ONGOING NEURONAL POPULATION ACTIVITY DYNAMICS IN THE NEOCORTEX - REPRESENTATIONAL DRIFT IN EXPERIMENT AND MODEL

DISSERTATION

zur Erlangung des Doktorgrades
der Naturwissenschaften

vorgelegt beim Fachbereich Physik
der Johann Wolfgang Goethe Universität
in Frankfurt am Main

von

JENS-BASTIAN EPPLER

aus

Frankfurt am Main

May 2022
Frankfurt am Main

(D30)

vom FACHBEREICH PHYSIK der
Johann Wolfgang Goethe Universität als Dissertation angenommen.

DEKAN:

Prof. Dr. Harald Appelshäuser
Institut für Kernphysik

GUTACHTER:

Prof. Dr. Matthias Kaschube
Frankfurt Institute for Advanced Studies

Prof. Dr. Jochen Triesch
Frankfurt Institute for Advanced Studies

DATUM DER DISPUTATION:

19 July 2022

Jens-Bastian Eppler: *Ongoing neuronal population activity dynamics in the neocortex - Representational drift in experiment and model* © May 2022

We are plastic and we are you
We're the future, we're the new
Nothing is new except that it's true

— Crash Tokio, 2006

PREFACE

CONTENTS

Preface v
ZUSAMMENFASSUNG xv
ABSTRACT xxi

I INTRODUCTION 1

II FUNDAMENTALS 9
1 BIOLOGICAL BACKGROUND 11
  1.1 Auditory pathway 11
  1.2 Auditory cortex 13
  1.3 Cell assemblies 15
  1.4 Drift 16
    1.4.1 Synaptic drift 16
    1.4.2 Representational drift 17
2 TWO-PHOTON IMAGING 19
3 MODELS IN NEUROSCIENCE 21
  3.1 Firing rate models 23
  3.2 Plasticity and drift 24
4 STABILITY AND FLEXIBILITY OF BRAIN FUNCTION 27

III ANALYSIS OF EXPERIMENTAL DATA FROM MOUSE AUDITORY CORTEX 29
5 ONGOING DYNAMICS OF SENSORY REPRESENTATIONS 31
  5.1 Introduction 32
  5.2 Results 32
    5.2.1 Chronic large-scale calcium imaging of neurons in the mouse auditory cortex 32
    5.2.2 Dynamic long-term remodeling of sound responses in individual neurons and populations of neurons 34
    5.2.3 A discrete set of response modes forms a population-level representation of the auditory world 41
    5.2.4 Ongoing recombination of sensory representations during basal conditions 45
    5.2.5 The impact of learning on the dynamics of sensory representations 47
    5.2.6 Fear conditioning drives the formation of associations between sensory representations and stabilizes them 52
    5.2.7 ACFC increases population responses for stimuli with representational similarity to the conditioned sound 53

5.2.8 Fear conditioning drives stimulus specific formation of associations between sensory representations, predictive of behavioral generalization 54
5.3 Discussion 58
5.4 Methods 60
5.4.1 Molecular cloning 60
5.4.2 rAAV production 60
5.4.3 Animal use 61
5.4.4 Stereotactic injection 61
5.4.5 Cranial window implantation 62
5.4.6 Habituation to awake chronic two-photon imaging 63
5.4.7 Sound presentation 63
5.4.8 Two-photon imaging 64
5.4.9 Auditory cued fear conditioning 64
5.4.10 Conditioning session 65
5.4.11 Memory test session 65
5.4.12 Quantitative analysis of freezing behavior . . . 66
5.4.13 Confocal imaging 66
5.4.14 Image processing of chronic two-photon data . 66
5.4.15 ROI inclusion criteria 67
5.4.16 Calculation of $\Delta F/F_0$ and deconvolution 68
5.4.17 Stimulus-evoked sound responsiveness of single cells 68
5.4.18 Sound response profiles of single cells 68
5.4.19 FOV inclusion criteria 68
5.4.20 Definition of response modes 69
5.4.21 Mode-associated responsiveness of single cells . 70
5.4.22 Sound decoding based on full response vectors 70
5.4.23 Sound decoding based on a 34-fold classifier . . 70
5.4.24 Dependence of decoding on number of modes and stimuli per mode 71
5.4.25 Statistics 72
6 A BASIS SET OF ELEMENTARY OPERATIONS 75
6.1 Introduction 75
6.2 Results 76
6.2.1 Transitions between response modes 76
6.2.2 A set of operations to describe response mode dynamics 80
6.2.3 Ongoing recombination of cell assemblies . . . 84
6.2.4 The impact of learning on cell assembly dynamics 87
6.3 Discussion 91
6.4 Methods 92
6.4.1 t-distributed Stochastic Neighbor Embedding . 93

IV A FIRING RATE MODEL OF POPULATION ACTIVITY IN MOUSE AUDITORY CORTEX 95
7 A MODEL OF CLUSTERING IN MOUSE AUDITORY CORTEX 97
  7.1 Introduction . . . 97
  7.2 Model setup . . . 99
    7.2.1 Construction of connectivity matrices . . . 99
    7.2.2 Construction of input . . . 101
    7.2.3 Clustered network responses . . . 102
    7.2.4 Normalizing the network's eigenvalue spectrum with respect to the network size . . . 104
  7.3 Results . . . 106
    7.3.1 A scan of parameter space reveals a clustering regime for strong inhibition and strong recurrent connections . . . 106
    7.3.2 Clusters are formed by weaker than average inhibitory synapses between participating neurons 116
  7.4 Discussion . . . 118
8 ABRUPT TRANSITIONS IN THE MODEL 121
  8.1 Introduction . . . 121
  8.2 Model setup . . . 122
    8.2.1 Modeling drift of excitatory synapses . . . 123
    8.2.2 Implementation of synaptic drift . . . 124
    8.2.3 Modeling drift of inhibitory synapses . . . 125
  8.3 Results . . . 126
    8.3.1 Drift of inhibitory synapses is predicted to be an order of magnitude slower than drift of excitatory synapses . . . 126
    8.3.2 Abrupt transitions of response patterns are caused by ongoing synaptic drift . . . 127
    8.3.3 The fixed point topology of a neural network . 131
    8.3.4 Finding fixed points numerically . . . 134
    8.3.5 Abrupt transitions coincide with qualitative changes of the fixed point topology . . . 136
  8.4 Discussion . . . 146

V DISCUSSION 149

VI APPENDIX 157
9 SUPPLEMENTARY FIGURES 159

BIBLIOGRAPHY 169
ACKNOWLEDGEMENTS 185
CURRICULUM VITAE 187

LIST OF FIGURES

Figure 5.1 Two-photon imaging 33
Figure 5.2 Example activity traces. 33
Figure 5.3 Spectrograms of stimuli 34
Figure 5.4 Preprocessing pipeline 35
Figure 5.5 Chronic imaging example cells 36
Figure 5.6 Balanced drifts in tuning 37
Figure 5.7 Balanced drifts statistics 38
Figure 5.8 Cell responsiveness 38
Figure 5.9 Dynamic population responses 39
Figure 5.10 Decoding analyses 40
Figure 5.11 Clustered population responses 41
Figure 5.12 Abrupt transitions delineate modes. 42
Figure 5.13 Response mode identification 43
Figure 5.14 Response mode statistics 44
Figure 5.15 Response mode correlations 44
Figure 5.16 Recombination of response modes 45
Figure 5.17 Response mode life times 46
Figure 5.18 Response mode flow diagram 47
Figure 5.19 Fear conditioning paradigm (ACFC) 48
Figure 5.20 ACFC increases decoding performance 49
Figure 5.21 A model of decoding 50
Figure 5.22 Improved decoding after ACFC 51
Figure 5.23 Response mode life times (ACFC) 51
Figure 5.24 Examples of associations 52
Figure 5.25 Increased associations through ACFC 53
Figure 5.26 Representational similarity 54
Figure 5.27 Examples of specific associations 55
Figure 5.28 Co-mapping with the CS+ 56
Figure 5.29 Conditioning session and control 57
Figure 6.1 Response mode changes across time 77
Figure 6.2 tSNE plots of response modes 78
Figure 6.3 Single day tSNE plots 80
Figure 6.4 Single stimulus tSNE plots 81
Figure 6.5 Operations schematic 82
Figure 6.6 Operations examples 85
Figure 6.7 Operations frequencies 86
Figure 6.8 Operations Venn diagram 86
Figure 6.9 Operations: mode number 88
Figure 6.10 Operations: mode size 88
Figure 6.11 Operations statistics 89
Figure 6.12 Operations: response modes 89

Figure 6.13 Operations: stimulus response 90
Figure 6.14 Operations: associations 90
Figure 6.15 Operations for sets schematic 92
Figure 7.1 Example connectivity matrix 101
Figure 7.2 Stimuli & responses 102
Figure 7.3 Response vectors & correlations 103
Figure 7.4 Example eigenvalue spectrum 104
Figure 7.5 Scalability of eigenvalues 105
Figure 7.6 Eigenvalues in parameter space 105
Figure 7.7 Parameter space: activity 106
Figure 7.8 Responses in parameter space 107
Figure 7.9 Correlation matrices in parameter space 108
Figure 7.10 Parameter space: overview 110
Figure 7.11 Comparison to experiment 112
Figure 7.12 Scalability of clustering 114
Figure 7.13 Distribution of cells in clustered responses . . 116
Figure 7.14 Distribution of synaptic connections 117
Figure 8.1 Connectivity correlation decay 125
Figure 8.2 Response correlation decay 126
Figure 8.3 Example response transitions 127
Figure 8.4 Example transitions: correlation 128
Figure 8.5 Correlation change distributions (model) . . . 129
Figure 8.6 Correlation change distributions (experiment) 130
Figure 8.7 Energy landscape schematic 132
Figure 8.8 Energy landscape changes schematic 134
Figure 8.9 Histogram of correlations (t vs. $t-1$) 137
Figure 8.10 Fixed point changes schematic 139
Figure 8.11 Histograms of non-moving FP 140
Figure 8.12 Ratio explained by non-moving FP 141
Figure 8.13 Ratio explained by *stable* FP 142
Figure 8.14 Ratio explained by *unstable* FP 143
Figure 8.15 Bar plots of ratio for non-moving FP 144
Figure 8.16 Bar plots of ratio for *stable* FP 144
Figure 8.17 Bar plots of ratio for *unstable* FP 145

LIST OF SUPPLEMENTARY FIGURES

Sup. Fig. 9.1 Additional dynamic population responses . . 159
Sup. Fig. 9.2 Additional response mode dynamics 160
Sup. Fig. 9.3 Additional examples of specific associations . 161
Sup. Fig. 9.4 Correlation across operations 162
Sup. Fig. 9.5 Correlations pre & post operations 163
Sup. Fig. 9.6 Mode sizes pre & post operations 164
Sup. Fig. 9.7 Additional operation statistics 165
Sup. Fig. 9.8 Complete set of stimuli 166
Sup. Fig. 9.9 Responses to complete set of stimuli 166
Sup. Fig. 9.10 Additional example response transitions . . . 167
Sup. Fig. 9.11 Histograms of *stable* FP 167
Sup. Fig. 9.12 Histograms of *unstable* FP 168

ACRONYMS

AAV	Adeno-Associated Virus
ACFC	Auditory Cued Fear Conditioning
ANOVA	Analysis of Variance
AP	Action Potential
CS+	Conditioned Stimulus
DAPI	4′,6-Diamidino-2-Phenylindole
FOV	Field of View
FP	Fixed Point
GABA	Gamma-Aminouutyric Acid
LASER	Light Amplification by Stimulated Emission of Radiation
MGN	Medial Geniculate Nucleus
NG	Neurogliaform
NND	Nearest Neighbour Distance
non-CS+	Non-Conditioned Stimulus
NSSI	Normalized Soma Signal Intensity
OFV	Objective Function Value
PDF	Probability Density Function
PV	Parvalbumin
rAAV	recombinant Adeno-Associated Virus
ROI	Region of Interest
SD	Standard Deviation
SEM	Standard Error of the Mean
SMP	Significant Motion Pixel
SPL	Sound Pressure Level
SSN	Soma Signal to Noise ratio
SST	Somatostatin

tSNE	t-distributed Stochastic Neighbor Embedding
US	Unconditioned Stimulus
VIP	Vasoactive Intestinal Peptide

ZUSAMMENFASSUNG

Die erfolgreiche Navigation in einer komplexen Umgebung erfordert eine zuverlässige Repräsentation derselben, um aus den gegebenen Informationen die richtigen Schlüsse für das Verhalten ziehen zu können. Allerdings ist es für Lebewesen aller Art auch wichtig sich schnell an Veränderungen der Umwelt anpassen und flexibel auf diese reagieren zu können. Das Gehirn muss also in der Lage sein Informationen stabil über lange Zeiträume zu speichern und gleichzeitig wichtige neue Informationen schnell zu integrieren.

Sensorische Reize werden im Gehirn als Aktivitätsmuster spezifischer Gruppen von Neuronen dargestellt. Über deren langfristige Dynamik ist wenig bekannt. Das vorherrschende Bild ist, dass das Gehirn diese Aufgabe auf eine ähnliche Art und Weise erfüllt wie Computer: wichtige Informationen werden fest eingespeichert und daran ändert sich nur etwas durch das Lernen neuer Informationen. Da davon ausgegangen wird, dass Informationen im Gehirn in den synaptischen Verbindungen zwischen Neuronen gespeichert werden, wären synaptische Verbindungen und damit auch die Aktivitätsmuster der Neuronen (ohne äußere Einflüsse) stabil, d.h. derselbe Stimulus würde stets dieselbe Aktivität als Antwort hervorrufen. Aktuelle Studien haben in den letzten Jahren gezeigt, dass dies nicht unbedingt der Fall zu sein scheint. Es gibt immer mehr Evidenz dafür, dass sich die synaptische Verschaltung biologischer neuronaler Netze konstant verändert, und das insbesondere auch ohne äußere Einflüsse. Diese Erkenntnisse sprechen dafür, dass sich neuronale Netze im Gehirn in einem intrinsisch dynamischen Zustand befinden. Zudem konnte gezeigt werden, dass auch neuronale Aktivitätsmuster in vielen cortikalen (und nicht-cortikalen) Arealen nicht so stabil sind, wie angenommen wurde, und der immer gleiche Stimulus, der zu einem Zeitpunkt stabil immer dieselbe Gruppe von Neuronen aktiviert, über die Zeit andere Aktivitätsmuster hervorruft. Während diese Änderungen innerhalb einzelner Messreihen nicht dokumentiert sind, die Aktivität also über Minuten bis Stunden stabil zu sein scheint, kann es vorkommen, dass sich Aktivitätsmuster von einem Tag zum anderen verändern.

Hier stellen sich uns die folgenden Fragen: Wie stabil sind diese Aktivitätsmuster auf der Ebene von neuronalen Populationen – bestehend aus mehreren hundert Neuronen – tatsächlich, wenn es keine äußeren Einflüsse gibt? Welchen Einfluss hat Lernen auf diese Dynamik? Und wie wirkt sich Veränderung der synaptischen Verschal-

tung eines Netzwerks generell auf die Aktivität des Netzwerks aus? Wir untersuchen diese Fragen im Zusammenspiel der Analyse experimenteller Daten und dem theoretischen Modell eines neuronalen Netzwerks.

Die verwendeten experimentellen Daten stammen aus chronischen 2-Photon-Mikroskopie-Experimenten unserer Kollaboratoren aus Mainz, aufgenommen im auditorischen Cortex (zuständig für die Verarbeitung von Tönen im Gehirn) wacher Mäuse. Hierbei wird zunächst mithilfe eines Virus ein Protein in den Neuronen exprimiert, das bei Aktivität der Neuronen fluoresziert. Diese Fluoreszenz kann dann – angeregt von einem LASER – durch ein Fenster im Schädel der Maus ausgelesen werden, während der Maus Töne vorgespielt werden. Der auditorische Cortex eignet sich gleich aus mehreren Gründen für die Untersuchung der Dynamiken von Aktivitätsmustern. Zum Einen erlaubt er eine sehr kontrollierte Präsentation der Stimuli, da diese definitiv wahrgenommen werden, anders als z.B. visuelle Stimuli, bei denen auch noch die Blickrichtung beachtet werden muss. Zum Anderen ist das auditorische System in der Maus als Fluchttier stark ausgeprägt und der auditorische Cortex nimmt einen relativ großen Teil des gesamten Neocortex ein.

Außerdem konnte in früheren Studien gezeigt werden, dass in einer lokalen Population mehrere unterschiedliche Stimuli häufig identische Aktivitätsmuster hervorrufen. Das limitiert die Anzahl der möglichen Aktivitätsmuster und reduziert die Komplexität, da nicht mehr komplexe hochdimensionale Muster betrachtet werden müssen, sondern lediglich jedem Stimulus aus einer handvoll Möglichkeiten ein Muster zugeordnet werden kann. Dass unterschiedliche Stimuli identische Antworten hervorrufen, scheint ersteinmal kontraintuitiv. Da diese Gruppierung der Stimuli aber in unterschiedlichen lokalen Populationen in ein und demselben Tier stets unterschiedlich ist, ergibt sich auf Ebene des kompletten auditorischen Cortex ein eindeutiges globales Aktivitätsmuster für jeden Stimulus. Diese lokalen Antwortmuster eignen sich zur quantitativen Analyse der Änderungen, da wir einfach klassifizieren können, ob ein Stimulus an einem Tag ein Muster A oder ein Muster B oder gar keine lokale Antwort hervorruft.

Unter Zuhilfenahme dieser nahezu diskreten lokalen Aktivitätsmuster sehen wir, dass sich diese im auditorischen Cortex von Mäusen über mehrere Tage hinweg – selbst unter stabilen äußeren Bedingungen – kontinuierlich neu zusammensetzen. Über einen Zeitraum von zwei Tagen evozieren lediglich 50% der Stimuli noch dieselbe Antwort, während 20% eine andere Antwort hervorrufen, und 30% der Stimuli, die eine Antwort hervorgerufen haben, diese verlieren. Dieser letzte

Teil wird ausgeglichen durch eine ungefähr ebenso große Anzahl von Stimuli, die zunächst keine Antwort evoziert haben, dies aber nun tun. Das bedeutet, dass über die Tage hinweg immer neue Neuronen in die Verabreitung von Sinneseindrücken involviert werden. Das System befindet sich aber in einem dynamischen Gleichgewichtszustand. Sowohl die Anzahl der aktiven Neuronen, als auch die Anzahl der Stimuli, die eine Antwort hervorrufen, und auch die Möglichkeit der Vorhersage des Stimulus aus den Aktivitätsmustern bleiben konstant. Während sich also die einzelnen Komponenten des Systems verändern, bleiben diese Größen erhalten.

Um zu untersuchen, welchen Einfluss Lernen auf diese intrinsische Dynamik hat, untersuchen wir experimentelle Daten von Mäusen, die klassischer Konditionierung unterzogen werden. Bei dieser gängigen experimentellen Methode lernen die Tiere einen der Stimuli mit einem elektrischen Schock zu verbinden. Der Lernprozess wird aus dem Verhalten der Mäuse ersichtlich, wenn dieser Stimulus nach der Konditionierung ohne Schock gespielt wird: Die Mäuse erstarren. Auf cortikaler Ebene sehen wir, dass nach der Konditionierung vermehrt Stimuli eine Antwort evozieren, die zuvor keine evoziert haben. Dieser Effekt betrifft insbesondere Stimuli, die von der Maus als ähnlich zum konditionierten Stimulus wahrgenommen werden. Außerdem aktivieren diese Stimuli häufiger dasselbe Aktivitätsmuster, das auch vom konditionierten Stimulus aktiviert wird. Die Zuordnung zweier Stimuli auf ein einziges Aktivitätsmuster wird häufig als kognitive Verknüpfung dieser Stimuli verstanden. Wir sehen also einen Anstieg der Verknüpfungen nicht-konditionierter Stimuli mit dem konditionierten Stimulus, und zwar abhängig davon, wie sehr die nicht-konditionierten Stimuli dem konditionierten Stimulus ähneln. Diese Beobachtung der erhöhten Verknüpfungen im Cortex ist außerdem prediktiv für eine Generalisierung im Verhalten der Mäuse. So zeigen die Tiere nicht nur für den konditionierten Stimulus ein Erstarren, sondern auch für ähnliche Stimuli – nicht jedoch für Stimuli, die dem konditionierten Stimulus nicht ähneln.

