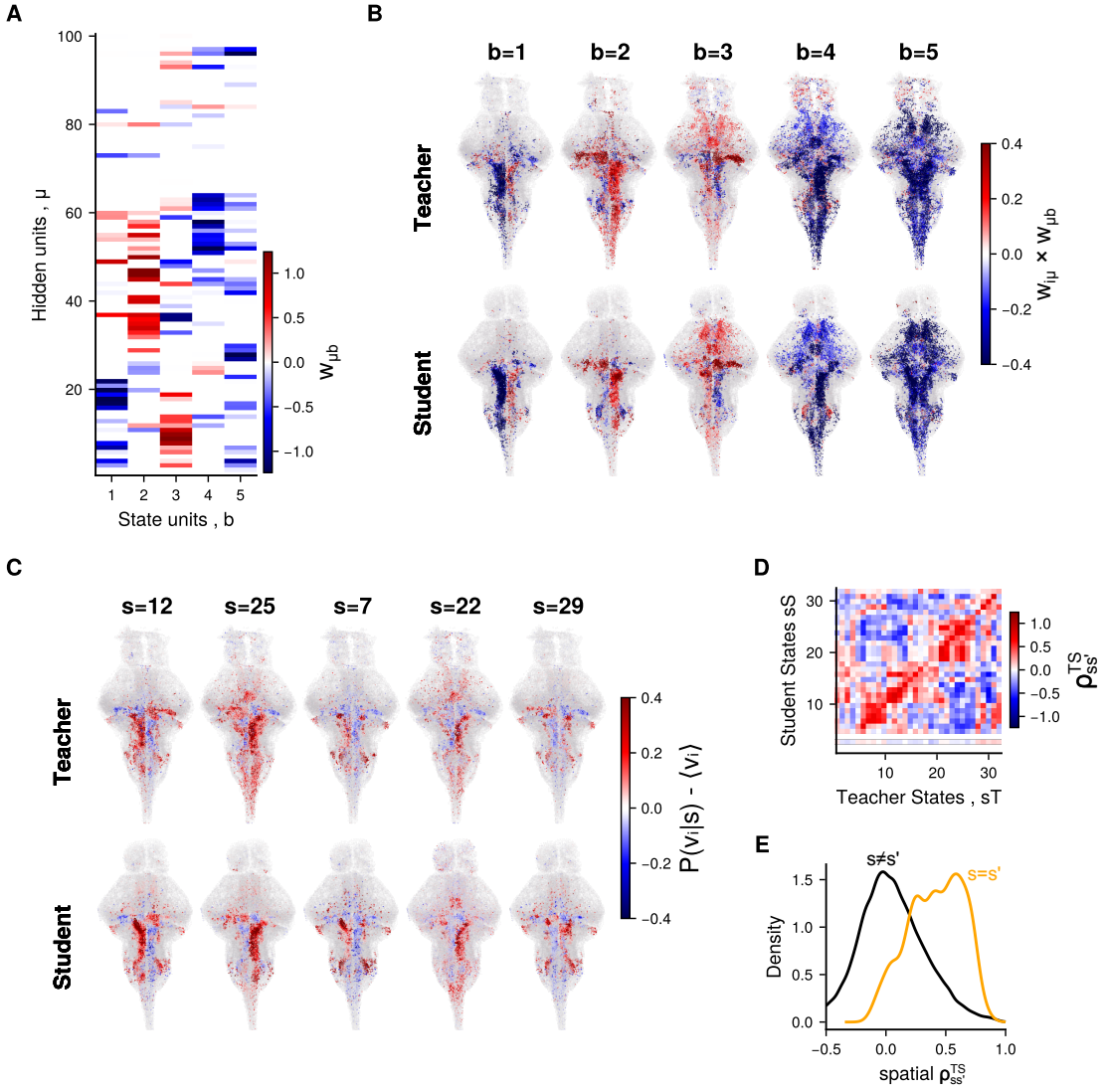


Figure 4.3: **Neuronal composition of states**. **A**: Example state RBM (sRBM) weight matrix $w_{\mu b}$ for S=32 states. **B**: Receptive fields of sRBM binary units projected into neuronal space via the bRBM hidden layer, computed as $w_{i\mu}\, w_{\mu b}$, for an example teacher–student pair with S=32. **C**: Expected deviation from mean neuronal activity, $\hat{\mathbf{v}}_s - \langle \mathbf{v} \rangle$, where $\hat{\mathbf{v}}_s$ is the Monte Carlo estimator of $P(\mathbf{v} \mid s)$ (Eq. 4.5), shown for four example states. **D**: Spatial correlation matrix $\rho^{TS}_{ss'}$ comparing teacher and student state maps, i.e., $\hat{\mathbf{v}}_s - \langle \mathbf{v} \rangle$ in the teacher versus $\hat{\mathbf{v}}_{s'} - \langle \mathbf{v} \rangle$ in the student (see Methods 3.4.8). **E**: Distributions of $\rho^{TS}_{ss'}$ pooled over all teacher–student pairs for an sRBM with S=32 states. Black: off-diagonal pairs $s \neq s'$. Orange: matched state $s = s'$ across teacher and student.

analogous to how the bRBM's hidden layer aggregates neuronal-population activity.

Projecting binary units into neuronal space through the hidden layer with $w_{i\mu} \times w_{\mu b}$ reveals that each binary unit is coupled to large neuronal populations composed of multiple functional assemblies (Fig. 4.3 B). These maps are highly conserved between teacher and student RBMs, as expected from the bi-training procedure described previously.

Together with the previous section, these observations indicate that sRBM states are *compositions* of co-activation patterns in hidden space. Because each state s corresponds to a configuration of the binary layer (converted from binary to decimal; Eq. 4.1), the stacked bRBM–sRBM architecture yields a hierarchical, physiologically interpretable representation culminating in a discrete state label.

States map onto conserved neuronal activity patterns. In principle, one could obtain $P(\mathbf{v} \mid s)$ by sampling $P\big(\mathbf{v} \mid P(\mathbf{h} \mid s)\big)$. However, because the sRBM does not fully preserve pairwise hidden interactions $\langle hh \rangle$ (Fig. 4.2 D–E), the backward sampling $P(\mathbf{h} \mid s)$ would fail. We therefore estimate $P(\mathbf{v} \mid s)$ directly from data via Monte-Carlo sampling. Specifically, for each observed neuronal configuration $\mathbf{v}_t$ ($t = 1, \ldots, T$) we draw R independent state samples (typically 30-50)

$$S_t^{(r)} \sim P(s \mid v_t) \qquad r = 1, \ldots, R, \tag{4.3}$$

From these samples we can count the number of time the state s was obtain as :

$$N_s = \sum_{t=1}^{T} \sum_{r=1}^{R} \mathbf{1}\{\, S_t^{(r)} = s \,\} \tag{4.4}$$

where $\mathbf{1}\{\,\cdot\,\}$ is the indicator function.
The Monte Carlo estimator $\hat{\mathbf{v}}_s$ of the state-conditioned neuronal map is then

$$\hat{\mathbf{v}}_s = \frac{\displaystyle\sum_{t=1}^{T} \sum_{r=1}^{R} v_t \, \mathbf{1}\{\, S_t^{(r)} = s \,\}}{\displaystyle\sum_{t=1}^{T} \sum_{r=1}^{R} \mathbf{1}\{\, S_t^{(r)} = s \,\}} = \frac{1}{N_s} \sum_{t=1}^{T} \sum_{r=1}^{R} v_t \, \mathbf{1}\{\, S_t^{(r)} = s \,\} \tag{4.5}$$

We first assess whether sRBM-defined states capture salient features of population activity. Movie 1 shows the sequence of S=32 states aligned to observed neuronal configurations for an example fish. Qualitatively, state maps summarize large-scale patterns of activity. Movie 2 repeats this analysis for different values of S. For most time frames, descriptions are consistent across S, with higher S refining coarser states, confirming our prior intuition that states are hierarchically structured.

Next, we test whether states describe comparable neuronal activity across fish. Because maps $\hat{\mathbf{v}}_s$ are influenced by the mean activity $\langle \mathbf{v} \rangle$ of neurons, we compare *de-meaned* maps $\hat{\mathbf{v}}_s - \langle \mathbf{v} \rangle$.
Example teacher–student maps are shown in Fig. 4.3 C, and we find that many states, especially those with large sample counts $N_s \gg 1$, appear conserved across individuals. To test

this quantitatively, we compute the spatial correlation $\rho^{TS}_{ss'}$ between the maps of state s in the teacher and state s' in the student (Fig. 4.3 D; see Methods 3.4.8). On average, matching states $s = s'$ are significantly correlated across teacher and student (Fig. 4.3 E), although a notable subset performs no better than shuffled pairs ($s \neq s'$). We attribute this to three effects.
First, some states differ only subtly in hidden space (e.g., by one or two hidden units; *cf.* Fig. 4.2 F), leading to small, spatially localized neuronal differences that spatial correlation may miss (see Supp. 4.9). Importantly, these subtleties are generally preserved across individuals, even if their global correlation is modest. This explains the wide distribution of $\rho^{TS}_{s\neq s'}$.
Second, the model may allocate “catch-all” or "garbage" states, with weak or mixed patterns that are effectively unconstrained. Such states are unlikely to be stereotyped across fish, and some may even be absent in some individuals (Fig. 4.2 F). This, in part, explains the negative tail of the distribution $\rho^{TS}_{s=s'}$.
Third, maps $\hat{\mathbf{v}}_s$ are the result of two projections in two different models, leading to an accumulation of error/noise. Furthermore, these maps were obtained from the averaging of empirical configurations, and are therefore subject to finite-size effects (some states where only observed a couple of times). This explains both the large width of the $\rho^{TS}_{s\neq s'}$ and the relatively small values of $\rho^{TS}_{s=s'}$.

In summary, sRBM states are composed of coordinated patterns of bRBM hidden-unit co-activation, yielding physiologically interpretable maps in neuronal space. These states capture prominent features of Spontaneous Brain Activity and provide a compact sequence representation that is comparable across fish. Importantly, we don't expect these states to capture the complexity of neuronal activity. Indeed, this stacked-RBM approach performs aggressive dimensionality reduction: from a space of roughly $2^{40,000}$ possible neuronal configurations to only 2^3–2^7 discrete states. Beyond serving as a classifier, these results further support the existence of a shared latent space of neuronal dynamics across individuals.

4.2.4 Markovian dynamics of neuronal states

The previous sections established a mapping from neuronal activity to sequences of discrete states that summarize coordinated cell-assembly activations during Spontaneous Brain Activity.
We now characterize the *dynamics* of these state sequences using first-order Markov chains, and therefore under the assumption that activity at time t depends solely on the configuration at time $t-1$. Our central question is whether such Markovian descriptions are stereotyped across individuals.

Markov Chain : transition matrices. Given the conditional state probabilities $P(s \mid \mathbf{v}_t)$, we define the one-step transition matrix between states s and s' as :

$$P_{ss'} = P(s \to s') = \frac{\sum_{t=1}^{T-1} P(s \mid \mathbf{v}_t) \cdot P(s' \mid \mathbf{v}_{t+1})}{\sum_{t=1}^{T-1} P(s \mid \mathbf{v}_t)} \quad \text{with} \quad \sum_{s'=1}^{S} P_{ss'} = 1. \tag{4.6}$$

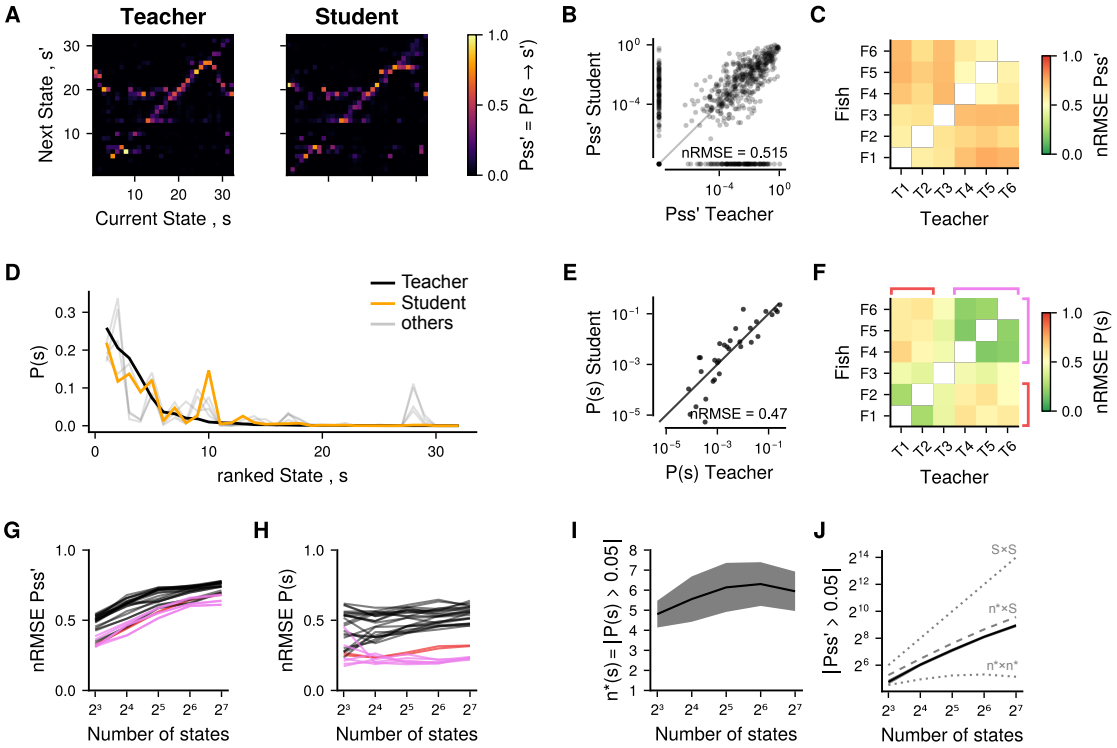


Figure 4.4: **Markovian state dynamics are highly conserved across fish**. **A**: Transition matrices $P_{ss'} = P(s \to s')$ for two example fish (see Eq 4.6). **B**: Identity plot comparing the two example transition matrices from panel A. **C**: nRMSE (see Materials 3.4.7) of $P_{ss'}$ for every fish pair. **D**: Steady state probability $P(s)$ for the two example fish (black=teacher, orange=student), and all others students (gray) (see Eq 4.7). **E**: Identity plot comparing the two example steady-state distributions from panel D. **F**: nRMSE of $P(s)$ for every fish pair. **G**: nRMSE of $P_{ss'}$ for sRBMs with different number of states. **H**: nRMSE of $P(s)$ for sRBMs with different number of states. **G-H**: Each line corresponds to a fish pair, red and purple highlight pairs within groups $[F1, F2]$ and $[F4, F5, F6]$, respectively; black denotes other pairs. **I**: Number of states with substantial occupancy : $n^*(S) = |P(s) > 0.05|$. **J**: Number of state transitions with substantial probability, $P_{ss'} > 0.05$. Gray doted lines corresponds to the number of possible transitions between all states S^2 and significantly occupied states n^{*2}. Gray dashed line corresponds to the number of possible transitions from substantially occupied states $n^* \times S$. **I-J**: Black line is the mean across all fish pairs, and gray band is the standard deviation.

Because our mapping from $\mathbf{v}_t$ to states is probabilistic, this weighted estimator replaces naive transition counts.

As shown for an example pair of fish in Fig. 4.4 A–B, corresponding entries of $P_{ss'}$ are in good agreement whenever a given transition $s \to s'$ is observed in both fish (typical $\mathrm{nRMSE} \approx 0.5$). This agreement extends across all fish pairs and is robust to the choice of teacher RBM (Fig. 4.4 C). Notably, this stereotypy is a property of the data rather than the model. Indeed, RBMs are time-agnostic and capture only configuration statistics, thus similarities in $s \to s'$ necessarily reflect similarities in $\mathbf{v}_t \to \mathbf{v}_{t+1}$.

Markov Chain : steady states. We also compare state occupancies

$$P(s) = \frac{1}{T} \sum_{t=1}^{T} P(s \mid \mathbf{v}_t) \tag{4.7}$$

which coincide with the empirical steady-state distribution of the observed sequence under our estimator.
For the same example fish-pair (Fig. 4.4 D–E), we find that $P(s)$ is highly conserved. This result generalizes across all pairs and regardless of the chosen teacher (Fig. 4.4 F).

Phenotypic similarities. Two fish groups emerge consistently from these analyses: $[F1, F2]$ and $[F4, F5, F6]$. Within-group transition matrices and, even more so, steady-state occupancies are significantly more similar than across groups (Fig. 4.4 C,F–H). These groups correspond to different clutches imaged on different weeks.
While we cannot affirm anything definitely due to a lack of statistical power, this pattern suggests that the model captures phenotypic differences and/or systematic variation in experimental conditions. Although an experienced observer might notice subtle distinctions in *raw* activity movies (Movie 1), the state-space formalism makes these differences explicit and quantifiable.

Effect of the number of states. All results above used S=32 states. Varying S (Fig. 4.4 G–H) reveals that the stereotypy of transitions $P_{ss'}$ depends only weakly on S, with larger state spaces producing slightly less stereotyped transition matrices. In contrast, the stereotypy of occupancies $P(s)$ is essentially unaffected by S, indicating that state *usage* is a stable feature.

Consistently, the number of substantially occupied states, $n^*(S) = |\{s : P(s) > 0.05\}|$, remains nearly constant (about 5–6; Fig. 4.4 I), meaning that only a small set of distinguishable states accounts for most of the observed activity. We also find that the number of substantial transitions $|\{(s, s') : P_{ss'} > 0.05\}|$ grows with $n^* \times S$ (rather than S^2 or n^{*2}, Fig. 4.4 J). This can be explained by a transition structure where those significantly occupied states act like *hubs* for the whole state-space.

Analysis of the transition structure. We present on Fig. 4.5 a graphical representation of the transition structure between the states identified by a $S = 8$ sRBM, averaged for

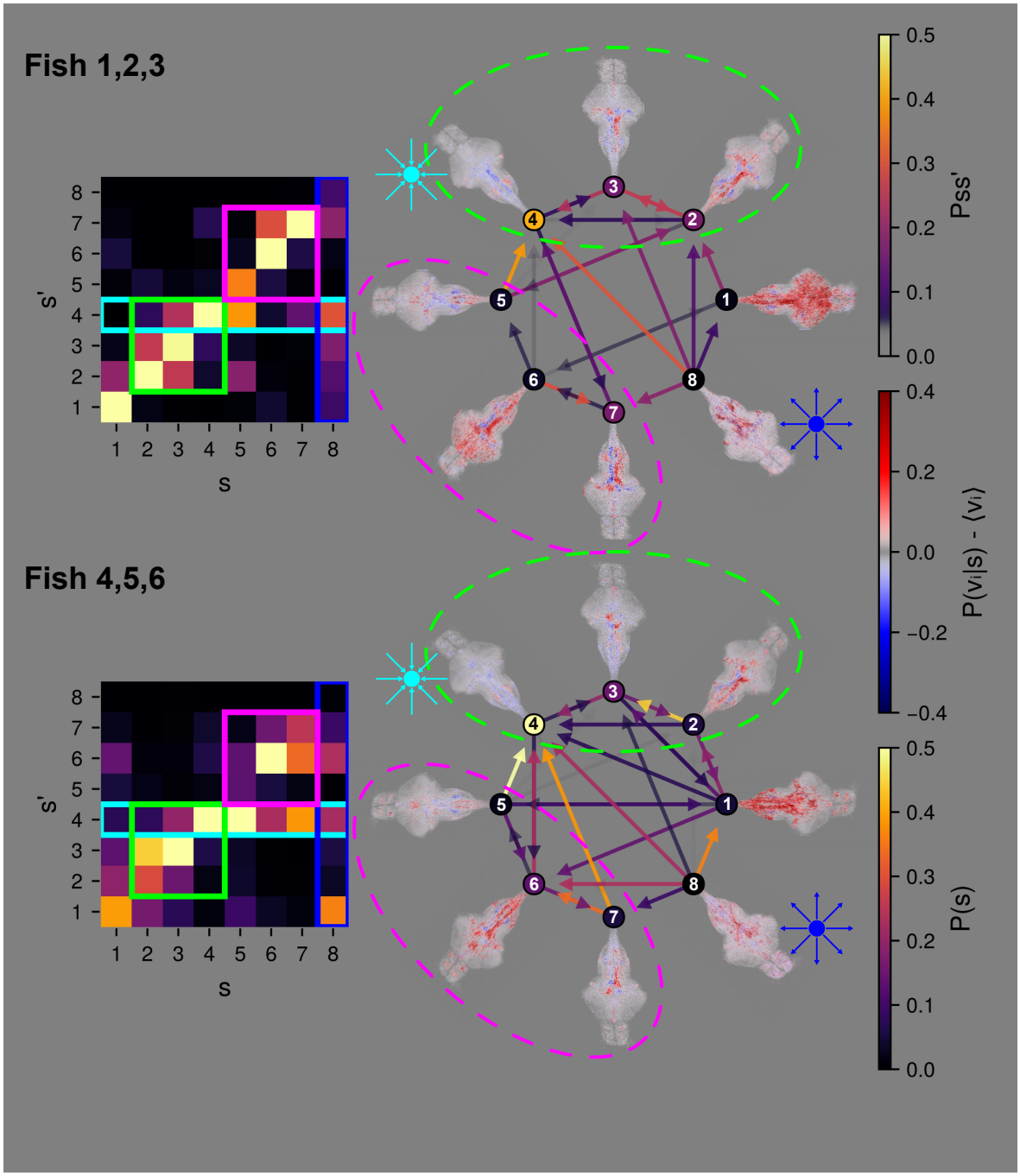


Figure 4.5: **Analysis of the markovian transition structure**. Graphical representation of the transition structure between states identified by a $S = 8$ sRBM, averaged by fish groups : $[F1, F2, F3]$ (top), and $[F4, F5, F6]$ (bottom). Neuronal maps represent the de-meaned neuronal activity conditioned on each state $\hat{\mathbf{v}}_s - \langle \mathbf{v} \rangle$ (see Eq. 4.5), combined for all fish in each group. See Supp. 4.10 for the individual transitions structures.

each fish group identified above (see Supp.4.10 for individual transition structures).

We find consistently that state 4 (the most probable) is a global *sink* of the system, receiving transitions from virtually all other states. It corresponds to a left-right symmetrical neuronal configuration where most neurons are equally or less active than on average (*inhibited*). This suggests that global inhibition of the hindbrain is stable configuration of the system. On the other hand, state 8 (low probability) is a *source*, projecting to most states, suggesting that strong and symmetrical activation of the hindbrain is an instable configuration.

We also find that the structure between states 2, 3, and 4 is very consistent across the two groups. These states correspond respectively to right-ARTR with motor neurons active, right-ARTR with motor neurons inactive, and global inhibition of the hindbrain. While transitions $2 \leftrightarrow 3$ and $3 \leftrightarrow 4$ are somewhat reversible, the transition $2 \rightarrow 4$ is very improbable in reverse. This suggests that a *right* tail bout requires first a right-activation of the ARTR, or, as indicated by transitions $\{1, 5, 8\} \rightarrow 2$, a state where motor neurons are already recruited.

A main difference between the two fish groups is the structure around state 1. This state consistently corresponds to an intense left-right symmetrical activation of most neurons, akin to what is observed during *struggle* tail movements. In the first group (fish 1, 2 and 3), state 1 projects weakly on states 2 and 6 which most likely correspond respectively to right and left swim bouts. In the second group (fish 4, 5 and 6), state 1 projects weakly to most other states.

While a detailed analysis of these states in reference to standard anatomical atlases (e.g. Zbrain [6]) is missing, these results illustrate how such a state-space combined with a Markovian analysis could shed light on whole-brain spontaneous dynamics.

In summary, the Markovian dynamics of Spontaneous Brain Activity are strongly conserved across individuals, especially within two fish groups. We consistently find 5-6 significantly occupied states which act as hubs. These results support the view that a shared, interpretable state space paired with simple Markov dynamics provides a robust framework for comparing spontaneous whole-brain activity across animals.

4.3 Discussion

In this chapter we introduced a cross-animal framework that (i) embeds whole-brain spontaneous activity from multiple zebrafish larvae into a shared latent space using bi-trained RBMs, (ii) segments that space into a finite alphabet of discrete, interpretable states via a stacked sRBM, and (iii) characterizes the resulting state sequences with a first-order Markov description.
We have shown that, although individuals access a stereotyped region of the latent space, they *explore* that region with distinct yet partially conserved motifs. The sRBM groups recurrent patterns of hidden co-activation into states which map to stereotypical neuronal

patterns, enabling direct comparisons across animals.
Finally, state occupancy, and to a lesser extent the transition structure between states, are partially conserved across fish, especially within two fish groups, suggesting phenotypic, environmental, or experimental-conditions effects on spontaneous dynamics.

From assemblies to symbols: a compositional view. A central thread of this work is compositionality: bRBM hidden units capture distributed neuronal assemblies, and the sRBM composes those assemblies into discrete states. This is based on previous theoretical and empirical analyses showing that RBMs can enter a compositional regime in which sparse weights and appropriate nonlinearities lead to a description of data points in terms of a limited number of latent features [141, 142, 11]. Our stacked bRBM–sRBM design leverages these properties to move from graded assembly activations to a symbolic description of brain-wide activity that is naturally aligned across subjects. Beyond interpretability, the generative nature of the model permits bidirectional mappings among neurons, assemblies, and states, supporting mechanistic hypotheses about how coordinated neuronal populations are recruited in concert.

State dynamics, hierarchy, and metastability. Modeling the state sequence as a Markov chain reveals robust cross-animal regularities. The consistency of steady-state probabilities with respect to the number of available states, along with a modest dependence of transition stereotypy, suggests a small core of frequently occupied states surrounded by a larger halo of rarer, more specific refinements. This pattern matches observations that large-scale brain activity alternates among a small number of recurrent configurations whose organization appears hierarchical, with faster "sub-states" nested within slower dynamics [203, 204]. It is also consistent with the view that, across scales and species, brain activity exhibits *metastability* (*i.e.* long-lived but transient regimes punctuated by transitions) [205, 206, 207, 208].