Zur detaillierteren Beschreibung dieser Dynamiken definieren wir zehn grundlegende Operationen, die alle möglichen Übergänge von Antworten auf Stimuli umfassen. Diese Übergänge beinhalten eine Neuzuordnung von Stimuli auf andere Aktivitätsmuster, sowie das Entstehen neuer Aktivitätsmuster oder das Verschwinden von Aktivitätsmustern. Mithilfe dieser Operationen können wir die Dynamik weiter aufschlüsseln und finden, dass während und nach der Konditionierung vermehrt Operationen gefunden werden, die Stimuli miteinander verknüpfen, sowie weniger Operationen, die Verknüpfungen zwischen Stimuli aufbrechen. Das lässt sich interpretieren als eine erhöhte Bildung von Assoziationen und eine Stabilisierung bereits existierender Assoziationen.

Zusammengefasst zeigt die Analyse experimenteller Daten, dass Aktivitätsmuster im auditorischen Cortex von Mäusen kontinuierlicher Veränderung unterliegen. Diese kontinuierliche Veränderung wird beeinflusst durch klassische Konditionierung hin zu einer Verknüpfung ähnlicher Stimuli und einer erhöhten Bildung von Assoziationen.

Um zu untersuchen, auf welche Art und Weise Veränderungen der synaptischen Verbindungen zu Veränderungen der Aktivität eines neuronalen Netzes führen, implementieren wir ein Feuerratenmodell aus excitatorischen und inhibitorischen Neuronen. Um möglichst wenig Annahmen treffen zu müssen, sind sowohl die Verbindungsmatrix als auch die Stimuli zufällig. Wir variieren die Stärke der rekurrenten Verbindungen und die Stärke der Inhibition im Vergleich zur Excitation. Neben einem unistabilen Input-dominierten Regime für schwache rekurrente Verbindungen, in welchem jeder Stimulus eine Antwort unabhängig vom Netzwerk evoziert, und einem unistabilen Netzwerk-dominierten Regime für starke rekurrente Verbindungen und schwache Inhibition, in welchem jedes Netzwerk eine Antwort unabhängig vom Stimulus generiert, finden wir ein multistabiles Regime für starke rekurrente Verbindungen und starke Inhibition. In diesem Regime ist das Modell in der Lage Eigenschaften der Populationsaktivität im auditorischen Cortex der Maus zu reproduzieren, wie z.B. spärliche Aktivität, eine breite Verteilung der Feuerraten und auch die typische Zuordnung mehrerer Stimuli auf eines von wenigen möglichen Aktivitätsmustern.

In diesem Modell einer lokalen Population im auditorischen Cortex untersuchen wir den Effekt synaptischer Veränderung auf die Aktivität des Netzwerks. Wir verändern die synaptischen Verbindungen zufällig mit einer multiplikativen Version eines Ornstein-Uhlenbeck-Prozesses, der an experimentelle Daten angepasst wurde und so die log-normale Verteilung der Synapsenstärken in einem Gleichgewichtszustand belässt. Dieser kontinuierliche Umbau der Synapsen führt zu Perioden mit stabilen Aktivitätsmustern als Antwort auf Stimuli, die immer wieder unterbrochen werden von abrupten Übergängen zu neuen Aktivitätsmustern. Um diese abrupten übergänge besser zu verstehen, untersuchen wir die Fixpunkte des Systems. Da das Modell im multistabilen Regime operiert, haben wir stets mehrere *stabile* und *instabile* Fixpunkte. Für einen gegebenen Stimulus bewegt sich die Trajektorie der Netzwerkaktivität typischerweise entlang mehrerer instabiler Fixpunkte, bis sie in einem stabilen Fixpunkt konvergiert. Eine Veränderung des Netzwerks könnte auf zwei Arten zu abrupten Übergängen führen. Zum Einen könnten leichte Verschiebungen, also quantitative Veränderungen, z.B. eines instabilen Fixpunkts zu einer Umleitung der Trajektorie führen, die dann in einem anderen stabilen Fixpunkt endet, der aber auch schon vorher existiert hat. Zum anderen könnte sich die Fixpunktlandschaft qualitativ ändern, also Fixpunkte verschwinden oder entstehen, was zu einer Umleitung der Trajektorie auf diese neuen Fixpunkte führen könnte.

Wir verwenden eine numerische Methode um die Fixpunkte des Systems zu finden, indem wir das Quadrat der ersten Ableitung der Energielandschaft minimieren. Dadurch wird jeder Fixpunkt, an dem die Ableitung per Definition 0 ist, zu einem Minimum der Funktion und gängige Optimierungsalgorithmen können verwendet werden. Wir klassifizieren die Fixpunkte in stabile und instabile und untersuchen Veränderungen der Fixpunktlandschaft von einem Zeitpunkt zum nächsten. Fixpunkte können identisch bleiben, sie können sich verschieben oder sie können verschwinden, bzw. neue können entstehen. Wir finden, dass abrupte Übergänge häufig zusammenfallen mit stabilen Fixpunkten, die in der Nähe bereits existierender Fixpunkte entstehen oder verschwinden und mit instabilen Fixpunkten die entstehen oder verschwinden – unabhängig von der Distanz dieser zu bereits existierenden Fixpunkten. Neu entstehende oder verschwindende stabile Fixpunkte in der Nähe bereits existierender Fixpunkte führen zu einer Umleitung der Trajektorie auf diese, wohingegen neu entstehende oder verschwindende instabile Fixpunkte zu einer Umleitung früher in der Trajektorie führen. Beides sind qualitative Änderungen der Fixpunktlandschaft und beides führt zu einem abrupten Übergang der Aktivitätsmuster.

Zusammenfassend können wir in einem Feuerratenmodell Eigenschaften des auditorischen Cortex reproduzieren und beobachten, dass kontinuierliche Veränderung der synaptischen Verbindungen zu Perioden stabiler Aktivität führt – unterbrochen von abrupten Übergängen von einem Zeitpunkt zum nächsten, die einhergehen mit qualitativen und nicht qunatitativen Veränderungen der Fixpunkte des Systems.

Insgesamt sehen wir, dass neuronale Aktivitätsmuster im auditorischen Cortex weniger stabil sind als angenommen. Es ist nach wie vor unklar, wie aus diesen instabilen Repräsentationen wahrgenommene Stabilität entstehen kann. Homöostatische Mechanismen könnten eine Rolle spielen. Auch unklar ist, ob Stabilität überhaupt unbedingt erstrebenswert ist, oder ob das Gehirn vielmehr von einer höheren Flexibilität profitiert. Vorteile könnten sein die verbesserte Fähigkeit zu verallgemeinern oder die Vermeidung von Overfitting. Demnach wäre das Resultat der Evolution nicht ein Organ zuständig für maximal gute Erinnerung, sondern vielmehr ein Organ, das in der Lage ist, sich optimal an immer neue Situationen anzupassen.

Lernen, insbesondere klassische Konditionierung, beeinflusst diese Dynamik hin zu einer differentiellen Generalisierung, sodass Stimuli, die dem konditionierten Stimulus ähneln, ähnliche cortikale Aktivität und ähnliches Verhalten hervorrufen. Eine solche Generalisierung kann auch bei Konditionen wie z.B. post-traumatischer Belastungsstörung beobachtet werden. Wir beschreiben einen möglichen Mechanismus auf der Ebene lokaler neuronaler Populationsaktivitäten.

Ein Feuerratenmodell ist in der Lage Aktivitätsstatistiken des auditorischen Cortex zu reproduzieren in einem Regime, das dominiert wird von starken rekurrenten Verbindungen und einer noch stärkeren Inhibition. Interessanterweise wird gemeinhin angenommen, dass auch der auditorische Cortex im Vergleich zu anderen sensorischen Arealen inhibitorisch geprägt ist, was z.B. an der relativ schwachen und spärlichen Aktivität ersichtlich ist.

Zufällige kontinuierliche Veränderungen der synaptischen Verbindungen im Modell führen zu Phasen mit wenig Veränderung der Aktivitätsmuster und abrupten Übergängen dazwischen. Das Auftreten dieser komplexen Dynamik in einem solch simplen System lässt darauf schließen, dass ähnliche Dynamiken wahrscheinlich auch im Cortex zu finden sind. Vielleicht sind sie das zugrundeliegende Prinzip einer Art von Lernen, bei der wir Lernstoff wiederholen und kaum merklich vorankommen, bis es zu einem Moment der Erkenntnis kommt und wir plötzlich etwas verstehen, was wir vorher nicht verstanden haben.

ABSTRACT

Navigating a complex environment is assumed to require stable cortical representations of environmental stimuli. Previous experimental studies, however, show substantial ongoing remodeling at the level of synaptic connections, even under behaviorally and environmentally stable conditions. It remains unclear, how these changes affect sensory representations on the level of neuronal populations during basal conditions and how learning influences these dynamics.

Our approach is a joint effort between the analysis of experimental data and theory. We analyze chronic neuronal population activity data – acquired by out collaborators in Mainz – to describe population activity dynamics during basal dynamics and during learning (fear conditioning). The data analysis is complemented by the analysis of a circuit model investigating the link between a neural network's activity and changes in its underlying structure.

Using chronic two-photon imaging data recorded in awake mouse auditory cortex, we reproduce previous findings that responses of neuronal populations to short complex sounds typically cluster into a near discrete set of possible responses. This means that different stimuli evoke basically the same response and are thus grouped together into one of a small set of possible *response modes*. The near discrete set of response modes can be utilized as a sensitive and robust means to detect and track changes in population activity over time. Doing so we find that sound representations are subject to a significant ongoing remodeling across the timespan of days under basal conditions. The mapping of sound stimuli onto response modes can undergo changes, while at the same time some repsonse modes disappear or new response modes emerge. Auditory cued fear conditioning introduces a bias into these ongoing dynamics, resulting in a differential generalization both on the level of neuronal populations and on the behavioral level. This means that sounds that are perceived similar to the conditioned stimulus (CS+) show an increased co-mapping to the same response mode the CS+ is mapped to. This differential generalization is also observed in animal behavior, where sounds similar to the CS+ result in the same freezing behavior as the CS+, whereas dissimilar sounds do not.

Further insight into these dynamics is gained by identifying a set of ten operations capturing all possible transitions the response to a stimulus can undergo between imaging sessions. These operations

capture the remapping of stimuli between persistent response modes as well as disappearing or emerging response modes. Deploying this framework, we can further dissect the effects of auditory cued fear conditioning. We observe an increase in operations mapping stimuli onto the same response mode accompanied by a decrease in operations separating stimuli that were previously evoking the same response mode. This can be interpreted as an increase in the formation of associations, as well as a stabilization of existing associations. Together with the differential generalization our observations could provide a potential mechanism of stimulus generalization, which is one of the most common phenomena associated with post-traumatic stress disorder, on the level of neuronal populations.

To investigate how the aforementioned changes in neuronal population activity are linked to changes in the underlying synaptic connectivity, we devised a circuit model of excitatory and inhibitory neurons. We studied this firing rate model to investigate the effect of gradual changes in the network's connectivity on its activity. Apart from an input dominated uni-stable regime (one response per stimulus independent of the network) and a network dominated uni-stable regime (one response per network independent of the stimulus), we also find a multi-stable regime for strong recurrent connectivity and a high ratio of inhibition to excitation. In this regime the model reproduces properties of neural population activity in mouse auditory cortex, including sparse activity, a broad distribution of firing rates, and clustering of stimuli into a near discrete set of response modes. This clustering in the multi-stable regime means that, not only can identical stimuli evoke different responses, depending on the network's initial condition, but different stimuli can also evoke the same response.

Applying gradual drift to the network connectivity we find periods of stable responses, interrupted by abrupt transitions altering the stimulus response mapping. We study the mechanism underlying these transitions by analyzing changes in the fixed points of this network model, employing a method to numerically find all the fixed points of the system. We find that such abrupt transitions typically cannot be explained by the mere displacement of existing fixed points, but involve qualitative changes in the fixed point structure in the vicinity of the response trajectory. We conclude that gradual synaptic drift can lead to abrupt transitions in stimulus responses and that qualitative changes in the network's fixed point topology underlie such transitions.

In summary we find that cortical networks display ongoing representational drift under basal conditions that is biased towards a differential generalization during fear conditioning. A circuit model is able to reproduce key characteristics of auditory cortex, including a clustering of stimulus responses into a near discrete set of response modes. Implementing synaptic drift into this model leads to periods of stable responses interrupted by abrupt transitions towards new responses.

The observed instability of representations in auditory cortex is in contrast to our perceived stability of the environment. It remains a matter of debate where this stability arises or what this instability might be useful for. Advantages could include an increased ability to generalize or a prevention of overfitting.

Abrupt representational transitions were found in a firing rate model subject to synaptic drift. The occurence of such complex dynamics in relatively simple systems let us conclude that something similar might also be found in cortical systems. We speculate that abrupt transitions might underlying cognitive processes as sudden insights or even creativity.

Part I

INTRODUCTION

Sensory-evoked activity patterns at the level of sensory cortices are believed to serve as neural correlate of a percept. In light of our daily experience that the perception of the world around us is stable, i.e. that the same sensory stimulus evokes the same percept from day to day, sensory representations in the brain have been thought to be stable, too. This is in stark contrast, however, with recent findings showing that neuronal tuning to sensory stimuli is subject to ongoing remodeling even under stable conditions. This remodeling of functional properties of neurons has been reported in different cortical and non-cortical areas of the mouse brain – areas as diverse as hippocampus (Mankin et al., 2012, Ziv et al., 2013, Clopath et al., 2017, Hainmueller and Bartos, 2018), barrel cortex (Margolis et al., 2012), visual cortex (Deitch et al., 2021), motor cortex (Rokni et al., 2007, Huber et al., 2012, Clopath et al., 2017), and posterior parietal cortex (Driscoll et al., 2017, Rule et al., 2019). It remains unclear, however, how these changes to the function of individual neurons affect functional properties of sensory representations at the level of neuronal populations and how such ongoing changes are associated to to those expected to occur during learning. This gap in knowledge is due to both the shortage of data and the difficulty to establish appropriate frameworks to describe complex dynamics of neuronal populations.

The increasing body of evidence for changing neuronal representations, even in the absence of an apparent learning paradigm – also called *representational drift* – is in line with recent findings about the synaptic structure of the brain. Synapses have been shown to be plastic during learning, in the absence of learning and even in the absence of neuronal activity (Yasumatsu et al., 2008, Loewenstein et al., 2011, Rubinski and Ziv, 2015, Dvorkin and Ziv, 2016, Villa et al., 2016). Synapses appear to be in a constant state of flux. While learning induced plasticity is known since Hebb (1949), recent studies highlight seemingly random structural changes, often termed *synaptic drift*.

This drift – both on the level of synapses and on the level of neuronal representations – raises the question, at what level this perceived stability does arise. Functional stability (in our case seemingly stable perception) despite structural changes (synaptic and representational drift) is nothing special and seen in many fields, be it ecological sciences, where ecosystems (like the rain forest) are stable although single components (e.g. individual animals) turn over at a high rate compared to the entire system, social sciences, where a society continues to function even though individuals performing certain functions (e.g officials, politicians) fluctuate, or biology, where the proteins making up an organism have half life times of minutes to hours, but organisms maintain their function for years. This seems to be analogous to the brain and many theoretical publications have focused on maintaining functional stability despite structural remodeling (e.g. Vogels et al., 2011, Litwin-Kumar and Doiron, 2014, Mongillo et al., 2017, Fauth and Rossum, 2019), mostly suggesting homeostatic plasticity mechanisms to maintain stability.

Mouse auditory cortex is a well suited system for the study of longevity of sensory representations. Previous work has shown that at the scale of local neuronal populations sound stimuli lead to the non-linear activation of groups of neurons (Bathellier et al., 2012, Atencio and Schreiner, 2013, See et al., 2018). A specific local group of neurons is typically activated by different sounds and thus different stimuli evoke near identical response patterns. We called these stereotypical response patterns, evoked by a subset of stimuli in a local population of imaged neurons a *response mode* (Bathellier et al., 2012). The combination of different stimuli that are mapped to a given response mode locally varies across auditory cortex, resulting in distinct sound representations at the scale of the entire auditory cortex through a combinatorial pattern of multiple local response modes. This global description of neuronal population activity based on local response modes is sufficient to predict spontaneous categorization behavior in mice trained to discriminate pairs of sound stimuli (Bathellier et al., 2012). The discrete nature of response modes at the level of neuronal populations provides a sensitive and robust readout for representational changes occurring over the time course of days and allows for a systematic analysis of their dynamics.

We utilize these response modes to define a set of operations to systematically assess and deconstruct representational drift into its basic constituents. With a discrete mapping of stimulus responses onto response modes, we are able to identify and classify changes, like the re-mapping of a stimulus from one response mode to another or disappearing and emerging response modes. A complete set of unique operations helps us to better describe ongoing representational drift during basal conditions (i.e. without any explicit learning paradigm) and during learning.

Cell assemblies were first postulated by Hebb (1949) as a group of neurons responsible for a certain task (e.g. a movement in motor cortex, a sensory impression in sensory cortex). Many interesting aspects of cell assemblies – like associative memory (Hopfield, 1982), multistability (e.g. Stern et al., 2014), or stability towards synaptic drift (e.g. Kossio et al., 2021) – have been investigated in theoretical studies. In experimental work cell assemblies are harder to grasp, however, due to their distributed nature. In auditory cortex the above mentioned combinatorial code suggests cell assemblies that are distributed across the entire auditory cortex. As experimental data is limited to the simultaneous recording of several hundred cells in a single field of view (FOV), we find it difficult to call these local population response patterns cell assemblies and instead talk about response modes. Response modes are thus only a part of cell assemblies and might not capture their functional relevance in its entirety, they might, however, be of importance to putative readout neurons, which would only ever receive input from a subset of all neurons in auditory cortex.

To gain a mechanistic understanding of response modes we implemented a firing rate model to model neuronal population responses in auditory cortex. Firing rate models of neuronal networks were initially used to model the behavior of entire populations of neurons (Wilson and Cowan, 1972, Wilson and Cowan, 1973). One excitatory and one inhibitory population of neurons were assumed to consist of largely random connections within, but precise connections between and could thus be described by two coupled non-linear differential equations of population activity. While initially proposed as populations of neurons the units of this model can also be interpreted as single neurons – assuming exact timing of spikes is negligible. Regardless of the interpretation as populations or single neurons, already these very simple models show very complex behavior like hysteresis (Wilson and Cowan, 1972), different dynamic regimes (Wilson and Cowan, 1973), bifurcations (Borisyuk and Kirillov, 1992, Beer, 1995), chaotic behavior (Pasemann, 2002), spontaneous symmetry breaking (Fasoli et al., 2016).

In larger firing rate networks we typically find three dynamic regimes:

(a) a uni-stable regime, where the system has one attractor state, it converges to,

(b) a multi-stable regime, where the system has multiple stable attractor states, and

(c) a chaotic regime.

The system can typically switch between dynamic regimes by changing parameters such as synaptic gain (e.g. Wilson and Cowan, 1972), recurrent connection strength (e.g. Ostojic, 2014), the ratio of inhibition to excitation (e.g. Zhang and Saggar, 2020). These different dynamic regimes and the complexity of observed behaviors make firing rate models well suited to find a regime able to reproduce key characteristics of neuronal representations in mouse auditory cortex, most importantly the observed clustering of stimulus responses into response modes.

The firing rate model of auditory cortex is a tool to simultaneously investigate both synaptic and representational drift. Due to no available experimental method up to now the simultaneous recording of a biological network's synapses and its neuronal activity is impossible. It is, however, possible to study them in a model, as both can be read out. Synaptic drift can be implemented in different ways, ideally leaving the synapse size distribution in a steady state via a random process fit to experimental recordings (e.g. Loewenstein et al., 2011). At the same time the network's responses to stimulation can be recorded and analyzed. In the recent past a lot of studies have focused on, how representational stability can be achieved despite synaptic changes (e.g. Vogels et al., 2011), but not so much on the link between synaptic and representational drift. Instead of leaving representations stable, ongoing synaptic drift can lead to changes in neuronal activity patterns.

The dynamics of representational drift in this firing rate model subject to synaptic drift can be quantified for static stimuli and linked to changes in the underlying network connectivity via an analysis of the energy landscape of the system. This energy landscape is shaped by the connectivity matrix and determines the representational dynamics in a way, that the stimulus trajectory is guided through this landscape into a well of attraction, where the activity converges to a local minimum. The energy landscape can be described to a large extent by knowledge of the system's fixed points. Apart from multiple stable fixed points – local minima of the energy landscape – dynamics are also shaped by unstable fixed points, i.e. local maxima and (high-dimensional) saddle points between local minima. As the network connectivity is drifting the energy landscape is also changed and different phenomena could lead to response changes: stable fixed points could slowly move leading to slowly drifting network representations or unstable fixed points could move slowly, which could have no effect on the response or cause an abrupt change, whenever the local maximum between two local minima is shifted in a way that leads to a rerouting of the activity trajectory from one of the minima to the other. Of course qualitative changes, i.e. emerging or disappearing fixed points could have a major effect on the network activity, too.