Relation to alternative state-space approaches. There is a rich literature on discrete-state models of neuroimaging data, including Hidden Markov Models (HMMs) [208] and switching linear dynamical systems (SLDS), which parse complex time series into recurrent dynamical modes with Markovian switching [203, 209]. Our framework is complementary: instead of inferring states directly from raw activity or pairwise connectivity, we first construct a shared, generative latent space grounded in assembly structure and only then discretize it. This ordering offers two advantages. First, the bRBM provides a biologically interpretable basis (in term of cell assemblies) that maps states back to neurons and functional networks. Second, the bi-training procedure aligns subjects in the latent space, enabling *direct* cross-animal state comparisons without post hoc alignment. Recent work shows that population activity often resides in low-dimensional manifolds that can be shared across subjects or tasks [210, 180, 211]. Our results add that a layered, generative approach can yield a compact symbolic representation on top of such manifolds while preserving interpretability.

Biological interpretation and limitations. The observation that a small number of core states accounts for a large fraction of spontaneous activity suggests that whole-brain dynamics repeatedly recruits a limited repertoire of mesoscale configurations [183]. In zebrafish larvae, one main contributor is the ARTR and associated hindbrain networks, which display pseudo-periodic, anti-phasic dynamics and influence turning behavior [79, 167, 171, 173]. The cross-individual conservation of transition structure and state occupancy points to constraints in circuit organization and/or common internal drives, even as individuals exhibit idiosyncratic exploration of the accessible latent subspace.

However, the method we present here favors in priority the most structured activity, at the expense of other, less obvious, dynamics. This is exemplified in our case by the domination of hindbrain/midbrain dynamics (particularly the ARTR) over the forebrain. This effect shows the limits of whole-brain analysis, and might be alleviated by analyzing brain regions separately. For example, one might imagine fitting different RBMs on forebrain, midbrain and hindbrain data, and recombining them into a whole-brain model subsequently.

Methodological considerations. Several methodological limitations frame the interpretation of these findings.
(i) **Temporal sampling and time scales.** Light-sheet acquisition rates on the order of 2 Hz constrain the resolvable dynamics to seconds and above [125, 126]. Our emphasis on assembly-level dynamics and state persistence is thus well matched to the measurement regime but omits millisecond-scale phenomena.
(ii) **Biased neuronal recordings.** The six brain recordings presented in this chapter were acquired on a 1-photon scanning light-sheet microscope with no LASER-eye blocker. This means that the fish's eye on the side of the illumination objective were constantly stimulated during the whole recording. This has two major effects. First, asymmetrical eye illumination induces a phototactic response, producing asymmetrical brain activity [167, 171]. Second, constant visual stimulation inhibits activity in most of the optic tectum.
(iii) **Markovian assumption.** First-order Markov models are tractable and capture much of the structure, but they may miss higher-order dependencies. Further work with this type of modeling should test for memory effects, particularly when comparing SA with EA.
(iv) **State granularity.** Increasing the number of states refines transitions but leaves core occupancies largely unchanged. The probabilistic formulation of the present model lends itself nicely to more advanced hierarchical methods and quantification of uncertainty.
(v) **Back-mapping limitations.** Because the sRBM compresses hidden-layer dependencies, backward sampling $P(\mathbf{h} \mid s)$ does not fully reconstruct pairwise structure. We therefore estimated $P(\mathbf{v} \mid s)$ empirically, a pragmatic choice that preserves interpretability but is not a substitute for a fully invertible generative model.

Implications and future directions. The combination of functional alignment, symbolic state discovery, and Markovian dynamics yields a general language for comparing whole-brain spontaneous activity across individuals. This framework could be extended by :
(i) **Task/perturbation experiments** learning shared state spaces that encompass Spontaneous Brain Activity and Evoked Brain Activity datasets, testing how evoked states relate

to or reorganize the spontaneous vocabulary and dynamics.
(ii) **Behavioral experiments** coupling state transitions to motor dynamics to test causal hypotheses about sensorimotor loops.
(iii) **Development and phenotype** tracking individuals across multiple days to quantify maturation of state repertoires and testing for heritable or environmental-dependent signatures.
Finally, the hierarchical organization suggested by our results invites models that learn multiscale state trees explicitly, bridging assemblies, mesoscale networks, and whole-brain dynamics. Potentially linking mechanistic circuit models with the statistical structure of spontaneous activity.

Methodological Details

Data and Code Availability

All code, models, and post-processed data used in the present article are available in a centralized repository at : `https://github.com/EmeEmu/MultiDynRBM/tree/v.PhD.1`. Raw data is available upon request (approximately 400GB).

Datasets and models

Throughout this chapter, we used the same six recordings and the same teacher and student RBMs introduced in the previous chapter (Chap. 3). Briefly, student RBMs were initialized by spatially interpolating the teacher's weight matrix onto the student fish and copying the teacher's hidden-unit potentials. They were then *bi-trained* to reproduce both the fish's neuronal statistics and the teacher's hidden-layer statistics (Fig. 4.1 A). This procedure yielded 30 teacher–student pairs, and we showed that the initial choice of teacher had negligible impact on model performance. For consistency, we continue to illustrate results with the same example teacher–student pair as before, noting that all results generalize across teachers.

State RBM training

We trained the sRBM separately from the bRBM. For each fish, we drew $\approx 32{,}000$ hidden configurations from its bRBM using Gibbs sampling (roughly ten times the number of time frames), we then concatenated all fish-specific samples into a single training set for the sRBM. Training was performed with the following hyperparameters :

- $B \in \{3, \dots, 7\}$ binary units;
- L_1^2 regularization with coefficient $\lambda_{21} = 0.5$;

- $100{,}000$ gradient updates with 15 Markov chain steps between updates;
- mini-batches of 256 configurations;
- learning rate 10^{-3} for the first quarter of training, then geometrically annealed to 10^{-4} by the end.

Supplementary Information

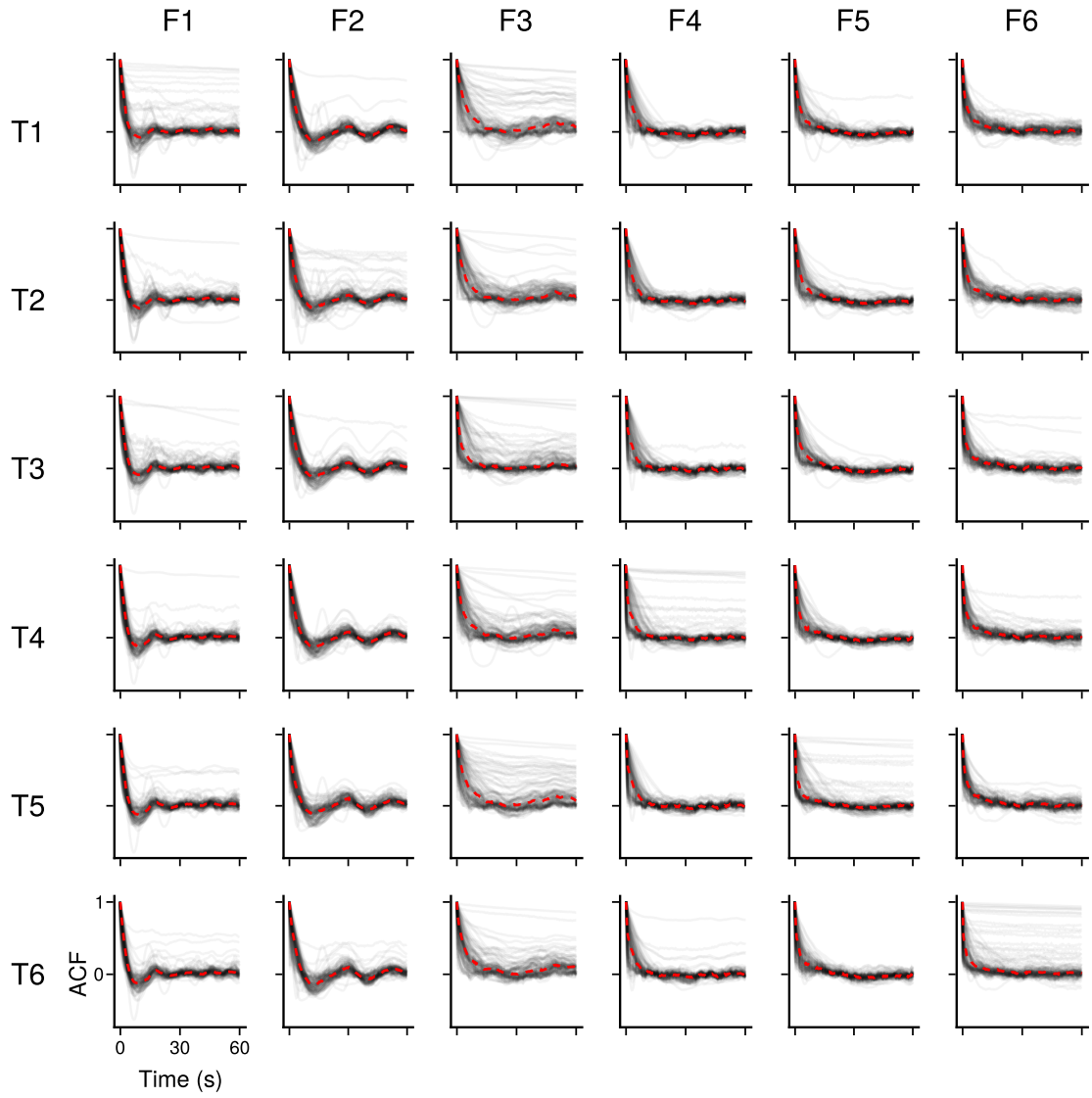


Figure 4.6: **Supplementary for Fig. 4.1 C**. Autocorrelation for each student, from every teacher, and for every hidden unit (gray lines). Average autocorrelation $\left\langle \mathrm{ACF}\big(h_\mu(t)\big)\right\rangle_\mu$ are shown as dashed red lines.

Movie Movies 1-2 and their captions can be found here `https://doi.org/10.5281/zenodo.16886970`.

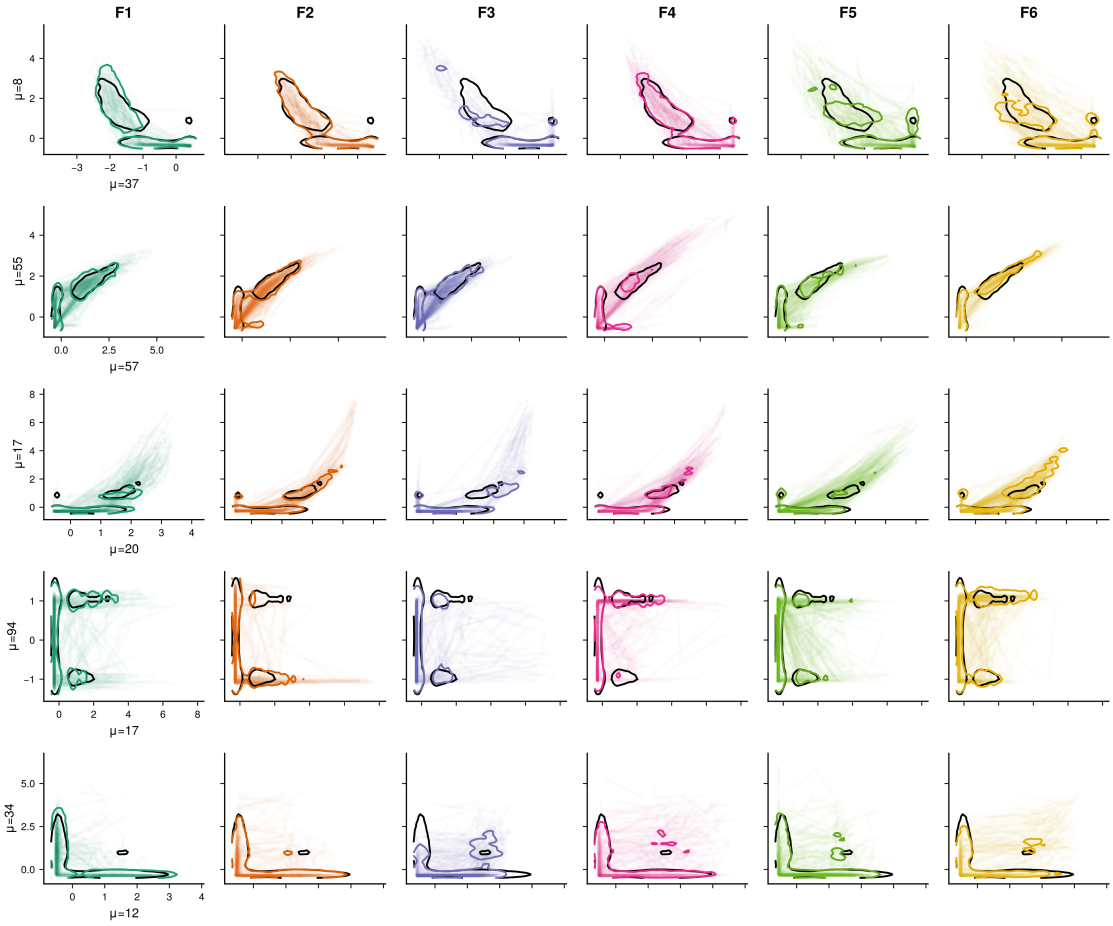


Figure 4.7: **Supplementary for Fig. 4.1 D**. Example 2D slices of the 100D latent space. Each column corresponds to a fish (F3 teacher and the rest students), and each line corresponds to a 2D slice in hidden space. Same caption as Fig 4.1D, with black contour representing the joint distribution over all fish, and colored contours representing the individual distributions.

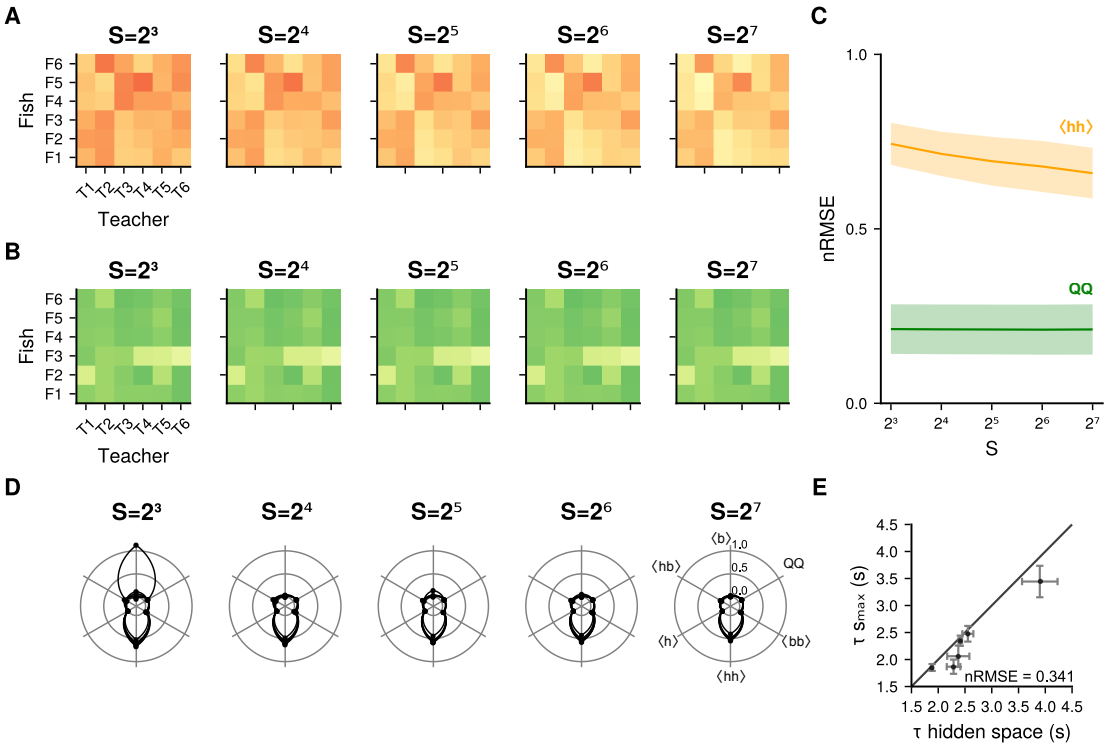


Figure 4.8: **Supplementary for Fig. 4.2.** **A**: Same as Fig. 4.2 E for sRBMs with different numbers of states S. **B**: Same as Fig. 4.2 C for sRBMs with different numbers of states S. **C**: Mean (lines) and standard deviation (bands) of nRMSE matrices in panels A (orange) and B (green) per number of states S. **D**: nRMSE (see Methods 3.4.7) of all five moments (see Methods 3.4.4.4) used to evaluate sRBM convergence. One polar plot per RBM, one line per fish. **E**: Characteristic autocorrelation decay time τ of hidden configurations vs. most probable state sequence $s_{max}(t)$. One point per fish. Horizontal error bars where computed as standard error of the mean (sem) over multiple teacher RBMs. Vertical error bars where computed as sem over 50 samples of $s_{max}(t)$ and over all 5 sRBMs.

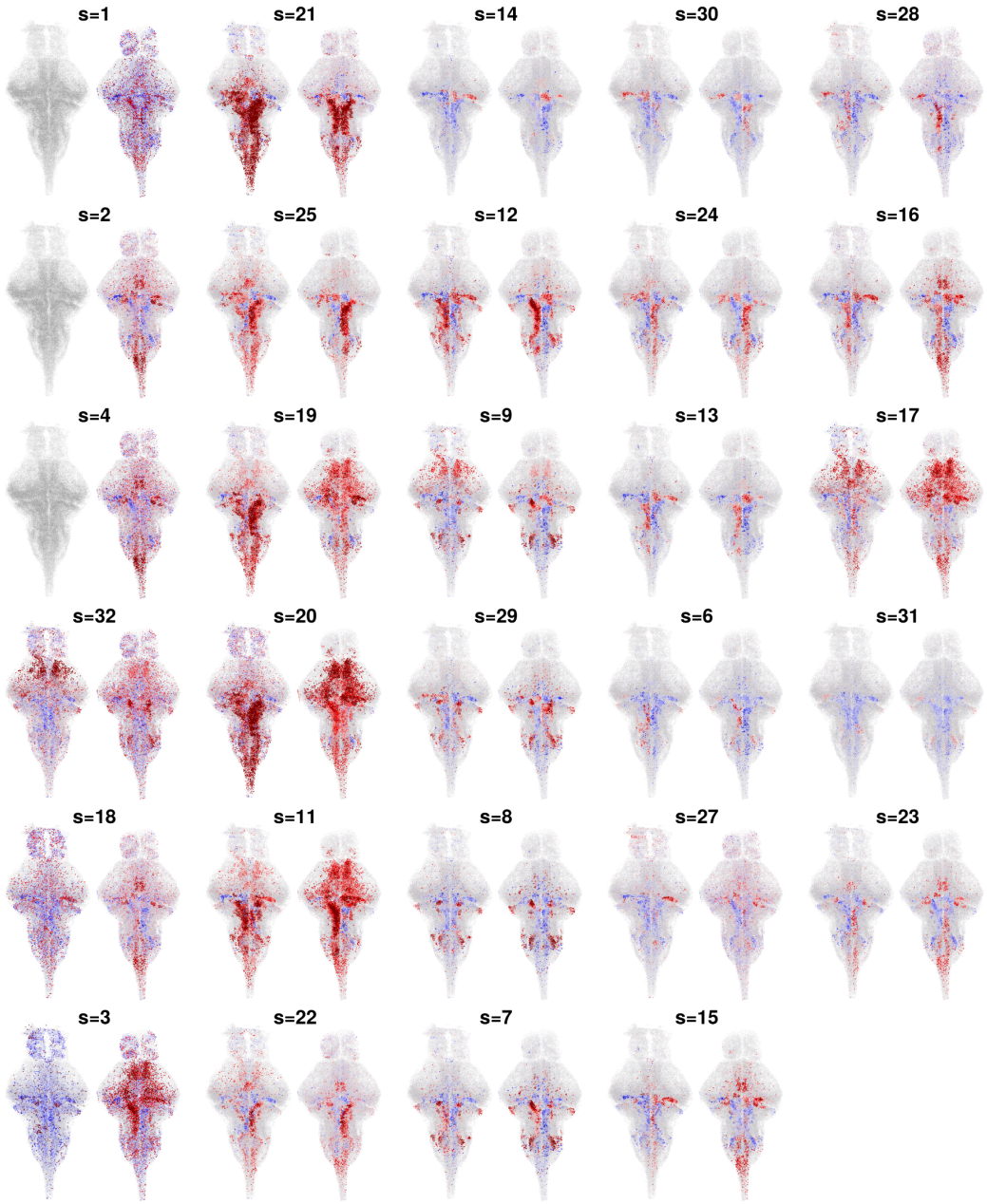


Figure 4.9: **Supplementary for Fig 4.3C**. Same as panel Fig 4.3C, for every state of a 32-states-sRBM, sorted manually. For each state, left is the teacher and right is the student.

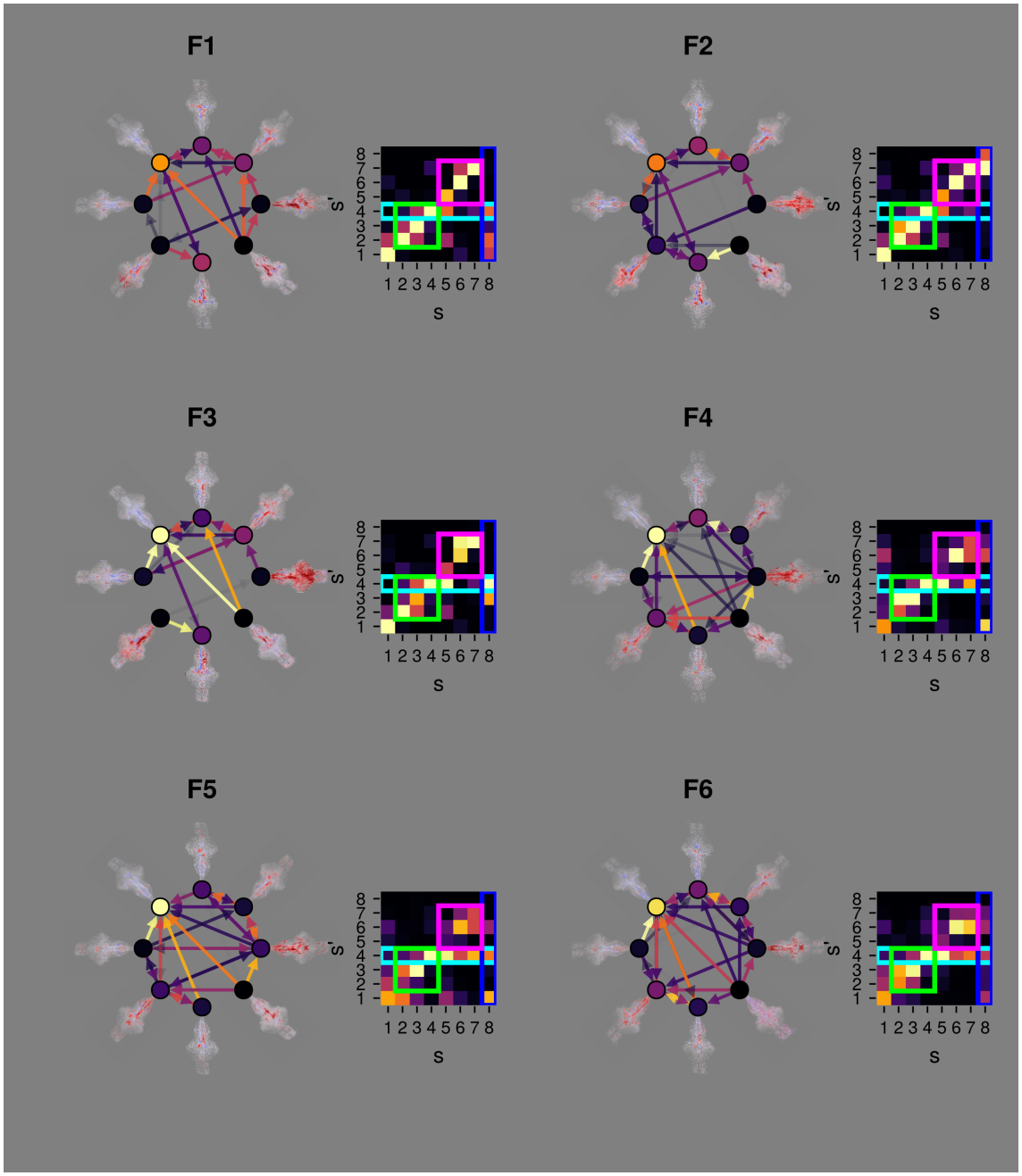


Figure 4.10: **Supplementary for Fig 4.5**. Same as Fig 4.5 for each fish individually.

Chapter 5

Discussion and Perspectives

5.1 Summary of main results

Across the last three chapters, we developed an operational framework in which Spontaneous Brain Activity (SA) can be compared across individuals. Building on the idea that SA is an intrinsic and relatively stable feature of brain activity, as well as the fact that SA both structures and is structured by the neuronal network, we hypothesized the existence of a representational level at which spontaneous neuronal activity is comparable between individuals of the same species.

Chapter 2. We used the same Hidden Markov Model (HMM) architecture to model freely swimming trajectories and Anterior Rhombencephalic Turning Region (ARTR) population activity in zebrafish larvae. A three-state model was sufficient to reveal bout-type persistence underestimated by manual segmentation, yielding a more Markovian description of reorientation behavior. From this simple model, individual fish could be identified from their Markovian statistics. We further showed that ARTR neurons are well captured by the same three-state HMM. After a simple temporal rescaling, we showed that transition rates in behavior and in the ARTR align: left-/right-dominant ARTR map to left/right turns, and balanced ARTR activity maps to forward bouts. This suggests that a single state space bridges spontaneous locomotion and its turning circuit.

Chapter 3. Using Restricted Boltzmann Machines (RBMs), we introduced two methods to build a shared latent space for spontaneous, whole-brain, single-neuron recordings of multiple zebrafish larvae. First, voxelizing each brain on a common grid and training a single RBM on concatenated activity produced a latent representation that is more stereotyped across fish. Second, we proposed a bi-training paradigm that constrains RBMs trained on different individuals to share hidden-layer priors and spatial organization, aligning cell assemblies across animals. Translating neuronal activity from one fish to another through this latent space preserved key spatial and statistical structure, indicating that SA admits a unified representation across individuals.

Chapter 4. Leveraging this shared latent space, we compared SA dynamics across individuals. By stacking a state RBM on top of the latent space we segmented brain activity into a finite neuronal "vocabulary". Modeling state sequences with first-order Markov chains showed that transition rates and steady-state occupancies are partially conserved, especially within two groups of fish, suggesting that this analysis could reveal phenotypic or internal-state differences. These findings establish shared rules governing how individuals explore spontaneous state space.