In this thesis we want to address several aspects of synaptic drift, its link to representational drift, the nature and statistics of representational drift, and the difference between representational drift and learning induced changes of representations. This is done in two parts: Part iii describes results of the analysis of experimental data and Part iv consists of modeling studies.

We start with a brief recapitulation of some basic concepts useful for the understanding of the subsequent analyses (Part ii) giving detailed background on the auditory system (Chapter 1), the experimental recording of neuronal activity (Chapter 2), models of neuronal networks (Chapter 3) and the trade-off between flexibility and stability in general (Chapter 4). In Chapter 5 we describe representational stability (or the lack thereof) in mouse auditory cortex making use of response modes and find substantial representational drift under basal conditions that is biased towards a differential increase of associations during learning. To reveal the functional relevance of these dynamics and deconstruct them into their essential parts we then devise a set of ten elementary operations as a tool to further investigate representational drift and disentangle the increase of associations into both the stabilization of existing associations and the formation of new associations in Chapter 6.

The experimental findings of those two chapters will be complemented by model studies in the next two chapters. In Chapter 7 we formulate a firing rate model to reproduce findings from mouse auditory cortex and describe a regime with similar population statistics for strong recurrent connections and strong inhibition. In this regime stimulus responses are clustered randomly into response modes. In order to analyze the effect of synaptic drift on network resposnes, synaptic drift is added to this model in Chapter 8 leading to periods of stable stimulus responses, which are sometimes interrupted by abrupt transitions towards new activity patterns. These abrupt transitions coincide with qualitative changes in the topology of the fixed points of the network. Finally, these results are discussed in Part v. Supplementary figures can be found in the appendix (Chapter 9).

Part II

FUNDAMENTALS

In this part we provide fundamental background knowledge on several topics we deem essential for this thesis. We discuss the biological background (Chapter 1), specifically the auditory pathway, the auditory cortex as well as cell assemblies and synaptic and representational drift. In Chapter 2 we discuss the experimental method of two-photon imaging. We then introduce computational models (Chapter 3) used in theoretical neuroscience in general and focus on firing rate models and the distinction between plasticity and drift in models, before we close this part with some more general considerations on change and stability (Chapter 4).

1

BIOLOGICAL BACKGROUND

Here, we briefly describe the auditory system, as all data analyzed and modeled in this thesis has been recorded in auditory cortex. We give a quick introduction to the auditory pathway (Section 1.1) and the auditory cortex (Section 1.2), before we briefly talk about cell assemblies (Section 1.3) and finally discuss evidence for synaptic and representational drift in the neocortex (Section 1.4).

1.1 AUDITORY PATHWAY

In general, input from our sensory organs (except olfaction) is processed along its specific sensory tract into the thalamus, where dedicated thalamic subregions pre-process and relay information to the respective primary sensory cortex (Sherman, 2007, Guillery and Sherman, 2011, Sherman, 2012, Metzger et al., 2013). For the auditory modality this pathway begins at the tympanic membrane in the middle ear, where sound is transferred to the cochlea via the *malleus*, *incus*, and *stapes* (Ades and Engström, 1974). In the cochlea the organ of Corti transforms the mechanic sound wave into a nerve signal, tonotopically endoding the sound into frequency bands (Robles and Ruggero, 2001, Theunissen and Elie, 2014). The auditory pathway is considered rather complex compared to other sensory modalities and the auditory signal is relayed and processed multiple times, before it reaches auditory cortex. It remains a highly debated matter, what parts of processing are performed at which station along the processing pipeline, but many feats of auditory perception are processed subcortically, like spatial origin or temporal characteristics of sounds (Middlebrooks and Knudsen, 1984, Brainard and Knudsen, 1993, Frisina, 2001, Palmer and Kuwada, 2005, Singheiser et al., 2012, Pannese et al., 2015). The signal passes through the cochlear nuclei (one for each ear), before it is merged with signal from the other ear in the trapezoid body. It is further processed in the superior olivary complex, generally thought of as being responsible for stereo hearing (Gray, 1997). From the superior olivary complex the signal travels on through the inferior colliculus, before it reaches thalamus in different parts of the medial geniculate nucleus (MGN): the ventral MGN belongs to the so-called lemniscal pathway. It is tonotopically organized and mainly projects to auditory cortex (Morel and Kaas, 1992, Guillery and Sherman, 2011, Metzger et al., 2013). The so-called non-lemniscal pathway passes through the medial MGN, which also receives input from other sensory modalities and is thus not tonotopically organized. It projects to auditory cortex, too, but also to parts of the limbic system, e.g. the amygdala (Ma and Suga, 2009).

This entire chain of sound processing prior to auditory cortex is often modeled in a specific class of *cochlea models*. These cochlea models model anything from fluid coupling and micromechanics over cochlear amplification to cochlear non-linearities and electrical coupling of the cochlea to the auditory nerve (for an overview see Ni et al., 2014 or Rudnicki et al., 2015). Interesting from our point of view would be a model that takes any sound as an input and preprocesses it in the way the auditory pathway does, leading to an output similar to the signal the auditory cortex would get for this specific stimulus. For any realistic (i.e. frequency modulated) sound this would be a sort of a time-frequency representation of this sound, similar to a spectrogram, where frequency is on the y-axis, time on the x-axis and power at each time and frequency is color coded. As frequency is time dependent, spectrograms (as well as any other time-frequency representations) look qualitatively different depending on the applied temporal resolution.

Two other commonly used time-frequency representations of sounds, the *cochleogram* and the *correlogram* are described in Chaurasiya (2020). While the cochleogram is computed similarly to a spectrogram with subject specific filters for audible frequency and with varying bandwidths (the term cochleogram is sometimes also referring to this filter instead of the filtered function, see e.g. Linss et al., 2007), the correlogram is computed by splitting the signal into a time function and a correlation function. Both have the same problem as spectrograms, that the output depends heavily on the temporal resolution. And while all of these models work nicely for pure tone stimuli, that are mapped correctly onto the tonotopic map, they have a hard time with frequency modulated complex stimuli, that elicit responses in neurons, which are silent to all inherent frequencies of these stimuli (Theunissen and Elie, 2014). Additionally, recent work showcases the difficulties in reconstructing the stimulus from population activity in ferret auditory cortex and even more so the population activity from a spectrogram of the stimulus (Lubda, 2021). Nevertheless, this remains an interesting area of research, especially regarding cochlear implants and their performance in speech recognition (Pan, 2018, Russo et al., 2019) or the perception of music (Gauer et al., 2019). So, while modeling of the cochlea alone can be achieved, especially for cochlear implants, the following pre-processing steps prior to auditory cortex are not yet clear.

While these models capture pure tone frequency stimuli nicely and map them onto a tonotopic map similar to auditory cortex, they fail to predict responses to complex stimuli. This is linked to the second problem, which is the representation of complex sounds in general. It is not straightforward to describe complex stimuli in time-frequency representations. Due to these reasons we chose to follow a different approach, assuming a very general, random input with some correlations in our model in Chapter 7 and Chapter 8.

1.2 AUDITORY CORTEX

Auditory cortex is the part of neocortex tasked with the processing of auditory stimuli. Processing of the auditory scene is performed by its own dedicated region of the cortex as is the case for most other sensory modalities (e.g. Mendez and Geehan, 1988, Kentridge et al., 1999), before higher regions – located mostly in the frontal lobes – deal with more complex tasks like abstraction, planning, problem solving, and the coordination of motor and sensory functions (Waxman, 2017).

The main fraction of cortex is made up of neurons. In total there are 14 to 16 billion neurons in human cortex (Saladin, 2011) and, as this study is mostly concerned with mouse auditory cortex, around 14 million neurons in mouse cortex (Herculano-Houzel et al., 2013). Each neuron forms connections to on average $1,000$ to $10,000$ other neurons (Herculano-Houzel, 2009), however, the connectivity is still sparse: a neuron has connections to no more than 10% of neurons in its proximity and generally less to neurons further away (Gerstner et al., 2014). Neurons can be distinguished into two main classes, depending on their effect on other neurons, which is either *excitatory* (i.e. facilitating the response of consecutive neurons) or *inhibitory* (i.e. suppressing the response of consecutive neurons). In general, neurons are either fully excitatory or fully inhibitory, which is known as Dale's law (Dale, 1934). The ratio of excitatory to inhibitory neurons varies throughout cortex, but is roughly 80% excitatory and 20% inhibitory neurons (Hendry et al., 1987, Gentet et al., 2000, Sahara et al., 2012, Keller et al., 2018). While excitatory neurons are mostly so-called pyramidal or principal cells, inhibitory neurons (also called interneurons) come in many different varieties (Cajal, 1911, Jones, 1975), distinguishable by the neurotransmitters they express. Interneurons, i.e. neurons expressing GABA (gamma-Aminobutyric acid), which can be used to identify neurons as interneurons (opposed to pyramidal cells which express glutamate as their principal neurotransmitter), can be grouped into four main groups (Kawaguchi and Kubota, 1997, Harris and Mrsic-Flogel, 2013): parvalbumin (PV) expressing cells, somatostatin expressing cells (SST), vasoactive intestinal peptide (VIP) expressing cells, and neurogliaform cells (NG). Excitatory neurons are mostly innervated by PV cells, but apart from that it is currently a matter of intensive research to identify each subtype's functional role in cortical circuits. Even though inhibitory interneurons are outnumbered by excitatory pyramidal cells, they play an important role to keep to system in balance and avoid epileptic cortical activity (Dichter and Ayala, 1987), but also to maintain function like stimulus selectivity (Sillito, 1975).

Despite this apparent difference in numbers, excitation seems to be balanced by inhibition in cortical circuits, keeping the membrane potential of indiviual neurons just below threshold for the firing of an action potential, leading to the observed spiking statistics (Shadlen and Newsome, 1994, Shadlen and Newsome, 1998). The difference in numbers seems to be counterbalanced by synaptic strength and firing rates of inhibitory neurons being substantially higher.

Neurons in neocortex are organized in six layers, anatomically and physiologically distinguishable by cell densities and connectivity patterns (Noback et al., 2005, Kandel, 2013). The first, i.e. outermost, layer mostly consists of apical dendrites of neurons from lower layers. The second and third layer consist mostly of pyramidal neurons, connected to other cortical neurons. Layer four and layer five are the main input and output layers, respectively. Thus in sensory cortex, layer four is typically more pronounced, receiving input from e.g. sensory regions of thalamus. In motor cortex the output layer five is more pronounced. Layer six again, mostly consists of incoming and outgoing corticothalamic connections. In other words, input from thalamus arrives at the cortical level in layer four, is then processed in layers two/three, before the signal is transferred to different cortical and non-cortical nodes via layer five. The recordings, we are going to discuss in this thesis come from layers two and three of mouse auditory cortex. They are thus thought to be recordings of recurrent processing of auditory stimuli.

Compared to other cortical areas auditory cortex is characterized by sparser neuronal activity (Hromádka et al., 2008, Hromádka and Zador, 2009, Liang et al., 2019). This is generally believed to be due to a higher influence of inhibition (Hromádka et al., 2008, Zhao et al., 2015, Liang et al., 2019, Studer and Barkat, 2022). Neurons in auditory cortex are organized in a tonotopical way (Reale and Imig, 1980, Romani et al., 1982, Tsukano et al., 2017). This means neurons are sorted along a one-dimensional axis by their preferred pure tone stimulus. Other organizing principles are not known to date. It has, however, been shown in recent years, that at the local circuit scale sounds lead to a non-linear activation of neuronal assemblies, where a given assembly is typically activated by a set of different stimuli (Bathellier et al., 2012, Atencio and Schreiner, 2013, See et al., 2018). This leads to a very reduced set of responses as each population of neurons responds to a multitude of stimuli with only a handful of possible activity patterns. Sounds that are indistinguishable on this population scale can be decoded on the global scale, as the local grouping of stimuli is different in each imaged population of neurons. This local reduction to a small set of responses makes auditory cortex well suited to study different phenomena, as this small set is relatively easy to keep track of.

In our model of auditory cortex we want to model very local population activity and thus do not apply any tonotopical organization. We vary the ratio of inhibition to excitation to reveal a dynamic regime similar to auditory cortex.

1.3 CELL ASSEMBLIES

The term *cell assembly* is used to describe a group of neurons that is repeatedly activated together. Cell assemblies are believed to be the underlying neural substrate of cognititve and behavioral function. A cell assembly can then be understood to be the single functional unit behind an action, a percept, or a more abstract concept. Cell assemblies were first postulated by Hebb (1949) and they are thought to be the result of synaptic connections between cells that are repeatedly coactivated. Thus they can be understood as a group of neurons with strong synaptic connections among each other and weaker synaptic connections to neurons that do not belong to this cell assembly. Strong synaptic connections between neurons in a specific cell assembly compared to neurons that are not part of this specific cell assembly have a number of interesting properties, which are also present in the brain, as has been illustrated by theoretical studies: they play a role in modeling associative memory (Hopfield, 1982), where a partial memory state is enough to retrieve the full state. They can explain multistability as can be observed on the level of perception (Beer, 1995, Stern et al., 2014, Fasoli et al., 2016), a transition from winner-takes-all like dynamics to multiple attractors (Miller, 2016, Chen and Miller, 2020), or correction of synaptic drift (Acker et al., 2019, Kossio et al., 2021).

In biological neural circuits cell assemblies are notoriously hard to study because of their extent in space. Due to limited techniques it is near impossible to be certain that all cells of a specific assembly are recorded. Nevertheless, the recent development of imaging techniques has made it possible to record from several hundred cells simultaneously and capture groups of simultaneously activated cells. Thus, over the course of the last years evidence for the existence of cell assemblies has been gathered in different areas of the brain (e.g. Harris, 2005, Buzsaki, 2010, Yuste, 2015, Holtmaat and Caroni, 2016).

They could also be linked to cognitive function, specifically to memory recall (Tonegawa et al., 2015).

Cell assemblies are relevant both for the data analysis part and for the modeling part of this thesis as they are classically used to describe collective behavior of neurons and we discuss our results compared to them.

1.4 DRIFT

Drift is known from physical reaction diffusion systems as the linear term of a stochastic process. In a more general context it is often employed synonymously for random change in a system. In neuroscience the term *drift* is used to describe any truly random changes (opposed to plasticity, which is following underlying rules). We discuss *synaptic* drift, i.e. random changes of synapses, in Section 1.4.1 and *representational* drift, i.e. random changes of neuronal (population) activity in Section 1.4.2. For a more detailed discussion on drift and plasticity in a modeling context see Section 3.2.

1.4.1 *Synaptic drift*

It is widely believed that all cortical function is stored in synaptic connectivity. Synaptic connectivity, however, has been shown to undergo constant remodeling in recent years. Synapses display changes in strength, they emerge and disappear (Rumpel and Triesch, 2016). For a long time, learning induced changes of synapses, as first postulated by Hebb (1949), have been studied (for a recent review, see Humeau and Choquet, 2019). In recent years, changes in synaptic connectivity have also been found to be present in the absence of any explicit learning paradigm (Loewenstein et al., 2011, Loewenstein et al., 2015) and even during a pharmacological blockade of neuronal activity (Yasumatsu et al., 2008, Rubinski and Ziv, 2015, Dvorkin and Ziv, 2016, Nagaoka et al., 2016). This synaptic drift seems to be a fundamental feature of neuronal networks going beyond Hebb's famous rule *"fire together, wire together"*. It has been found in both excitatory (Yasumatsu et al., 2008, Loewenstein et al., 2011, Loewenstein et al., 2015, Berry and Nedivi, 2017, Ziv and Brenner, 2018) and inhibitory synapses (Rubinski and Ziv, 2015, Dvorkin and Ziv, 2016, Villa et al., 2016).

The reasons for synaptic drift can only be speculated about. One obvious reason might be lifetimes of synaptic proteins that typically range from hours to days, although some are surprisingly long lived (Cohen and Ziv, 2019). Other reasons might be the competition for limited resources (Triesch et al., 2018) or a more explorative version of a Hebbian mechanisms, where neurons would have to sample multiple partners in order to figure out, which they fire together with.

Recent theoretical modeling approaches have focused on compensatory mechanisms for synaptic drift (Kappel et al., 2015, Kappel et al., 2018, Mongillo et al., 2018, Humble et al., 2019, Susman et al., 2019, Kossio et al., 2021). Various synaptic plasticity mechanisms can compensate some amount of synaptic drift. Above a certain threshold, however, this modeling work has shown that synaptic drift can lead to drift on the level of cell assemblies.

We modeled synaptic drift using a multiplicative Ornstein-Uhlenbeck-process fitted to experimental data by Loewenstein et al. (2011). This process, which changes synapse sizes dependent on the current synapse size, keeps the system in a steady state with a log-normal distribution of synapse sizes.

1.4.2 *Representational drift*

Similarly to changes in the synaptic connectivity, changes in neuronal population activity have long been researched in the context of learning, starting with Hebb (1949). Apart from learning, the neuronal responses were thought to remain stable in order to maintain stable function. However, recently, long-term remodeling of population activity has been reported in the mouse hippocampus and barrel, olfactory, visual, motor and posterior parietal cortex (Rokni et al., 2007, Huber et al., 2012, Mankin et al., 2012, Margolis et al., 2012, Ziv et al., 2013, Clopath et al., 2017, Driscoll et al., 2017, Hainmueller and Bartos, 2018, Rule et al., 2019, Deitch et al., 2021, Schoonover et al., 2021). All of these report changes in population patterns across days for the same stimuli or tasks, even in the absence of an apparent learning paradigm.

Of course this representational drift is challenging the idea that stable behavior, as can be observed on timescales from days to years, is rooted in cortex. But how can this instability be overcome to finally result in stable behavior? Compensatory mechanisms for this representational drift have been proposed based on a constant re-learning of changing representations in potential read-out neurons (e.g. Acker et al., 2019, Kossio et al., 2021). While apparently somewhere along the processing path some sort of stabilization has to happen, this instability through representational drift might also have its advantages. It could be used as a time stamping mechanism, so each instance of a memory has a different representational instance. It could prevent the brain from overfitting by randomly altering memories, which has been shown useful in artificial neuronal networks. Another idea would be that the main advantage of a big brain during evolution is not its storage capacity, but rather its ability to quickly adapt to changes in environment. This is achieved easier in a drifting brain. As the precise nature and function of representational drift are still unknown (Chambers and Rumpel, 2017), different cognitive processes could be linked to it, like spontaneous associations (Wallas, 1926) or their forgetting (Richards and Frankland, 2017).

2

TWO-PHOTON IMAGING

One of conventional light – and for these matters also fluorescent – microscopy's main disadvantages is its inability to image through tissue. While surfaces and structures beneath optically transparent materials are visible, everything beyond an intransparent surface is invisible. Photons are reflected by intransparent tissue, making the object both visible and opaque. Inside opaque tissue the light intensity decreases with $1/r^2$ for a given depth r; the rest of the light is scattered in all directions. This scattering is the reason, why it is hard to image through tissue, even with high light intensity levels and fluorescence imaging: photons reach deeper levels inside the tissue and are reflected back, but they are not distinguishable from the photons refelected back by the tissue above.

Two-photon fluorescent microscopy is able to solve this issue to some extent. Instead of one photon from the LASER being absorbed by an electron in the tissue and shifting this electron from one energy state to another, the energy of the photons is determined in a way, that the electron needs the energy of exactly two photons to switch states. As this excitation requires two photons hitting the same electron simultaneously, it can only occur at regions of very high light intensity, i.e. the focus point of the LASER, and hardly ever outside of this focus point. This leads to negligible scattering from surrounding tissue and a good signal-to-noise ratio. Additionally, the single photon emitted by the electron falling back to its ground state is of a different wave length and thus easily distinguishable from the light emitted by the LASER. A third advantage is that this excitation requires two lower frequency photons instead of one with higher frequency and low energy light has a higher absorption length in tissue. For an in detail description of two-photon imaging methods see Schmitt et al. (2013), the quantum mechanics of two-photon imaging have been nicely summarized by Shi et al. (2015).

Two-photon absorption was first predicted by Göppert-Mayer (1931) and confirmed experimentally shortly after the development of LASERs (Kaiser and Garrett, 1961). The first two-photon LASER scanning fluorescence microscope was developed by Denk et al. (1990) and since then it has become an important tool in many fields of biology (König, 2018), among others also the neurosciences (Svoboda and Yasuda, 2006).