5.2 Future directions

5.2.1 Applications

Sampling the space of brain-states.
Many neural processes unfold over long timescales, from hours (e.g., circadian rhythms and slow internal-state fluctuations) to days (e.g., learning and development). Because brain-state probabilities are not uniform, exhaustive sampling of the space of possible brain states would, in principle, require very long recordings. In practice, functional brain recordings are short, so the state space is underexplored. The framework developed in this thesis suggests an alternative: pool brain states across animals. Under the working hypothesis that individuals explore the same underlying state space, such pooling is statistically analogous to extending a single recording, but also naturally accounts for individual variabilities.

Analysis of Phenotypes. Beyond studying Spontaneous Brain Activity to better understand brain function and organization, the methods in Chapters 3 and 4 could be used to compare brain activity across phenotypes and experimental conditions. The proposed framework offers a natural way to detect alteration in functional organization and dynamics, in mutant fish or following pharmacological treatments.

The same methodology could also be used to compare the brain activity of a single individual over multiple sessions, particularly in paradigms where the same neurons cannot be imaged across sessions.

Single-neuron functional fingerprints. While most of our analyses focused on comparing brain activity across individuals at the assemblies or brain-wide scales, it might be possible to extend the same framework to comparisons at the single-neuron scale. In sensorimotor paradigms, the *functional fingerprint* of individual neurons can be identified as their responses to controlled sensory inputs and motor variables (e.g., tuning curves, latency profiles, gain modulation). Much as transcriptomic studies assign genetic types, such fingerprints assign *functional types* that are comparable across animals at coarse-grained scales (e.g. response maps to vestibular stimulation [116]), but might also be predictable at the single-neuron level. Indeed, the shared latent space introduced in Chap. 3 maps neurons from different animals to shared functional ensembles. We could then test whether we can infer the fingerprint of a neuron in fish B from the brain activity of fish A. Concretely, we

could translate evoked activity patterns from fish A through the shared latent space and predict the response profile of target neurons in fish B. These predictions could then be validated by presenting the same stimuli to fish B and comparing predicted vs. measured fingerprints. Success would establish that functional identity is at least partially conserved across individuals and that our generative framework can *transfer* single-neuron function, not just assembly-level structure.

Joint *in vivo* and *in silico* optogenetic perturbations. Multi-fish RBMs can guide hypotheses about functional organization that are testable with optogenetic activation/ablation. With precise perturbations of opsin-expressing neurons, it is now feasible to alter the structure and activity of targeted populations (e.g. an ongoing project in the lab uses 3D holographic two-photon optogenetics based on acousto-optic deflectors [212]). The physiological interpretability of RBM-identified assemblies could be tested by activating a small set of neurons and checking for pattern completion of the assembly. Likewise, the RBM-predicted couplings between ensembles could be assessed statistically by repeated stimulation of an assembly or its constituent neurons. Although RBMs produce degenerate coupling models, this degeneracy is largely lifted at coarse-grained scales of $\approx 20\,\mu\mathrm{m}$ (see Chap. 3). Finally, because RBMs are generative, one could constrain a subset of neurons and predict effects on the rest of the circuit, then test these predictions optogenetically, effectively performing joint *in vivo*/*in silico* optogenetics.

Practically, such experiments would require RBMs trained on the same fish being stimulated: a brief recording, rapid training, and then optogenetic testing. A whole-brain, single-neuron RBM currently takes $\sim$2 h to train from scratch, which is impractical. Using the teacher–student paradigm from Chap. 3, we reduced the number of gradient steps by one order of magnitude, lowering training time to $\sim$20 min. We expect that a finetuning of training hyperparameters could further reduce this time, enabling *quasi-online* RBM modeling.

5.2.2 Methodological improvements and extensions

This thesis presented three main methods applied to three problems:

- Joint brain–behavior description with HMMs.
- A shared latent space of whole-brain activity built from bi-trained RBMs.
- A stacked RBM with a Markov chain to study sequences of shared neuronal states.

We outline several possible improvements.

Co-trained brain RBMs. A limitation of the bi-training method (Chap. 3) is its reliance on a single reference fish (the teacher) to which all other fish (students) are constrained. Although the choice of teacher has little measured impact on student accuracy, this design

may bias results, especially when fish differ markedly (e.g., across phenotypes or behavioral states). Moreover, while the method maps multiple fish into a shared space, it does not yield a unified model applicable to any subsequent fish.

A potential solution would be to train multiple RBMs jointly, one per fish, but in parallel, with shared hidden-unit potentials $\mathcal{U}_\mu$, repurposing the constraints introduced for bi-training. To ensure that hidden units describe the same ensembles across fish, we could introduce a weight field $w_{\vec{x}\mu}$ that connects any position $\vec{x}$ in a shared anatomical space to a hidden unit μ. This field would be learned alongside the RBMs and sampled per fish to produce weights matrices $w_{i\mu}$. The pair $\{w_{\vec{x}\mu}, \mathcal{U}_\mu\}$ would then define a generic RBM, and new fish could be incorporated by training only a few parameters (e.g., visible fields g_i). This would enable very fast training and simplify cross-fish comparisons.

A functional atlas of the zebrafish brain. Such a model could support functional atlases of the zebrafish brain, analogous to anatomical atlases (e.g., ZBrain [6]). One could build generic RBMs for different phenotypes and share them publicly to map new recordings into a common latent space, enabling quantitative comparisons of functional organization and dynamics across conditions and laboratories.

Mixing RBMs and HMMs. In Chap. 4, we segmented the latent space into a fixed set of states, then analyzed the resulting sequences with Markov chains. As shown in Chap. 2, such a two-step approach can introduce biases because state segmentation is learned independently of temporal structure. A natural improvement would be to replace the state RBM with an Hidden Markov Model (HMM), a project already initiated by Guillaume Faye-Bedrin.

This approach could also address the generation of temporally faithful dynamics. Gibbs sampling from RBMs matches per-frame statistics but not long-timescale dynamics. Conditioning RBM-generated configurations on an HMM state space could yield activity that better matches empirical temporal structure, in the spirit of what has already been done with Temporal RBMs by Monnens et al. [213].

Split-brain RBMs. As discussed in Chap. 4, RBMs tend to prioritize the most structured activity. In our data, ARTR and related circuits dominate the whole-brain activity, therefore many hidden units map to midbrain/hindbrain ensembles. A possible improvement would be to train separate RBMs on distinct brain regions and, when needed, recombine them into a whole-brain model. This might enforce a more uniform assembly tiling of the whole brain, but also allow for faster and parallelized trainings.

Anatomical priors. This type of anatomical priors could be extended by adding further constraints during training, for example by including information on the neurotransmitter type of each neuron, or by initializing the RBM weight matrix based on structural connectivity data. Indeed, van der Plas et al. [11] showed that the region-wise functional connectivity inferred from trained RBMs reflects the structural connectivity. With the recent publica-

tions of *C. elegans* and *Drosophila* whole-brain connectomes at cellular resolution [22, 131], and the soon to be published zebrafish connectomes, we could compare neuron-to-neuron structural and functional connectivities, or use it as an initialization and/or training constraint.

Sensory-motor RBMs. All recordings analyzed in this thesis were obtained in paralyzed animals and without explicit sensory stimulation. Extending the analysis to sensorimotor paradigms by including behavioral and/or stimulation variables directly in the visible layer is a natural next step. For example, during my Master's internship with head-embedded, tail-free larvae, I observed instances of clear behavioral switching: several fish alternated minutes-long epochs of sustained tail activity with long passive periods. An RBM with an additional binary visible unit encoding active vs. passive behavior would effectively split the energy landscape into two regions. Sampling the model while constraining this unit would generate neuronal configurations specific to each state.

5.3 On the importance of interpretable brain models.

Finally I would like to discuss one last subject : the interpretability of neuronal models.

With the fast improvements in machine learning over the last decade, and particularly in the last 4 years with the rise of Large Language Models, a recent trend has emerged in computational neuroscience where more and more complex models are being used to analyze and model brain activity. A representative example is Azabou et al. [214], who used a transformer architecture with tokenized neuronal activity to learn a model that maps neural data from multiple animals into a shared behavioral space predictive of behavior. Impressively, their model generalizes across individuals, tasks, recording methods (including different temporal resolutions), and laboratories, and can be applied to new data with minimal retraining.

While this and related work [215, 216, 217] achieve impressive mappings from neural activity to behavior, such models are largely black boxes and provide limited mechanistic insight. In particular, they make it difficult to map from the latent space back to neuronal activity in a way that clarifies how individual neurons and assemblies contribute to the representation. By contrast, probabilistic models like RBMs are less impressive but far more interpretable. Their two-layer bipartite architecture is simple, and probabilistic modeling provides a natural pathway from data to latent representation and back. In Chapters 3 and 4, we showed that this simple, interpretable architecture can be shared across individuals, supports transfer of activity from one individual to another, and preserves a mechanistic view of brain organization and function.

This work testifies that small-scale, interpretable models can still advance our understanding of the brain, encouraging *slow science* and *slow machine learning*.

Bibliography

[1] Kandel Eric Richard, Jessell Thomas M., and Schwartz James Harris. "Principles of Neural Science / Edited by Eric R. Kandel,... James H. Schwartz,... Thomas M. Jessell,..." In: *Principles of Neural Science*. 3rd edition. Norwalk London: Appleton & Lange Prentice-Hall international, 1991. ISBN: 0-8385-8068-8.

[2] Olaf Sporns, Giulio Tononi, and Rolf Kötter. "The Human Connectome: A Structural Description of the Human Brain". In: *PLOS Computational Biology* 1.4 (Sept. 30, 2005), e42. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.0010042. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.0010042 (visited on 08/09/2025).

[3] Danielle S. Bassett and Olaf Sporns. "Network Neuroscience". In: *Nature Neuroscience* 20.3 (3 Mar. 2017), pp. 353–364. ISSN: 1546-1726. DOI: 10.1038/nn.4502. URL: https://www.nature.com/articles/nn.4502 (visited on 04/07/2023).

[4] Olaf Sporns. *Networks of the Brain*. Cambridge, MA, USA: MIT Press, Feb. 12, 2016. 424 pp. ISBN: 978-0-262-52898-6.

[5] Maria Antonietta Tosches et al. "Evolution of Pallium, Hippocampus, and Cortical Cell Types Revealed by Single-Cell Transcriptomics in Reptiles". In: *Science* 360.6391 (May 25, 2018), pp. 881–888. DOI: 10.1126/science.aar4237. URL: https://www.science.org/doi/10.1126/science.aar4237 (visited on 08/09/2025).

[6] Owen Randlett et al. "Whole-Brain Activity Mapping onto a Zebrafish Brain Atlas". In: *Nature Methods* 12.11 (11 Nov. 2015), pp. 1039–1046. ISSN: 1548-7105. DOI: 10.1038/nmeth.3581. URL: https://www.nature.com/articles/nmeth.3581 (visited on 04/07/2023).

[7] Philipp Schlegel et al. "Whole-Brain Annotation and Multi-Connectome Cell Typing of Drosophila". In: *Nature* 634.8032 (Oct. 2024), pp. 139–152. ISSN: 1476-4687. DOI: 10.1038/s41586-024-07686-5. URL: https://www.nature.com/articles/s41586-024-07686-5 (visited on 08/09/2025).

[8] Fabian Svara et al. “Automated Synapse-Level Reconstruction of Neural Circuits in the Larval Zebrafish Brain”. In: *Nature Methods* 19.11 (Nov. 2022), pp. 1357–1366. ISSN: 1548-7105. DOI: 10.1038/s41592-022-01621-0. URL: https://www.nature.com/articles/s41592-022-01621-0 (visited on 08/09/2025).

[9] Ashwin Vishwanathan et al. “Predicting Modular Functions and Neural Coding of Behavior from a Synaptic Wiring Diagram”. In: *Nature Neuroscience* 27.12 (Dec. 2024), pp. 2443–2454. ISSN: 1546-1726. DOI: 10.1038/s41593-024-01784-3. URL: https://www.nature.com/articles/s41593-024-01784-3 (visited on 08/09/2025).

[10] C. J. Honey et al. “Predicting Human Resting-State Functional Connectivity from Structural Connectivity”. In: *Proceedings of the National Academy of Sciences* 106.6 (Feb. 10, 2009), pp. 2035–2040. DOI: 10.1073/pnas.0811168106. URL: https://www.pnas.org/doi/10.1073/pnas.0811168106 (visited on 08/09/2025).

[11] Thijs L van der Plas et al. “Neural Assemblies Uncovered by Generative Modeling Explain Whole-Brain Activity Statistics and Reflect Structural Connectivity”. In: *eLife* 12 (Jan. 17, 2023). Ed. by Peter Latham and Laura L Colgin, e83139. ISSN: 2050-084X. DOI: 10.7554/eLife.83139. URL: https://doi.org/10.7554/eLife.83139 (visited on 04/05/2023).

[12] Paul Dean and John Porrill. “The Importance of Marr’s Three Levels of Analysis for Understanding Cerebellar Function”. In: *Computational Theories and Their Implementation in the Brain: The Legacy of David Marr*. Ed. by Lucia M. Vaina and Richard E. Passingham. Oxford University Press, Nov. 3, 2016, p. 0. ISBN: 978-0-19-874978-3. DOI: 10.1093/acprof:oso/9780198749783.003.0004. URL: https://doi.org/10.1093/acprof:oso/9780198749783.003.0004 (visited on 08/09/2025).

[13] Peter Dayan and L. F. Abbott. *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. Computational Neuroscience. Cambridge, Mass: Massachusetts Institute of Technology Press, 2001. 460 pp. ISBN: 978-0-262-04199-7.

[14] Wulfram Gerstner and Werner M. Kistler. *Spiking Neuron Models: Single Neurons, Populations, Plasticity*. Cambridge: Cambridge University Press, 2002. ISBN: 978-0-521-89079-3. DOI: 10.1017/CBO9780511815706. URL: https://www.cambridge.org/core/books/spiking-neuron-models/76A3FC77EC2D24CDD (visited on 08/09/2025).

[15] L.F. Abbott. “Theoretical Neuroscience Rising”. In: *Neuron* 60.3 (Nov. 2008), pp. 489–495. ISSN: 08966273. DOI: 10.1016/j.neuron.2008.10.019. URL: https://linkinghub.elsevier.com/retrieve/pii/S0896627308008921 (visited on 08/09/2025).

[16] David J. Anderson and Pietro Perona. "Toward a Science of Computational Ethology". In: *Neuron* 84.1 (Oct. 1, 2014), pp. 18–31. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2014.09.005. pmid: 25277452. URL: https://www.cell.com/neuron/abstract/S0896-6273(14)00793-4 (visited on 08/09/2025).

[17] Joshua M. Mueller et al. "Drosophila Melanogaster Grooming Possesses Syntax with Distinct Rules at Different Temporal Scales". In: *PLOS Computational Biology* 15.6 (June 26, 2019), e1007105. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1007105. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1007105 (visited on 07/08/2025).

[18] Andrew M Seeds et al. "A Suppression Hierarchy among Competing Motor Programs Drives Sequential Grooming in Drosophila". In: *eLife* 3 (Aug. 19, 2014), e02951. ISSN: 2050-084X. DOI: 10.7554/eLife.02951. pmid: 25139955. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4136539/ (visited on 08/09/2025).

[19] Gordon J. Berman et al. "Mapping the Stereotyped Behaviour of Freely Moving Fruit Flies". In: *Journal of The Royal Society Interface* 11.99 (Oct. 6, 2014), p. 20140672. DOI: 10.1098/rsif.2014.0672. URL: https://royalsocietypublishing.org/doi/full/10.1098/rsif.2014.0672 (visited on 07/08/2025).

[20] Alexander B. Wiltschko et al. "Mapping Sub-Second Structure in Mouse Behavior". In: *Neuron* 88.6 (Dec. 16, 2015), pp. 1121–1135. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2015.11.031. URL: https://www.sciencedirect.com/science/article/pii/S0896627315010375 (visited on 12/10/2023).

[21] John Graham White et al. "The Structure of the Nervous System of the Nematode Caenorhabditis Elegans". In: *Philosophical Transactions of the Royal Society of London. B, Biological Sciences* 314.1165 (Jan. 1997), pp. 1–340. DOI: 10.1098/rstb.1986.0056. URL: https://royalsocietypublishing.org/doi/10.1098/rstb.1986.0056 (visited on 08/09/2025).

[22] Steven J. Cook et al. "Whole-Animal Connectomes of Both Caenorhabditis Elegans Sexes". In: *Nature* 571.7763 (July 2019), pp. 63–71. ISSN: 1476-4687. DOI: 10.1038/s41586-019-1352-7. URL: https://www.nature.com/articles/s41586-019-1352-7 (visited on 08/09/2025).

[23] George F. Striedter. *Principles of Brain Evolution*. Principles of Brain Evolution. Sunderland, MA, US: Sinauer Associates, 2005, pp. xii, 436. xii, 436. ISBN: 978-0-87893-820-9.

[24] A. Arieli et al. "Dynamics of Ongoing Activity: Explanation of the Large Variability in Evoked Cortical Responses". In: *Science (New York, N.Y.)* 273.5283 (Sept. 27, 1996), pp. 1868–1871. ISSN: 0036-8075. DOI: 10.1126/science.273.5283.1868. pmid: 8791593.

[25] Artur Luczak, Peter Barthó, and Kenneth D. Harris. "Spontaneous Events Outline the Realm of Possible Sensory Responses in Neocortical Populations". In: *Neuron* 62.3 (May 14, 2009), pp. 413–425. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2009.03.014. URL: https://www.sciencedirect.com/science/article/pii/S0896627309002372 (visited on 03/24/2025).

[26] Pietro Berkes et al. "Spontaneous Cortical Activity Reveals Hallmarks of an Optimal Internal Model of the Environment". In: *Science* 331.6013 (Jan. 7, 2011), pp. 83–87. DOI: 10.1126/science.1195870. URL: https://www.science.org/doi/10.1126/science.1195870 (visited on 05/11/2025).

[27] J. S. Damoiseaux et al. "Consistent Resting-State Networks across Healthy Subjects". In: *Proceedings of the National Academy of Sciences* 103.37 (Sept. 12, 2006), pp. 13848–13853. DOI: 10.1073/pnas.0601417103. URL: https://www.pnas.org/doi/10.1073/pnas.0601417103 (visited on 08/09/2025).

[28] Emily S. Finn et al. "Functional Connectome Fingerprinting: Identifying Individuals Using Patterns of Brain Connectivity". In: *Nature Neuroscience* 18.11 (Nov. 2015), pp. 1664–1671. ISSN: 1546-1726. DOI: 10.1038/nn.4135. URL: https://www.nature.com/articles/nn.4135 (visited on 08/09/2025).

[29] Timothy O. Laumann et al. "Functional System and Areal Organization of a Highly Sampled Individual Human Brain". In: *Neuron* 87.3 (Aug. 5, 2015), pp. 657–670. ISSN: 1097-4199. DOI: 10.1016/j.neuron.2015.06.037. pmid: 26212711.

[30] Evan M. Gordon et al. "Precision Functional Mapping of Individual Human Brains". In: *Neuron* 95.4 (Aug. 16, 2017), 791–807.e7. ISSN: 1097-4199. DOI: 10.1016/j.neuron.2017.07.011. pmid: 28757305.

[31] Eyal Bergmann et al. "Individual Variability in Functional Connectivity Architecture of the Mouse Brain". In: *Communications Biology* 3.1 (1 Dec. 4, 2020), pp. 1–10. ISSN: 2399-3642. DOI: 10.1038/s42003-020-01472-5. URL: https://www.nature.com/articles/s42003-020-01472-5 (visited on 04/07/2023).

[32] Caterina Gratton et al. "Functional Brain Networks Are Dominated by Stable Group and Individual Factors, Not Cognitive or Daily Variation". In: *Neuron* 98.2 (Apr. 18, 2018), 439–452.e5. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2018.03.035. URL: https://www.sciencedirect.com/science/article/pii/S0896627318302411 (visited on 04/07/2023).

[33] Kevin M. Anderson et al. "Heritability of Individualized Cortical Network Topography". In: *Proceedings of the National Academy of Sciences* 118.9 (Mar. 2, 2021), e2016271118. DOI: 10.1073/pnas.2016271118. URL: https://www.pnas.org/doi/10.1073/pnas.2016271118 (visited on 08/09/2025).

[34] Marcus E. Raichle. “Two Views of Brain Function”. In: *Trends in Cognitive Sciences* 14.4 (Apr. 1, 2010), pp. 180–190. ISSN: 1364-6613. DOI: 10.1016/j.tics.2010.01.008. URL: https://www.sciencedirect.com/science/article/pii/S136466131000029X (visited on 04/11/2023).

[35] P T Fox and M E Raichle. “Focal Physiological Uncoupling of Cerebral Blood Flow and Oxidative Metabolism during Somatosensory Stimulation in Human Subjects.” In: *Proceedings of the National Academy of Sciences* 83.4 (Feb. 1986), pp. 1140–1144. DOI: 10.1073/pnas.83.4.1140. URL: https://www.pnas.org/doi/10.1073/pnas.83.4.1140 (visited on 08/06/2025).

[36] Anastasia Dimakou et al. “The Predictive Nature of Spontaneous Brain Activity across Scales and Species”. In: *Neuron* (Mar. 17, 2025). ISSN: 0896-6273. DOI: 10.1016/j.neuron.2025.02.009. URL: https://www.sciencedirect.com/science/article/pii/S0896627325001278 (visited on 04/07/2025).

[37] Marcus E. Raichle and Abraham Z. Snyder. “A Default Mode of Brain Function: A Brief History of an Evolving Idea”. In: *NeuroImage* 37.4 (Oct. 1, 2007), 1083–1090, discussion 1097–1099. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2007.02.041. pmid: 17719799.

[38] Vinod Menon. “20 Years of the Default Mode Network: A Review and Synthesis”. In: *Neuron* 111.16 (Aug. 16, 2023), pp. 2469–2487. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2023.04.023. pmid: 37167968. URL: https://www.cell.com/neuron/abstract/S0896-6273(23)00308-2 (visited on 07/22/2025).

[39] Gordon B. Smith et al. “Distributed Network Interactions and Their Emergence in Developing Neocortex”. In: *Nature Neuroscience* 21.11 (Nov. 2018), pp. 1600–1608. ISSN: 1546-1726. DOI: 10.1038/s41593-018-0247-5. URL: https://www.nature.com/articles/s41593-018-0247-5 (visited on 05/11/2025).

[40] Dillan J. Newbold et al. “Plasticity and Spontaneous Activity Pulses in Disused Human Brain Circuits”. In: *Neuron* 107.3 (Aug. 5, 2020), 580–589.e6. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2020.05.007. URL: https://www.sciencedirect.com/science/article/pii/S0896627320303536 (visited on 04/08/2025).

[41] Anoopum S. Gupta et al. “Hippocampal Replay Is Not a Simple Function of Experience”. In: *Neuron* 65.5 (Mar. 11, 2010), pp. 695–705. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2010.01.034. pmid: 20223204. URL: https://www.cell.com/neuron/abstract/S0896-6273(10)00060-7 (visited on 05/11/2025).

[42] Alexei V. Egorov and Andreas Draguhn. "Development of Coherent Neuronal Activity Patterns in Mammalian Cortical Networks: Common Principles and Local Hetereogeneity". In: *Mechanisms of Development*. Neural Development 130.6 (June 1, 2013), pp. 412–423. ISSN: 0925-4773. DOI: 10.1016/j.mod.2012.09.006. URL: https://www.sciencedirect.com/science/article/pii/S0925477312000913 (visited on 07/26/2025).

[43] Francisco J. Martini et al. "Spontaneous Activity in Developing Thalamic and Cortical Sensory Networks". In: *Neuron* 109.16 (Aug. 18, 2021), pp. 2519–2534. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2021.06.026. pmid: 34293296. URL: https://www.cell.com/neuron/abstract/S0896-6273(21)00467-0 (visited on 07/26/2025).

[44] Lowry A. Kirkby et al. "A Role for Correlated Spontaneous Activity in the Assembly of Neural Circuits". In: *Neuron* 80.5 (Dec. 4, 2013), pp. 1129–1144. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2013.10.030. URL: https://www.sciencedirect.com/science/article/pii/S0896627313009343 (visited on 07/26/2025).

[45] Tomokazu Ohshiro, Shaista Hussain, and Michael Weliky. "Development of Cortical Orientation Selectivity in the Absence of Visual Experience with Contour". In: *Journal of Neurophysiology* 106.4 (Oct. 2011), pp. 1923–1932. ISSN: 0022-3077. DOI: 10.1152/jn.00095.2011. URL: https://journals.physiology.org/doi/full/10.1152/jn.00095.2011 (visited on 07/27/2025).

[46] S.A. Budick and D.M. O'Malley. "Locomotor Repertoire of the Larval Zebrafish: Swimming, Turning and Prey Capture". In: *Journal of Experimental Biology* 203.17 (Sept. 1, 2000), pp. 2565–2579. ISSN: 0022-0949. DOI: 10.1242/jeb.203.17.2565. URL: https://doi.org/10.1242/jeb.203.17.2565 (visited on 04/07/2023).

[47] Ruth M. Colwill and Robbert Creton. "Imaging Escape and Avoidance Behavior in Zebrafish Larvae". In: *Reviews in the neurosciences* 22.1 (Feb. 1, 2011), pp. 63–73. ISSN: 0334-1763. DOI: 10.1515/RNS.2011.008. pmid: 21572576. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3092434/ (visited on 04/07/2023).

[48] Isaac H. Bianco and Florian Engert. "Visuomotor Transformations Underlying Hunting Behavior in Zebrafish". In: *Current Biology* 25.7 (Mar. 30, 2015), pp. 831–846. ISSN: 0960-9822. DOI: 10.1016/j.cub.2015.01.042. URL: https://www.sciencedirect.com/science/article/pii/S0960982215000743 (visited on 07/26/2025).

[49] Thomas Pietri et al. "The Emergence of the Spatial Structure of Tectal Spontaneous Activity Is Independent of Visual Inputs". In: *Cell Reports* 19.5 (May 2, 2017), pp. 939–948. ISSN: 2211-1247. DOI: 10.1016/j.celrep.2017.04.015. URL:

https://www.sciencedirect.com/science/article/pii/S2211124717304904 (visited on 07/27/2025).

[50] Lilach Avitan et al. "Spontaneous Activity in the Zebrafish Tectum Reorganizes over Development and Is Influenced by Visual Experience". In: *Current Biology* 27.16 (Aug. 21, 2017), 2407–2419.e4. ISSN: 0960-9822. DOI: 10.1016/j.cub.2017.06.056. pmid: 28781054. URL: https://www.cell.com/current-biology/abstract/S0960-9822(17)30793-5 (visited on 07/27/2025).

[51] Feng Han, Natalia Caporale, and Yang Dan. "Reverberation of Recent Visual Experience in Spontaneous Cortical Waves". In: *Neuron* 60.2 (Oct. 23, 2008), pp. 321–327. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2008.08.026. pmid: 18957223. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3576032/ (visited on 05/11/2025).

[52] Carsen Stringer et al. "High-Dimensional Geometry of Population Responses in Visual Cortex". In: *Nature* 571.7765 (7765 July 2019), pp. 361–365. ISSN: 1476-4687. DOI: 10.1038/s41586-019-1346-5. URL: https://www.nature.com/articles/s41586-019-1346-5 (visited on 04/07/2023).

[53] Sophie Aimon et al. "Global Change in Brain State during Spontaneous and Forced Walk in Drosophila Is Composed of Combined Activity Patterns of Different Neuron Classes". In: *eLife* 12 (Apr. 17, 2023). Ed. by Damon A Clark and Claude Desplan, e85202. ISSN: 2050-084X. DOI: 10.7554/eLife.85202. URL: https://doi.org/10.7554/eLife.85202 (visited on 08/06/2025).

[54] Sebastián A. Romano et al. "Spontaneous Neuronal Network Dynamics Reveal Circuit's Functional Adaptations for Behavior". In: *Neuron* 85.5 (Mar. 4, 2015), pp. 1070–1085. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2015.01.027. URL: https://www.sciencedirect.com/science/article/pii/S0896627315000537 (visited on 04/07/2023).

[55] Saul Kato et al. "Global Brain Dynamics Embed the Motor Command Sequence of Caenorhabditis Elegans". In: *Cell* 163.3 (Oct. 22, 2015), pp. 656–669. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2015.09.034. pmid: 26478179. URL: https://www.cell.com/cell/abstract/S0092-8674(15)01196-4 (visited on 08/06/2025).

[56] William E. Allen et al. "Thirst Regulates Motivated Behavior through Modulation of Brainwide Neural Population Dynamics". In: *Science* 364.6437 (Apr. 19, 2019), eaav3932. DOI: 10.1126/science.aav3932. URL: https://www.science.org/doi/10.1126/science.aav3932 (visited on 08/06/2025).

[57] Yoav Livneh and Mark L. Andermann. "Cellular Activity in Insular Cortex across Seconds to Hours: Sensations and Predictions of Bodily States". In: *Neuron* 109.22 (Nov. 17, 2021), pp. 3576–3593. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2021.08.036. pmid: 34582784. URL: https://www.cell.com/neuron/abstract/S0896-6273(21)00653-X (visited on 08/06/2025).

[58] Ian R. Kleckner et al. “Evidence for a Large-Scale Brain System Supporting Allostasis and Interoception in Humans”. In: *Nature Human Behaviour* 1.5 (Apr. 24, 2017), p. 0069. ISSN: 2397-3374. DOI: 10.1038/s41562-017-0069. URL: https://www.nature.com/articles/s41562-017-0069 (visited on 08/06/2025).

[59] Damiano Azzalini, Ignacio Rebollo, and Catherine Tallon-Baudry. “Visceral Signals Shape Brain Dynamics and Cognition”. In: *Trends in Cognitive Sciences* 23.6 (June 1, 2019), pp. 488–509. ISSN: 1364-6613, 1879-307X. DOI: 10.1016/j.tics.2019.03.007. pmid: 31047813. URL: https://www.cell.com/trends/cognitive-sciences/abstract/S1364-6613(19)30089-0 (visited on 08/06/2025).

[60] Lisa Feldman Barrett and W. Kyle Simmons. “Interoceptive Predictions in the Brain”. In: *Nature Reviews Neuroscience* 16.7 (July 2015), pp. 419–429. ISSN: 1471-0048. DOI: 10.1038/nrn3950. URL: https://www.nature.com/articles/nrn3950 (visited on 08/06/2025).

[61] Lilach Avitan and Geoffrey J. Goodhill. “Code Under Construction: Neural Coding Over Development”. In: *Trends in Neurosciences* 41.9 (Sept. 1, 2018), pp. 599–609. ISSN: 0166-2236. DOI: 10.1016/j.tins.2018.05.011. URL: https://www.sciencedirect.com/science/article/pii/S0166223618301589 (visited on 07/26/2025).

[62] Giovanni Pezzulo, Marco Zorzi, and Maurizio Corbetta. “The Secret Life of Predictive Brains: What’s Spontaneous Activity for?” In: *Trends in cognitive sciences* 25.9 (Sept. 2021), pp. 730–743. ISSN: 1364-6613. DOI: 10.1016/j.tics.2021.05.007. pmid: 34144895. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8363551/ (visited on 02/22/2025).

[63] Andrew Ng and Michael Jordan. “On Discriminative vs. Generative Classifiers: A Comparison of Logistic Regression and Naive Bayes”. In: *Advances in Neural Information Processing Systems*. Vol. 14. MIT Press, 2001. URL: https://papers.nips.cc/paper_files/paper/2001/hash/7b7a53e239400a13bd6be6c91c4f
Abstract.html (visited on 08/27/2025).

[64] Mahta Ramezanian-Panahi et al. “Generative Models of Brain Dynamics”. In: *Frontiers in Artificial Intelligence* 5 (July 15, 2022). ISSN: 2624-8212. DOI: 10.3389/frai.2022.807406. URL: https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2022.807406/full (visited on 08/06/2025).

[65] Rishikesan Maran, Eli J. Müller, and Ben D. Fulcher. “Analyzing the Brain’s Dynamic Response to Targeted Stimulation Using Generative Modeling”. In: *Network Neuroscience* 9.1 (Mar. 5, 2025), pp. 237–258. ISSN: 2472-1751. DOI: 10.1162/netn_a_00433. URL: https://doi.org/10.1162/netn_a_00433 (visited on 08/06/2025).

[66] Marcus A. Triplett et al. "Model-Based Decoupling of Evoked and Spontaneous Neural Activity in Calcium Imaging Data". In: *PLoS computational biology* 16.11 (Nov. 2020), e1008330. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1008330. pmid: 33253161.

[67] James V. Haxby et al. "A Common, High-Dimensional Model of the Representational Space in Human Ventral Temporal Cortex". In: *Neuron* 72.2 (Oct. 20, 2011), pp. 404–416. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2011.08.026. URL: https://www.sciencedirect.com/science/article/pii/S0896627311007811 (visited on 07/27/2025).

[68] Alexis Thual et al. *Aligning Individual Brains with Fused Unbalanced Gromov-Wasserstein*. Nov. 22, 2022. DOI: 10.48550/arXiv.2206.09398. arXiv: 2206.09398 [q-bio, stat]. URL: http://arxiv.org/abs/2206.09398 (visited on 04/23/2023). Pre-published.

[69] Po-Hsuan Chen et al. "A Reduced-Dimension fMRI Shared Response Model". In: *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1*. Vol. 1. NIPS'15. Cambridge, MA, USA: MIT Press, Dec. 7, 2015, pp. 460–468.

[70] Bruno Müller and Ueli Grossniklaus. "Model Organisms–A Historical Perspective". In: *Journal of Proteomics* 73.11 (Oct. 10, 2010), pp. 2054–2063. ISSN: 1876-7737. DOI: 10.1016/j.jprot.2010.08.002. pmid: 20727995.

[71] Rachel A. Ankeny and Sabina Leonelli. "What's so Special about Model Organisms?" In: *Studies in History and Philosophy of Science Part A*. Model-Based Representation in Scientific Practice 42.2 (June 1, 2011), pp. 313–323. ISSN: 0039-3681. DOI: 10.1016/j.shpsa.2010.11.039. URL: https://www.sciencedirect.com/science/article/pii/S0039368110001184 (visited on 08/10/2025).

[72] Christiane Nüsslein-Volhard. "The Zebrafish Issue of Development". In: *Development (Cambridge, England)* 139.22 (Nov. 2012), pp. 4099–4103. ISSN: 1477-9129. DOI: 10.1242/dev.085217. pmid: 23093421.

[73] M. Westerfield and UNIVERSITY OF OREGON USA. *The Zebrafish Book : A Guide for the Laboratory Use of Zebrafish (Danio Rerio)*. 5 ed. University of Oregon Press, 2007. 5ed.

[74] C. B. Kimmel et al. "Stages of Embryonic Development of the Zebrafish". In: *Developmental Dynamics: An Official Publication of the American Association of Anatomists* 203.3 (July 1995), pp. 253–310. ISSN: 1058-8388. DOI: 10.1002/aja.1002030302. pmid: 8589427.

[75] Christian Lawrence. "The Husbandry of Zebrafish (*Danio Rerio*): A Review". In: *Aquaculture* 269.1 (Sept. 14, 2007), pp. 1–20. ISSN: 0044-8486. DOI: 10.1016/j.aquaculture.2007.04.077. URL: https://www.sciencedirect.com/science/article/pii/S0044848607004012 (visited on 08/10/2025).

[76] Ruben Portugues and Florian Engert. “The Neural Basis of Visual Behaviors in the Larval Zebrafish”. In: *Current Opinion in Neurobiology* 19.6 (Dec. 2009), pp. 644–647. ISSN: 1873-6882. DOI: 10.1016/j.conb.2009.10.007. pmid: 19896836.

[77] Michael B. Orger and Gonzalo G. de Polavieja. “Zebrafish Behavior: Opportunities and Challenges”. In: *Annual Review of Neuroscience* 40.1 (2017), pp. 125–147. DOI: 10.1146/annurev-neuro-071714-033857. pmid: 28375767. URL: https://doi.org/10.1146/annurev-neuro-071714-033857 (visited on 04/07/2023).

[78] Ruben Portugues et al. “Whole-Brain Activity Maps Reveal Stereotyped, Distributed Networks for Visuomotor Behavior”. In: *Neuron* 81.6 (Mar. 19, 2014), pp. 1328–1343. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2014.01.019. URL: https://www.sciencedirect.com/science/article/pii/S0896627314000506 (visited on 08/10/2025).

[79] Timothy W Dunn et al. “Brain-Wide Mapping of Neural Activity Controlling Zebrafish Exploratory Locomotion”. In: *eLife* 5 (Mar. 22, 2016). Ed. by Ronald L Calabrese, e12741. ISSN: 2050-084X. DOI: 10.7554/eLife.12741. URL: https://doi.org/10.7554/eLife.12741 (visited on 11/22/2024).

[80] Sophia Karpenko et al. “From Behavior to Circuit Modeling of Light-Seeking Navigation in Zebrafish Larvae”. In: *eLife* 9 (Jan. 2, 2020). Ed. by Gordon J Berman, Ronald L Calabrese, and Gordon J Berman, e52882. ISSN: 2050-084X. DOI: 10.7554/eLife.52882. URL: https://doi.org/10.7554/eLife.52882 (visited on 04/07/2023).

[81] Yu Mu et al. “Glia Accumulate Evidence That Actions Are Futile and Suppress Unsuccessful Behavior”. In: *Cell* 178.1 (June 27, 2019), 27–43.e19. ISSN: 0092-8674. DOI: 10.1016/j.cell.2019.05.050. URL: https://www.sciencedirect.com/science/article/pii/S009286741930621X (visited on 04/07/2023).

[82] James A. Lister et al. “Nacre Encodes a Zebrafish Microphthalmia-Related Protein That Regulates Neural-Crest-Derived Pigment Cell Fate”. In: *Development* 126.17 (Sept. 1, 1999), pp. 3757–3767. ISSN: 0950-1991. DOI: 10.1242/dev.126.17.3757. URL: https://doi.org/10.1242/dev.126.17.3757 (visited on 08/10/2025).

[83] Richard Mark White et al. “Transparent Adult Zebrafish as a Tool for in Vivo Transplantation Analysis”. In: *Cell stem cell* 2.2 (Feb. 7, 2008), pp. 183–189. ISSN: 1934-5909. DOI: 10.1016/j.stem.2007.11.002. pmid: 18371439. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2292119/ (visited on 08/10/2025).

[84] Koichi Kawakami et al. “A Transposon-Mediated Gene Trap Approach Identifies Developmentally Regulated Genes in Zebrafish”. In: *Developmental Cell* 7.1 (July 1, 2004), pp. 133–144. ISSN: 1534-5807. DOI: 10.1016/j.devcel.2004.06.

005. pmid: 15239961. URL: https://www.cell.com/developmental-cell/abstract/S1534-5807(04)00205-9 (visited on 08/10/2025).

[85] Kazuhide Asakawa and Koichi Kawakami. "The Tol2-mediated Gal4-UAS Method for Gene and Enhancer Trapping in Zebrafish". In: *Methods (San Diego, Calif.)* 49.3 (Nov. 2009), pp. 275–281. ISSN: 1046-2023. DOI: 10.1016/j.ymeth.2009.01.004. pmid: 19835787. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2764541/ (visited on 08/10/2025).

[86] Thomas Mueller et al. "The Dorsal Pallium in Zebrafish, Danio Rerio (Cyprinidae, Teleostei)". In: *Brain Research* 1381 (Mar. 24, 2011), pp. 95–105. ISSN: 1872-6240. DOI: 10.1016/j.brainres.2010.12.089. pmid: 21219890.

[87] Julián Yáñez et al. "The Organization of the Zebrafish Pallium from a Hodological Perspective". In: *Journal of Comparative Neurology* 530.8 (2022), pp. 1164–1194. ISSN: 1096-9861. DOI: 10.1002/cne.25268. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/cne.25268 (visited on 08/10/2025).

[88] Gregory D. Marquart et al. "High-Precision Registration between Zebrafish Brain Atlases Using Symmetric Diffeomorphic Normalization". In: *GigaScience* 6.8 (Aug. 1, 2017), gix056. ISSN: 2047-217X. DOI: 10.1093/gigascience/gix056. URL: https://doi.org/10.1093/gigascience/gix056 (visited on 04/07/2023).

[89] Michael Kunst et al. "A Cellular-Resolution Atlas of the Larval Zebrafish Brain". In: *Neuron* 103.1 (July 3, 2019), 21–38.e5. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2019.04.034. pmid: 31147152. URL: https://www.cell.com/neuron/abstract/S0896-6273(19)30391-5 (visited on 08/10/2025).

[90] Antoine Légaré et al. "Zebrafish Brain Atlases: A Collective Effort for a Tiny Vertebrate Brain". In: *Neurophotonics* 10.4 (Oct. 2023), p. 044409. ISSN: 2329-423X. DOI: 10.1117/1.NPh.10.4.044409. pmid: 37786400. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10541682/ (visited on 08/10/2025).

[91] L. Saint-Amant and P. Drapeau. "Time Course of the Development of Motor Behaviors in the Zebrafish Embryo". In: *Journal of Neurobiology* 37.4 (Dec. 1998), pp. 622–632. ISSN: 0022-3034. DOI: 10.1002/(sici)1097-4695(199812)37:4<622::aid-neu10>3.0.co;2-s. pmid: 9858263.

[92] Louis Saint-Amant and Pierre Drapeau. "Motoneuron Activity Patterns Related to the Earliest Behavior of the Zebrafish Embryo". In: *Journal of Neuroscience* 20.11 (June 1, 2000), pp. 3964–3972. ISSN: 0270-6474, 1529-2401. DOI: 10.1523/JNEUROSCI.20-11-03964.2000. pmid: 10818131. URL: https://www.jneurosci.org/content/20/11/3964 (visited on 08/10/2025).

[93] Erica Warp et al. “Emergence of Patterned Activity in the Developing Zebrafish Spinal Cord”. In: *Current Biology* 22.2 (Jan. 24, 2012), pp. 93–102. ISSN: 0960-9822. DOI: 10.1016/j.cub.2011.12.002. URL: https://www.sciencedirect.com/science/article/pii/S0960982211013741 (visited on 08/10/2025).

[94] Laura D. Knogler et al. “A Hybrid Electrical/Chemical Circuit in the Spinal Cord Generates a Transient Embryonic Motor Behavior”. In: *Journal of Neuroscience* 34.29 (July 16, 2014), pp. 9644–9655. ISSN: 0270-6474, 1529-2401. DOI: 10.1523/JNEUROSCI.1225-14.2014. pmid: 25031404. URL: https://www.jneurosci.org/content/34/29/9644 (visited on 08/10/2025).

[95] Adrián Ponce-Alvarez et al. “Whole-Brain Neuronal Activity Displays Crackling Noise Dynamics”. In: *Neuron* 100.6 (Dec. 19, 2018), 1446–1459.e6. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2018.10.045. pmid: 30449656. URL: https://www.cell.com/neuron/abstract/S0896-6273(18)30953-X (visited on 04/05/2023).

[96] Mahdi Zarei et al. “High Activity and High Functional Connectivity Are Mutually Exclusive in Resting State Zebrafish and Human Brains”. In: *BMC biology* 20.1 (Apr. 11, 2022), p. 84. ISSN: 1741-7007. DOI: 10.1186/s12915-022-01286-3. pmid: 35410342.

[97] Antoine Légaré et al. “Structural and Genetic Determinants of Zebrafish Functional Brain Networks”. In: *Science Advances* 11.28 (July 11, 2025), eadv7576. DOI: 10.1126/sciadv.adv7576. URL: https://www.science.org/doi/10.1126/sciadv.adv7576 (visited on 08/10/2025).

[98] A. L. Hodgkin and A. F. Huxley. “A Quantitative Description of Membrane Current and Its Application to Conduction and Excitation in Nerve”. In: *The Journal of Physiology* 117.4 (Aug. 28, 1952), pp. 500–544. ISSN: 0022-3751. DOI: 10.1113/jphysiol.1952.sp004764. pmid: 12991237. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1392413/ (visited on 08/10/2025).

[99] Donald O. Hebb. *The Organization of Behavior: A Neuropsychological Theory*. New York: Wiley, June 15, 1949. ISBN: 0-8058-4300-0.

[100] Kenneth D. Harris. “Neural Signatures of Cell Assembly Organization”. In: *Nature Reviews Neuroscience* 6.5 (5 May 2005), pp. 399–407. ISSN: 1471-0048. DOI: 10.1038/nrn1669. URL: https://www.nature.com/articles/nrn1669 (visited on 04/07/2023).

[101] György Buzsáki. “Neural Syntax: Cell Assemblies, Synapsembles, and Readers”. In: *Neuron* 68.3 (Nov. 4, 2010), pp. 362–385. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2010.09.023. URL: https://www.sciencedirect.com/science/article/pii/S0896627310007658 (visited on 08/10/2025).

[102] Jan Mölter, Lilach Avitan, and Geoffrey J. Goodhill. "Detecting Neural Assemblies in Calcium Imaging Data". In: *BMC Biology* 16.1 (Nov. 28, 2018), p. 143. ISSN: 1741-7007. DOI: 10.1186/s12915-018-0606-4. URL: https://doi.org/10.1186/s12915-018-0606-4 (visited on 06/06/2025).

[103] Christophe Gardella, Olivier Marre, and Thierry Mora. "Modeling the Correlated Activity of Neural Populations: A Review". In: *Neural Computation* 31.2 (Feb. 1, 2019), pp. 233–269. ISSN: 0899-7667. DOI: 10.1162/neco_a_01154. URL: https://doi.org/10.1162/neco_a_01154 (visited on 04/07/2023).

[104] Duncan J. Watts and Steven H. Strogatz. "Collective Dynamics of 'Small-World' Networks". In: *Nature* 393.6684 (June 1998), pp. 440–442. ISSN: 1476-4687. DOI: 10.1038/30918. URL: https://www.nature.com/articles/30918 (visited on 08/10/2025).

[105] Danielle S. Bassett et al. "Adaptive Reconfiguration of Fractal Small-World Human Brain Functional Networks". In: *Proceedings of the National Academy of Sciences* 103.51 (Dec. 19, 2006), pp. 19518–19523. DOI: 10.1073/pnas.0606005103. URL: https://www.pnas.org/doi/10.1073/pnas.0606005103 (visited on 08/10/2025).

[106] Michael Stobb et al. "Graph Theoretical Model of a Sensorimotor Connectome in Zebrafish". In: *PLoS ONE* 7.5 (May 18, 2012), e37292. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0037292. pmid: 22624008. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3356276/ (visited on 08/10/2025).

[107] Ryohei Shibue and Fumiyasu Komaki. "Deconvolution of Calcium Imaging Data Using Marked Point Processes". In: *PLOS Computational Biology* 16.3 (Mar. 12, 2020), e1007650. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1007650. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1007650 (visited on 08/10/2025).

[108] Erwin Neher and Bert Sakmann. "Single-Channel Currents Recorded from Membrane of Denervated Frog Muscle Fibres". In: *Nature* 260.5554 (Apr. 1976), pp. 799–802. ISSN: 1476-4687. DOI: 10.1038/260799a0. URL: https://www.nature.com/articles/260799a0 (visited on 08/10/2025).

[109] O. P. Hamill et al. "Improved Patch-Clamp Techniques for High-Resolution Current Recording from Cells and Cell-Free Membrane Patches". In: *Pflügers Archiv* 391.2 (Aug. 1, 1981), pp. 85–100. ISSN: 1432-2013. DOI: 10.1007/BF00656997. URL: https://doi.org/10.1007/BF00656997 (visited on 08/10/2025).

[110] S. Ogawa et al. "Brain Magnetic Resonance Imaging with Contrast Dependent on Blood Oxygenation". In: *Proceedings of the National Academy of Sciences of the United States of America* 87.24 (Dec. 1990), pp. 9868–9872. ISSN: 0027-8424. DOI: 10.1073/pnas.87.24.9868. pmid: 2124706.

[111] J. Nakai, M. Ohkura, and K. Imoto. "A High Signal-to-Noise Ca(2+) Probe Composed of a Single Green Fluorescent Protein". In: *Nature Biotechnology* 19.2 (Feb. 2001), pp. 137–141. ISSN: 1087-0156. DOI: 10.1038/84397. pmid: 11175727.

[112] Lin Tian et al. "Imaging Neural Activity in Worms, Flies and Mice with Improved GCaMP Calcium Indicators". In: *Nature Methods* 6.12 (Dec. 2009), pp. 875–881. ISSN: 1548-7105. DOI: 10.1038/nmeth.1398. pmid: 19898485.

[113] Jasper Akerboom et al. "Optimization of a GCaMP Calcium Indicator for Neural Activity Imaging". In: *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 32.40 (Oct. 3, 2012), pp. 13819–13840. ISSN: 1529-2401. DOI: 10.1523/JNEUROSCI.2601-12.2012. pmid: 23035093.

[114] Tsai-Wen Chen et al. "Ultrasensitive Fluorescent Proteins for Imaging Neuronal Activity". In: *Nature* 499.7458 (July 18, 2013), pp. 295–300. ISSN: 1476-4687. DOI: 10.1038/nature12354. pmid: 23868258.

[115] Yan Zhang et al. "Fast and Sensitive GCaMP Calcium Indicators for Imaging Neural Populations". In: *Nature* 615.7954 (Mar. 2023), pp. 884–891. ISSN: 1476-4687. DOI: 10.1038/s41586-023-05828-9. pmid: 36922596.

[116] Geoffrey Migault et al. "Whole-Brain Calcium Imaging during Physiological Vestibular Stimulation in Larval Zebrafish". In: *Current Biology* 28.23 (Dec. 3, 2018), 3723–3735.e6. ISSN: 0960-9822. DOI: 10.1016/j.cub.2018.10.017. URL: https://www.sciencedirect.com/science/article/pii/S0960982218313460 (visited on 04/07/2023).

[117] Jérôme Tubiana et al. "Blind Deconvolution for Spike Inference from Fluorescence Recordings". In: *Journal of Neuroscience Methods* 342 (Aug. 1, 2020), p. 108763. ISSN: 0165-0270. DOI: 10.1016/j.jneumeth.2020.108763. URL: https://www.sciencedirect.com/science/article/pii/S0165027020301862 (visited on 04/07/2023).

[118] M. Minsky. "Memoir on Inventing the Confocal Scanning Microscope". In: *Scanning* 10.4 (1988), pp. 128–138. ISSN: 1932-8745. DOI: 10.1002/sca.4950100403. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/sca.4950100403 (visited on 08/10/2025).

[119] José-Angel Conchello and Jeff W. Lichtman. "Optical Sectioning Microscopy". In: *Nature Methods* 2.12 (Dec. 2005), pp. 920–931. ISSN: 1548-7091. DOI: 10.1038/nmeth815. pmid: 16299477.

[120] Winfried Denk, James H. Strickler, and Watt W. Webb. "Two-Photon Laser Scanning Fluorescence Microscopy". In: *Science* 248.4951 (Apr. 6, 1990), pp. 73–76. DOI: 10.1126/science.2321027. URL: https://www.science.org/doi/10.1126/science.2321027 (visited on 08/10/2025).

[121] Fritjof Helmchen and Winfried Denk. "Deep Tissue Two-Photon Microscopy". In: *Nature Methods* 2.12 (Dec. 2005), pp. 932–940. ISSN: 1548-7105. DOI: 10.1038/nmeth818. URL: https://www.nature.com/articles/nmeth818 (visited on 08/10/2025).

[122] Sébastien Wolf et al. "Whole-Brain Functional Imaging with Two-Photon Light-Sheet Microscopy". In: *Nature Methods* 12.5 (5 May 2015), pp. 379–380. ISSN: 1548-7105. DOI: 10.1038/nmeth.3371. URL: https://www.nature.com/articles/nmeth.3371 (visited on 04/07/2023).

[123] Jan Huisken et al. "Optical Sectioning Deep Inside Live Embryos by Selective Plane Illumination Microscopy". In: *Science* 305.5686 (Aug. 13, 2004), pp. 1007–1009. DOI: 10.1126/science.1100035. URL: https://www.science.org/doi/10.1126/science.1100035 (visited on 08/10/2025).

[124] Philipp J. Keller et al. "Reconstruction of Zebrafish Early Embryonic Development by Scanned Light Sheet Microscopy". In: *Science (New York, N.Y.)* 322.5904 (Nov. 14, 2008), pp. 1065–1069. ISSN: 1095-9203. DOI: 10.1126/science.1162493. pmid: 18845710.