As tissue samples are rarely fluorescent by themselves, dyes are used to highlight relevant structures. In the case of two-photon fluorescence microscopy these dyes are mostly fluorescent proteins. A lot of research goes into the development of ever new fluorescent proteins in different colors with different absorption and fluorescence spectra, binding to different biomarkers (for a recent overview see Xu et al., 2020). These proteins then bind to e.g. specific cell types or various intracellular structures like synaptic scaffolding proteins. Especially useful for neuroscience are fluorescent proteins in neurons (sometimes also specific neuron types, like PV-, SST-, or VIP-interneurons) and proteins that are only fluorescent, when a neuron is active. Two elegant ways exist in order to get cells to express these fluorescent proteins. Cells can either be made to express specific proteins via viral transduction, where a virus causes the production of fluorescent markers by specific cells or there are entire mouse lines genetically engineered to express fluorescent proteins in their cells.

For the experimental data described and analyzed in Chapter 5 and Chapter 6 mice were transduced with two fluorescent markers, one structural marker identifying somata of neurons (H2B::mCherry, Nathanson et al., 2009), and one functional marker as read-out of neuronal activity reacting to calcium influx into a neuron (GCaMP6, Chen et al., 2013).

Apart from the aforementioned advantages compared to single photon imaging two-photon imaging allows for chronic imaging and tracking of individual cells across several days – even silent cells due to structural markers. Individual synapses and other subcellular structures can be investigated due to a spatial resolution in the sub-micrometer regime. The trade-off is a temporal resolution which is not on the level of single neuronal spikes, especially a slow decay after the spike onset, mostly due to the fact that calcium dynamics are measured and not membrane voltage. Another disadvantage is the persisting limited imaging depth, which can be tackled by using three (or more)-photon imaging, which requires even higher light intensities. Both problems are currently tackled by the ongoing development of new fluorescent markers.

3

MODELS IN NEUROSCIENCE

Neurons are typically modeled as integrators of inputs, becoming active, when the accumulated input passes a threshold. The simplest form of this idea might be (leaky) integrate and fire neurons (Lapicque, 1907), which collect incoming action potentials, accumulating charge, increasing their membrane potential until a threshold is crossed and then themselves fire an action potential. Typically, a slow leakage current is included, requiring accumulation of input over a certain time. This captures the behaviour of biological neurons quite well and does a great job in disentangling the comparably fast dynamics involved in generating an action potential (as modeled by e.g. Hodgkin and Huxley, 1952) from the rather slow dynamics of the network. Depending on the input weights to these neurons they are already able to perform simple logical tasks, inspiring further abstractions to networks of neurons that only represent logical gates (McCulloch and Pitts, 1943). Togehter with Hebbian inspired learning rules (*"fire together, wire together"*, Hebb, 1949) these ideas led to the *perceptron* (Rosenblatt, 1958), which is able to learn input weights to produce a wanted outcome and perform logical operations.

Networks made of spiking neurons, especially leaky integrate and fire neurons, are able to capture synchronization and chaos, synchronous and asynchronous network states and some more features of biological neuronal networks like contrast invariance or oscillations (Hansel and Sompolinsky, 1992, Gerstner, 1995, Hansel and Sompolinsky, 1996, Hansel et al., 1998, Brunel and Vincent, 1999, Brunel, 2000, Gerstner, 2000, Brunel et al., 2001 Hansel and Vreeswijk, 2002, Mattia and Giudice, 2002). An in depth overview is given in Gerstner and Kistler (2002). Adding plasticity rules for synaptic connections between the neurons, networks are able to self organize to achieve these tasks (Amit and Brunel, 1997, Turrigiano et al., 1998, Desai et al., 1999, Zhang and Linden, 2003, Steil, 2007, Schrauwen et al., 2008, Lazar et al., 2009, Watt and Desai, 2010, Zenke and Gerstner, 2017, Zenke et al., 2017). They can furthermore become robust against noise (Toutounji and Pipa, 2014), autonomously form and maintain cell assemblies (Litwin-Kumar and Doiron, 2014), and reproduce experimentally measured synapse distributions (Zheng et al., 2013, Miner and Triesch, 2016). Apart from that, spiking neural networks are also used in the field of artificial inteligence (for a review, see Ponulak and Kasinski, 2011).

Despite this huge success, however, networks of spiking neurons are often cast aside for a further simplification to firing rate models. Spiking neural networks are biologically motivated and they are able to account for the various experimentally observed phenomena. Also, features depending on the exact time of an action potential are lost in rate models (e.g. synchronization and spike timing dependent plasticity). Nevertheless, this further simplification to firing rate networks comes with many advantages. Firing rate models are less computationally expensive, because they are ignoring the fast time scales of single spikes. Thus the time scales of an individual unit and the entire system move closer together. Apart from fewer free parameters in rate models and the possibility to perform some calculations analytically, one major difference between the two is, that firing rate networks can easily be scaled down to few units (which can then be interpreted as populations, Wilson and Cowan, 1972, Wilson and Cowan, 1973), otherwise leading to unrealistic synchronization in spiking neural networks. For a modern comparison of spiking neural networks and firing rate networks, see e.g. Brette (2015). As we use a firing rate model for our investigations, we want to give a more detailed introduction to firing rate models in Section 3.1.

Most interesting dynamics can be found in models of neuronal networks independent of the exact implementation. Typically three dynamic regimes are described in literature:

(a) a uni-stable regime with one attractor state, the network activity converges to,

(b) a multi-stable regime, where the system can display different dynamics – e.g. multiple attractor states, bump attractors, line attractos etc. – and

(c) a chaotic regime.

These regimes can be found in random networks and transitions between regimes can be found by changing parameters as synaptic gain (e.g. Wilson and Cowan, 1972, Sompolinsky et al., 1988), strength of stimulation (Wilson and Cowan, 1972, Rubin et al., 2015b), ratio of inhibition to excitation (e.g. Rost et al., 2018, Zhang and Saggar, 2020), recurrent connection strength (e.g. Ostojic, 2014, Stern et al., 2014).

The different dynamic regimes (mostly within the above mentioned multi-stable regime) which are comparable to cortical dynamics often arise in balanced systems, i.e. systems with both strong excitation and strong inhibition. This means that network dynamics are governed by strong recurrent connections. For these systems, in a ground state, single neurons receive strong excitatory and inhibitory inputs that cancel out to first order, but when they receive only little further activation they are easily pushed from below threshold to above threshold and start to fire action potentials. Balance between excitation and inhibition emerges automatically in large networks with strong synapses (Vreeswijk and Sompolinsky, 1996) and can account for different regimes ranging from a single attractor state via multiple attractor states to truly chaotic dynamics (Vreeswijk and Sompolinsky, 1996, Vreeswijk and Sompolinsky, 1997, Jahnke et al., 2009).

In this balanced state many features are typically attributed to different forms of inhibition, be it a stabilization via fast feedback inhibition in the so-called stabilized supralinear network (Ahmadian et al., 2013, Rubin et al., 2015b), symmetry breaking and multiple stable solutions mitigated by recurrent inhibition in small networks (Fasoli et al., 2016), winner-takes-all like dynamics (Miller, 2016, Chen and Miller, 2020), sub- and supra-linear summation and balanced amplification (Murphy and Miller, 2009, Ahmadian et al., 2013, Rubin et al., 2015b) or stable dynamics around a single attractor accounting for the difference in evoked and spontaneous activity (Hennequin et al., 2018).

In the next section (Section 3.1) we give a more detailed description of firing rate networks, in Section 3.2 we will then give some background on our implementation of synaptic drift and the difference between synaptic plasticity and drift.

3.1 FIRING RATE MODELS

Firing rate models of neuronal networks can be derived from spiking neural networks in a multitude of ways (e.g. Wilson and Cowan, 1972, Ermentrout, 1994, Aviel and Gerstner, 2006, Ostojic and Brunel, 2011), but in general they are given by one of two equations (Equation 3.1 and Equation 3.2), proven to be equivalent by Miller and Fumarola (2012):

$$\tau\frac{\partial}{\partial t}\mathbf{v} = -\mathbf{v} + \tilde{\mathbf{I}} + \mathbf{W}\mathbf{f}(\mathbf{v}) \tag{3.1}$$

and

$$\tau\frac{\partial}{\partial t}\mathbf{r} = -\mathbf{r} + \mathbf{f}\left(\mathbf{W}\mathbf{r} + \mathbf{I}\right), \tag{3.2}$$

where $\mathbf{v}$ and $\mathbf{r}$ are vectors of firing rates of individual neurons, $\mathbf{W}$ is the recurrent connectivity matrix, $\mathbf{f}(\mathbf{x})$ is a nonlinearity acting on the individual entries of $\mathbf{x}$, τ is the characteristic time constant of the system, and $\tilde{\mathbf{I}}$ and $\mathbf{I}$ are inputs to the respective networks.

We use the element wise version of Equation 3.2, given by Equation 3.3:

$$\tau \frac{\partial}{\partial t} r_i = -r_i + f\left(\sum_{j=1}^{N} W_{ij} r_j + s_i(t) \right), \tag{3.3}$$

where r_i is the firing rate of neuron i, W_{ij} is the connection strength from neuron j to neuron i, τ is the time constant, $f(x)$ a non-linearity, and $s_i(t)$ is the time dependent input to neuron i.

Firing rate models have been used to study neuronal networks and produce insight into the rich dynamics of neural processing. Already small networks of two to three units display interesting phenomena such as different dynamic regimes and hysteresis between those different dynamic regimes (Wilson and Cowan, 1972, Wilson and Cowan, 1973), input and connectivity dependent bifurcations (Borisyuk and Kirillov, 1992, Beer, 1995), fixed points, periodic, quasi-periodic and chaotic behavior (Wilson and Cowan, 1973, Pasemann, 2002), and spontaneous symmetry breaking (Fasoli et al., 2016). Larger networks are able to reproduce these rich dynamics (e.g. Cessac, 1995, Rajan et al., 2010, Mastrogiuseppe and Ostojic, 2017) and gain even more insight into cortical dynamics. Supra- and super-linear summation of inputs has been recorded depending on the context in the so-called stabilized supra linear network (Murphy and Miller, 2009, Ahmadian et al., 2013, Rubin et al., 2015b), both of which have also been observed in experimental recordings. These networks are mostly governed by some form of balance between excitation and inhibition with generally strong recurrent connections. Firing rate models are used for many different tasks, e.g. pattern completion (Curto and Morrison, 2016) and underlie most artificial neuronal networks.

We use a generic firing rate network to understand a clustering of stimuli into a small group of possible responses as observed in mouse auditory cortex and apply synaptic drift to investigate the representational drift of the system.

3.2 PLASTICITY AND DRIFT

The two words *plasticity* and *drift* are used in various ways, sometimes even synonymously, however, mostly they are employed to differentiate between rule based changes (plasticity) and random changes (drift) in a system. Here, we want to follow this convention.

The idea that synapses change through learning was first brought forward by Hebb (1949). The general idea is that neurons that are active together form a stronger connection between each other. This can be further refined to spike timing dependent plasticity, where a synapse is strengthened, if it goes from neuron A to neuron B and neuron A fires an action potential in a short time window, before neuron B fires an action potential. The general concept makes sense intuitively and it is widely believed that a mechanism like this is in charge of forming synaptic connections. The exact biological implementation of such a mechanism, however, remains a matter of debate, as this strengthening of synapses has to be counterbalanced by some other mechanism to avoid diverging synapses. In models this is typically achieved via some sort of homeostatic normalization (e.g. Lazar et al., 2009, Watt and Desai, 2010). This homeostatic normalization can be enforced by a constraint on the sum of all synapse weights, for example, which works perfectly in models, but is not biologically plausible, as it would require global knowledge of all synapse sizes in each individual neuron. So, this idea of synaptic plasticity makes sense intuitively and there is evidence for some sort of it in experimental data, but it remains unclear, how exactly it might be implemented in biological circuits.

Synaptic drift, on the other hand, is observed in a lot of biological systems (e.g. Loewenstein et al., 2011, Statman et al., 2014, Loewenstein et al., 2015, Ziv and Brenner, 2018). Synapses appear and disappear and those remaining show substantial fluctuations in their strengths, all seemingly randomly. This drift is easy to model (e.g. Loewenstein et al., 2011), keeping the system in a steady state, but contrary to synaptic plasticity there is so far no straightforward interpretation of its function or apparent reason for its presence. It might be employed as a homeostatic mechanism for synaptic plasticity, play an important role for the adaptability to ever new situations, or it might be utilized to avoid overfitting and allow for generalization. For considerations about its biological relevance, see Section 1.4.1 for synaptic drift and Section 1.4.2 for representational drift. The general issue of the trade-off between stability and flexibility for brain function will be addressed in Chapter 4.

We model synaptic drift following a fit to experimental data by Loewenstein et al. (2011) as a stochastic process, more precisely a multiplicative version of an Ornstein-Uhlenbeck process (Uhlenbeck and Ornstein, 1930). This changes individual synapse sizes randomly dependent on their strength while maintaining the log-normal distribution of synapse weights in a steady state.

4

STABILITY AND FLEXIBILITY OF BRAIN FUNCTION

Brains are able to maintain memories throughout a lifetime, while they are also able to form new memories in an instant. We are, for example, able to remember friends and family and at the same time form a new memory of a person, we just met. This combination of abilities without catastrophic forgetting, which would be caused by just overwriting old memories with new ones, displays the need for both stability and flexibility (Grossberg, 1980). As the brain has for a long time been understood as the place of memory storage, similar to a hard disk, there exists a vast body of literature focussing on the stability of certain aspects of the brain (e.g. Bliss and Collingridge, 1993, McGaugh, 2000, Kandel, 2001, Poo et al., 2016). In recent years this focus has slowly been shifting towards a discussion of unstable aspects, too, and thus there is a growing body of evidence for instability in the brain (e.g. Frankland et al., 2013, Hardt et al., 2013, Berry and Davis, 2014).

The idea here is that the aim of the brain is not the accumulation of knowledge, but rather making the best decision in the presence based on past experience (Dudai and Carruthers, 2005, Schacter et al., 2007, Richards and Frankland, 2017). To do so, forgetting (i.e. transience of memory) might be as important as remembering (i.e. persistence of memory). This is evident from anecdotal stories and case studies of patients with super-human memory (e.g. Luria, 1968). Patients are not able to perform well in everyday life despite (or because of) having (near) perfect memory of everything they encounter. Remembering in great detail every moment of their lifes is associated with an inability to perform seemingly simple tasks. This is hypothesized to be due to a lack of flexibility or a lack of generalizability by Richards and Frankland (2017). The brain needs both stability and flexibility, which means both memory and forgetting.

The problem remains: How can the brain be flexible and stable at the same time? A lot of theoretical work is focussed on the question how to maintain stability in the presence of synaptic or representational drift (Vogels et al., 2011, Litwin-Kumar and Doiron, 2014, Mongillo et al., 2017, Acker et al., 2019, Kossio et al., 2021). This is mostly achieved by homeostatic plasticity. Random drift, however, might be useful, too, both from a psychological point of view to overcome trauma (Richards and Frankland, 2017) and from a theoretical point of view to prevent overfitting (Aitken et al., 2021). The need for stability and flexibility depends on the situation. It is very helpful to remember the location of one's favorite restaurant. This does not change too often and is thus stored in memory in a stable manner. It might also be useful to have a general idea of the currently trendy area downtown, where all the cool bars are. The exact location of each bar might not be so important, and thus generalization comes in handy. Flexibility is needed, when the favorite restaurant moves to a new location. The old location is best forgotten and a new one quickly memorized.

Our brain has to be able to perform all these various tasks and be very stable – sometimes for decades – on the one hand and on the other hand be able to rapidly update and flexibly change its memories. How this is achieved in detail remains unclear.

Part III

ANALYSIS OF EXPERIMENTAL DATA FROM MOUSE AUDITORY CORTEX

This part consists of two chapters. In Chapter 5 we discuss analyses performed on imaging data of populations of neurons from mouse auditory cortex. We find that neuronal representations of stimuli undergo constant remodeling and that learning biases this ongoing remodeling towards a differential generalization. In Chapter 6 we define a set of operations to deconstruct these cortical population dynamics and find that learning is linked to an increased formation of associations.

5

FORMATION OF ASSOCIATIONS BY LEARNING-INDUCED BIASES IN THE ONGOING DYNAMICS OF SENSORY REPRESENTATIONS.

This chapter is in large parts based on the publication

Aschauer*, D. F., Eppler*, J. B., Ewig, L., Chambers, A., Pokorny, C., Kaschube§, M., and Rumpel§, S. (2022). "Learning-induced biases in the ongoing dynamics of sensory representations predict stimulus generalization.", *Cell Reports*.

* shared first authors, § shared last authors.

Author contributions: Conceptualization: MK, SR; Methodology: DFA, JBE, CP; Software: JBE; Formal Analysis: JBE, DFA, LE, AC; Investigation: DFA; Writing-Original Draft: SR, MK, DFA, JBE; Writing-Review & Editing: LE, AC; Supervision: MK, SR; Funding Acquisition: SR

All experiments were performed by Dominik Aschauer. Data preprocessing and formal analyses (except for Figure 5.6, Figure 5.7 and Figure 5.8) were performed by Jens-Bastian Eppler.

This chapter starts with an introduction to ongoing activity dynamics in the brain in general and in mouse auditory cortex specifically (Section 5.1). Next, we present our results (Section 5.2), namely that we found ongoing changes of response modes during basal conditions that were modified towards the formation of associations during learning. This is followed by a brief discussion of these results (Section 5.3). We close this chapter with the methods section (Section 5.4).

5.1 INTRODUCTION

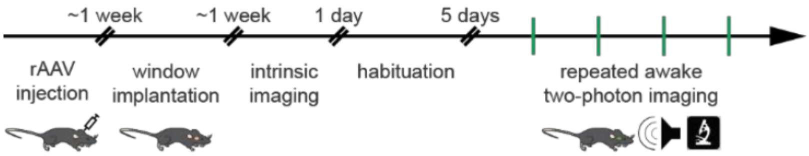
We want to study representational dynamics in mouse auditory cortex under basal conditions (i.e. representational drift) and during learning. We make use of response modes found in mouse auditory cortex by earlier work (Bathellier et al., 2012). Response modes are observable in local populations of neurons in mouse auditory cortex. Multiple stimuli evoke the near same response and these responses can thus be clustered to form a so-called response mode.

We used chronic two-photon calcium imaging in the mouse auditory cortex over several days to monitor sound-evoked activity patterns forming sensory representations. We find that even in behaviorally habituated mice, auditory representations display significant plasticity involving the remapping of stimuli to response modes, the creation of new response modes and their elimination. We analyzed these ongoing changes, exploiting the discrete nature of response modes, finding near stationary dynamics on the level of both single cells and population dynamics. Applying the same analysis to data from mice undergoing auditory cued fear conditioning, we observed specific biases in the dynamics of response modes that explain an increase in efficacy of sound encoding during learning and reveal an increased rate of the formation of new associations among sensory stimuli.

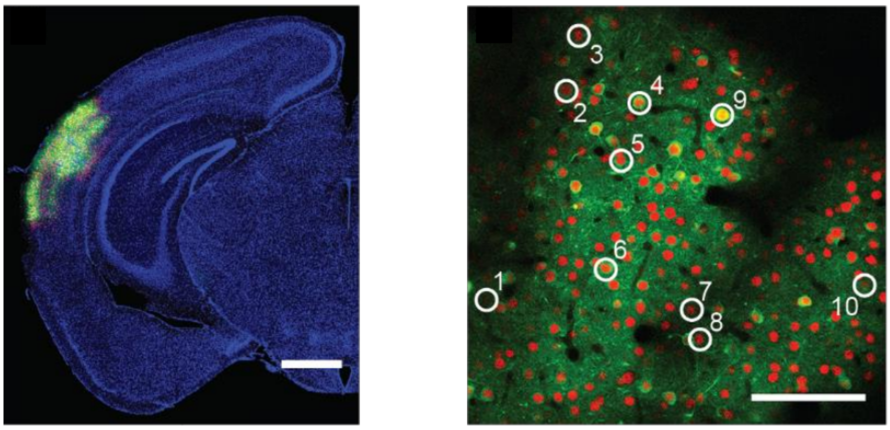
5.2 RESULTS

5.2.1 *Chronic large-scale calcium imaging of neurons in the mouse auditory cortex*

In order to assess the long-term dynamics of auditory representations, we transduced cells in the mouse auditory cortex with a co-injection of two rAAV8-vectors to drive stable expression of two fluorescent proteins under the control of the Synapsin promoter: The genetically encoded calcium indicator GCaMP6m (Chen et al., 2013), to chronically record neural activity, and the fusion protein H2B::mCherry, as a structural marker to distinctively label the nuclei of transduced neurons (Figure 5.1, Figure 5.2, Nathanson et al., 2009).