[125] Misha B. Ahrens et al. "Whole-Brain Functional Imaging at Cellular Resolution Using Light-Sheet Microscopy". In: *Nature Methods* 10.5 (May 2013), pp. 413–420. ISSN: 1548-7105. DOI: 10.1038/nmeth.2434. URL: https://www.nature.com/articles/nmeth.2434 (visited on 07/27/2025).

[126] Thomas Panier et al. "Fast Functional Imaging of Multiple Brain Regions in Intact Zebrafish Larvae Using Selective Plane Illumination Microscopy". In: *Frontiers in Neural Circuits* 7 (2013). ISSN: 1662-5110. URL: https://www.frontiersin.org/articles/10.3389/fncir.2013.00065 (visited on 04/07/2023).

[127] Carsen Stringer et al. "High-Dimensional Geometry of Population Responses in Visual Cortex". In: *Nature* 571.7765 (July 2019), pp. 361–365. ISSN: 1476-4687. DOI: 10.1038/s41586-019-1346-5. pmid: 31243367.

[128] João D. Semedo et al. "Cortical Areas Interact through a Communication Subspace". In: *Neuron* 102.1 (Apr. 3, 2019), 249–259.e4. ISSN: 1097-4199. DOI: 10.1016/j.neuron.2019.01.026. pmid: 30770252.

[129] Nicholas A. Steinmetz et al. "Distributed Coding of Choice, Action and Engagement across the Mouse Brain". In: *Nature* 576.7786 (Dec. 2019), pp. 266–273. ISSN: 1476-4687. DOI: 10.1038/s41586-019-1787-x. pmid: 31776518.

[130] Zhihao Zheng et al. "A Complete Electron Microscopy Volume of the Brain of Adult Drosophila Melanogaster". In: *Cell* 174.3 (July 2018), 730–743.e22. ISSN: 00928674. DOI: 10.1016/j.cell.2018.06.019. URL: https://linkinghub.elsevier.com/retrieve/pii/S0092867418307876 (visited on 08/19/2025).

[131] Sven Dorkenwald et al. “Neuronal Wiring Diagram of an Adult Brain”. In: *Nature* 634.8032 (Oct. 2024), pp. 124–138. ISSN: 1476-4687. DOI: 10.1038/s41586-024-07558-y. URL: https://www.nature.com/articles/s41586-024-07558-y (visited on 08/19/2025).

[132] Simona Cocco et al. “Functional Networks from Inverse Modeling of Neural Population Activity”. In: *Current Opinion in Systems Biology*. • Mathematical Modelling • Mathematical Modelling, Dynamics of Brain Activity at the Systems Level • Clinical and Translational Systems Biology 3 (June 1, 2017), pp. 103–110. ISSN: 2452-3100. DOI: 10.1016/j.coisb.2017.04.017. URL: https://www.sciencedirect.com/science/article/pii/S2452310017300215 (visited on 04/07/2023).

[133] Gasper Tkacik et al. *Spin Glass Models for a Network of Real Neurons*. Dec. 30, 2009. DOI: 10.48550/arXiv.0912.5409. arXiv: 0912.5409 [q-bio]. URL: http://arxiv.org/abs/0912.5409 (visited on 08/19/2025). Prepublished.

[134] Anthony G. Hudetz, Colin J. Humphries, and Jeffrey R. Binder. “Spin-Glass Model Predicts Metastable Brain States That Diminish in Anesthesia”. In: *Frontiers in Systems Neuroscience* 8 (Dec. 11, 2014). ISSN: 1662-5137. DOI: 10.3389/fnsys.2014.00234. URL: https://www.frontiersin.org/journals/systems-neuroscience/articles/10.3389/fnsys.2014.00234/full (visited on 08/19/2025).

[135] J J Hopfield. “Neural Networks and Physical Systems with Emergent Collective Computational Abilities.” In: *Proceedings of the National Academy of Sciences* 79.8 (Apr. 1982), pp. 2554–2558. DOI: 10.1073/pnas.79.8.2554. URL: https://www.pnas.org/doi/abs/10.1073/pnas.79.8.2554 (visited on 04/07/2023).

[136] David H. Ackley, Geoffrey E. Hinton, and Terrence J. Sejnowski. “A Learning Algorithm for Boltzmann Machines”. In: *Cognitive Science* 9.1 (Jan. 1, 1985), pp. 147–169. ISSN: 0364-0213. DOI: 10.1016/S0364-0213(85)80012-4. URL: https://www.sciencedirect.com/science/article/pii/S0364021385800124 (visited on 08/08/2025).

[137] Jérôme Tubiana. “Restricted Boltzmann Machines : From Compositional Representations to Protein Sequence Analysis”. These de doctorat. Paris Sciences et Lettres (ComUE), Nov. 29, 2018. URL: https://www.theses.fr/2018PSLEE039 (visited on 04/07/2023).

[138] Miguel A Carreira-Perpinan and Geoffrey E Hinton. “On Contrastive Divergence Learning”. In: ().

[139] W. Bialek. *Biophysics: Searching for Principles*. Princeton University Press, 2012. ISBN: 978-0-691-13891-6. URL: https://books.google.fr/books?id=5In_FKA2rmUC.

[140] Geoffrey E. Hinton. "Training Products of Experts by Minimizing Contrastive Divergence". In: *Neural Computation* 14.8 (Aug. 1, 2002), pp. 1771–1800. ISSN: 0899-7667, 1530-888X. DOI: 10.1162/089976602760128018. URL: https://direct.mit.edu/neco/article/14/8/1771-1800/6687 (visited on 08/19/2025).

[141] J. Tubiana and R. Monasson. "Emergence of Compositional Representations in Restricted Boltzmann Machines". In: *Physical Review Letters* 118.13 (Mar. 28, 2017), p. 138301. DOI: 10.1103/PhysRevLett.118.138301. URL: https://link.aps.org/doi/10.1103/PhysRevLett.118.138301 (visited on 04/07/2023).

[142] Jérôme Tubiana, Simona Cocco, and Rémi Monasson. "Learning Compositional Representations of Interacting Systems with Restricted Boltzmann Machines: Comparative Study of Lattice Proteins". In: *Neural Computation* 31.8 (Aug. 1, 2019), pp. 1671–1717. ISSN: 0899-7667. DOI: 10.1162/neco_a_01210. URL: https://doi.org/10.1162/neco_a_01210 (visited on 04/07/2023).

[143] Geoffrey E. Hinton. "A Practical Guide to Training Restricted Boltzmann Machines". In: *Neural Networks: Tricks of the Trade: Second Edition*. Ed. by Grégoire Montavon, Geneviève B. Orr, and Klaus-Robert Müller. Berlin, Heidelberg: Springer, 2012, pp. 599–619. ISBN: 978-3-642-35289-8. DOI: 10.1007/978-3-642-35289-8_32. URL: https://doi.org/10.1007/978-3-642-35289-8_32 (visited on 07/29/2025).

[144] Jérôme Tubiana, Simona Cocco, and Rémi Monasson. "Learning Protein Constitutive Motifs from Sequence Data". In: *eLife* 8 (Mar. 12, 2019). Ed. by Lucy J Colwell and Detlef Weigel, e39397. ISSN: 2050-084X. DOI: 10.7554/eLife.39397. URL: https://doi.org/10.7554/eLife.39397 (visited on 04/07/2023).

[145] Tijmen Tieleman. "Training Restricted Boltzmann Machines Using Approximations to the Likelihood Gradient". In: *Proceedings of the 25th International Conference on Machine Learning*. ICML '08. New York, NY, USA: Association for Computing Machinery, July 5, 2008, pp. 1064–1071. ISBN: 978-1-60558-205-4. DOI: 10.1145/1390156.1390290. URL: https://doi.org/10.1145/1390156.1390290 (visited on 08/19/2025).

[146] H. Chau Nguyen, Riccardo Zecchina, and Johannes Berg. "Inverse Statistical Problems: From the Inverse Ising Problem to Data Science". In: *Advances in Physics* 66.3 (July 3, 2017), pp. 197–261. ISSN: 0001-8732. DOI: 10.1080/00018732.2017.1341604. URL: https://doi.org/10.1080/00018732.2017.1341604 (visited on 08/09/2025).

[147] Jérôme Tubiana et al. "Funneling Modulatory Peptide Design with Generative Models: Discovery and Characterization of Disruptors of Calcineurin Protein-Protein Interactions". In: *PLOS Computational Biology* 19.2 (Feb. 2, 2023), e1010874. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1010874. URL: https://

journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1010874 (visited on 08/09/2025).

[148] Cyril Malbranke et al. "Improving Sequence-Based Modeling of Protein Families Using Secondary-Structure Quality Assessment". In: *Bioinformatics* 37.22 (Nov. 18, 2021), pp. 4083–4090. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btab442. URL: https://doi.org/10.1093/bioinformatics/btab442 (visited on 08/09/2025).

[149] Kai Shimagaki and Martin Weigt. "Selection of Sequence Motifs and Generative Hopfield-Potts Models for Protein Families". In: *Physical Review E* 100.3 (Sept. 19, 2019), p. 032128. DOI: 10.1103/PhysRevE.100.032128. URL: https://link.aps.org/doi/10.1103/PhysRevE.100.032128 (visited on 08/09/2025).

[150] Cyril Malbranke et al. "Computational Design of Novel Cas9 PAM-interacting Domains Using Evolution-Based Modelling and Structural Quality Assessment". In: *PLOS Computational Biology* 19.11 (Nov. 17, 2023), e1011621. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1011621. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1011621 (visited on 08/09/2025).

[151] Aurélien Decelle, Beatriz Seoane, and Lorenzo Rosset. "Unsupervised Hierarchical Clustering Using the Learning Dynamics of Restricted Boltzmann Machines". In: *Physical Review E* 108.1 (July 7, 2023), p. 014110. DOI: 10.1103/PhysRevE.108.014110. URL: https://link.aps.org/doi/10.1103/PhysRevE.108.014110 (visited on 08/09/2025).

[152] Jorge Fernandez-de-Cossio-Diaz. "Generative Modeling of RNA Sequence Families with Restricted Boltzmann Machines". In: *RNA Design: Methods and Protocols*. Ed. by Alexander Churkin and Danny Barash. New York, NY: Springer US, 2025, pp. 163–175. ISBN: 978-1-07-164079-1. DOI: 10.1007/978-1-0716-4079-1_11. URL: https://doi.org/10.1007/978-1-0716-4079-1_11 (visited on 08/09/2025).

[153] Barbara Bravi et al. "A Transfer-Learning Approach to Predict Antigen Immunogenicity and T-cell Receptor Specificity". In: *eLife* 12 (Sept. 8, 2023). Ed. by Anne-Florence Bitbol and Michael B Eisen, e85126. ISSN: 2050-084X. DOI: 10.7554/eLife.85126. URL: https://doi.org/10.7554/eLife.85126 (visited on 08/09/2025).

[154] Barbara Bravi. "Development and Use of Machine Learning Algorithms in Vaccine Target Selection". In: *npj Vaccines* 9.1 (Jan. 20, 2024), p. 15. ISSN: 2059-0105. DOI: 10.1038/s41541-023-00795-8. URL: https://www.nature.com/articles/s41541-023-00795-8 (visited on 08/09/2025).

[155] Barbara Bravi et al. “Probing T-cell Response by Sequence-Based Probabilistic Modeling”. In: *PLOS Computational Biology* 17.9 (Sept. 2, 2021), e1009297. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1009297. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1009297 (visited on 08/09/2025).

[156] R. Devon Hjelm et al. “Restricted Boltzmann Machines for Neuroimaging: An Application in Identifying Intrinsic Networks”. In: *NeuroImage* 96 (Aug. 1, 2014), pp. 245–260. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2014.03.048. URL: https://www.sciencedirect.com/science/article/pii/S1053811914002080 (visited on 04/07/2023).

[157] L.R. Rabiner. “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition”. In: *Proceedings of the IEEE* 77.2 (Feb. 1989), pp. 257–286. ISSN: 1558-2256. DOI: 10.1109/5.18626. URL: https://ieeexplore.ieee.org/document/18626 (visited on 07/19/2025).

[158] Philip N Lehner. *Handbook of Ethological Methods*. Cambridge University Press, 1998.

[159] Kiran Girdhar, Martin Gruebele, and Yann R. Chemla. “The Behavioral Space of Zebrafish Locomotion and Its Neural Network Analog”. In: *PLOS ONE* 10.7 (July 1, 2015), e0128668. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0128668. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0128668 (visited on 07/09/2025).

[160] Greg J. Stephens, Leslie C. Osborne, and William Bialek. “Searching for Simplicity in the Analysis of Neurons and Behavior”. In: *Proceedings of the National Academy of Sciences* 108 (supplement_3 Sept. 13, 2011), pp. 15565–15571. DOI: 10.1073/pnas.1010868108. URL: https://www.pnas.org/doi/full/10.1073/pnas.1010868108 (visited on 07/08/2025).

[161] Jeanne Altmann. “Observational Study of Behavior: Sampling Methods”. In: *Behaviour* 49.3/4 (1974), pp. 227–267. ISSN: 0005-7959. JSTOR: 4533591. URL: https://www.jstor.org/stable/4533591 (visited on 07/08/2025).

[162] João C. Marques et al. “Structure of the Zebrafish Locomotor Repertoire Revealed with Unsupervised Behavioral Clustering”. In: *Current Biology* 28.2 (Jan. 22, 2018), 181–195.e5. ISSN: 0960-9822. DOI: 10.1016/j.cub.2017.12.002. URL: https://www.sciencedirect.com/science/article/pii/S0960982217316044 (visited on 07/08/2025).

[163] Robert Evan Johnson et al. “Probabilistic Models of Larval Zebrafish Behavior Reveal Structure on Many Scales”. In: *Current biology : CB* 30.1 (Jan. 6, 2020), 70–82.e4. ISSN: 0960-9822. DOI: 10.1016/j.cub.2019.11.026. pmid: 31866367. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6958995/ (visited on 07/08/2025).

[164] Edoardo Fazzari et al. “Animal Behavior Analysis Methods Using Deep Learning: A Survey”. In: *Expert Systems with Applications* 289 (Sept. 15, 2025), p. 128330. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2025.128330. URL: https://www.sciencedirect.com/science/article/pii/S0957417425019499 (visited on 07/08/2025).

[165] Merlin Lange et al. “Inter-Individual and Inter-Strain Variations in Zebrafish Locomotor Ontogeny”. In: *PLOS ONE* 8.8 (Aug. 9, 2013), e70172. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0070172. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0070172 (visited on 07/08/2025).

[166] Guillaume Le Goc et al. “Thermal Modulation of Zebrafish Exploratory Statistics Reveals Constraints on Individual Behavioral Variability”. In: *BMC Biology* 19.1 (Sept. 21, 2021), p. 208. ISSN: 1741-7007. DOI: 10.1186/s12915-021-01126-w. URL: https://doi.org/10.1186/s12915-021-01126-w (visited on 10/23/2023).

[167] Sophia Karpenko. “Light-Seeking Navigation in Zebrafish Larva: From Behavior to Neural Circuits”. In: ().

[168] Angel-Carlos Román et al. “Histone H4 Acetylation Regulates Behavioral Inter-Individual Variability in Zebrafish”. In: *Genome Biology* 19 (Apr. 25, 2018), p. 55. ISSN: 1474-7596. DOI: 10.1186/s13059-018-1428-y. pmid: 29695303. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5922312/ (visited on 07/08/2025).

[169] Harold A. Burgess, Hannah Schoch, and Michael Granato. “Distinct Retinal Pathways Drive Spatial Orientation Behaviors in Zebrafish Navigation”. In: *Current Biology* 20.4 (Feb. 23, 2010), pp. 381–386. ISSN: 0960-9822. DOI: 10.1016/j.cub.2010.01.022. pmid: 20153194. URL: https://www.cell.com/current-biology/abstract/S0960-9822(10)00061-8 (visited on 07/09/2025).

[170] Xiuye Chen and Florian Engert. “Navigational Strategies Underlying Phototaxis in Larval Zebrafish”. In: *Frontiers in Systems Neuroscience* 8 (Mar. 25, 2014). ISSN: 1662-5137. DOI: 10.3389/fnsys.2014.00039. URL: https://www.frontiersin.org/journals/systems-neuroscience/articles/10.3389/fnsys.2014.00039/full (visited on 07/10/2025).

[171] Sébastien Wolf et al. “Sensorimotor Computation Underlying Phototaxis in Zebrafish”. In: *Nature Communications* 8.1 (1 Sept. 21, 2017), p. 651. ISSN: 2041-1723. DOI: 10.1038/s41467-017-00310-3. URL: https://www.nature.com/articles/s41467-017-00310-3 (visited on 04/07/2023).

[172] Eva A. Naumann et al. “From Whole-Brain Data to Functional Circuit Models: The Zebrafish Optomotor Response”. In: *Cell* 167.4 (Nov. 3, 2016), 947–960.e20. ISSN: 0092-8674. DOI: 10.1016/j.cell.2016.10.019. URL: https://www.sciencedirect.com/science/article/pii/S0092867416314027 (visited on 06/29/2023).

[173] Sebastien Wolf et al. “Emergence of Time Persistence in a Data-Driven Neural Network Model”. In: *eLife* 12 (Mar. 14, 2023). Ed. by Tatyana O Sharpee, e79541. ISSN: 2050-084X. DOI: 10.7554/eLife.79541. URL: https://doi.org/10.7554/eLife.79541 (visited on 04/24/2023).

[174] Emiliano Marachlian et al. “Principles of Functional Circuit Connectivity: Insights From Spontaneous Activity in the Zebrafish Optic Tectum”. In: *Frontiers in Neural Circuits* 12 (June 21, 2018). ISSN: 1662-5110. DOI: 10.3389/fncir.2018.00046. URL: https://www.frontiersin.org/journals/plant-science/articles/10.3389/fncir.2018.00046/full (visited on 07/27/2025).

[175] György Buzsáki and Andreas Draguhn. “Neuronal Oscillations in Cortical Networks”. In: *Science* 304.5679 (June 25, 2004), pp. 1926–1929. DOI: 10.1126/science.1099745. URL: https://www.science.org/doi/10.1126/science.1099745 (visited on 07/27/2025).

[176] J. S. Damoiseaux et al. “Consistent Resting-State Networks across Healthy Subjects”. In: *Proceedings of the National Academy of Sciences* 103.37 (Sept. 12, 2006), pp. 13848–13853. DOI: 10.1073/pnas.0601417103. URL: https://www.pnas.org/doi/10.1073/pnas.0601417103 (visited on 07/27/2025).

[177] Lucina Q. Uddin, F. Xavier Castellanos, and Vinod Menon. “Resting State Functional Brain Connectivity in Child and Adolescent Psychiatry: Where Are We Now?” In: *Neuropsychopharmacology* 50.1 (Jan. 2025), pp. 196–200. ISSN: 1740-634X. DOI: 10.1038/s41386-024-01888-1. URL: https://www.nature.com/articles/s41386-024-01888-1 (visited on 07/27/2025).

[178] Hanna K. Hausman et al. “The Role of Resting-State Network Functional Connectivity in Cognitive Aging”. In: *Frontiers in Aging Neuroscience* 12 (June 12, 2020). ISSN: 1663-4365. DOI: 10.3389/fnagi.2020.00177. URL: https://www.frontiersin.org/journals/aging-neuroscience/articles/10.3389/fnagi.2020.00177/full (visited on 07/27/2025).

[179] Edgar Canario, Donna Chen, and Bharat Biswal. “A Review of Resting-State fMRI and Its Use to Examine Psychiatric Disorders”. In: *Psychoradiology* 1.1 (May 11, 2021), p. 42. DOI: 10.1093/psyrad/kkab003. pmid: 38665309. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC10917160/ (visited on 07/27/2025).

[180] Connor Brennan and Alexander Proekt. "A Quantitative Model of Conserved Macroscopic Dynamics Predicts Future Motor Commands". In: *eLife* 8 (July 11, 2019). Ed. by Ronald L Calabrese, William S Ryu, and Elizabeth Cropper, e46814. ISSN: 2050-084X. DOI: 10.7554/eLife.46814. URL: https://doi.org/10.7554/eLife.46814 (visited on 04/07/2023).

[181] Charles Fieseler et al. *An Intrinsic Neuronal Manifold Underlies Brain-Wide Hierarchical Organization of Behavior in C. Elegans*. Mar. 11, 2025. DOI: 10.1101/2025.03.09.642241. URL: https://www.biorxiv.org/content/10.1101/2025.03.09.642241v1 (visited on 07/27/2025). Pre-published.

[182] Adam A. Atanas et al. "Brain-Wide Representations of Behavior Spanning Multiple Timescales and States in C. Elegans". In: *Cell* 186.19 (Sept. 14, 2023), 4134–4151.e31. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2023.07.035. pmid: 37607537. URL: https://www.cell.com/cell/abstract/S0092-8674(23)00850-4 (visited on 07/27/2025).

[183] João C. Marques et al. "Internal State Dynamics Shape Brainwide Activity and Foraging Behaviour". In: *Nature* 577.7789 (7789 Jan. 2020), pp. 239–243. ISSN: 1476-4687. DOI: 10.1038/s41586-019-1858-z. URL: https://www.nature.com/articles/s41586-019-1858-z (visited on 04/07/2023).

[184] Luigi Petrucco et al. "Neural Dynamics and Architecture of the Heading Direction Circuit in Zebrafish". In: *Nature Neuroscience* 26.5 (May 2023), pp. 765–773. ISSN: 1546-1726. DOI: 10.1038/s41593-023-01308-5. URL: https://www.nature.com/articles/s41593-023-01308-5 (visited on 07/27/2025).

[185] James V Haxby et al. "Hyperalignment: Modeling Shared Information Encoded in Idiosyncratic Cortical Topographies". In: *eLife* 9 (June 2, 2020). Ed. by Chris I Baker and Floris P de Lange, e56601. ISSN: 2050-084X. DOI: 10.7554/eLife.56601. URL: https://doi.org/10.7554/eLife.56601 (visited on 07/27/2025).

[186] Alexander G. Huth et al. "Natural Speech Reveals the Semantic Maps That Tile Human Cerebral Cortex". In: *Nature* 532.7600 (Apr. 2016), pp. 453–458. ISSN: 1476-4687. DOI: 10.1038/nature17637. URL: https://www.nature.com/articles/nature17637 (visited on 07/27/2025).

[187] Alexander G. Huth et al. *PrAGMATiC: A Probabilistic and Generative Model of Areas Tiling the Cortex*. Apr. 14, 2015. DOI: 10.48550/arXiv.1504.03622. arXiv: 1504.03622 [q-bio]. URL: http://arxiv.org/abs/1504.03622 (visited on 07/27/2025). Pre-published.

[188] Po-Hsuan Chen et al. *A Convolutional Autoencoder for Multi-Subject fMRI Data Aggregation*. Aug. 17, 2016. DOI: 10.48550/arXiv.1608.04846. arXiv: 1608.04846 [stat]. URL: http://arxiv.org/abs/1608.04846 (visited on 07/27/2025). Pre-published.

[189] Xinke Shen et al. “Contrastive Learning of Shared Spatiotemporal EEG Representations across Individuals for Naturalistic Neuroscience”. In: *NeuroImage* 301 (Nov. 1, 2024), p. 120890. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2024.120890. URL: https://www.sciencedirect.com/science/article/pii/S1053811924003872 (visited on 07/27/2025).

[190] G. E. Hinton and R. R. Salakhutdinov. “Reducing the Dimensionality of Data with Neural Networks”. In: *Science* 313.5786 (July 28, 2006), pp. 504–507. DOI: 10.1126/science.1127647. URL: https://www.science.org/doi/10.1126/science.1127647 (visited on 07/27/2025).

[191] Simona Cocco et al. “Functional Networks from Inverse Modeling of Neural Population Activity”. In: *Current Opinion in Systems Biology* 3 (June 2017), pp. 103–110. ISSN: 24523100. DOI: 10.1016/j.coisb.2017.04.017. URL: https://linkinghub.elsevier.com/retrieve/pii/S2452310017300215 (visited on 08/22/2025).

[192] Lorenzo Posani et al. “Functional Connectivity Models for Decoding of Spatial Representations from Hippocampal CA1 Recordings”. In: *Journal of Computational Neuroscience* 43.1 (Aug. 2017), pp. 17–33. ISSN: 0929-5313, 1573-6873. DOI: 10.1007/s10827-017-0645-9. URL: http://link.springer.com/10.1007/s10827-017-0645-9 (visited on 08/22/2025).

[193] Sebastian Quiroz Monnens et al. “The Recurrent Temporal Restricted Boltzmann Machine Captures Neural Assembly Dynamics in Whole-Brain Activity”. In: *eLife* 13 (Sept. 30, 2024). DOI: 10.7554/eLife.98489.2. URL: https://elifesciences.org/reviewed-preprints/98489 (visited on 07/27/2025).

[194] Nicholas J. Tustison et al. “The ANTsX Ecosystem for Quantitative Biological and Medical Imaging”. In: *Scientific Reports* 11.1 (Apr. 27, 2021), p. 9068. ISSN: 2045-2322. DOI: 10.1038/s41598-021-87564-6. URL: https://www.nature.com/articles/s41598-021-87564-6 (visited on 07/29/2025).

[195] Jorge Fernandez-de-Cossio-Diaz, Simona Cocco, and Rémi Monasson. “Disentangling Representations in Restricted Boltzmann Machines without Adversaries”. In: *Physical Review X* 13.2 (Apr. 5, 2023), p. 021003. DOI: 10.1103/PhysRevX.13.021003. URL: https://link.aps.org/doi/10.1103/PhysRevX.13.021003 (visited on 07/29/2025).

[196] Nikita Vladimirov et al. “Brain-Wide Circuit Interrogation at the Cellular Level Guided by Online Analysis of Neuronal Function”. In: *Nature Methods* 15.12 (Dec. 2018), pp. 1117–1125. ISSN: 1548-7105. DOI: 10.1038/s41592-018-0221-x. URL: https://www.nature.com/articles/s41592-018-0221-x (visited on 07/29/2025).

[197] Jorge Fernandez-de-Cossio-Diaz et al. *Designing Molecular RNA Switches with Restricted Boltzmann Machines*. May 12, 2023. DOI: 10.1101/2023.05.10.540155. URL: https://www.biorxiv.org/content/10.1101/2023.05.10.540155v1 (visited on 07/29/2025). Pre-published.