(a) Experimental timeline.



(b) Coronal section.

(c) Example Field of View (FOV).

Figure 5.1: Two-photon imaging of neuronal activity in mouse auditory cortex. (a) Experimental timeline. (b) Confocal image of coronal section of a mosue brain sacrificed 46 days after stereotactic injection of two rAAVs leading to the expression of GCaMP6m (green) and H2B::mCherry (red) in auditory cortex. Counterstain DAPI (blue). Scale bar 1 mm. (c) *In vivo* image of a local population of neurons in layer 2/3 of auditory cortex showing expression of GCaMP6m (green) and H2B::mCherry (red). Scale bar 1 µm. White circles and digits represent the neurons in Figure 5.2.

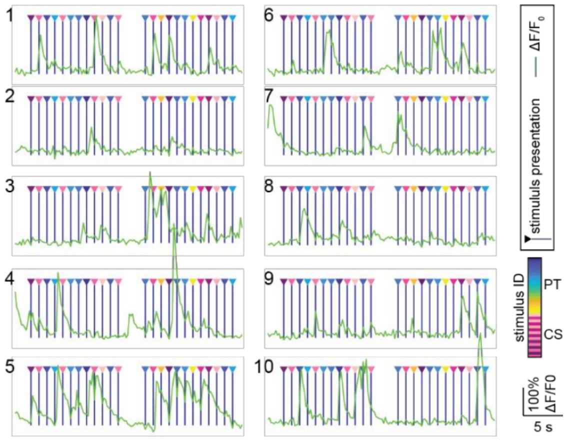


Figure 5.2: Simultaneously recorded activity traces of ten example neurons from Figure 5.1c. (green: $\Delta F/F_0$, blue: stimulus presentation, arrowheads: stimulus, PT: pure tones, CS: complex stimuli).

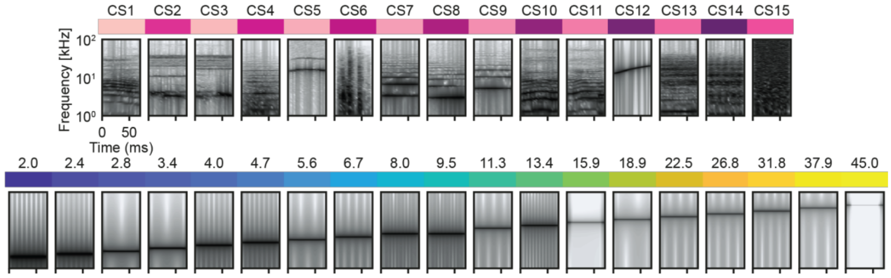


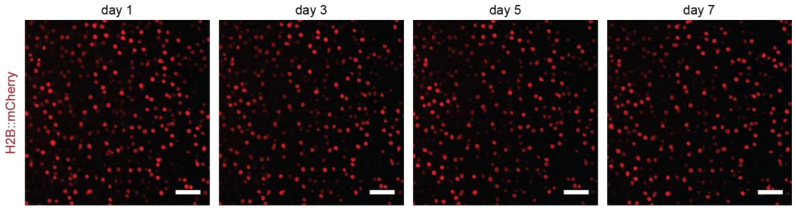
Figure 5.3: Spectrograms of the complex and pure tone stimuli used for *in vivo* two-photon experiments.

We used intrinsic signal imaging in response to a set of pure-tone stimuli of varying frequency in order to guide subsequent two-photon imaging. For calcium imaging experiments in awake, head-fixed, passively listening mice, we used a stimulus set of brief ($50\,\text{ms}$ - $70\,\text{ms}$) sounds containing 19 sinusoidal pure-tones and 15 complex sounds characterized by temporally modulated power in multiple frequency bands delivered free-field using a calibrated speaker at $74\,\text{dB}$ SPL (Figure 5.3).

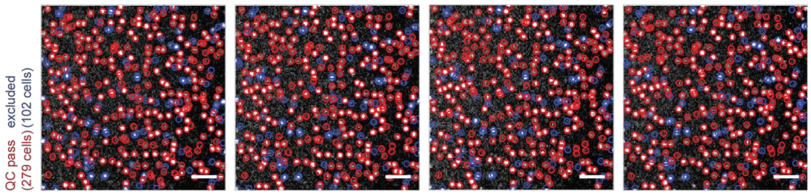
Mice were habituated to head-fixation and pre-exposed to the set of sound stimuli for at least five days to ensure that adaption to novel sensory responses has largely completed (Kato et al., 2015) and that data acquisition occurred under behaviorally and environmentally familiar and constant conditions. The red nuclear marker enabled high-fidelity re-identification and registration of individual local populations that were re-imaged for four time points at a two-day interval (Figure 5.4a to Figure 5.4d). We imaged neuronal activity in response to the 34 sound stimuli ($20-30$ repetitions each, presented in a random order) in a total number of $21,506$ neurons in 97 different fields of view ($100-300$ neurons per FOV) in cortical layer $2/3$ of 12 mice (Figure 5.2; Figure 5.4e to Figure 5.4g). When assessing trial-averaged calcium responses to pure-tones and complex sounds over this period, we observed that many neurons that were responsive within a FOV showed essentially stable responses to the sound stimuli over the course of several days. Others, however, showed substantial re-tuning involving the gain or the loss of responses with substantial signal amplitudes (Figure 5.5).

5.2.2 *Dynamic long-term remodeling of sound responses in individual neurons and populations of neurons*

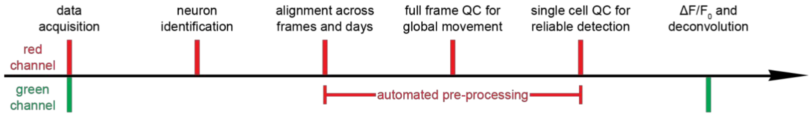
In light of the re-tuning of sound responsiveness in a substantial fraction of neurons, we next asked how these changes would affect the ability of the auditory cortex to form a stable representation of the



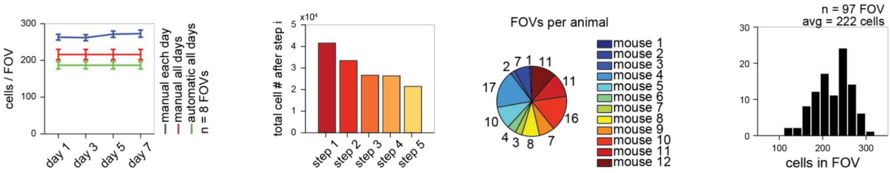
(a) Raw images.



(b) Detected and tracked cells.



(c) Preprocessing pipeline.



(d) Found cells. (e) Kept cells. (f) Animals. (g) Cells/FOV.

Figure 5.4: An automated image processing pipeline for high-fidelity tracking of neurons. (a) H2B::mCherry signal of an example FOV on all four imaging days. The distinct labeling in the red channel allows high-fidelity tracking of individual neurons using a signal that is independent of neuronal activity (scale bar, 50 µm). (b) Same example FOV as in (a). Cells passing quality control criteria based on signal intensity, signal/noise ratio and nearest neighbor distance in the red channel on all time points are marked in red. Excluded cells are marked in blue. (c) Data preprocessing pipeline. All steps regarding cell identification and tracking are based on the red channel. Only in the last step the green channel was used to read out neuronal activity. (d) Quantification of cells with different tracking methods. Blue: Cells were manually identified on each day individually; Red: Manually identified cells that could be reliably identified on all four imaging days; Green: Cells manually identified on day 1 and automatically tracked on the subsequent days. The automated procedure was applied for the full dataset of this study. (e) Total number of cells in dataset during basal conditions passing the different image pre-processing steps. Step 1: All manually identified cells on day 1; Step 2: All cells from step 1 with good signal quality on all time points after automated image alignment; Step 3: All cells from step 2 with sufficient distance from nearest neighbor; Step 4: All cells from step 3 from FOVs with at least 100 cells; Step 5: All cells from step 4 from FOVs with reliable sound-evoked population responses. (f) Number of imaged FOVs per animal in dataset capturing dynamics during basal conditions. (g) Histogram of the number of cells in each FOV in dataset of dynamics during basal conditions.

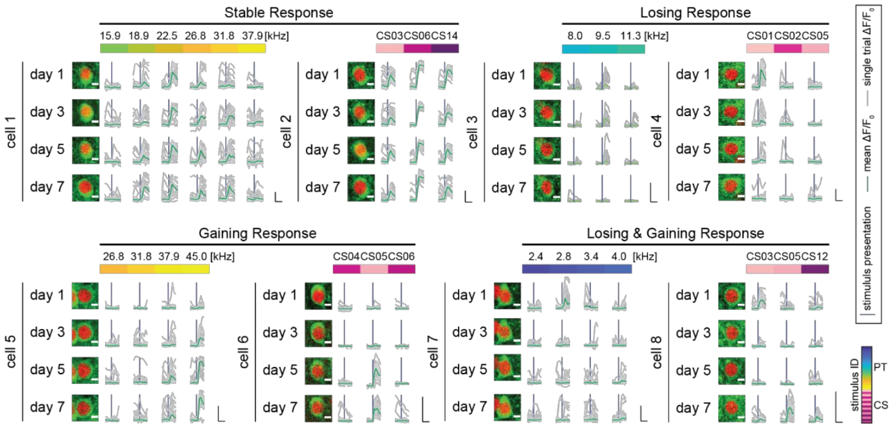
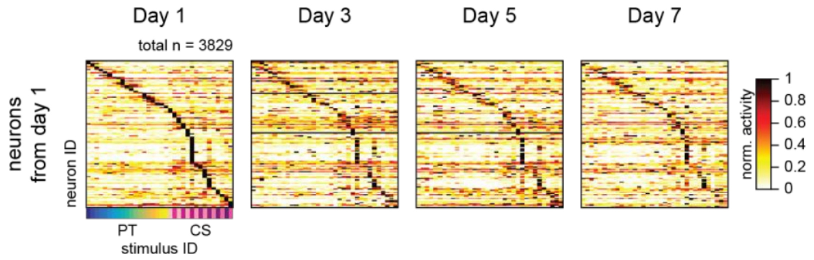


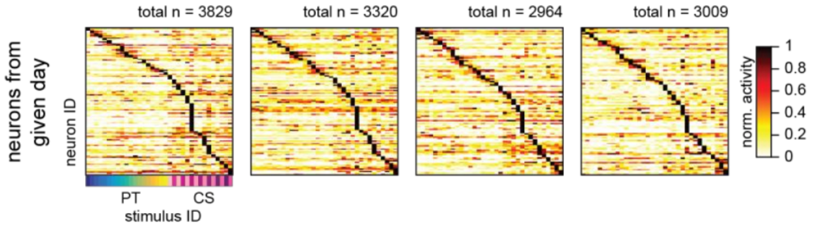
Figure 5.5: Responses to auditory stimuli in single neurons monitored over multiple days (green: mean $\Delta F/F_0$; gray: single trial $\Delta F/F_0$; blue line: stimulus presentation; insets show image of cell on different days; image scale bar 5 µm; trace scale bar 1 s, 250% $\Delta F/F_0$).

auditory world. When pooling the data obtained from all mice and FOVs, we identified on any given day a comparable number of neurons showing significant responses for any of the 34 stimuli used in this study (Figure 5.6, Figure 5.7, controls: Figure 5.8). Moreover, the distribution of preferred stimuli remained stable, as can be seen by comparing the curves of maximal response across days (black 'trace' from top left to bottom right in each panel). However, when considering only those neurons that displayed significant sound-evoked responses on the first imaging day, and following them across days, we observed a progressive blurring of the response profile with time, reflecting the fact that some neurons changed their preferred stimulus or became unresponsive. This process was largely mirrored when considering only those cells with sound-evoked responses on the last imaging day, highlighting neurons that gained responsiveness to sounds during the course of the experiment. We quantified the degree of instability of responses across days by computing the averaged, normalized response amplitude for the preferred stimulus (i.e. average along the dark trace in each panel). When computed for the stimulus that was driving the neurons' maximal response on that given day, this quantity is 1 and stable, by definition. However, if this analysis is performed on each day for the preferred stimulus from the first day, we found a substantial and continuous loss of average response amplitudes. Again, a symmetric observation was made when normalizing to the preferred stimulus amplitude of the last day. These observations suggest that the ability of the auditory cortex to form representations of sounds is maintained in a dynamic equilibrium at a global level.

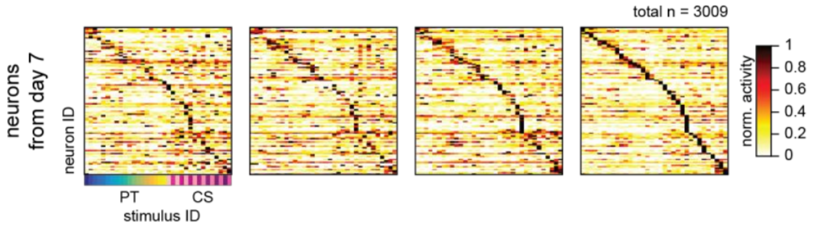
Next, we asked how these changes, observed on the level of individual cells, become manifest on the level of population activity. We



(a) Neurons from day 1.

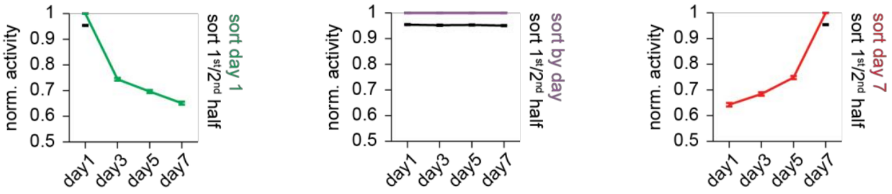


(b) Neurons from day i.



(c) Neurons from day 7.

Figure 5.6: Balanced drifts in tuning at the single-cell level. (a) Normalized response profiles of neurons with a significant response to at least one stimulus on day 1 sorted by stimulus with highest response amplitude. Sorting from day 1 is applied to the subsequent days. N is total number of significantly sound responsive cells on each day. For illustrative purposes, only every thirtieth cell is shown (PT: pure tones, CS: complex sounds). (b) Same as (a), for cells with a significant response on a given day. Sorting is done for each day individually. (c) Same as (a), for cells with a significant response on day 7. Sorting from day 7 is applied to the previous days.



(a) Sorted day 1. (b) Sorted day i. (c) Sorted day 7.

Figure 5.7: Average (mean± SEM) normalized activity to the stimulus with highest response amplitude on day 1 (Figure 5.6a) plotted across days. Estimation of best stimulus is robust against sub-sampling of trials (black; Figure 5.8). (b) Same as (a), for cells with a significant response on a given day. (c) Same as (a), for cells with a significant response on day 7.

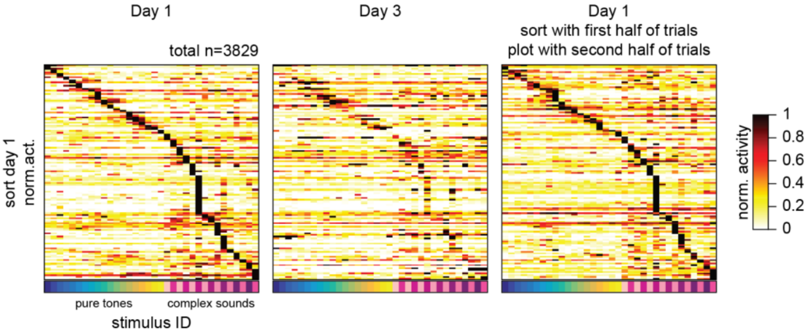
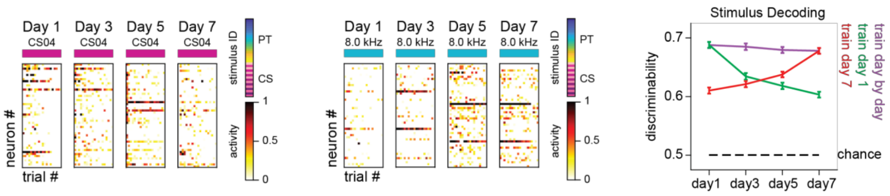


Figure 5.8: Robust categorization of significant sound responsiveness in single cells. Normalized sound response profiles of individual cells responsive on day 1 (left), day 3 (middle) as shown in Figure 5.6a. The panel on the right shows an analogous plot for day 1, however, considering a subsample of the trials. The high degree of similarity between the left and right panels indicates a robust estimation of the best stimulus and the considerate drifts with time cannot be simply explained by noise. N is total number of significantly sound responsive cells on day 1. For illustrative purposes, only every thirtieth cell is shown.

observed that population responses were often stable across several days. However, consistent with our single cell analysis above, sometimes the set of neurons responding to a given stimulus changed from one imaging day to the next (Figure 5.9a, Figure 5.9b, further examples Sup. Fig. 9.1). To assess whether these changes affect the ability of neural populations to stably distinguish between auditory stimuli, we trained a linear classifier to discriminate single-trial activity patterns elicited by different sound stimuli in a given FOV (see Section 5.4.22). When training and testing with activity patterns recorded on the same day, we observed similar performances across different imaging days (Figure 5.9c). However, the impact of the changes in the population response on stimulus discrimination became particularly evident when training the classifier on sound responses from the first day and testing the performance with activity patterns from the following imaging days. The decoding performance decreased monotonically with an increasing interval between training and testing. Again, we made a symmetrical observation when training the classifier on the data from the last imaging day and testing with sound responses from previous imaging days. Consistent observations were made using an alternative, multi-class decoding approach (Figure 5.10a). Note that the representations of sounds at the level of local populations of neurons varied across the FOVs imaged within a mouse, such that a robust representation emerged at a global scale (Figure 5.10b). Thus, extending the analysis to the level of neuronal populations, we found that the ability to decode sounds from the auditory cortex was largely robust against the ongoing remodeling of sound-evoked response patterns.

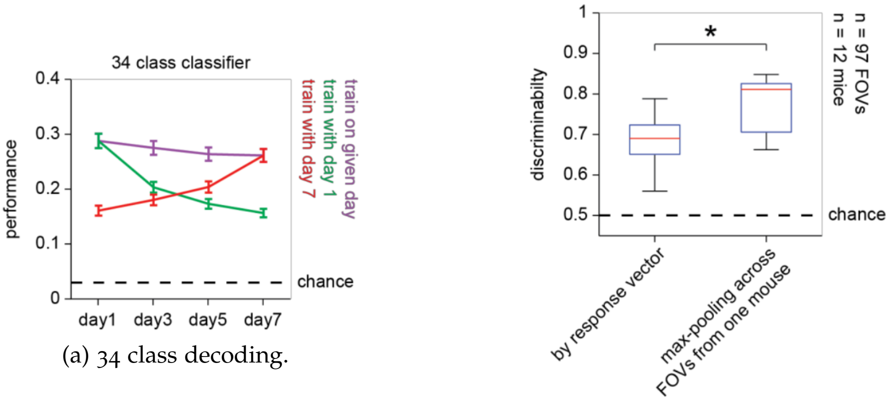


(a) Example response A.

(b) Example response B.

(c) Decoding.

Figure 5.9: Population responses to auditory stimuli are dynamic under basal conditions. (a)-(b) Single trial population response vectors acquired from a given FOV. Examples shown are for two stimuli and two different FOVs over the time course of seven days. For illustrative purposes, only the fifty most active cells are shown and trials are sorted by descending mean activity (PT: pure tones, CS: complex sounds). (c) Linear discriminability calculated by logistic regression averaged across all possible sound pairs and FOVs (mean ± SEM) plotted across days. The classifier was trained with data from either first (green), last (red), or given (purple) imaging day. Dashed line indicates chance level.



(a) 34 class decoding.

(b) Max-pool. decoding.

Figure 5.10: Further decoding analyses. (a) Linear discriminability per field of view (34-fold classifier, SVM, corresponding to the 34 stimuli used) plotted across days, averaged over FOVs (mean ± SEM). The classifier was trained with data from either first (green), last (red), or given (purple) imaging day. Dashed line indicates chance level. This result is comparable to that obtained with a pairwise decoding approach shown in Figure 5.9c. (b) Max-pooling of decoding performance across FOVs within individual animals significantly increases the decoding performance. This indicates that a high-fidelity sound representation can be obtained at the global level. Red line indicates median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively, whiskers represent range (* $p = 0.0111$).