[198] Gregoire Montavon, Mikio Braun, and Klaus-Robert Muller. "Deep Boltzmann Machines as Feed-Forward Hierarchies". In: *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*. Ed. by Neil D. Lawrence and Mark Girolami. Vol. 22. Proceedings of Machine Learning Research. La Palma, Canary Islands: PMLR, Apr. 21–23, 2012, pp. 798–804. URL: https://proceedings.mlr.press/v22/montavon12.html.

[199] Jan Melchior, Asja Fischer, and Laurenz Wiskott. "How to Center Deep Boltzmann Machines". In: *Journal of Machine Learning Research* 17.99 (2016), pp. 1–61. URL: http://jmlr.org/papers/v17/14-237.html.

[200] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *CoRR* (Dec. 22, 2014). URL: https://www.semanticscholar.org/paper/Adam%3A-A-Method-for-Stochastic-Optimization-Kingma-Ba/a6cb366736791bcccc5c8639de5a8f9636bf87e8 (visited on 07/29/2025).

[201] Kelsey M Hallinen et al. "Decoding Locomotion from Population Neural Activity in Moving C. Elegans". In: *eLife* 10 (July 29, 2021). Ed. by Ronald L Calabrese, e66135. ISSN: 2050-084X. DOI: 10.7554/eLife.66135. URL: https://doi.org/10.7554/eLife.66135 (visited on 03/19/2025).

[202] Mostafa Safaie et al. "Preserved Neural Dynamics across Animals Performing Similar Behaviour". In: *Nature* 623.7988 (Nov. 2023), pp. 765–771. ISSN: 1476-4687. DOI: 10.1038/s41586-023-06714-0. URL: https://www.nature.com/articles/s41586-023-06714-0 (visited on 08/12/2025).

[203] Diego Vidaurre, Stephen M. Smith, and Mark W. Woolrich. "Brain Network Dynamics Are Hierarchically Organized in Time". In: *Proceedings of the National Academy of Sciences* 114.48 (Nov. 28, 2017), pp. 12827–12832. DOI: 10.1073/pnas.1705120114. URL: https://www.pnas.org/doi/10.1073/pnas.1705120114 (visited on 08/04/2025).

[204] Diego Vidaurre et al. "Discovering Dynamic Brain Networks from Big Data in Rest and Task". In: *Neuroimage* 180 (Pt B Oct. 15, 2018), pp. 646–656. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2017.06.077. pmid: 28669905. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6138951/ (visited on 08/04/2025).

[205] B. A. W. Brinkman et al. "Metastable Dynamics of Neural Circuits and Networks". In: *Applied Physics Reviews* 9.1 (Mar. 2022), p. 011313. ISSN: 1931-9401. DOI: 10.1063/5.0062603. pmid: 35284030. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8900181/ (visited on 08/04/2025).

[206] Fran Hancock et al. “Metastability Demystified — the Foundational Past, the Pragmatic Present and the Promising Future”. In: *Nature Reviews Neuroscience* 26.2 (Feb. 2025), pp. 82–100. ISSN: 1471-0048. DOI: 10.1038/s41583-024-00883-1. URL: https://www.nature.com/articles/s41583-024-00883-1 (visited on 08/04/2025).

[207] James A. Roberts et al. “Metastable Brain Waves”. In: *Nature Communications* 10.1 (Mar. 5, 2019), p. 1056. ISSN: 2041-1723. DOI: 10.1038/s41467-019-08999-0. URL: https://www.nature.com/articles/s41467-019-08999-0 (visited on 08/21/2025).

[208] Stefano Recanatesi et al. “Metastable Attractors Explain the Variable Timing of Stable Behavioral Action Sequences”. In: *Neuron* 110.1 (Jan. 5, 2022), 139–153.e9. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2021.10.011. pmid: 34717794. URL: https://www.cell.com/neuron/abstract/S0896-6273(21)00779-0 (visited on 04/07/2023).

[209] Joshua I. Glaser et al. *Recurrent Switching Dynamical Systems Models for Multiple Interacting Neural Populations*. Oct. 22, 2020. DOI: 10.1101/2020.10.21.349282. URL: http://biorxiv.org/lookup/doi/10.1101/2020.10.21.349282 (visited on 08/04/2025). Pre-published.

[210] Hayoung Song, Won Mok Shim, and Monica D Rosenberg. “Large-Scale Neural Dynamics in a Shared Low-Dimensional State Space Reflect Cognitive and Attentional Dynamics”. In: *eLife* 12 (July 3, 2023). Ed. by Shella Keilholz et al., e85487. ISSN: 2050-084X. DOI: 10.7554/eLife.85487. URL: https://doi.org/10.7554/eLife.85487 (visited on 08/04/2025).

[211] Steffen Schneider, Jin Hwa Lee, and Mackenzie Weygandt Mathis. “Learnable Latent Embeddings for Joint Behavioural and Neural Analysis”. In: *Nature* 617.7960 (7960 May 2023), pp. 360–368. ISSN: 1476-4687. DOI: 10.1038/s41586-023-06031-6. URL: https://www.nature.com/articles/s41586-023-06031-6 (visited on 08/28/2023).

[212] Antoine Hubert et al. “Random-Access Two-Photon Holographic Optogenetic Stimulation Combined with Brain-Wide Functional Light-Sheet Imaging in Larval Zebrafish”. In: *Advances in Microscopic Imaging IV*. Advances in Microscopic Imaging IV. Vol. 12630. SPIE, Sept. 1, 2023, pp. 11–14. DOI: 10.1117/12.2671030. URL: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/12630/1263007/Random-access-two-photon-holographic-optogenetic-stimulation-combined-with-brain/10.1117/12.2671030.full (visited on 09/07/2023).

[213] Sebastian Quiroz Monnens et al. “The Recurrent Temporal Restricted Boltzmann Machine Captures Neural Assembly Dynamics in Whole-brain Activity”. In: *eLife* 13 (June 27, 2024). DOI: 10.7554/eLife.98489.1. URL: https://elifesciences.org/reviewed-preprints/98489 (visited on 09/12/2024).

[214] Mehdi Azabou et al. *A Unified, Scalable Framework for Neural Population Decoding*. Oct. 24, 2023. DOI: 10.48550/arXiv.2310.16046. arXiv: 2310.16046 [cs]. URL: http://arxiv.org/abs/2310.16046 (visited on 08/16/2025). Pre-published.

[215] Joel Ye and Chethan Pandarinath. "Representation Learning for Neural Population Activity with Neural Data Transformers". In: *Neurons, Behavior, Data analysis, and Theory* 5.3 (Aug. 11, 2021). ISSN: 2690-2664. DOI: 10.51628/001c.27358. arXiv: 2108.01210 [q-bio]. URL: http://arxiv.org/abs/2108.01210 (visited on 08/16/2025).

[216] Ran Liu et al. *Seeing the Forest and the Tree: Building Representations of Both Individual and Collective Dynamics with Transformers*. Oct. 20, 2022. DOI: 10.48550/arXiv.2206.06131. arXiv: 2206.06131 [q-bio]. URL: http://arxiv.org/abs/2206.06131 (visited on 08/16/2025). Pre-published.

[217] Ran Liu et al. *Drop, Swap, and Generate: A Self-Supervised Approach for Generating Neural Activity*. Nov. 3, 2021. DOI: 10.48550/arXiv.2111.02338. arXiv: 2111.02338 [cs]. URL: http://arxiv.org/abs/2111.02338 (visited on 08/16/2025). Pre-published.

Annex

On the Ethics of Animal Research

Mattéo Dommanget-Kott

August 2025

What follows is a discussion on the ethics of using non-human animals in scientific research. It will be partial and opinionated. While I will try to reference scientific publications when possible, it is important to note that this subject has not been studied extensively by the community. My goal here is to document, for myself and for others, how I formalize this question and attempt to formulate an opinion.

1 Introduction

Animal experimentation is often defended with a simple moral calculus: the benefit to human knowledge and health is said to outweigh the harm to animal subjects. In this view, researching on live animals is portrayed as necessary for medical progress, and the suffering caused, while unfortunate, is justified by the greater good of curing diseases and saving human lives. This reasoning is fundamentally utilitarian, weighing consequences on a societal scale.

Utilitarianism is the ethical theory that an action is right if it produces the greatest overall happiness or benefit (and minimal suffering) for the greatest number. In the context of animal research, a utilitarian would argue that if using animals leads to significant net benefits for humanity (or for other animals), then it can be morally permissible despite harm to animals.

This contrasts with other moral frameworks like deontological ethics, which focus on duties and rights and not consequences (for example, a deontologist might contend that beings who can suffer have certain inviolable rights, making it wrong to use them purely as means to an end, regardless of outcome), or virtue ethics, which asks what a compassionate and just person would do. While utilitarianism dominates the scientific narrative around animal research, other perspectives would challenge the assumption that a favorable cost-benefit ratio alone makes animal experimentation ethical.

Proponents of the status quo often imply that because there are arguments on both sides, the ethical "balance" is nuanced and therefore current practices must be a reasonable middle ground. The idea is that, since animal research brings both positives and negatives, our current regulatory system (which permits animal use with some welfare controls) has struck an acceptable compromise. In this opinion piece, I argue that taking an informed position on animal experimentation requires a much more critical and honest assessment of all the forces acting on this moral scale. Simply defaulting to the status quo because both benefits and harms exist can be a convenient way to avoid uncomfortable truths. Instead, scientists have a responsibility to weigh each factor with clear eyes. We should ask ourselves: *Are the benefits as great as we assume ? Is the harm minimized and justified ?* We must confront the facts openly, rather than accept easy reassurances.

In the sections that follow, I will examine the major considerations that should inform where the ethical balance truly lies. In particular, I will explore four key questions central to the utilitarian justification for animal research:

1. **Who actually benefits from animal research ?** Are the fruits of animal experiments truly benefiting patients and society at large, or do they primarily serve narrower interests ? And certainly, what about the animals themselves ?

2. **Does animal research advance medical research ?** How necessary and effective are animal experiments for scientific and therapeutic progress ?

3. **What is the extent of animal suffering caused by research ?** What kinds of harm do lab animals endure, and how do we define and recognize that suffering ?

4. **What is the impact of animal research on the scientists involved ?** Does using and harming animals carry emotional or psychological costs for researchers, and what does that tell us about the ethics of the practice ?

After discussing these points, I will examine how the structure of academia and research culture helps to maintain a system in which animal use is often seen as necessary or even unquestionable. Factors such as scientific training, peer pressure, hierarchical lab dynamics, the composition of ethics committees, and the "publish or perish" climate all influence researchers' choices and moral comfort levels. Understanding these forces can explain why the practice persists and how change is systemically discouraged.

Finally, I will step beyond the utilitarian cost-benefit framework to consider arguments beyond the numbers: issues of speciesism, rights, and the problem of

consent. Even if one accepts some balancing of harm and benefit, there are ethical red lines we might be unwilling to cross. I will discuss whether and why those lines are drawn differently for non-human animals, and what that means for the morality of animal experiments.

In conclusion, I will offer my personal perspective on where, after examining the evidence and ethical arguments, I believe the balance lies. My aim is not to impose a final answer, but to encourage fellow scientists to revisit their own intuitions and to approach this debate with full knowledge of the facts and the ethical stakes.

2 Where Does the Balance Lie ?

2.1 Who Actually Benefits from Animal Research ?

Advocates of animal experimentation typically assert that society at large benefits from the practice, especially patients who may be cured of illnesses thanks to discoveries in animal models. It is true that much of our knowledge in biology comes from animals, and that many medical advances have been attributed to animal research. For example, the development of insulin therapy, vaccines, and some chemotherapy drugs involved animal studies. However, a critical examination shows that the distribution of benefits is more complex and perhaps more limited than the straightforward narrative suggests.

Firstly, it must be acknowledged that the animals themselves (as a collective) do not benefit. In fact, by the very nature of experimentation, the animal's welfare is sacrificed for others' sake. This is obvious, but it's worth stating because it frames the moral situation: the intended beneficiaries are humans, either directly, or through animals we use as food, workforce, companions, or raw material. Whereas the costs are borne almost entirely by the animal subjects. Even if an experiment yields a successful treatment for a disease, the animals used in those studies gain nothing. At best they are alleviating future suffering of others while enduring suffering themselves, or are helping species-conservation efforts which would be unnecessary if not for humans' tendency of destroying ecosystems. This asymmetry is at the heart of the ethical concern. In this context, the human-animal relationship is one of exploitation.

Secondly, we should ask which humans benefit from animal research. The usual answer is "patients" (e.g. through treatments) or "society" (e.g. through knowledge). In many cases this is true, but it can be overstated. A large portion of animal experimentation is fundamental research that may not translate into any

clinical application (we will examine this problem in the next section). Moreover, even when therapies do result, they often target diseases that affect a subset of the population (sometimes a very small subset, like for rare diseases). That isn't to say those lives are unimportant, nor is the knowledge itself, just that "societal benefit" is diffuse.

Beyond patients and society, there are other stakeholders who benefit from the continuation of animal research, in ways that are rarely highlighted in ethical discussions. The enterprise of animal experimentation is deeply embedded in our scientific and economic system. One analysis described biomedical research using animals as "big business", noting that it generates substantial financial profit and career incentives for various actors: research institutions receive grants and prestige, scientists build careers, companies supply laboratory animals and specialized equipment, pharmaceutical and biotech firms develop products to sell [19]. It is understandably more appealing to talk about curing Alzheimer's or cancer than to acknowledge that people's livelihoods or companies' earnings depend on the continuation of animal experimentation. When weighing the ethical balance, we should be honest that some benefits are private or institutional, and not driven by compassion or thirst for knowledge.

None of this is to say that no societal benefits come from animal research. Clearly, many people are healthier today because of past animal-based studies. However, the ethical balance should account for the full picture of benefits. If we find that certain forms of animal research primarily benefit a small community of researchers or companies (by helping them secure money and prestige) while yielding little of tangible value, then the utilitarian justification weakens considerably. Even for promising research programs, we should remain aware that benefits are often probabilistic and long-term, whereas the suffering imposed on animals is certain and immediate. In the next section, we will scrutinize just how effective animal experiments have been in delivering on the promise of medical advancement, because if that promise is overstated, the claimed benefits might not truly outweigh the costs.

2.2 Does Animal Research Advance Medical Research ?

The central premise for using animals in research is that they are excellent models for understanding human biology and diseases, leading to treatments or cures. Scientists often emphasize that humans share fundamental physiological and genetic similarities with other animals (especially mammals), making animals useful proxies for ourselves. This reasoning has merit. Animals can recapitulate aspects of human biology, and research on them has indeed yielded insights. However,

a growing body of evidence shows that the utility of animal models is far more limited than one might expect, and sometimes animal studies mislead or fail to translate to human benefits.

One analysis was published in 2024 [11], reviewing how well animal experiments translate into human therapies across many fields of biomedicine. The authors found that about 50% of interventions tested in animals proceed to human trials, 40% to clinical trials, and only 5% of animal-tested interventions ultimately achieve regulatory approval as licensed treatments. In other words, out of all the drugs, procedures or treatments that show promise in animal studies, roughly 1 in 20 ends up benefiting human patients in the form of an approved treatment. It is important to note that this percentage is most likely overestimated, as it doesn't take into account unpublished negative results which are more rare in human trials due to more regulated reporting standards.

Failure and semi-blind exploration is an inherent part of science, and animals involved in "unsuccessful" experiments might still have provided important incites, even indirectly. However, if 95% of animal-studies do not translate to any direct medical benefit, can we still justify these experiments purely on utilitarian grounds? What does that say about the animal research paradigm or its underlying methodology?

The lack of animal-to-human translation is not just caused by the vagaries of the scientific method, instead, it often points to systemic shortcomings in the animal research paradigm itself. For example, animals and humans can respond very differently to the same substance: classic cases include the corticosteroids which cause birth defects in many animal species but not in humans, or the thalidomide, a tranquilizer that was harmless to several lab animal species yet caused severe birth defects in humans [7]. Even more dramatically, in 2006 a drug trial (TGN1412) nearly killed healthy human volunteers despite having been apparently safe in monkeys [7, 23], a reminder that even our closest genetic relatives can be poor predictors for human reactions.

Why do animal studies so often fall short when moved to the clinic ? One reason is simply biological differences [1]. A mouse is not a tiny human, a drug that lowers blood pressure in a dog might not do the same in a person, etc. But another reason, which is within our control, is that many animal studies are simply not designed or reported with the rigor we expect in human trials. A 2009 review by Michael Bracken [7] noted that a great number of animal experiments are *poorly designed, conducted, or analyzed*, which can produce false or irreproducible findings. For example, the under-representation of female animals in research [5] has

severe consequences on drug development [27]. Another example of methodological issues comes from Avey et al. [3] who found that animal experiments often fail to report critical details about their methods and results. In a sample of 47 preclinical studies, less than 50% of the necessary methodological details were reported on average. Important elements like randomization, blinding, or sample size calculation were frequently missing. The authors concluded that such incomplete reporting "will impede attempts to replicate research findings and maximize the value of preclinical studies". In other words, if animal studies are not conducted and reported with high standards, they may produce unreliable knowledge, which in a utilitarian perspective would correspond to an unnecessary waste of animal suffering. The scientific community has recognized this problem, leading to guidelines like ARRIVE (Animal Research: Reporting of In Vivo Experiments) [12] to improve reporting. Adherence to these guidelines is improving, but remains very inconsistent.

It's worth noting that some scientists remain optimistic about the predictive value of animal research. For example, Ineichen et al. [11] report a high level of concordance between animal and human studies (0.86), meaning that the fraction of positive results is mostly comparable between animal and clinical trials on the same treatments. However, this concordance refers to a relatively late stage in the scientific process, and is subject to publication and reporting biases. It is very probable that many ideas fail between the early animal experiment stage and the clinic, either because they never yielded positive results, or simply because animal studies tend to focus on mechanisms while human studies focus primarily on the effectiveness of treatments [11]. This implies that we should be cautious in assuming that every animal experiment is crucial to research. In fact, some critics argue that an over-reliance on animal models can slow progress, either by directing resources into avenues that don't pan out, or causing potentially good treatments to be abandoned because they didn't work in animals [10].

In summary, The utilitarian argument assumes a clear and strong benefit to humans. While it is true that virtually every major medical treatment can be traced back to animal research, we have shown not only that very few research project lead to real treatments, but also that animals might not be the human-models we often expect them to be. Importantly, very few publications assess the potential benefits of animal research, which illustrates on overall disinterest of the community towards animals, except as biological models.

It may be that we need to accept a high failure rate as part of scientific exploration. Yet from an ethical perspective, a practice that produces a lot of animal suffering for relatively sparse or vague successes should be scrutinized. The ulti-

mate goal for many is to replace animal models entirely with more human-relevant and ethical tools. But until that goal is reached, we need to be very honest about the limitations of the animal research paradigm when we weigh its moral justification.

2.3 What is the Extent of Animal Suffering Caused by Research ?

Balancing a moral equation means not only measuring benefits, but also fully understanding the costs. In ethical terms, the "cost" here is the harm, pain, distress, and death inflicted on animals in the course of research. It's easy to speak abstractly about "animal suffering" without appreciating what that entails, so let's break it down: *How many animals are we talking about ? What do they experience ? And how do scientists define and mitigate suffering ?*

Scale of animal use. A recent estimate indicated that around 192.000.000 animals were used for scientific purposes worldwide in a single year (2015) [24]. The EU alone reports about 10 – 12 million uses of animals in experiments per year in the last decade. This includes mammals like mice, rats, guinea pigs, rabbits, dogs, cats, non-human primates, as well as birds, fish, and others. The vast majority are rodents and fish, with mice alone being the most common laboratory animal by far. Each of these animals is a sentient being capable of suffering. When we talk about the ethics, we must keep in mind that we are not dealing with a few cases, but tens of millions of animals each year that we purposely expose to distressing or harmful conditions.

Procedures and suffering. Modern regulations classify experiments in terms of *severity*, which ranges widely depending on the experiment. Some animals are used in relatively mild procedures (like a brief injection or blood draw under anesthesia). Others experience significant pain or distress being subjected to surgery, toxic substances, induced diseases, or psychological stress (such as inducing fear, isolation, or depression-like states in behavioral research). Many animals are euthanized at the end of experiments. In some fields, death is the required endpoint to gather tissues or because the study observes how long animals survive a condition. Modern regulations classify experiments by severity: "non-recovery" (under full anesthesia, doesn't wake up), "mild", "moderate", or "severe" suffering. The "severe" category includes procedures expected to cause a high degree of pain, suffering, or distress, or lasting impairment of well-being. In the EU's 2017 report, for example, about 12% of uses were classified as severe, and roughly 26% as moderate, 50% mild, and the rest under anesthesia with no recovery. This means that

millions of animals experience significant suffering each year. Beyond this classification, it is worth questioning whether we, as humans, would find these levels of suffering acceptable if inflicted on us. If the answer is no, we need to question why it would then be acceptable to inflict it upon an animal. If the answer is yes, we need to ask ourselves why we inflict it upon animals and not humans. We will discuss this point in section 4.

Recognizing pain and distress. One might think scientists always relieve animal pain as much as possible. Certainly, animal welfare regulations and guidelines compel researchers to minimize pain when compatible with the study. This is commonly known as *Refinement* in the 3Rs [25]: minimize pain, suffering, distress, etc. And indeed, many researchers are compassionate toward their animals and do provide pain relief and "humane" endpoints. However, implementing refinement is not always straightforward. Studies have found that scientists and even veterinarians sometimes struggle to recognize animal pain or distress reliably. Animals, especially prey species like rodents, often hide outward signs of pain (an evolutionary adaptation). In an interview study of Canadian scientists and vets [9], many participants acknowledged that recognizing when lab animals are in pain is challenging and prone to subjectivity. Some felt that only experience gave them a "feel" for an animal's discomfort, while others pointed out that even experienced observers can miss or misinterpret pain signs. Other experiments have shown that even trained, experienced lab animal professionals are not better than novices at detecting pain [14]. There is also an observed tendency for caregivers to underestimate the suffering of animals under their care, perhaps due to a form of bias or wishful thinking. For example, one review noted that observers who are close to the animals often "see" them as doing better than they objectively are [16], maybe as a coping mechanism. All of this suggests that animals in labs may be suffering more than we fully realize.

Even when pain is recognized, it may not always be alleviated. Guidelines typically state that if an animal would suffer significant pain that cannot be relieved, the experiment should be ended early or not done at all. Yet in practice, what is deemed "significant" or "unrelievable" is somewhat subjective and left to ethics committees and investigators to decide. The ethical framework assumes researchers will balance the suffering against the scientific gains. But given that animal subjects cannot consent or speak, they rely on humans to decide when their suffering is "too much". It is then legitimate to ask: *Are we good proxies for our animals' well-being ?* And more importantly: *Are we any good at identifying when not to perform an experiment ?*

Institutional evaluation of suffering vs. benefits. One review [17] of animal ethics committees noted that harm-benefit analyses sometimes focus heavily on minimizing harms but pay less attention to whether the harms are truly justified by likely benefits. An ironic situation which could lead to reduced but pointless suffering. Importantly, this review found that the hierarchy of the 3Rs (Replacement > Reduction > Refinement) are given inverse priorities during discussions, with Replacement arriving last. As noted by the authors, this is expected from the composition of committees (scientists in majority), the training of committee members, and the biases present in the field. We will discuss this point in more detail in section 3.

It is important to note this does not mean committee members are devoid of empathy. These are systemic problems relative to training, information availability, and more importantly little to no guidelines on how to assess the potential benefits of a research project. Nevertheless, it illustrates the failing nature of this fail-safe institution.

In summary, the suffering of animals in research is real and significant. It ranges from mild distress to severe pain, and from short-lived to lifelong. We try to contain it with ethical guidelines, but given the numbers of animals, even a small percentage experiencing severe suffering equates to a very large absolute number of individuals in severe distress. Furthermore, as humans we are not always (and perhaps rarely) able to recognize the signs of distress in our animal subjects, and the institutions whose role is to regulate the use of animals in research are not functioning as intended.

2.4 What is the impact of animal research on the scientists involved ?

An often overlooked aspect of animal experimentation is its emotional and psychological impact on the scientists, technicians, and students who carry it out. If using animals were an unequivocally noble endeavor yielding clear benefits, perhaps those involved would feel only pride. In reality, many people who work with research animals experience conflicted feelings, stress, or even trauma related to harming animals as part of their job. This phenomenon is sometimes referred to as "moral distress": the pain of doing something that one feels is morally troubling [18].

Surveys and studies have begun to shed light on this issue. Mamzer et al. [15] surveyed 150 Polish scientists who use animals in experiments, aiming to identify negative psychological effects of their work. They found that 72% of the researchers

surveyed said that performing experiments on animals was emotionally burdening for them, and 63% reported feeling stress when conducting animal procedures. Nearly half (47%) even admitted that they felt many experiments on animals are unnecessary, suggesting that a significant number of scientists harbor doubts about the justification of what they're doing.

Qualitative research and anecdotal reports echo these statistics. Laboratory animal technicians often form attachments to the animals yet are also the ones who must euthanize them or witness their suffering. This can create *compassion fatigue*[21]. Some techs describe becoming numb or using euphemisms (like saying an animal was "sacrificed" rather than "killed") to cope with the tasks of ending animals' lives. Scientists themselves, especially younger researchers and students, can face a moral shock the first time they have to, for example, kill a mouse or dog for an experiment. There are documented cases of students abandoning research careers because they couldn't reconcile with animal suffering. Others continue but report a lingering remorse or guilt. Mamzer et al. [15] noted that younger researchers in particular felt *remorse, emotional tension, and helplessness* in the face of what they were doing. This suggests that over time, people might adapt or use psychological strategies to reduce *cognitive dissonance* [8], perhaps becoming desensitized or convincing themselves what they do is necessary. Indeed, more senior researchers often report being more at ease, which could indicate that those who couldn't handle it have left, and those remaining have found ways to rationalize or compartmentalize the harm (a kind of "survivor bias" in the profession). It could also be a symptom that senior researcher tend to delegate experiments to their students and postdocs, and therefore become disconnected from day-to-day lab life and its consequences. Or it could be that the "emotional load" has worsened during the last few decades, either because of changes in procedures or because newer generations are more aware of the issue. In any case, this creates distance and misunderstanding between senior and junior researchers, which only adds to the psychological effects described above.

Ethically, why does this matter ? After all, one could say the emotional state of researchers is secondary to animal suffering. However, the fact that hurting animals can hurt humans too underscores that there is something ethically troubling going on. It's a signal that a moral line is being crossed, triggering natural empathy and distress in some individuals. The psychological burden on researchers can also have practical consequences: chronic stress or moral conflict can lead to burnout, mental health struggles, or a decline in the quality of science (a distracted, emotionally drained researcher may not perform optimally).

Moral disengagement. The causes of this distress are not hard to fathom. Many people enter science because they love animals or nature. To then find that to advance science or career, one must intentionally harm animals, creates a moral dissonance. Researchers might cope through what psychologist Albert Bandura called the moral disengagement mechanism [4]. For example, by euphemistic labeling ("sacrifice" instead of kill, "animal model" instead of individual), diffusion of responsibility ("the ethics committee approved this, it's legal, so it's not just my doing"), advantageous comparison ("if I didn't do this, someone else would", or "think of the worse suffering in the wild or in factory farming"), and de-individualizing the animals (referring to them by number, or as material). These strategies can dampen empathy.

"For Science". A social psychology experiment by Bègue and Vezirian [6] showed that when people are put in a "pro-scientific mindset" and instructed by an authority, they became more willing to administer what they believed were harmful doses of a chemical to a fish, compared to people who weren't given that scientific rationalization. In essence, invoking "science" and "for knowledge" can induce ordinary people to override their compassionate instincts. This can be compared with the classic Milgram obedience experiments, which showed that ordinary people can exceed their moral limits by entering an *agentic* state where, prompted by authority, a person can view themselves as an instrument of that authority, and therefore feel as though they are not responsible for their actions. This finding suggests that the culture of science, which valorizes discovery and may sometimes encourage a rather cold, rationalistic approach, enables moral disengagement when it comes to animals. People tell themselves they are doing good for a higher cause (science and human health), which helps them cope with doing harm in the moment.

In summary, the *human cost* of animal research is non-negligible. Many scientists carry an emotional burden, some become desensitized as a coping mechanism, which has its own ethical implications. The presence of moral distress among researchers indicates that, on some level, many recognize a moral conflict between causing animal suffering and their own moral values. This internal conflict adds another crack in the idea of "balance" that supposedly justifies animal experimentation.

3 Academic Culture and the Perpetuation of Animal Use

Given the significant ethical challenges we've outlined: questionable translation to human benefits, clear harm to animals, and even harm to researchers' well-being, one might wonder: *why does the practice of animal experimentation persist so pervasively ?* I believe that part of the answer lies in the culture and structure of academia and scientific research. Over decades, a complex system has evolved that in many ways locks-in the use of animals, making it difficult for individuals to deviate even if they have moral misgivings or see scientific alternatives. Let's examine several facets of this system.

Education and Early Training. The professional socialization of scientists often begins in university or even earlier. Biology and medical students are typically introduced to animal use as a matter-of-fact necessity. Dissecting animals in anatomy class, performing simple experiments on animals in physiology or pharmacology labs. This early exposure, usually without deep ethical reflection, sends a message that using animals is a normal, accepted aspect of becoming a scientist. Graduate training reinforces this: young researchers learn techniques from mentors, and if the lab's established methods involve animals, a newcomer will adopt them without necessarily questioning the ethical dimension. In fact, learning to handle and experiment on animals is considered a valuable skill. There is also a psychological aspect: the first time a student must harm or kill an animal is often the hardest, subsequent times become easier as one becomes desensitized or rationalizes it. Thus, by the time scientists are independent, many have gone through a kind of moral acclimatization: what once might have been emotionally difficult becomes routine. The idea that "this is just how science is done" can become deeply ingrained.

Mentorship and Hierarchy. Academia is hierarchical. Senior scientists (professors, principal investigators) have tremendous influence over the direction of research in their field and the training of new scientists. If senior figures built their careers on animal research, they are likely to champion its continued use. This is sometimes called the "old man effect" (though senior scientists are not all old or male, the term captures the inertia of established authority). These established researchers sit on grant panels, head departments, and serve on ethics committees. Their beliefs and biases shape institutional policies. There may also be survivor bias at play: those who were comfortable with animal experimentation (or at least not opposed to it) are the ones who advanced in the field, whereas those who strongly objected likely left or never achieved influential positions. Therefore,

the leadership in many scientific domains may disproportionately consist of people who are firmly convinced that animal research is necessary and ethical, because if they weren't, they wouldn't have risen to the top in the current system. Junior scientists learn quickly that openly challenging the paradigm can be career suicide. If you're a postdoc who says "I refuse to do animal experiments on ethical grounds", you may not last long in a lab whose funding depends on animal work. Peer pressure, whether explicit or implicit, keeps people in line.

Peer Culture and Norms. Even aside of direct hierarchy, there is a powerful peer norming in science. The default assumption is that animal experiments are the gold standard for certain questions [13]. For example, in drug development, you must show efficacy in an animal model before a journal will consider your therapy paper, or before a grant agency will fund a clinical trial. This norm is even encoded in regulations (e.g., the FDA requires animal toxicity studies before human trials). So a scientist who opts not to use animals might be seen as doing inferior work or skipping important steps, even in cases where alternatives exist or might even be more appropriate. There can also be a fear of being labeled as "too emotional" or "unscientific" if one voices concerns about animal suffering, as if that is a sentimental distraction rather than a legitimate ethical issue. Thus, the system encourages scientists to keep any doubts to themselves, which in turn encourages others to do the same.

Beyond these passive incentives are more direct ones. Publishing novel results is the key to academic survival ("publish or perish"), and animal experiments often provide a more straightforward route to a publication. New in vivo results (especially in prestigious journals) can be seen as more "exciting" or convincing than in vitro or computational studies, which means researchers may choose animal experiments to increase their chances of publication. Indeed, a recent survey [13] on publishing bias found that many scientists and reviewers perceive a "methods bias" in favor of animal studies: research using non-animal methods (e.g. cell culture, computer modeling) is taken less seriously and sometimes faces pressure during peer review to add animal data. Such bias in the publication process reinforces the idea that to be a successful scientist, one *must* use animals, discouraging the development and/or use of alternative methods.

Ethics Committees. Animal Ethics Committees are supposed to be the safeguard for animal welfare and ethical balance. One of their key roles is to evaluate the scientific necessity of research projects, and to verify that the experimental protocols minimize harm as much as possible (example in french legislation [2]). However, their structure often builds on a conflict of interest as many committee members are researchers who use animals themselves. The law usually requires

some non-affiliated members (e.g. community representatives or ethicists) to be on these committees, but they are typically a minority. Milford et al. [17] found that committees can be numerically but also socially "imbalanced in favor of researchers and veterinarians", affecting, among other things, the independence of the review, with the effect of skewing discussions "towards harm reduction rather than ethical evaluation". It has been suggested that this imbalance could be due to difficulties in recruitment, reluctance to raise ethical concerns with colleagues, or even institutional pressure [20]. This is coherent with other studies which suggest that successful funding applications can influence committee members towards positive reviews [26], suggesting that committee tend to "outsource" the evaluation of *scientific necessity* (*i.e.* reduction of animal research) in favor of harm minimization (*i.e.* refinement of protocols).

This illustrates the fact that *institutionalization* can create a dilution of responsibility: researchers can say "the committee approved my study, so it must be ethically acceptable", while committees might assume the researcher wouldn't propose it if it was not important, especially when this study received prior funding. Essentially, no single person feels fully responsible, which, psychologically, makes it easier for everyone to allow ethically dubious practices to continue.

Fundamentals of Science This dilution of responsibility also lies much deeper into the foundations of the scientific method. Indeed, scientific breakthroughs are very rarely due to a single researcher or project. Science is an inherently collective endeavor where a dense web of experiments and theoretical works build on each other. This idea leads to the claim (often heard in discussions with fellow scientists) that "the value of animal research cannot be assessed at the level of a single experiment, we must instead look at the cumulative progress over years or decades".

Judging the necessity of animal research in a particular field could then only be done over long time scales ? Then how long should we wait before passing judgment ? Five years ? Twenty ? One hundred ? How widely should we search the literature for signs of usefulness ? What if we don't find any ? Should we then wait/search more ? With different criteria ?

It seems that this line of reasoning cannot lead us anywhere because this claim is irrefutable. *What would be conclusive evidence that a research project is not or will not be useful ?*

One could then argue that science doesn't have to be "useful" or applied, even on long time scales. Science can just provide knowledge for knowledge's sake. But then, *how do we weigh knowledge compared to animal suffering ?* Or more precisely: *how do we weigh potential future knowledge compared to present, tangible suffering ?* One would either need a high degree of faith in "scientific progress"

to claim that the potential benefits outweigh the present costs, or consider the current costs negligible.

Consequently, the scientific method is built in such a way that each scientist downplays the significance of their own animal use in the grand scheme, even as all scientists together create a large system of harm. And as the formal harm-benefits analysis cannot be applied to fundamental research, virtually any project can be defended, regardless of the suffering it creates.

All these factors create a powerful inertia. Even a scientist who personally feels uneasy about animal suffering might conclude that "this is just how it has to be if I want to contribute to science or keep my job". This effect might be particularly strong when researchers feel their only skills lie in animal handling, surgery, etc. The structure does not readily reward individuals for cutting back on animal use (though there are now some grants for alternative methods, those are still niche). There can also be a fear that using alternative methods is risky. If they fail to produce clear results, one might have to fall back on animals anyway, thus having lost time. Given the competition in science, many stick to the devil they know.

Cultural change is slow. It requires not just individual awakening but systemic reform. This could mean, for instance, journals and funding agencies explicitly rewarding non-animal methods and not reflexively demanding animal data. It could mean ethics committees placing much heavier weight on the "necessity" of using animals. It might also involve better training in ethics for scientists, so that early-career researchers are at least equipped to think critically about the moral implications, rather than just inheriting the status quo mindset.

In conclusion, academia's current structure has many feedback loops that maintain animal experimentation. Beyond the potential benefits, understanding those loops helps explain why, despite ethical and scientific misgivings, the practice continues. It's not just about the science. It's also about careers, norms, money, and group dynamics. Any serious ethical discussion about the necessity of animal research needs to grapple with these institutional forces, because even the strongest moral arguments may not prevail if the system is rigged to ignore them.

Changing the trajectory will likely require concerted effort on multiple fronts: policy reform, cultural change, and continued development of alternative methods that can break the monopoly of the animal model.

4 Beyond Utilitarianism: Speciesism, Rights, and the Question of Consent

Up to now, we have been discussing animal experimentation largely in the framework its proponents prefer: a utilitarian cost-benefit analysis. We've tried to weigh the suffering versus the benefits and asked "is it worth it ?" But even this way of framing the issue can be challenged. *Are we in favor of animal research because we are utilitarian ? Or are we utilitarian because we are in favor of animal research ?*. There are compelling ethical arguments that even if animal research did produce great benefits and even if harms were minimized, something fundamental is wrong with the practice. These arguments invoke concepts of justice, rights, and equality rather than utility. In this final analytical section, let's consider some of these perspectives

4.1 Speciesism and Moral Equality

Ask the experimenters why they experiment on animals, and the answer is: 'Because the animals are like us.' Ask the experimenters why it is morally okay to experiment on animals, and the answer is: 'Because the animals are not like us.' Animal experimentation rests on a logical contradiction.
Philosopher Charles R. Magel (allegedly)

The term speciesism was popularized by psychologist Richard Ryder and later by philosopher Peter Singer to describe an arbitrary discrimination against beings based on their species membership. Singer, a utilitarian philosopher, argued that the capacity to suffer (to feel pain and joy) is the key trait that gives a being moral consideration. *If a being suffers, there can be no moral justification for refusing to take that suffering into account* he wrote [22]. By this logic, the suffering of a mouse or a dog matters as much in principle as the suffering of a human, to the extent that the suffering is of similar intensity and duration. This does not mean treating humans and mice identically in all respects, it means equal consideration of interests [22]. An animal has an interest in not feeling pain, and thus we are morally obligated to consider that interest rather than push it aside simply because it is not a human animal. Speciesism is compared to racism or sexism: a prejudice that favors one's own group (in this case, one's own species) without morally relevant justification.

We routinely do to animals things that we would never do to a human, even if there were great benefit to other humans. For example, imagine a research program that could potentially cure cancer but required deliberately infecting a

group of children with a deadly disease to test a treatment. Most people would not accept that because we recognize humans have rights and dignity that cannot be overridden by usefulness. Yet, we do this to animals because they are not members of our species. *Is that a morally relevant difference ?* Many people would argue that, as we are the smartest animals, we would "feel" the effects of experimentation more keenly. But then *what of human children or people with disabilities ?* This is known as the "argument from marginal cases" in ethics: if intelligence or autonomy is the criterion for protection, then some humans would fall short and be exploitable, which we reject. Therefore, we must be using a different criterion (basically, species membership) which is arbitrary.

4.2 Animal Rights and Deontological Perspectives

Beyond utilitarian weighing, there is the view that animals, or at least certain animals (like all mammals, or all vertebrates), have inherent rights: rights not to be killed or not to be caused to suffer for others' purposes. Philosopher Tom Regan famously argued that any being who is a "subject-of-a-life" (meaning an individual with perceptions, memories, desires, sense of future, and an emotional life) has inherent value and rights, in much the same way persons do. Under Regan's view, most animals used in labs qualify as subjects-of-a-life. Therefore, using them as mere means to an end (even a beneficial end) violates their rights. This is a stark contrast to utilitarianism: in a rights view, it's not about balancing pain versus pleasure outcomes, it's about respecting the individual. If it's wrong to kill a healthy human to harvest organs to save five other humans, then by the same logic, it's wrong to kill or hurt an animal to potentially save humans, because the animal's basic right to life and to not be harmed should not be transgressed even for good ends.

Even if one doesn't give "full rights" to animals, one might hold a deontological constraint that says: "No matter the benefit, you must not perform certain harmful acts on animals", akin to how we have absolute prohibitions on say, torture or non-consensual experimentation in human ethics. Certainly, in human research ethics, consent is paramount: we do not experiment on people without their informed consent (except in rare cases like emergency experimental treatments, and even those are tightly regulated).

4.3 Consent and Autonomy

In human medical ethics, the principle of autonomy is king. Even potentially life-saving research on a person requires that person's voluntary consent (or their legal surrogate's consent if they can't consent). We do not allow experimenting

on prisoners or the mentally disabled or children without consent, even if it could benefit others, because that would treat those individuals as mere means. Yet we routinely treat animals exactly as means to our ends.

Animals cannot consent. This is a fundamental ethical problem. We justify bypassing consent by saying animals aren't capable of consenting, but that's exactly why we perhaps shouldn't use them. They have no capacity to agree or refuse (or we don't have the capacity to understand when they do), so doing things to them is inherently coercive. If an experiment is of a kind that we believe morally requires consent when subjects are humans, performing it on non-consenting animals raises the question of double standards. *Should we presume "consent" on behalf of animals because we deem the research important ?* Not more or less than we should for humans. Even though we tend to value human lives more that animals', it does not logically follow that animals have less value.

The problem of consent also ties to the notion of trust and betrayal. Many lab animals (especially primates, dogs, pigs, ...) are social creatures capable of relationships with humans. In labs, animals often come to trust the humans that feed or handle them, at least to some degree. To then harm them can be seen as a kind of betrayal of that trust. This is more of a virtue ethics perspective: what does it say about our character that we can betray and harm the vulnerable who trust us? And what might that do to society's moral fiber? We know historically that seeing certain groups as less than fully human enabled atrocities. There is an uneasy parallel in how some talk about lab animals as "models" or "tools", stripping them of individual identity. One might argue that the practice of animal experimentation exists in an ethical gray zone largely because the victims cannot advocate for themselves. It is an abuse of the voiceless.

In summary, stepping outside the utilitarian framework, one can argue that even a perfectly efficient animal research paradigm would be morally problematic, because it rests on an assumption of human privilege that is impossible to justify logically. If we grant animals a significant moral status then harming them starts to look like tyranny. This line of thought doesn't necessarily provide easy solutions, it might demand a radical shift, like according legal rights to animals or ending all uses of animals.

5 Conclusion: Weighing the Balance and My Personal Opinion

It is for everyone to decide where, for them, the balance lies. But it is important to take position and to review the facts. "Moral neutrality" doesn't exist: every moral choice, conscious or not, holds moral value.

How it started. I was drawn to my thesis project, not because I wanted to work with animals, but because I was exceedingly curious and passionate about the subject. I had never worked with animals, nor had I ever been a proponent of animal research, quite the opposite. My prior opinion was that the use of animals is highly immoral, whether in research, industry, and as food. I had been of the opinion that we live in a speciesist society for a few years before the start of my PhD. But then, why did I enter this field ? I think there were 4 reasons:

1. **Curiosity and excitement for the research project**. Scientifically, the project was appealing, the lab friendly. I felt this was the project and place that would help me become a good scientist.

2. **Curiosity for the reality of animal research**. Although I was largely against animal research, I was curious how my beliefs would stack up against the reality of the field. While I would not have been able to work with mice or monkeys, I though that working with fish would cause me less moral distress. Thus it felt like an opportunity to understand, from the inside, how animals are used or abused in a scientific context.

3. **Belief that the project could help reduce animals suffering**. Contrary to most experiments in neuroscience, imaging zebrafish brains doesn't require opening their skulls or any other form of surgery. I thought that, by promoting the use of zebrafish in my field instead of mammals, I could help reduce the number of severe procedures, and therefore help reduce overall suffering. *A small evil for a greater good.*

4. **Belief that I could help change things from the inside**. By being there, seeing what was wrong, I though I could help change things for the better. Even if it was just at the small scale of a single lab.

While at the time those reasons where genuine, I think retrospectively that I was both naive and lying to myself, probably already rationalizing as a way of reducing cognitive dissonance.

How I tried to anchor myself to my moral beliefs. From the beginning, I knew there was a risk that I would become desensitized to the suffering I inflicted upon the fish, and this scared me immensely. To mitigate this risk, I decided that I would give a name to each fish as a way of constantly remembering that they are individuals and not *things*. However, given the number of eggs, larvae, and adult fish involved, this was impossible. Instead, I named (almost) every fish I used directly in an experiment ($\sim$ 300 over 4.5 years, even for tests, and regardless of the severity of the experiment). While this might seem ridiculous or futile, I think it really helped me.

I also decided I would try and get to know the adult fish individually, at least as much as possible. I watched them, took pictures, exercised my identification skills. To some extent this worked. For each generation ($\sim$ 1.5 years), I was able to identify most of them. However retrospectively, I don't know if this truly helped in any way.

Finally, I decided that every day, after work, I would try to purposefully recall the fish I interacted with during that day, to keep a written record of my thoughts and emotions. I tried to remember how they were, what I did to them, how many I killed, how I felt, etc. This was by far the hardest thing to do. It was hard because it forced me to pay a lot of attention, but also because it forced me to remember things I would have preferred to forget: shocking event, images, or mistakes that I had made. While this was sometimes painful, I believe it helped immensely.

However, in the end, I found that the only way to calm my "inner moral voice" was to avoid doing experiments altogether. I spent around two years in the middle of my PhD without doing any experiments, and the scientific quality of my research probably suffered from it. When I finally had to start experimenting again, it was psychologically very hard, but by then I also felt keenly the pressure to finish the project.

How I still got caught by the system. I am proud that during those 4.5 years, I never *forgot* that each fish was an individual deserving consideration and respect. However, and despite my best efforts, I did become somewhat desensitized to some aspects of their suffering. I sometimes got carried away by the scientific excitement of a particular experiment, and, in periods of publication-related stress, I sometimes did experiments which were retrospectively useless, or where the level of suffering could not be justified by the scientific need. Most of all, I repeatedly got caught by the sunk cost fallacy: *I started this series of experiments. Stopping now would mean that all the fish used previously were "waisted" suffering ! And so it continues.*

I am not saying this because I want to atone or because I am asking for some kind of forgiveness. I mention it because it is a good example of how the context

and human psychology can lead someone to do things they consider morally wrong, even when a person is motivated not to and was aware of the risks.

My personal conclusion. So, in the end, *has my opinion changed on animal research ?* The answer is *Yes*, but not in favor of it. I am more convinced now than I was before that animal experimentation is morally unjustifiable.

When considering the benefits of animal research, I agree with the prevalent opinion that it produces valuable knowledge and medical applications (at least to some degree). Whether we are talking about vaccines, cancer treatments, psychoactive drugs, etc. However, the scale at which we are organizing the suffering of million of animals is just too great. I see no reason to apply different moral value to a human than to another animal, and I certainly won't accept a system which organizes the suffering of millions of humans. Thus, and even in a utilitarian paradigm, the balance falls squarely against animal research.

I used to think that refinement and reduction of animal use could be a good solution to decrease animal suffering in labs. However, I am now convinced that the system is fundamentally rigged in a way that cannot be fixed. Refinement and reduction are used as political selling point to justify the continued used of animals in research, scientists and industrials have to much invested interests in the use of animals, and ethical committees are biased towards the "necessity" of animal research. If the goal is to reduce animal suffering, then only replacing or abandoning the use of animals in research would allow to reach this goal.

In fact, I believe there are so many reinforcing feedback loops in the system that it cannot be changed, even from the inside. My only regret in leaving this field is that I will participate in a survivor's bias, where only researcher who seem morally comfortable with experimenting on animal stay, and others, like me, abandon the animals in others' hands. May those who stay prove me wrong.

Acknowledgments

I would like to thank Antoine Hubert, Monica Coraggioso, Leonardo Demarchi, and Georges Debrégeas for reading this piece and providing useful incites and corrections.

References

[1] Aysha Akhtar. "The Flaws and Human Harms of Animal Experimentation". In: *Cambridge Quarterly of Healthcare Ethics* 24.4 (Oct. 2015), pp. 407–419. ISSN: 0963-1801, 1469-2147. DOI: 10.1017/S0963180115000079. URL:

https://www.cambridge.org/core/journals/cambridge-quarterly-of-healthcare-ethics/article/flaws-and-human-harms-of-animal-experimentation/78D1F5E6B65AE7157B7AA85FF3F06017#fn63 (visited on 08/17/2025).

[2] *Article 4 - Arrêté Du 1er Février 2013 Relatif à l'évaluation Éthique et à l'autorisation Des Projets Impliquant l'utilisation d'animaux Dans Des Procédures Expérimentales - Légifrance*. URL: https://www.legifrance.gouv.fr/loda/article_lc/LEGIARTI000027038889 (visited on 08/26/2025).

[3] Marc T. Avey et al. "The Devil Is in the Details: Incomplete Reporting in Preclinical Animal Research". In: *PLOS ONE* 11.11 (Nov. 17, 2016), e0166733. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0166733. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0166733 (visited on 08/17/2025).

[4] Albert Bandura. "Moral Disengagement in the Perpetration of Inhumanities". In: *Personality and Social Psychology Review* 3.3 (Aug. 1, 1999), pp. 193–209. ISSN: 1088-8683. DOI: 10.1207/s15327957pspr0303_3. URL: https://doi.org/10.1207/s15327957pspr0303_3 (visited on 08/17/2025).

[5] Annaliese K. Beery and Irving Zucker. "Sex Bias in Neuroscience and Biomedical Research". In: *Neuroscience & Biobehavioral Reviews* 35.3 (Jan. 1, 2011), pp. 565–572. ISSN: 0149-7634. DOI: 10.1016/j.neubiorev.2010.07.002. URL: https://www.sciencedirect.com/science/article/pii/S0149763410001156 (visited on 08/24/2025).

[6] Laurent Bègue and Kevin Vezirian. "Sacrificing Animals in the Name of Scientific Authority: The Relationship Between Pro-Scientific Mindset and the Lethal Use of Animals in Biomedical Experimentation". In: *Personality & Social Psychology Bulletin* 48.10 (Oct. 2022), pp. 1483–1498. ISSN: 1552-7433. DOI: 10.1177/01461672211039413. pmid: 34583579.

[7] Michael B Bracken. "Why Animal Studies Are Often Poor Predictors of Human Reactions to Exposure". In: *Journal of the Royal Society of Medicine* 102.3 (Mar. 1, 2009), pp. 120–122. ISSN: 0141-0768. DOI: 10.1258/jrsm.2008.08k033. pmid: 19297654. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2746847/ (visited on 08/17/2025).

[8] Robyn M. Engel et al. "Cognitive Dissonance in Laboratory Animal Medicine and Implications for Animal Welfare". In: *Journal of the American Association for Laboratory Animal Science* 59.2 (NaN/NaN/NaN), pp. 132–138. ISSN: 1559-6109. DOI: 10.30802/AALAS-JAALAS-19-000073. URL: https://aalas.kglmeridian.com/view/journals/72010024/59/2/article-p132.xml (visited on 08/17/2025).

[9] Nicole Fenwick, Shannon E. G. Duffus, and Gilly Griffin. "Pain Management for Animals Used in Science: Views of Scientists and Veterinarians in Canada". In: *Animals* 4.3 (Sept. 2014), pp. 494–514. ISSN: 2076-2615. DOI: 10.3390/ani4030494. URL: https://www.mdpi.com/2076-2615/4/3/494 (visited on 08/17/2025).

[10] Thomas Hartung. "The (Misleading) Role of Animal Models in Drug Development". In: *Frontiers in Drug Discovery* 4 (Apr. 8, 2024). ISSN: 2674-0338. DOI: 10.3389/fddsv.2024.1355044. URL: https://www.frontiersin.org/journals/drug-discovery/articles/10.3389/fddsv.2024.1355044/full (visited on 08/18/2025).

[11] Benjamin V. Ineichen et al. "Analysis of Animal-to-Human Translation Shows That Only 5% of Animal-Tested Therapeutic Interventions Obtain Regulatory Approval for Human Applications". In: *PLOS Biology* 22.6 (June 13, 2024), e3002667. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.3002667. URL: https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3002667 (visited on 08/17/2025).

[12] Carol Kilkenny et al. "Improving Bioscience Research Reporting: The ARRIVE Guidelines for Reporting Animal Research". In: *PLOS Biology* 8.6 (June 29, 2010), e1000412. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.1000412. URL: https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1000412 (visited on 08/17/2025).

[13] Catharine E. Krebs et al. "A Survey to Assess Animal Methods Bias in Scientific Publishing". In: *ALTEX - Alternatives to animal experimentation* 40.4 (Oct. 17, 2023), pp. 665–676. ISSN: 1868-8551. DOI: 10.14573/altex.2210212. URL: https://altex.org/index.php/altex/article/view/2568 (visited on 08/17/2025).

[14] Matthew C. Leach et al. "Are We Looking in the Wrong Place? Implications for Behavioural-Based Pain Assessment in Rabbits (Oryctolagus Cuniculi) and Beyond?" In: *PLOS ONE* 6.3 (Mar. 15, 2011), e13347. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0013347. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0013347 (visited on 08/18/2025).