5.2.3 *A discrete set of response modes forms a population-level representation of the auditory world*

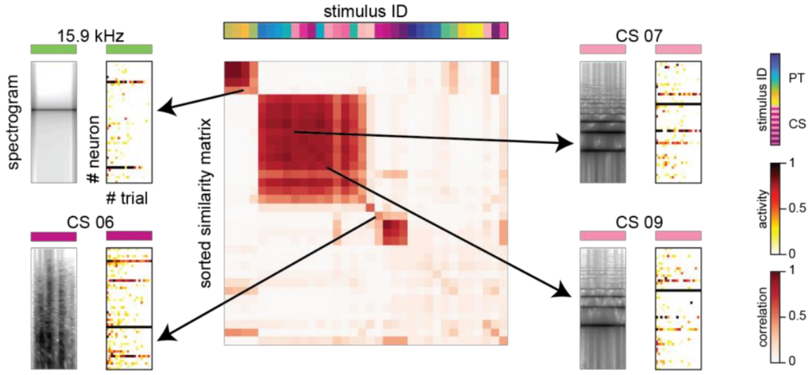
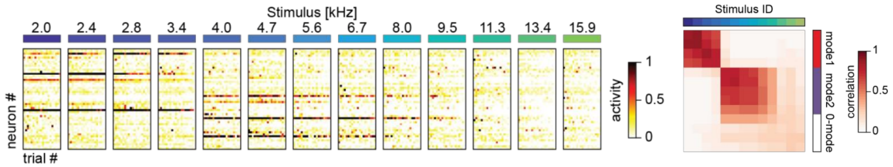


Figure 5.11: Population responses are clustered. Similarity matrix computed as average pairwise correlation of single-trial population response vectors sorted by hierarchical clustering for an example FOV. Diagonal entries show the average correlation across pairs of trials elicited by the same stimulus; off-diagonal entries show the average correlation across pairs of trials from different stimuli. Stimulus identity is shown above. Inlays for four example stimuli: spectrograms of stimuli and single trial population response vectors. In response vectors, for illustrative purposes, only fifty most active cells are shown, and trials are sorted by descending mean activity (PT: pure tones, CS: complex sounds).

Having observed a considerable degree of ongoing remodeling, not only of tuning properties in individual neurons, but also of neural representations on the population level, we next sought to capture the essence of these changes more efficiently. To this end, we exploited the fact that on a microcircuit scale sound-evoked activity patterns fall into a near-discrete set of response modes (Figure 5.11, Bathellier et al., 2012, See et al., 2018).

This phenomenon is illustrated for an example FOV in Figure 5.12a, where we probed the population response with a set of pure-tones with gradually changing frequency. Instead of a gradual change of the response pattern with frequency, we observed an abrupt, highly non-linear transition from one response pattern to another at 4 kHz and a rapid change towards no significant response at frequencies above 11.3 kHz. In most FOVs we observed a similar tendency of groups of stimuli to show stereotypic responses, while the composition of groups tended to vary across FOVs and often comprised a mix of both pure tones and complex sounds.

To assess objectively the response modes from the population activity vectors in a given FOV on a given day we extended the methods of Bathellier et al., 2012. In short, we used Pearson correlation (Galton,



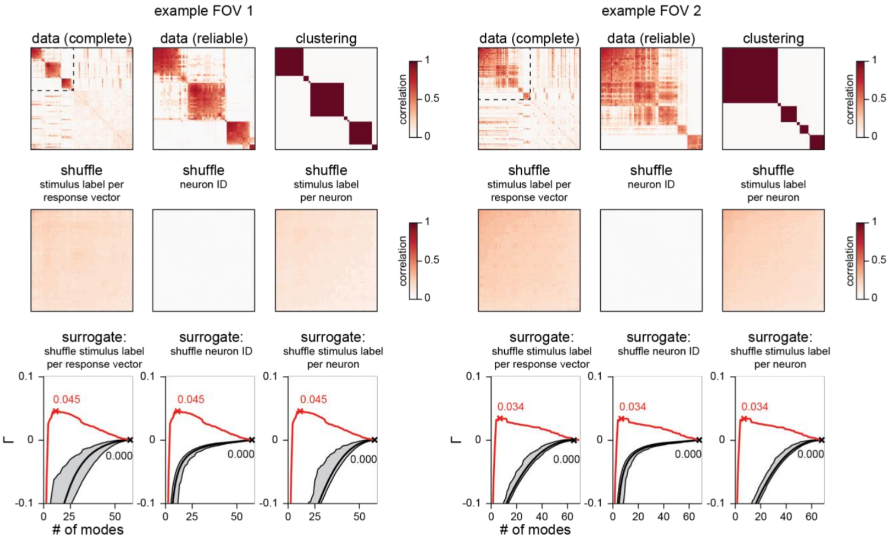
(a) Stimulus responses.

(b) Similarity matrix.

Figure 5.12: Abrupt transitions in population responses elicited by gradually changing stimuli delineate discrete response modes. (a) Example population activity from a FOV showing non-linear response mode transition to interpolation of pure-tone frequencies from 2 kHz to 15.9 kHz. Top: Stimulus identity; Bottom: Single trial population response vectors. For illustrative purposes, only fifty most active cells are shown and trials are sorted by descending mean activity. (b) Similarity matrix (Pearson correlation) from the example FOV shown in (a). Top: Stimulus identity; Right: Response mode identity, 0-mode: no significant population response.

1886, Pearson, 1895) as a measure of similarity between response vectors, which emphasizes the contribution of cells with large $\Delta F/F_0$ and is less affected by the many cells whose signal change is indistinguishable from noise. We constructed a similarity matrix by calculating pairwise correlations of single-trial response vectors (Figure 5.12b). The entries along the diagonal were calculated as the average correlation of all pairwise combinations of response vectors elicited by the same sound, thus reflecting the reliability of the response pattern elicited by a given sound. Correlation was low for sounds that did not evoke any activity in the particular FOV. Off-diagonal entries were calculated as the average correlation of all pairwise combinations of single-trial response vectors elicited by a pair of two different sounds. Then, the similarity matrix was sorted using hierarchical clustering, thereby grouping the sound stimuli by the similarity of their respective response patterns. We used Hubert's Γ statistics (Hubert and Baker, 1977) to estimate the number of clusters and to assess significance (see Section 5.4.20). Each cluster of response patterns, often elicited by different sets of sound stimuli, was defined a *response mode* (Figure 5.11, Figure 5.13). Stimuli that did not elicit a reliable response pattern in a given FOV were not considered for clustering and instead grouped together into the 0-*mode*.

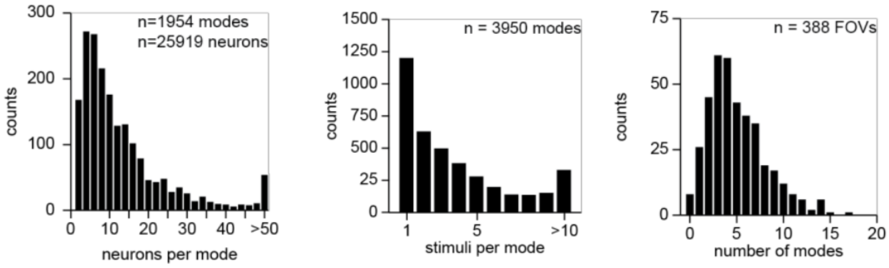
The composition of the subsets of sound stimuli mapped to a shared response mode varied across FOVs and often comprised a mix of both pure tones and complex sounds. In a given FOV we typically observed between 3 and 9 different response modes, on average associated with 2.40 ± 0.21 (mean ± SEM) sound stimuli and comprising up to 10 significantly active neurons (for histograms see Figure 5.14,



(a) Clustering example A.

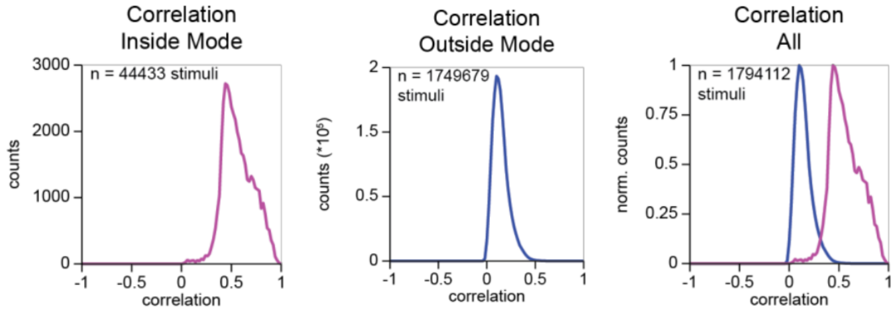
(b) Clustering example B.

Figure 5.13: Statistical identification of response modes using clustered similarity matrices of response vectors. (a), (b) Examples for complete (data pooled from all imaging days) clustered similarity matrices (left), magnifications of the part of the matrix representing reliable responses (center) and validated clusters (i.e. response modes) based on Hubert's Γ statistics (right; Hubert and Baker, 1977). Stimuli not evoking a reliable response on a given day were mapped to the 0-mode. Note that when pooling the data across days, each mode is now defined for all imaging days and may contain responses to the same stimulus from different days. This scheme allowed us not only to track how individual stimuli transition between different modes, but also to detect modes that become void of stimuli (referred to as disappearing modes) and modes that transition from void to containing stimuli (referred to as appearing modes). Middle three panels show clustered similarity matrices constructed from surrogate data generated by shuffling prior to the correlation analysis (shuffling from left to right: stimulus labels of population response vectors; neuron identity on individual trials; stimulus labels of individual neurons). Bottom three panels show Hubert's Γ assuming different number of modes for data (red lines, x denotes maximum Γ) and the three methods of shuffling the data (black lines, grey area indicates minimum and maximum, x denotes maximum Γ).



(a) Hist. neurons/mode.
(b) Hist. stim./mode.
(c) Hist. mode number.

Figure 5.14: Quantitative description of response modes. (a) Histogram of significantly active neurons per mode ($n = 1,954$ modes, $n = 25,919$ neurons). (b) Histogram of stimuli per mode in a given FOV ($n = 3,950$ modes). (c) Histogram of number of modes per FOV ($n = 388$ FOVs).



(a) Histogram inside.
(b) Histogram outside.
(c) Both histograms.

Figure 5.15: Histogram of average pairwise correlations of trials inside a response mode and outside a response mode. (a) and (b) show absolute counts, and (c) shows normalized counts for comparison.

Figure 5.15). Conveniently, response modes provide a massively reduced and simplified description of sound-evoked activity patterns and we will make use of this in the following sections. At first sight, this type of representation may seem to hamper the capacity of the cortex to discriminate among different stimuli. Note, however, that the mapping of the various sounds to particular response modes varied across different FOVs, supporting a combinatorial code at a more global scale (Figure 5.10b).

In summary, our analysis of the data based on individual imaging days corroborates previous reports suggesting a functional layout of the superficial layers of mouse auditory cortex with scattered and partially overlapping cell assemblies that are driven in a non-linear manner by different groups of sound stimuli (Bathellier et al., 2012, See et al., 2018).

5.2.4 *Ongoing recombination of sensory representations during basal conditions*

Having observed that the structure of auditory representations in a FOV can be well approximated by a small set of response modes, we exploited this highly reduced, non-linear description to capture main aspects of the ongoing representational changes during basal conditions. Following the response modes in the FOV from Figure 5.11 across time, we found evidence for stimuli to transition between

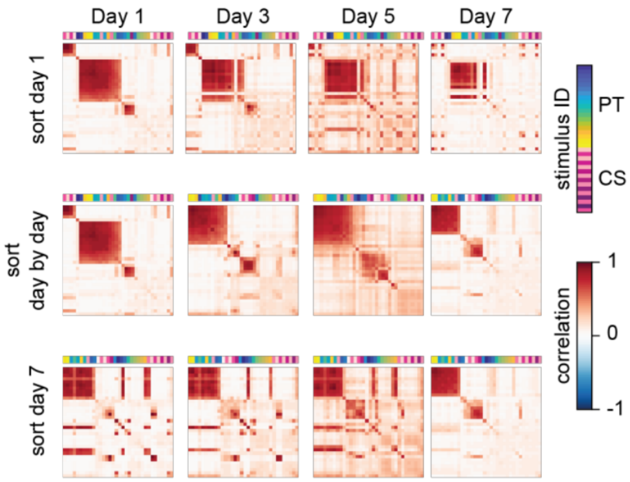
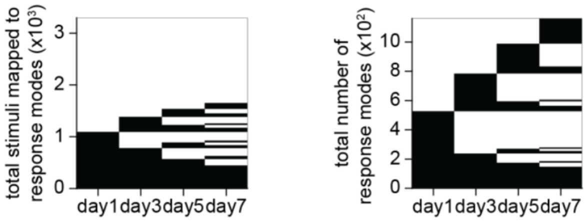


Figure 5.16: Recombination of response modes over the course of days. Top: Single day similarity matrices of sound evoked responses from an example FOV sorted by hierarchical clustering. Sorting from day 1 is applied to the subsequent days (PT: pure tones, CS: complex sounds). Middle: Same as above but sorted on each day individually. Bottom: Same as above but sorting from day 7 is applied to the previous days.

modes, sometimes leading to the disappearance of an existing or the occurrence of a new mode. The latter is illustrated in Figure 5.16 (middle row) by the similarity matrices of stimulus responses, sorted separately on each imaging day using hierarchical clustering (see Section 5.4.20): Here, three modes that were present during the first two imaging days transitioned to only two modes on the last two imaging days, differing in size and stimulus composition. This change in stimulus composition is better seen when applying an identical sorting to all matrices from different days (using either the sorting of the first day (top row) or of the last day (bottom row). While a considerable fraction of stimuli remained stable, some stimuli dropped out of their response mode, as, for instance, indicated by the 'white gaps' in the large response mode on day 3, 5, and 7 in the top row and by the 'red lines' on day 1, 3, and 5 in the bottom row. For more examples see Sup. Fig. 9.2.



(a) Stimulus responses.

(b) Response modes.

Figure 5.17: Response mode changes across the imaging time period. (a) Life time plot of the mapping of stimuli to a specific response mode. The mapping was assessed for each stimulus (34) for the response modes identified in each FOV (97) resulting in a maximal number of 3,298 mappings. A thin horizontal black line indicates a significant response on a given day, data from 12 mice. (b) Life time plot of the total number of response modes that were identified in a given FOV. A thin horizontal black line indicates presence of a response mode on a given day, data pooled over all 97 FOVs and 12 mice.

To systematically analyze these ongoing changes in the response mode structure, we computed the full set of relevant response modes in a FOV by clustering simultaneously all responses from all imaging days (see Section 5.4.20 for details). Note that each mode is now defined for all imaging days and may contain responses to the same stimulus from different days. This scheme allowed us not only to track how individual stimuli transition between different modes, but also to detect modes that become void of stimuli (referred to as disappearing modes) and modes that transition from void to containing stimuli (referred to as appearing modes). We then aggregated this information over all imaged populations. Figure 5.17a shows for each of the 34 stimuli probed in each of the 97 FOVs whether a stimulation elicited a response (black), or no response (white) ($34 \times 97 = 3,298$ stimulations in total). We found that almost half of the stimulations elicited a response over the course of a week, but only a third of them on any single day. Individual stimulations gained (transitions white to black), lost (black to white) and sometimes regained responses. Likewise, whereas the number of response modes stayed roughly the same across days, a considerable fraction of modes disappeared from one imaging day to the next, while others appeared (Figure 5.17b), and some previously present modes reappeared. Of all stimuli being mapped to specific response mode on a given day, only half remained in that mode on the following imaging day, whereas almost 20% moved to a different response mode and more than 30% to the 0-mode, i.e. did no longer elicit a population response (Figure 5.18). This dynamics was largely balanced, as almost 30% of stimuli mapped to a response mode on a given day, did not elicit a population response on the previous imaging day. In summary, we found changes in the mapping of sounds to response modes, as well as changes in the response modes themselves, while the average number of stimuli being mapped to a response mode and the total number of response modes remained fairly stable across imaging days.

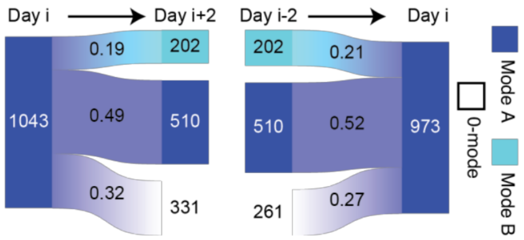
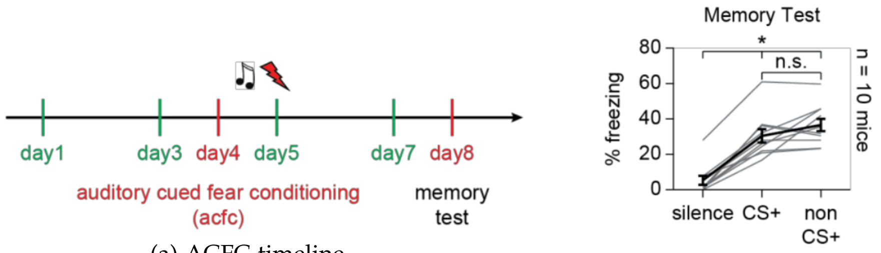


Figure 5.18: Flow chart of response mode dynamics. Left: Development of population responses present on day i (defined as Mode A) two days into the future (Mode A, Mode B, or 0-mode). Right: Development of population responses present on day i (defined as Mode A) from two days ago (Mode A, Mode B or 0-mode). Numbers are counts and fractions averaged across transitions.

5.2.5 *The impact of learning on the dynamics of sensory representations*

We next wondered to what extent behaviorally relevant experiences that trigger the formation of a memory to a sound would impact the long-term dynamics of sensory representations in mouse auditory cortex (Aschauer and Rumpel, 2018). To this end, we acquired a second dataset (10 mice; 74 FOVs; 16,882 neurons) with four imaging time points at a two-day interval using the same set of sound stimuli as before (Figure 5.19a). On the day between imaging sessions two and three, mice underwent an auditory cued fear conditioning paradigm, in which they learned to associate the presentation of a sound with the subsequent application of a mild foot shock. A complex sound from the stimulus set was chosen as conditioned stimulus (CS+). It has previously been shown that auditory cued fear conditioning and variants thereof induce specific changes in gene expression (Peter et al., 2012, Cho et al., 2017), affect the dynamics of synaptic connections (Maczulska et al., 2013, Yang et al., 2016, Lai et al., 2018) and induce changes in the tuning of neurons in mouse auditory cortex (Quirk et al., 1997, Weinberger, 2004, Gillet et al., 2017, Dalmay et al., 2019).

To test for the successful formation of a memory at the behavioral level, we exposed mice again to the sound cue used for conditioning in a neutral context after the last imaging session and scored freezing behavior as a readout of fear-related memory recall. As expected, we observed low freezing levels during silence and significantly increased freezing during presentation of the conditioned sound. Furthermore, we also observed high freezing levels during the presentation of a second sound stimulus that was not presented during the conditioning session (non-CS+), indicating a high level of generalization (Figure 5.19b; $n = 10$ mice; * $p \leqslant 0.0001$ for silence and CS+, * $p \leqslant 0.0001$ for silence and non-CS+). This high level of generalization is typically observed following classical fear conditioning unless specific differential conditioning paradigms are utilized (Letzkus et al., 2011).



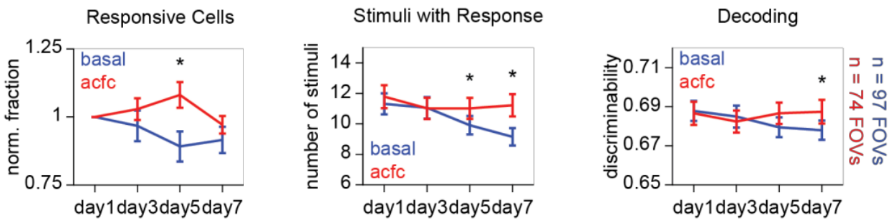
(a) ACFC timeline.

(b) Memory test.

Figure 5.19: Auditory cued fear conditioning paradigm. (a) Experimental timeline for cohort of mice undergoing auditory cued fear conditioning (ACFC). Dataset comprises $16,882$ cells from 10 mice. (b) Increase in freezing behavior for conditioned CS+ and high level of generalization to another non-conditioned sound in a memory test four days after fear conditioning. Gray lines depict behavior of individual animals and black line is mean ± SEM of all animals.

We compared the fraction of sound-responsive cells in both datasets and found only a transient increase in the first imaging session after conditioning (Figure 5.20a; control: $n = 97$ FOVs; ACFC: $n = 74$ FOVs; * $p \leqslant 0.0001$). In addition, we observed, that the average number of stimuli that evoked a population response in a given FOV, i.e.

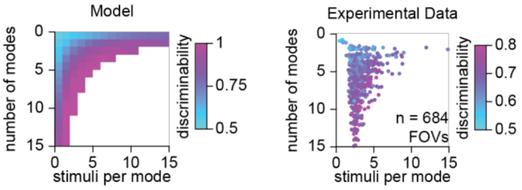
were mapped to a response mode, was significantly increased following conditioning compared to control. This suggests that sounds gain a longer-lasting, broader representation across the auditory cortex (Figure 5.20b; * $p = 0.0095$, day 7 * $p \leqslant 0.0001$). Parallel to this, we observed that the efficacy to decode sounds from population response vectors recorded within individual FOVs was increased following conditioning in comparison to the control group (Figure 5.20c; * $p \leqslant 0.05$). In accordance with previous studies, we observed a slight decrease in these measures during baseline periods for both groups, presumably due to continuing habituation (Kato et al., 2015). Interestingly, this effect was reversed for the conditioned cohort, showing a significant increase in decoding performance at time points after the conditioning session.



(a) Responsive cells. (b) Stimulus responses. (c) Decoding performance.

Figure 5.20: Auditory cued fear conditioning increases decoding performance. (a) Normalized fraction of sound responsive cells over the imaging days averaged over FOVs (mean ± SEM). (b) Average number of stimuli eliciting a population response in a given FOV (mean ± SEM). (c) Decoding performance (pairwise logistic regression, training and test data from same day) in baseline and ACFC dataset.

To better understand this improvement of decoding performance upon fear conditioning, we next studied the induced changes in neural activity in terms of the response mode dynamics introduced above. To gain intuition, we first studied a simple model, in which we considered the case that stimuli are encoded in a local neuronal population by distributing these stimuli evenly among a given number of modes (see Section 5.4.24 for details). In this model, both the total number of stimuli and number of response modes available determine how efficiently pairs of sounds can be decoded on average (Figure 5.21a). In the parameter regime consistent with our experiments, i.e. a considerably large 0-mode and about 5 response modes with $2-3$ stimuli per mode, the average decoding performance can improve not only by increasing the number of response modes, but also by increasing the number of stimuli per response mode. Note that the latter is a direct consequence of the fact that the 0-mode is relatively large; a stimulus, once it has left the 0-mode, can be distinguished from all stimuli in the 0-mode. In our data, we indeed observed a similar relationship between decoding performance and number of response modes and stimuli per mode (Figure 5.21b).

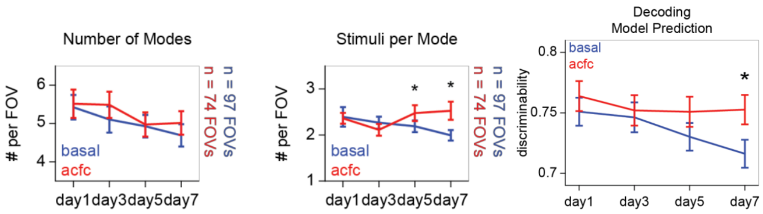


(a) Model. (b) Experiment.

Figure 5.21: Model of decoding performance based on response modes. (a) Model of decoding performance based on response modes suggests better performance with increased number of modes N_m as well as increased average number of stimuli per response mode L_m . White region of the matrix reflects the condition $N_m \times L_m \leqslant N_{stim}$ (total number of stimuli $N_{stim} = 34$). (b) Experimentally observed decoding performance from data of individual FOVs plotted as function of number of response modes and mean number of stimuli per response mode.

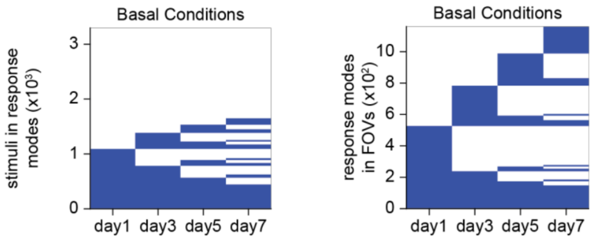
Therefore, we next asked whether the increase in decoding performance that we observed upon fear conditioning (Figure 5.20c) was due to an increase in the average number of response modes, or an increase in the average number of stimuli per mode, or both. Intriguingly, whereas the average number of modes per FOV was indistinguishable between the baseline and fear condition cohorts during all imaging days (Figure 5.22a; control: $n = 97$ FOVs; ACFC: $n = 74$), the average number of stimuli per response mode was significantly larger following conditioning (Figure 5.22b; day 5: * $p = 0.0021$; day 7: * $p \leqslant 0.0001$). Consistently, feeding the observed increase in stimulus number per response mode back into our model, it was able to account for the experimentally observed increase in sound decoding (Figure 5.22c).

Both datasets in our study were dominated by a substantial degree of representational changes (Figure 5.23). However, the analysis of the response mode dynamics revealed that learning induces a bias in the ongoing recombination of sensory representations compared to basal conditions. Specifically, changes that increase the number of stimuli being mapped to a response mode occurred more frequently, thereby mediating an improvement in sound decoding at the level of individual FOVs.

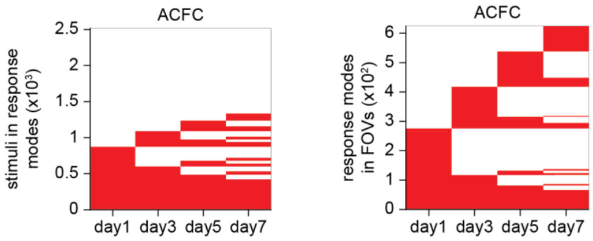


(a) Mode number. (b) Mode size. (c) Model prediction.

Figure 5.22: Improved decoding after auditory cued fear conditioning is mediated by an increase in the number of stimuli mapped to a response mode. (a) Mean number of response modes averaged over FOVs (mean± SEM). (b) Mean number of stimuli per response mode averaged over FOVs (mean ± SEM). (c) Predicted increase in decoding performance based on the model from Figure 5.21a when considering changes in number of modes and average number of stimuli per mode as observed in (a) and (b). Baseline: $n = 97$ FOVs, ACFC: $n = 74$ FOVs, Bootstrap test: * $p \leqslant 0.001$.



(a) Basal stim. responses. (b) Basal resp. modes.



(c) ACFC stim. responses. (d) ACFC resp. modes.

Figure 5.23: Further quantification of response mode dynamics. (a), (b) Reproductions of Figure 5.17a and Figure 5.17b, illustrating the dynamics in the data set acquired under basal conditions, shown here for comparison. (c), (d) Same as (a) and (b) for dataset from cohort of mice undergoing fear conditioning between day 3 and day 5. Note that the dynamics observed under basal conditions are also dominant in the fear conditioning dataset. The specific learning-induced changes weave into this ongoing dynamics.

5.2.6 *Fear conditioning drives the formation of associations between sensory representations and stabilizes them*

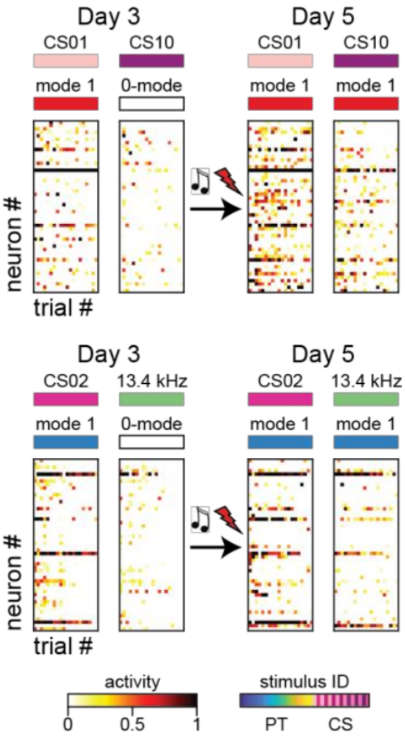
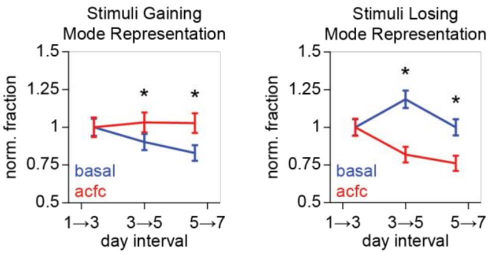


Figure 5.24: Examples from two FOVs showing the association of two stimuli following auditory cued fear conditioning by the co-mapping onto a shared response mode. For illustrative purposes, only the fifty most active cells are shown.

Notably, increasing the number of stimuli that map to a response mode may not only improve decoding, but could also indicate the formation of new associations (Figure 5.24). In fact, the activation of a shared cell assembly by different stimuli has been interpreted as the formation of an association among sensory representations at the microcircuit level (Cai et al., 2016, Grewe et al., 2017). Consistent with this idea, in the memory test session following fear conditioning, we observed a high level of behavioral generalization despite improved encoding (Figure 5.19b). Generalization suggests an association of the conditioned stimulus to other, non-conditioned stimuli during the memory test. We therefore wondered whether the above-described increase in response mode size upon fear conditioning is due to an enhanced stabilization of existing representations, or the genuine formation of new associations.

Under basal conditions, we observed that stimuli disappeared from a given response mode, by a rate that was almost balanced by new stimuli being added to it (Figure 5.18). Intriguingly, we found that during learning both of these rates were shifted: the rate of stimuli being added to a mode per day increased, relative to the baseline dynamics (Figure 5.25a; control: $n = 97$ FOVs; ACFC: $n = 74$ FOVs; transition $3 \rightarrow 5$: $^{*}\ p \leqslant 0.0001$; transition $5 \rightarrow 7$: $^{*}\ p \leqslant 0.0001$) while the rate of stimuli disappearing (and entering the 0-mode) decreased (Figure 5.25b; transition $3 \rightarrow 5$: $^{*}\ p \leqslant 0.0001$; transition $5 \rightarrow 7$: $^{*}\ p \leqslant 0.0001$). Note that both processes effectively increased the average number of stimuli that are mapped



(a) Response gain. (b) Response loss.

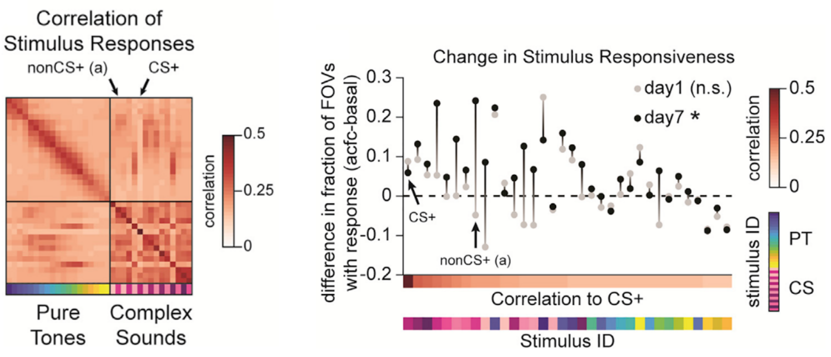
Figure 5.25: Increased mapping of individual stimuli to a shared response mode as a mechanism to form an association. (a) Normalized fraction of stimuli gaining a response mode representation averaged over FOVs (mean ± SEM). (b) Same as (a) for stimuli losing a response mode representation.

to a response mode. The former suggests an increase in the formation of new associations among stimuli, while the latter suggests the stabilization of existing associations. Thus, by the same mechanism, increasing the number of stimuli mapped to a response mode, fear conditioning can improve decoding performance and foster the formation of new associations as well as stabilizing existing ones.

5.2.7 *ACFC increases population responses for stimuli with representational similarity to the conditioned sound*

Next, we sought to investigate stimulus specific effects. We wondered to what extend the response increase observed in Figure 5.20b is specific to stimuli similar to the conditioned stimulus (CS+). Similarity between stimuli is easy to determine for pure tone stimuli, as they can be ordered along one dimension and through tonotopy in the auditory they are also perceived in that way. But what about complex sounds? We used representational similarity in the auditory cortex as a measure of stimulus similarity and followed an approach to take the correlation between stimulus evoked activity patterns as a proxy of perceived pairwise similarity (Kriegeskorte et al., 2008). To this aim we averaged the pairwise correlation of response vectors for all sound stimuli recorded on the first imaging day across all FOVs from both datasets (Figure 5.26a). As expected resonses to pure tone stimuli were similar along the diagonal indicating that close-by frequencies evoked similar responses, but this method also revealed the perceived similarity between any two complex stimuli and between complex stimuli and pure tones.

We used the representational similarity in order to sort stimuli by their similarity to the CS+ and counted the number of FOVs, where the respective stimuli evoked a response on day one, i.e. prior to fear



(a) Represent. similarity.

(b) Change in responsiveness.

Figure 5.26: Representational similarity to CS+ is predictive of increase in responsiveness. (a) Correlation matrix of sound response vectors for all stimuli averaged across all FOVs from both datasets on day 1. Arrows mark columns representing the stimulus used for fear conditioning (CS+) and the non-conditioned stimulus presented during the memory test session (non-CS+ (a)). (b) Difference between basal and fear conditioning groups in the fraction of FOVs in which a population response to a given stimulus was observed. Gray dots mark values for baseline (day 1) and black dots after conditioning (day 7). Stimuli are sorted on the x-axis by descending similarity to the stimulus used for fear conditioning. The correlation color bar represents the correlation of population response vectors to the CS+ (see CS+ column) in (a). Arrows mark the stimulus used for fear conditioning (CS+) and the non-conditioned stimulus (non-CS+ (a)) presented during the memory test session. * Spearman's rank correlation, $\rho = 0.66$, $p < 0.0001$.

conditioning, and on day seven, i.e. after fear conditioning. Prior to conditioning we found no significant difference between the ACFC and control datasets (Figure 5.26b, Spearman's $\rho = 0.32$, $p = 0.07$). In contrast, after conditioning, we found that stimuli were more likely to elicit a response in the conditionied group, the more similar they were to the CS+ (Figure 5.26b, Spearman's $\rho = 0.66$, $p < 0.0001$). Notably, responses to the CS+ itself seemed to be barely affected by conditioning.

5.2.8 *Fear conditioning drives stimulus specific formation of associations between sensory representations, predictive of behavioral generalization*

Generalization is believed to be linked to the association of the conditioned stimulus (CS+) and non-conditioned stimuli (non-CS+) during conditioning (Pavlov and Anrep, 1927, Dunsmoor and Paz, 2015). On the microcircuit the activation of a subgroup of neurons by different stimuli has been interpreted as such an association between those stimuli (Grewe et al., 2017). Having established a differential effect of fear conditioning on population activity evoked by stimuli other than the CS+ depending on the representational similarity of the respective stimulus to the CS+, we next wanted to leverage our response mode framework to investigate this some more. In our framework

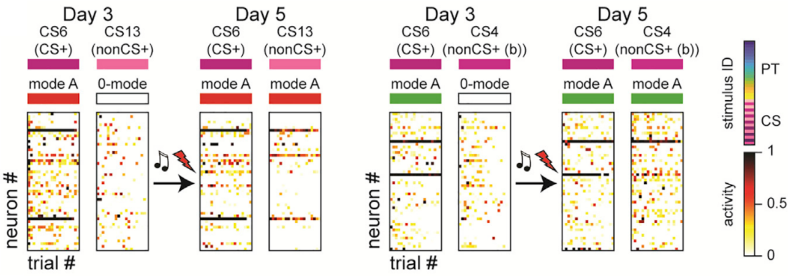
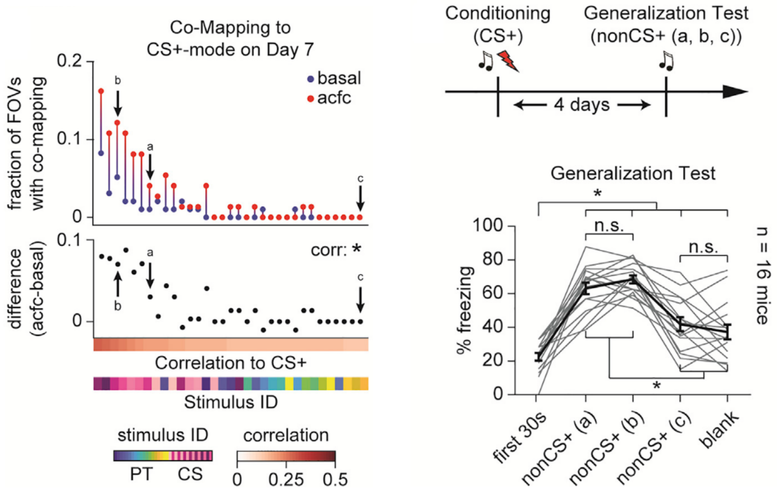


Figure 5.27: Examples from two FOVs showing the responses of the conditioned stimulus (CS+) and a non-conditioned stimulus (non-CS+). Prior to fear conditioning the non-CS+ did not elicit a significant response (0-mode), whereas after fear conditioning its response became similar to that of the CS+ (mode A). Top: stimulus identity; middle: mode identity; bottom: single trial population response vectors. For illustrative purposes, the 50 most active cells are shown in random order and trials are sorted by descending mean activity (PT, pure tones; CS, complex sounds). For further examples, see Sup. Fig. 9.3.

two stimuli evoking the same response mode can be interpreted as an association. As already shown above, an increasing number of stimuli evoking a response (Figure 5.20b), combined with no increase in the number of response modes (Figure 5.22a), and thus an increase of the number of stimuli being mapped to each mode (Figure 5.22b), clearly display an increase in the formation of associations. However, as we observed increased activity of stimuli similar to the CS+, we next investigated the co-mapping of stimuli, more precisely we assessed if increased co-mapping of stimuli following fear conditioning is specific to non-CS+ sound stimuli similar to the CS+. Exmples are shown in Figure 5.27 (for further examples see Sup. Fig. 9.3). In the basal control group we found that stimuli that were more similar to each other also tended to evoke the same response mode more often. This was no surprise, as both measures of population activity are related to each other. Indeed, the correlation was used to define the clustering into response modes. However, when comparing the two datasets, we found that the likelihood of co-mapping to the CS+ was more pronounced for non-CS+ with a higher representational similarity to the CS+ in the cohort of mice that underwent conditioning (Spearman rank correlation $\rho = 0.73$, $p < 0.0001$, Figure 5.28a). So, fear conditioning led to an increase of associations between the CS+ and those non-CS+ sounds that already showed a high level of representational similarity prior to the conditioning.

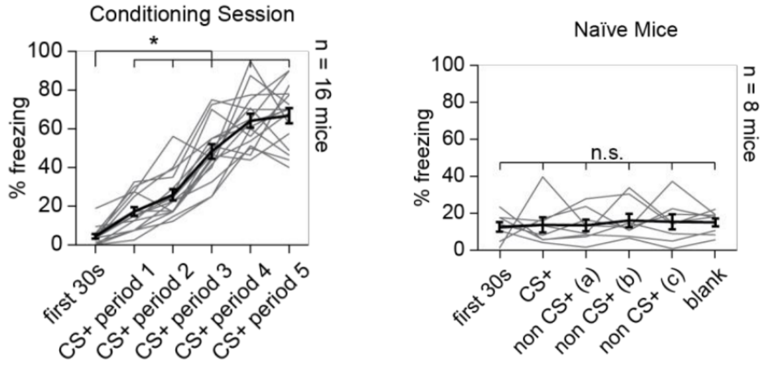


(a) Co-mapping with CS+.

(b) Generalization.

Figure 5.28: Co-mapping with the CS+ is predictive of behavioral generalization. (a) Top: for each stimulus, sorted on the x-axis by descending similarity to CS+, the plot shows the fraction of FOVs this stimulus is co-mapped to the response mode of the CS+ on day 7 in experimental groups with (ACFC, red) and without (basal, blue) fear conditioning. The correlation color bar represents average correlation of population response vectors to the CS+ (see Figure 5.26a). Arrows mark non-conditioned stimuli presented during the memory test session in (b) (non-CS+ (a, b, c)). Bottom: differences between fractions for ACFC and basal. * Spearman's rank correlation, $\rho = 0.73$, $p < 0.0001$. Note a strong increase in co-mapping after fear conditioning specifically for stimuli with larger representational similarity to the CS+ prior to conditioning. (b) Top: experimental time line of behavioral experiment. During conditioning the same complex sound was paired with the unconditioned stimulus. During the generalization test, mice were exposed to three non-conditioned sound stimuli. Bottom: increase in freezing behavior in a test session for three non-conditioned stimuli (two with high-response correlation to the CS+ (non-CS+ (a, b)) and one with low response correlation to the CS+ (non-CS+ (c)). Freezing to non-CS+ (c) is not different to time periods without the presentation of a stimulus (blank). Gray lines depict behavior of individual animals and the black line is the mean ± SEM of all animals. * One-way ANOVA with correction for multiple comparisons, first 30 s versus non-CS+ (a, b, c) and blank: $p < 0.05$, non-CS+ (a, b) versus non-CS+ (c) and blank: $p < 0.0001$.