[15] Hanna Mamzer et al. "Negative Psychological Aspects of Working with Experimental Animals in Scientific Research". In: *PeerJ* 9 (Apr. 20, 2021), e11035. ISSN: 2167-8359. DOI: 10.7717/peerj.11035. URL: https://peerj.com/articles/11035 (visited on 08/17/2025).

[16] Rebecca K. Meagher. "Observer Ratings: Validity and Value as a Tool for Animal Welfare Research". In: *Applied Animal Behaviour Science* 119.1 (June 1, 2009), pp. 1–14. ISSN: 0168-1591. DOI: 10.1016/j.applanim.2009.02.026. URL: https://www.sciencedirect.com/science/article/pii/S0168159109000690 (visited on 08/18/2025).

[17] Aoife Milford et al. "How Animal Ethics Committees Make Decisions – a Scoping Review of Empirical Studies". In: *PLOS ONE* 20.3 (Mar. 17, 2025), e0318570. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0318570. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0318570 (visited on 08/17/2025).

[18] Georgina Morley et al. "What Is 'Moral Distress'? A Narrative Synthesis of the Literature". In: *Nursing Ethics* 26.3 (May 2019), pp. 646–662. ISSN: 0969-7330. DOI: 10.1177/0969733017724354. pmid: 28990446. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6506903/ (visited on 08/17/2025).

[19] Kay Peggs. "An Insufferable Business: Ethics, Nonhuman Animals and Biomedical Experiments". In: *Animals* 5.3 (Sept. 2015), pp. 624–642. ISSN: 2076-2615. DOI: 10.3390/ani5030376. URL: https://www.mdpi.com/2076-2615/5/3/376 (visited on 08/17/2025).

[20] Denise Russell. "Why Animal Ethics Committees Don't Work". In: (Jan. 1, 2012). URL: https://ro.uow.edu.au/articles/journal_contribution/Why_animal_ethics_committees_don_t_work/27778056/1 (visited on 08/17/2025).

[21] Rebekah L. Scotney, Deirdre McLaughlin, and Helen L. Keates. "A Systematic Review of the Effects of Euthanasia and Occupational Stress in Personnel Working with Animals in Animal Shelters, Veterinary Clinics, and Biomedical Research Facilities". In: *Journal of the American Veterinary Medical Association* 247.10 (Nov. 15, 2015), pp. 1121–1130. DOI: 10.2460/javma.247.10.1121. URL: https://avmajournals.avma.org/view/journals/javma/247/10/javma.247.10.1121.xml (visited on 08/17/2025).

[22] Peter Singer, ed. *Animal Liberation*. Avon Books, 1977.

[23] Richard Stebbings et al. ""Cytokine Storm" in the Phase I Trial of Monoclonal Antibody TGN1412: Better Understanding the Causes to Improve PreClinical Testing of Immunotherapeutics". In: *The Journal of Immunology* 179.5 (Sept. 1, 2007), pp. 3325–3331. ISSN: 0022-1767. DOI: 10.4049/jimmunol.179.5.3325. URL: https://doi.org/10.4049/jimmunol.179.5.3325 (visited on 08/17/2025).

[24] Katy Taylor and Laura Rego Alvarez. "An Estimate of the Number of Animals Used for Scientific Purposes Worldwide in 2015". In: *Alternatives to laboratory animals: ATLA* 47.5-6 (2019), pp. 196–213. ISSN: 0261-1929. DOI: 10.1177/0261192919899853. pmid: 32090616.

[25] *The 3Rs | NC3Rs*. URL: https://nc3rs.org.uk/who-we-are/3rs (visited on 08/20/2025).

[26] Margaret Waltz, Jill A. Fisher, and Rebecca L. Walker. "Mission Creep or Mission Lapse? Scientific Review in Research Oversight". In: *AJOB Empirical Bioethics* 14.1 (Jan. 2, 2023), pp. 38–49. ISSN: 2329-4515. DOI: 10.1080/23294515.2022.2123868. pmid: 36125845. URL: https://doi.org/10.1080/23294515.2022.2123868 (visited on 08/17/2025).

[27] Irving Zucker and Brian J. Prendergast. "Sex Differences in Pharmacokinetics Predict Adverse Drug Reactions in Women". In: *Biology of Sex Differences* 11.1 (June 5, 2020), p. 32. ISSN: 2042-6410. DOI: 10.1186/s13293-020-00308-5. URL: https://doi.org/10.1186/s13293-020-00308-5 (visited on 08/24/2025).

For the Lore of Science : The positive impacts of opening our labs to schools

I remember the first time we welcomed a group of students into our lab. They arrived, some wide-eyed and curious, others sullen and hesitant, but all with a mix of excitement and apprehension written on their faces. For most, it was their first time in a lab. But for some, it was more than that—it was their first glimpse into a world they never thought they could belong to.

As a PhD student at the Laboratoire Jean Perrin in Paris, I have organized lab visits for school students, particularly those from underprivileged backgrounds. Over 2023-2024, we've hosted more than 300 students, and the impact has been profound—not just for the students but for us, the scientists, as well. This project is now handled by others, and I hope it will survive a long long time, at the LJP and/or elsewhere.

One thing we've learned in organizing those visits is that context matters. Students who feel out of place or intimidated in a lab setting might never fully engage with the experience. That's why we decided to combine traditional lab visits with in-classroom visits. Before students even set foot in our lab, we go to their schools and meet them on their own turf. This small step makes the biggest of differences. In the classroom, we get to know the students, answer their questions, and introduce some basic concepts that will make their lab visit more meaningful. But mostly, the goal is to break down the barrier of formality. By the time they arrive at our lab, the students are more comfortable and curious, ready to dive into the world of science with open minds.

Our focus on underprivileged schools is driven by the desire to break down the socio-economic barriers and elitism too often observed in the scientific community. An example which strikes me is when labs from my university only invite students from elite high-schools from the center of Paris, because : "At least these students know what an electron is !".

In our lab, we specifically target schools where students might not have regular access to such experiences. We use the Indice de Position Sociale (Social Positioning Index) provided by the government to identify schools with students from lower socio-economic backgrounds.

These students, many from neighborhoods with limited resources, bring with them a hunger for knowledge and a fresh perspective. Their questions are often challenging, their curiosity endless, and that's exactly what we want.

What follows is an article I wrote in 2024 detailing the motivations and methodology I used to organize those visits. I hope it will be useful to other labs, to replicate and improve this type of visits program.

For the Lore of Science :
The positive impacts of opening our labs to schools

Mattéo Dommanget-Kott
matteo.dommanget-kott@sorbonne-universite.fr

Abstract

Science remains largely the domain of the socio-economically privileged, with social background often serving as an unspoken barrier to entry. In this article, we advocate for a combined approach to science outreach that integrates both in-school and lab visits to engage students from underprivileged backgrounds. We share our experience organising lab visits for over 300 students, highlighting the positive impacts on students, teachers, and lab members. Our goal is to help other scientists in implementing similar initiatives. Finally, we call for broader adoption of these practices and advocate for mentality and policy changes that support and reward outreach efforts within the scientific community.

Introduction

Science is a privilege.

It might be one of the greatest human endeavours, a collective journey of curiosity that has yielded an incredible corpus of knowledge. As scientists, we contribute to this ever-growing body of understanding, providing invaluable insights and methods that advance society, even if these contributions are sometimes underappreciated. We are privileged

to work in a field that allows us to pursue our passion, a luxury not afforded to many.

Yet, despite its profound benefits, science remains largely the domain of the privileged. Historically, scientists (or natural philosophers, as they were known back then) were often nobles, individuals with the time and resources to pursue knowledge for knowledge's sake without any expectation of material gain. While the landscape of science has evolved, it still reflects this legacy of privilege [1].

Today, the scientific community is marked by the under-representation of women, people of colour, LGBTQ+ individuals [13], those with disabilities (whether physical, mental, or psychological), and those from economically disadvantaged backgrounds [6, 12]. This disparity is evident not only in the demographics of who becomes a scientist, but is even more pronounced when considering who among them becomes *influential* or *successful* [11, 9, 10, 7]. However, diversity in science should not only be welcomed; it is urgently needed.

Indeed, it has been reported many times that scientists from underrepresented groups often focus on topics that are particularly relevant to their communities [9], bringing fresh perspectives and new areas of study into the scientific fold. Moreover, diverse teams have been consistently found to be more creative [7] and foster better working environments [13], qualities that are sorely needed in the scientific community.

While the community seems to have acknowledged the gender disparity and has begun addressing it [8, **?**, 5], social background remains almost a taboo subject. From a very young age, science is often perceived as an elite career. This perception only strengthens with age, especially for children from disadvantaged backgrounds who are frequently discouraged from pursuing scientific careers unless they are exceptionally brilliant. This divide is perpetuated not only by how we communicate science to the public but, more importantly, by whom we choose to engage with.

Science is still predominantly *consumed* by those who are privileged: individuals who have the time, the financial means to buy books or museum tickets, and the educational background to engage with complex topics. Thankfully, the accessibility of science has improved dramatically in recent decades. The rise of science communication through television, specialised journals, online platforms like Wikipedia, and especially social media, has given younger generations virtually free and unlimited access to scientific knowledge.

However, an unlimited accessibility does not mean that everyone will access equally. If we are to broaden the reach of science, we crucially need a more diverse array of scientific role models [5, 3]. And we need to actively choose to interact with disadvantaged populations.

Despite these advances, recent years have seen a growing defiance towards science and scientists. Our societies are increasingly polarised, with some individuals fervently supporting scientific findings, while others vehemently reject them. This polarisation has serious consequences. Climate scientists, sociologists, and other experts are often ignored by policymakers, despite the urgency of their work.

Opening up our research to public scrutiny and engagement could help bridge this divide. By being more transparent and accessible, we can foster a deeper public interest in science, offering a rich context to communicate not only our findings, but the processes behind them. We can demystify the inner workings of scientific inquiry, explain the structure of the scientific community, and promote critical thinking.

For us, as scientists, this openness can be deeply rewarding. It encourages us to embrace rather than fear criticism, using it to refine our methods and improve our research practices. Controversies such as those surrounding animal experimentation highlight the importance of this openness. Engaging with the public on these issues can lead to better understanding and potentially to more informed and compassionate policies.

In our view, science should be more open to the public. And what better way to achieve this, than by inviting people to visit our labs, to see for themselves the work that we do and to engage directly with the process of research?

1 The Importance of Lab Visits: Igniting curiosity and sparking a love for science

In the larger family of science outreach methodologies, lab visits hold a particular place, offering a tangible, immersive experience that can demystify the world of research, making it accessible and engaging. It is an invaluable opportunity to bridge the gap between the scientific community and the public.

Lab visits can address a multitude of challenges. They are not just about showcasing the work we do. They are about creating a connection with the public, fostering an interest and understanding of the scientific process, and encouraging young minds to consider careers in science. By focusing on school children, we can strategically target disadvantaged populations, which can hopefully create a positive feedback loop where some of those students could become the role models of the next generations. Pedagogically we provide the students with firsthand experience of how science is conducted, which can be far more effective than what they learn from textbooks or media.

There are many different lab visit formats, each with its unique advantages. From classical open-house to more immersive experiences where students can spend a day shadowing a researcher (for example the "Vie ma vie de chercheur·euse" program : https://www.vmvdc.cnrs.fr/). This kind of exposure helps to humanise the field, showing students that science is not an abstract concept but a dynamic, hands-on endeavour conducted by real people with whom they can relate.

In addition to traditional lab visits, in-school visits by scientists also offer significant benefits. These visits can bring science directly into the classroom, making it more accessible to students who might not have the opportunity to visit a lab. By interacting with scientists in their environment, students can ask questions and engage in discussions that might not happen in a more formal lab setting. This flexibility allows for a more personalised approach, catering to each group of student's specific interests and needs.

In this article, we advocate for a combined approach that integrates both in-school visits and open-house-style lab visits. We and others [4] have found that starting with classroom discussions allows us to tailor the lab experience to the specific interests and questions of the students. This step is particularly crucial for students from disadvantaged backgrounds, who often feel out of place in scientific environments, unlike their peers from higher socio-economic backgrounds. By first building a friendly rapport with these students in the familiar setting of their school, we can help them feel more comfortable and confident when they visit the lab. This creates a more inclusive environment that

fosters a deeper understanding of science and encourages a more diverse group of students to consider scientific careers.

2 Welcoming school visits: Our lab's approach and experience

We have focused our outreach to schools, particularly those serving students from non-privileged families. In this section, we present how we have organised lab visits for 300 students from 13 schools at our lab over 18 months since January 2023, what we have learned through this process, and how these visits have impacted students, teachers, and our lab. While we do not claim that our approach is the definitive method for organising lab visits, we hope sharing our experiences will provide valuable ideas for other labs and help scientists advocate for similar initiatives within their institutions.

2.1 Our lab

The Laboratoire Jean Perrin is a physics lab that operates at the intersection of biology and medicine. Our research focuses on studying complex biological systems and developing bio-inspired systems for fundamental research. The lab is relatively small, consisting of approximately 60 members, primarily physicists, but also including a few biologists, chemists, and technicians. Our work is predominantly experimental, although we maintain a strong theoretical component. The lab is located in Paris, France.

2.2 Creating the science workshops: Balancing engagement and education

The first step in creating our lab visits involved identifying the key workshops that would be presented to the visiting students. In our experience, organising workshops is a balancing act between the following constraints:

- **Foster Lab Participation:** Any lab member who wishes to volunteer should be encouraged to participate, regardless of their experience or preparedness. This fosters a collective enthusiasm that significantly increases the likelihood that the project will be viable and sustainable over time.

- **Minimise Workshop Preparation Efforts:** Participating in lab visits should be easy and require minimal time and effort from lab members. This encourages higher participation rates and repeated involvement. One effective approach is to build workshops around the ongoing scientific work of each participant. For instance, showcasing a current experiment with minimal modifications to make it accessible to the target audience.

- **Identify Common Themes:** By identifying one or two themes common to all workshops, we can create a coherent narrative that helps visitors feel less overwhelmed by the diversity of information. While not critical, this strategy has proven particularly helpful for older students (11 years and up), as it provides a unifying thread throughout their visit.

- **Hands-On Involvement:** Workshops should involve practical engagement for the visitors whenever possible. This can include activities like looking through a microscope, manipulating scientific equipment, or participating in discussions. Lecture-style presentations should be avoided, as they are less effective in capturing the students' interest.
- **The "Wow" Effect:** Workshops should aim to include something impressive that elicits a sense of wonder from the visitors, be it a striking visual, an intriguing fact, or a sophisticated tool. While not essential and sometimes difficult to find, we have found that a "wow" moment can be identified in many (if not all) scientific topics.
- **Safety:** Of course, ensuring the safety of the visitors is paramount. Workshops must be designed to be as safe as possible, considering both the physical environment and the activities involved.

The overarching goal is to balance the pedagogical effectiveness of each workshop with the enjoyment and minimal effort required from the participating lab members. In our experience, even workshops that are not meticulously planned or pedagogically optimised tend to be appreciated and beneficial for the visitors. Workshops can always be refined and improved for future visits based on experience and feedback. It is also important not to self-censor; ideas that seem too complex or unrefined can turn out to be highly engaging for students.

Through this process, we developed a list of workshops and their associated lab members. This list was then shared with schools and teachers, allowing them to choose the most relevant and interesting workshops for their students.

2.3 Choice of schools: Prioritising non-privileged students

Finding schools willing to visit the lab proved to be surprisingly easy, as there is a strong demand from teachers. In France, the government actively encourages teachers to explore scientific subjects "outside the walls of the classroom".

To connect with schools, we primarily used email, reaching out through a list provided by the university or by contacting schools directly. Additionally, teachers reached out to us, either through word of mouth or via web searches, as our visits are advertised on our lab's website.

When selecting schools for our visits, we focused on targeting students from non-privileged backgrounds. To achieve this, we used the *Indice de Position Sociale* (IPS, or Social Positioning Index), a publicly available metric that measures the socio-cultural background of students, provided by the French government for each school in the country. We established a hard threshold, declining visits from schools with an IPS greater than 130 (which represents approximately 10% of all schools in France; the national IPS mean is around 104, with a standard deviation of 16).

Given that Paris is, on average, more socio-culturally privileged than the rest of the country (with an IPS mean of approximately 123 and a standard deviation of 19), we declined visits from roughly 40% of the schools in Paris.

Furthermore, we prioritised schools that are part of the REP and REP+ programs (*Réseaux d'Éducation Prioritaire* and *Réseaux d'Éducation Prioritaire Renforcés*), which

are government-led initiatives focused on schools in areas facing significant socio-economic challenges.

Over the course of 2023 and 2024, we welcomed approximately 300 students from 13 different schools, with an average IPS of 98 (standard deviation of 12). This group included four schools from the REP and REP+ programs. The students ranged in age from 8 to 18 years old (with a mean age of 12 and a standard deviation of 3), and they travelled from various parts of the Greater Paris Region, with commute times ranging up to 1 hour and 20 minutes by public transport (with a mean travel time of around 40 minutes and a standard deviation of 20 minutes).

2.4 Creating lab visits: Our tailored approach

The organisation of each lab visit is carefully coordinated with the teachers to tailor the experience to their student's specific needs and interests.

After establishing initial contact, we schedule a detailed discussion with the teacher. During this meeting, we present the lab visit project, discuss the pedagogical goals and constraints of the teacher, and gain insights into the students' interests and backgrounds. This helps us select the most appropriate workshops for the visit. We encourage teachers to engage their students in this planning process, and most do so enthusiastically. Once the details are finalised, we set a date that accommodates both the teacher's and the lab members' schedules.

We strongly recommend conducting a pre-visit in the classroom, and so far, all teachers have agreed to this approach. Typically scheduled about a week before the lab visit, one or two scientists go to the school for approximately one hour. This pre-visit serves three primary purposes:

1. **Build a Relationship with the Students:** : We aim to establish a friendly rapport with the students, so they see the visiting scientist not just as an expert, but as a relatable adult. We have found that this personal connection is the most significant predictor of a successful lab visit.

2. **Ascertain student's interests:** Understanding the students' interests, aspirations, and any hesitations they may have toward science is crucial. This knowledge allows us to engage in more meaningful discussions and tailor the lab visit to address specific concerns and curiosities.

3. **Introduce Core Concepts:** We systematically introduce a few key scientific concepts that will be relevant during the lab visit. Typically, this involves explaining a common theme across the workshops. In our case, this is often the concept of emergent properties. To aid in this, we bring along a USB stick with images and videos, allowing us to illustrate these concepts in a flexible and engaging manner depending on the direction of the discussion.

The pre-visit discussion is largely driven by the students' questions. Many are curious about our jobs, academic backgrounds, and the specifics of our scientific work. Depending on their age, students often inquire about both scientific topics and the broader

organisation of the scientific community. Overall, we find that the pre-visit is highly effective in preparing students for the lab visit, regardless of their age, albeit in different ways.

On the day of the lab visit, we begin by welcoming all students in a single classroom-like setting for a brief 5-10 minute introduction. During this time, we gather feedback from the pre-visit, introduce the lab members who will be leading the workshops, and conduct a quick safety briefing.

Students are then divided into smaller groups that rotate through the various workshops. The optimal structure of the visit, including the duration, the number of workshops, and the size of the groups, depends on the students' age (see Table??) and is also influenced by the total number of students and the available space in the workshop rooms. Finding the optimal strategy can be challenging due to these constraints.

After the workshop rotations, we reconvene all the students in a single room for a 10-15 minute debriefing session. We systematically ask for feedback, starting with what the students did not like about the visit, followed by what they did like (we find that this order is important). At this stage, students often have many questions about what it's like to be a scientist. These questions are answered by the lab members, providing a diverse range of perspectives.

In the days following the visit, we send a questionnaire to the teacher, which is to be completed by the students. This questionnaire contains three questions:

- *What did I like about the visit ?*
- *What didn't I like about the visit ?*
- *What did I learn during the visit ?*

We also ask the teacher for a small debriefing. This feedback allows us to evaluate the impact of the visit on both teachers and students and to continuously improve our approach.

2.5 The direct impact of lab visits on students, teachers, and scientists

The lab visits have consistently been a success, receiving generally positive feedback from all involved. It is important to note that a *successful* visit is not necessary a *perfect* one, what matters is the engagement and learning that occurs during these visits.

The overall feedback from students has been positive. The majority express that they enjoyed the visits, and post-visit assessments indicate that they retain key facts and concepts even days after the experience. Many students also reported that the visits helped them gain a better understanding of what scientists do in their daily lives. Notably, a few students mentioned that the visit changed their perception of science, leading them to consider careers as scientists. We encountered instances where students initially thought they were "too dumb" or "too clumsy" to pursue scientific careers, but changed their minds after discovering the diversity of scientists and the variety of methods used in scientific research, such as computer simulations.

Student feedback is also a mirror of our own limitations and biases, with a few students remarking upon the lack of gender and ethnic diversity, as well as pertinent comments

on animal experimentation.

Feedback from teachers has been extremely positive as well. They have observed a noticeable increase in their students' interest in science following the visits, especially among students under 13 years old. Teachers report that the visits have made it easier for them to approach scientific lectures, with students showing less reluctance toward scientific subjects.

In France, at around 13 years old, students have to perform a professional internship. Usually, these internships are performed within the family's entourage, but a few teachers connected their students to us or the University.

Several teachers have also noted instances of students explicitly choosing scientific disciplines as a direct result of their experience in the lab.

The impact on the lab itself has also been largely positive. The vast majority of lab members who participated in the visits have expressed a willingness to continue. Over the first two years of organising these visits, the quality of the workshops has improved significantly, with notable advancements in how we communicate scientific concepts and engage with students. Participating in these visits has proven to be an excellent learning opportunity for lab members, enhancing their outreach and communication skills. We have also observed a growing interest in outreach activities among lab members, even among those who did not initially participate in the visits.

Overall, lab visits have proven to be a win-win situation for everyone involved. They create an enriching environment for learning and fostering meaningful interactions between students and the scientific community. We would like to emphasise that, in our experience, the pre-visit is a critical component of this interaction. While lab visits can certainly be successful without a pre-visit, we have found that this preliminary engagement significantly enhances the overall experience.

A call for action

Science outreach is crucial for modern societies, which increasingly depend on information and technology. However, a proper understanding of scientific concepts requires both interest and time—resources that are not equally available to all segments of the population. Access to scientific outreach remains largely confined to higher socio-economic classes, creating a gap in scientific literacy across different demographics.

At the same time, the scientific community is in dire need of greater diversity. We continue to see imbalances in the representation of genders, ethnicities, sexual orientations, disabilities, and socio-economic backgrounds. Addressing this issue requires the presence of diverse role models who can inspire and guide the next generation of scientists.

Despite its importance, science outreach is time-consuming, a resource already in short supply for the scientific community. However, we propose that lab visits offer an inexpensive and relatively low-effort solution to these challenges. In this article, we have shared how we have successfully implemented lab visits, specifically targeting school groups, and provided a framework for organising these visits. Our experience shows that, even with limited resources and experience, lab visits are highly effective in communicating scien-

tific knowledge and methods to students aged 8 to 18 years. These visits have proven to be rewarding not only for students and teachers but also for the lab members involved.

That being said, it is important to acknowledge that organising such visits, while cost-effective in terms of time and money, still requires a significant commitment from at least one dedicated lab member. In response to this need, France and other countries have introduced programs for PhD students that recognise scientific outreach as equivalent to teaching, complete with accompanying salary support. We encourage policymakers to extend this recognition to other types of scientists as well.

Moreover, a shift in mindset is needed within the scientific community. It is essential to recognise that science communication is just as important as teaching, particularly when it comes to grant applications and considerations for permanent positions. It is still considered as a care-oriented task of low importance, which might be why it is predominantly undertaken by woman [2]. Embracing this perspective will help to ensure that science outreach is valued and integrated into the broader missions of science.

References

References

[1] Is science only for the rich? 537(7621):466–470.

[2] Elaine Howard Ecklund, Sarah A. James, and Anne E. Lincoln. How Academic Biologists and Physicists View Science Outreach. 7(5):e36240.

[3] Lee Ferguson and Michael K. Seery. Role Models and Inspirations of LGBT+ Scientists. 99(1):444–451.

[4] Andrew J. Gall, Peter J. Vollbrecht, and Tristan Tobias. Developing outreach events that impact underrepresented students: Are we doing it right? 52(6):3499–3506.

[5] Susana González-Pérez, Ruth Mateos de Cabo, and Milagros Sáinz. Girls in STEM: Is It a Female Role-Model Thing? 11.

[6] Stephen Gorard and Beng Huat See. The impact of socio-economic status on participation and attainment in science. 45(1):93–129.

[7] Bas Hofstra, Vivek V. Kulkarni, Sebastian Munoz-Najar Galvez, Bryan He, Dan Jurafsky, and Daniel A. McFarland. The Diversity–Innovation Paradox in Science. 117(17):9284–9291.

[8] Paul Nnanyereugo Iwuanyanwu. Is science really for me? Gender differences in student attitudes toward science. 122(5):259–270.

[9] Diego Kozlowski, Vincent Larivière, Cassidy R. Sugimoto, and Thema Monroe-White. Intersectional inequalities in science. 119(2):e2113067119.

[10] Allison C. Morgan, Nicholas LaBerge, Daniel B. Larremore, Mirta Galesic, Jennie E. Brand, and Aaron Clauset. Socioeconomic roots of academic faculty. 6(12):1625–1633.

[11] John N. Parker, Christopher Lortie, and Stefano Allesina. Characterizing a scientific elite: The social characteristics of the most highly cited scientists in environmental science and ecology. 85(1):129–143.

[12] Catherine A. Rushworth, Regina S. Baucom, Benjamin K. Blackman, Maurine Neiman, Maria E. Orive, Arun Sethuraman, Jessica Ware, and Daniel R. Matute. Who are we now? A demographic assessment of three evolution societies. 75(2):208–218.

[13] Dario Sansone and Christopher S. Carpenter. Turing's children: Representation of sexual minorities in STEM. 15(11):e0241596.

Acknowledgments

We would like to acknowledge the great work performed by the many members of the Laboratoire Jean Perrin who participated in these lab visits. We would also like to thank the members of the *Direction des relations Sciences Culture Société* of Sorbonne University, particularly Thibaut Valette and Léa Sorli, who have provided useful advise and contacts. Finally, we would like to thank all the teachers with whom we organised those visits and the students who came.