Can these associations on the level of cortical populations be linked to generalization on the behavioral level? To test this, we performed another fear conditioning experiment, during which mice were conditioned to the same complex sound stimulus as above, before we used three different non-CS+ to probe for a differential stimulus generalization 4 days later. In addition to the sound used in our previous experiment (non-CS+ (a)), we chose two more sounds from our 34 stimuli, one with a high similarity to the CS+ and a high increase of co-mapping (non-CS+ (b)), and one with little similarity to the CS+ and no change in co-mapping (non-CS+ (c)) (Figure 5.28a). None of the stimuli induced freezing in naive mice (Figure 5.29). After conditioning we found significantly increased freezing levels for both the non-CS+ (a) and the non-CS+ (b), but not for the non-CS+ (c) (Figure 5.28b, * $p < 0.0001$ for first 30 s and all other groups, * $p < 0.0001$ for non-CS+ (a, b) and non-CS+ (c) and blank). Together, this indicates that representational similarity is predictive of the level of ACFC induced co-mapping of non-CS+ stimuli onto the same response as the CS+. And this co-mapping in turn is correlated with the degree of behavioral generalization to non-conditioned stimuli.



(a) Conditioning session.

(b) Without ACFC.

Figure 5.29: Freezing during the conditioning session and in naive animals. (a) Freezing behavior of the experimental animals from the memory test in Figure 5.28b during the conditioning session with five consecutive pairings of the CS+ and the US (mild electric shock; * one-way ANOVA with correction for multiple comparisons, first 30 s vs. CS+ period 1: $p < 0.05$, first 30 s vs. CS+ periods 2 to 4: $p < 0.0001$). (b) Freezing behavior in a control experiment with naive mice which were exposed to all stimuli used in the behavior experiments without having previously experienced a CS-US pairing. Freezing levels are low for all stimuli and not different to periods of silence at the beginning of the session (first 30 s) and at interspersed time points throughout the session (blank; n.s.: one-way ANOVA with correction for multiple comparisons).

5.3 DISCUSSION

We studied the long-term dynamics of auditory representations in the cortex. Chronic monitoring of sound-evoked population activity over the course of several days revealed that sensory representations undergo substantial recombination even under environmentally and behaviorally stable conditions. In order to capture structure in the parallel recordings of hundreds to thousands of neurons, changes in population activity are often described using a single, albeit rather abstract metric, such as decoding power. Here, we developed a description of population activity at an intermediate and biologically more interpretable level, specifically at the level of cell assemblies whose non-linear activation forms sensory representations. The identification of response modes based on the non-linear activation of distinct cell assemblies provided a highly reduced and efficient description of the network dynamics. Response modes represent the association of a set of sensory stimuli with the prototypical activation of a specific cell assembly and thus reflect non-linear properties of auditory perception (Liberman et al., 1967). Breaking down the ongoing changes of sensory representations into response modes and changes between response modes we were able to capture the major remodeling, albeit stationary dynamics of population response changes.

What are the driving forces underlying the recombination of sensory representations during basal conditions? As the pattern of connectivity is considered a major determinant for the patterns of activity that can arise in neuronal networks, it appears plausible that ongoing remodeling of synaptic connections could underlie the plasticity we observed in our experiments. Indeed, such basal dynamics in connectivity are observed in the mouse auditory cortex during behaviorally stable conditions without need for adaptation (Loewenstein et al., 2011, Loewenstein et al., 2015). Interestingly, such spontaneous dynamics in synaptic connections even persist during pharmacological blockade of neuronal activity (Yasumatsu et al., 2008, Rubinski and Ziv, 2015, Dvorkin and Ziv, 2016, Nagaoka et al., 2016), and thus appear to represent a fundamental feature of neuronal circuits. More recently, theoretical modeling has been used to investigate how ongoing synaptic plasticity, as it is observed during basal conditions, affects the long-term stability of activity patterns in a network (Kappel et al., 2015, Kappel et al., 2018, Mongillo et al., 2018, Humble et al., 2019, Susman et al., 2019).

Increases in local population coding efficacy induced by behavioral training were observed in several sensory cortical systems (Huber et al., 2012, Poort et al., 2015). We observed a similar increase in coding efficacy in our dataset (Figure 5.20c) that could be explained by a selective increase in the number of stimuli per response mode (Figure 5.22b). However, global coding efficacy is typically sufficiently high to explain sensory discrimination even before training, hinting that this increase could reflect also other representationally relevant processes induced by learning. Our analysis indicates that auditory fear conditioning also specifically biases the dynamics of sensory representations leading to an increased mapping of different stimuli onto a shared cell assembly. This increased formation of associations between sensory representations is consistent with the observed high level of behavioral generalization.

Auditory cued fear conditioning led to a differential co-mapping of stimuli onto the same local response as the conditioned stimulus (CS+) dependent on their a priori representational similarity to the CS+. This co-mapping was correlated to a behavioral generalization. While the balance of discrimination and generalization of stimuli is essential for a successful navigation in a complex and changing environment, a surplus of generalization has been linked to diseases such as post-traumatic stress disorder (Besnard and Sahay, 2016). In contrast to earlier work, where mostly simple stimuli (like pure tones) were used, which can be modified in a one-dimensional way, we used complex stimuli and the degree of their representational similarity as a proxy for perceived similarity. This allowed us to infer the learning-induced changes of sound evoked activity patterns. This is in line with previous findings in which the representational similarity at the level of the auditory cortex was used as a neurometric measure to predict behavioral categorization of stimuli in a discrimination task (Bathellier et al., 2012). The generalization from CS+ to non-CS+ stimuli is believed to be based on perceptual features of the stimuli, but it is to date unclear, what is the involved circuit mechanism (Dunsmoor and Paz, 2015, Boddez et al., 2021). We observed an increased co-mapping of non-conditioned stimuli to the same local groups of neurons that are evoked by the CS+, consistent with the role auditory cortex plays in stimulus generalization (Thompson, 1962, Armony et al., 1997, Aizenberg and Geffen, 2013). Increased co-activation of shared neuronal subgroups by different stimuli has been reported at the level of the amygdala (Grewe et al., 2017) as well es in hippocampal ensembles (Cai et al., 2016). Together with our findings this suggests a close link between behavioral generalization and representational associations.

Interestingly, the recombination of representations, i.e. their association as well as their dissociation, ensues to a substantial degree also during basal conditions and it remains a matter of research to investigate their functional relevance (Chambers and Rumpel, 2017). We speculate that this ongoing dynamics of cell assemblies could support cognitive processes that occur without explicit mental engagement, like the spontaneous creation of associations (Wallas, 1926) or their forgetting (Richards and Frankland, 2017).

5.4 METHODS

5.4.1 *Molecular cloning*

For the generation of a recombinant AAV (rAAV) genome encoding for GCaMP6m under the human SynapsinI promoter (phSyn), a plasmid containing the inverted terminal repeats (ITRs) of AAV, phSyn (Addgene plasmid 26973), Woodchuck Hepatitis Prottranscriptional Regulatory Element (WPRE), and a human Growth Hormone polyadenylation site (hGH-pA site) was digested using BamHI and AccIII and the gene coding for GCaMP6m was PCR amplified from a commercially available plasmid (Addgene plasmid 40754) and inserted. Finally, the plasmid was digested with AccIII and HindIII to excise the original transgene and 3'overhangs were blunted and 5'overhangs were filled in using Klenow fragment.

For the generation of a recombinant AAV genome encoding for H2B-mCherry fusion protein under the phSyn, a gene coding for mCherry was PCR amplified and inserted into a plasmid containing a gene for H2B directly after its coding sequence using ClaI and SpeI to produce a fusion gene. The H2B-mCherry fusion gene was PCR amplified and inserted into a plasmid containing ITRs, phSyn, WPRE, and hGH-pA using KpnI and HindIII. Finally, the WPRE was removed using HindIII and XhoI, and 3'overhangs were blunted and 5'overhangs were filled in using Klenow fragment.

5.4.2 *rAAV production*

All rAAV vectors described were produced in HEK293 cells by using a helper virus free, two-plasmid based production method (Grimm et al., 2003) based on a commercially available AAV helper free system (Agilent Technologies, CA, USA; catalog# 240071). Briefly, HEK293 cells were transfected by using the calcium phosphate method. $72\,\mathrm{h}$ post transfection, cells were harvested and collected by centrifugation ($2500 \times g$, $20\,\mathrm{min}$ at $4\,^{\circ}\mathrm{C}$). Cell pellets were resuspended in resuspension buffer and lysed by three consecutive freeze/thaw cycles. For removal of genomic DNA, cell lysates were incubated with benzonase ($50\,\mathrm{u}\,\mathrm{ml}^{-1}$) for one hour at $37\,^{\circ}\mathrm{C}$. Subsequently, rAAV particles were precipitated with CaCl2 ($25\,\mathrm{mmol}$) followed by PEG precipitation (8% PEG-8000, $500\,\mathrm{mmol}$ NaCl). After resuspension of PEG precipitates in $50\,\mathrm{mmol}$ HEPES, $150\,\mathrm{mmol}$ NaCl, $25\,\mathrm{mmol}$ EDTA, pH 7.4 overnight at $4\,^{\circ}\mathrm{C}$, rAAV particles were further purified by CsCl density gradient centrifugation. Fractions from CsCl density gradients were analyzed by measuring the refractory index. Samples within a refractory index ranging from 1.3774 to 1.3696 were pooled and dialyzed against PBS for removal of CsCl by using dialysis cassettes with a molecular weight cutoff of 20 kDa (Thermo Scientific, MA, USA; catalog# 87738). Finally, rAAV preparations were concentrated by using ultrafiltration units with a molecular weight cutoff of 50 kDa (Millipore, MA, USA; catalog# UFC905024). After addition of glycerol to a final concentration of 10%, rAAV preparations were sterile filtered with Millex-GV filter units (Millipore, MA, USA; catalog# SLGV013SL), frozen in liquid nitrogen, and subsequently stored in aliquots at −80 °C. Genomic titers of purified rAAV stocks were determined by isolation of viral DNA (Viral Xpress DNA/RNA Extraction Reagent, Millipore, MA, USA; catalog# 3095) and subsequent qPCR analysis using primers specific for phSyn.

5.4.3 *Animal use*

Experimental subjects were were male CB57BL/6J mice of eight-twelve weeks of age from Jackson laboratory (strain #000664). Before surgical procedures, mice were kept in groups of five, and housed in 530 cm^2 cages on a 12 h light/dark cycle with unlimited access to dry food and water. Experiments were carried out during the light period. All animal experiments were performed in accordance with the Austrian laboratory animal law guidelines for animal research and had been approved by the Viennese Magistratsabteilung 58 (Approval M58/00236/2010/6).

5.4.4 *Stereotactic injection*

All surgical equipment was sterilized with 70% v/v ethanol before use. Animals were deeply anesthetized with a mixture of ketamine and medetomidine (KM; 2.5 mg ketamine-HCl and 0.02 mg medetomidine-HCl/25 g mouse weight) injected intraperitoneally, and positioned in a stereotaxic frame (Kopf Instruments, Tujunga, CA, USA; Stereotaxic System Kopf 1900). The eyes were protected from dehydration and intensive light exposure using sterile eye gel (Alcon Pharma, Novartis, CHE; Thilo-Tears Gel) and a piece of aluminum foil. Lidocaine was applied as local anesthetic subcutaneously before exposure of the skull. The scalp was washed with a 70% v/v ethanol in water solution and a cut along the midline revealed the skull. A small hole was drilled into the skull above the auditory cortex using a stereotaxic motorized drill (Kopf Instruments, Tujunga, CA, USA; Model 1911 Stereotaxic Drilling Unit) leaving the dura mater intact. Injections were performed perpendicular to the surface of the skull. Virus solution consisted of a mixture of two different recombinant AAV viruses (rAAV2/8 ITR-phSyn-GCaMP6m-WPRE-hGHpolyA-ITR; titer: 1.75×10^{11} viral genomes$(\mathrm{VG})/\mathrm{ml}$; rAAV2/8 ITR-phSyn-H2BmCherry-hGHpolyA-ITR; titer: $2 \times 10^{13}\,\mathrm{VG}/\mathrm{ml}$) in PBS. The virus mixture was loaded into a thin glass pipette and $150\,\mathrm{nl}$ were injected at a flow rate of $20\,\mathrm{nl\,min^{-1}}$ (World Precision Instruments, Sarasota, FL, USA; Nanoliter 2000 Injector) in five locations along the anterior-posterior axis, resulting in a total injection volume of $750\,\mathrm{nl}$. Stereotactic coordinates were: $4.4, -2, 5/-2.75/-3/-3.25/-3.5, 2.5$ (in mm, caudal, lateral, and ventral in reference to Bregma). Glass pipettes (World Precision Instruments, Sarasota, FL, USA; Glass Capillaries for Nanoliter 2000; Order# 4878) had been pulled with a long taper and the tip was cut to a diameter of 20 to $40\,\mathrm{\mu m}$. After the injection, the pipette was left in place for three minutes, before being slowly withdrawn and moved to the next coordinate. After completion of the injection protocol, the skin wound was sealed using tissue adhesive (3M Animal Care Products, St. Paul, MN, USA; 3M Vetbond Tissue Adhesive), and anesthesia was neutralized with $0.02\,\mathrm{ml}$ atipamezole. Mice were monitored daily and intraperitoneal injections of carprofen ($0.2\,\mathrm{ml}$ of $0.5\,\mathrm{mg\,ml^{-1}}$ stock) were applied on the first days after surgery.

5.4.5 *Cranial window implantation*

Two weeks after stereotactic injections, animals were anesthetized using isoflurane (Abbott Animal Health, IL, USA; IsoFlo). All surgical equipment and glass cover slip were sterilized with 70% v/v ethanol before use. Anesthesia was initialized in a glass desiccator filled with an isoflurane/air mixture. Anesthetized animals were mounted on a stereotaxic frame (Kopf Instruments, Tujunga, CA, USA; Stereotaxic System Kopf 1900) and the head was positioned using ear, teeth, and a custom-made v-shaped head holder. Anesthesia was maintained by delivery of a 1.5 to 2.4% isoflurane/air mixture with a vaporizer (High Precision Instruments, MT; Univentor 400 Anaesthesia Unit) at a flow rate of around $200\,\mathrm{ml\,min^{-1}}$ to the snout. $0.02\,\mathrm{ml}$ dexamethasone ($4\,\mathrm{mg\,ml^{-1}}$) was administered intramuscularly to the quadriceps, as well as $0.02\,\mathrm{ml}$ ml carprofen ($0.5\,\mathrm{mg\,ml^{-1}}$) intraperitoneally. The eyes were protected from dehydration and intensive light exposure using sterile eye gel (Alcon Pharma, Novartis, CHE; Thilo-Tears Gel) and a piece of aluminum foil. A local anesthetic (lidocaine/epinephrine (Gebro Pharma, Austria)) was applied subcutaneously before exposure of the skull. The scalp was washed with a 70% v/v ethanol in water solution and a flap of skin covering temporal, both parietal regions and part of the occipital bone was removed. The musculus temporalis was injected with lidocaine/epinephrine (Gebro Pharma, Austria) as an additional anesthetic and to minimize bleeding. Subsequently, the muscle was partly removed with a surgical scalpel and forceps to expose the right temporal bone. Using a fine motorized drill, the bones of the skull were smoothened, and part of the zygomatic process was removed. The surface was cleaned using cortex buffer and a two percent v/v hydrogen peroxide in water solution, and covered with a thin layer of one component-instant glue (Carl Roth, Germany; Roti coll). A thin layer of dental cement (Lang Dental, IL, USA; Ortho-Jet) was applied onto the skull, sparing the area of the temporal bone above the auditory cortex. A rectangular groove of about 2 mm by 3 mm was carefully drilled into the skull above the auditory cortex, and the bone was carefully lifted using scalpel and forceps. The exposed area was carefully cleaned and kept moist using sterile sponge (Pfizer, NY, USA; Gelfoam) and cortex buffer. The craniotomy was covered with a small circular cover glass (Electron Microscopy Sciences, PA, USA; five mm diameter, catalogue# $72195-05$), and sealed with 1.2% low-melting agarose (Sigma Aldrich, MO, USA; Agarose Type IIIA). The cover glass was finally set in place with one component-instant glue and dental cement. In order to position the animal under the microscope with the objective facing the window plane perpendicularly, a custom-made titanium head post was mounted on the implant above the window and embedded with dental cement. After dental cement had dried, animals were placed back in a pre-warmed cage. After the surgical procedure, animals recovered for at least one week before further handling.

5.4.6 *Habituation to awake chronic two-photon imaging*

Following mesoscopic imaging, animals were habituated to handling at the two-photon microscope. Therefore, animals were mildly water deprived and fixated under the objective in a custom-made acrylic glass tube, using a custom-made head post implant. The mouse head was laterally tilted such that the surface of the auditory cortex aligns approximately with the horizontal plane. During habituation, head fixation lasted for a minimum of 30 min each day, and animals were given access to a five percent m/v sucrose in water solution. This was repeated for at least five days until animals accommodated to the head fixation apparatus, showed reduced signs of stress and less body movements (typically consisting of few second long running bouts). The full sound stimulus set later used for recording of sound-evoked activity, was repeatedly presented resulting in animal subjects having experienced all sensory stimulations before any data acquisition.

5.4.7 *Sound presentation*

All sounds were delivered free field at 192 kHz sampling rate in a soundproof booth by a custom-made system consisting of a linear amplifier and a ribbon loudspeaker (Audiocomm, Austria) placed in

25 cm distance to the mouse head. The transfer function between the loudspeaker and the location of the mouse ear was measured using a probe microphone (Brüel&Kjær, Bremen, Germany; 4939-L-002) and compensated numerically by filtering the sound files with the inverse transfer function to obtain a flat frequency response at the mouse ear (between 0.5 kHz and 64 kHz ± 4 dB). Sound control and equalization was performed by a custom Matlab program running on a standard personal computer equipped with a Lynx 22 sound card (Lynx Studio Technology, CA, USA). The stimulus set consisted of 34 sound stimuli (19 pure-tone pips (50 ms; 2 to 45 kHz separated by a quarter octave) and 15 complex sounds (70 ms)) separated by one-second-intervals and played at 80 dB sound pressure level. The complex sounds in the stimulus set were characterized by broad frequency content and temporal modulations, generated from arbitrary samples of music pieces or animal calls replayed at fourfold speed. All stimulus on- and offsets were smoothened with a ten-ms-long half-period cosine function.

5.4.8 *Two-photon imaging*

The two-photon microscope (Prairie Technologies, WI, USA; Ultima IV) was comprised of a 20×-objective (Olympus, Tokyo, Japan; XLUMPlan Fl, NA = 0.95) and a pulsed laser (Coherent, CA, USA; Chameleon Ultra). Both fluorophores (GCaMP6m and mCherry) were co-excited at 920 nm wavelength, and separated by emission using a fluorescence filter cube (filter one: BP 480 −550 nm; filter two: LP 590 nm; dichromatic mirror: DM 570 nm; Olympus, Tokyo, Japan; U-MSWG2). Full frame imaging was performed using a field of view of 367 µm × 367 µm (pixel size: 256 × 128) and images were acquired at five Hertz frame rate (sampling period: 196.86 ms). In the last habituation session, several field of views (FOVs) at different xy-positions in layer 2/3 (about 150 −300 µm depth from cortical surface) were screened for the presence of reliable sound responses. FOVs where neuronal populations displayed reliable sound responses were repeatedly imaged at a two-day interval, using the stimulus set described above. Each stimulus was presented for at least 20 repetitions per FOV in pseudo-randomized order. Next, the focal plane was moved 50 µm in the z-axis and data was acquired for a second FOV with the same xy-coordinates. Between imaging periods, animals were given access to few drops of a five percent w/v sucrose in water solution.

5.4.9 *Auditory cued fear conditioning*

The behavioral setup was controlled by a personal computer with WINDOWS XP Professional, Version 2002, SP2 (Microsoft, Redmond, WA, USA) operating system running custom Matlab R2007a software

(MathWorks, Natick, MA, USA). All behavioral experiments were performed in an isolation cubicle (H10-24, Coulbourn Instruments, Whitehall, PA, USA) which was equipped with white LEDs as house light, a microphone and a CCD KB-R3138 camera with infrared LEDs (LG Electronics Austria, Vienna, Austria) which was connected to a Cronos frame grabber (Matrox, Dorval, Quebec, Canada). The conditioning chamber ($25\,\mathrm{cm} \times 25\,\mathrm{cm} \times 42\,\mathrm{cm}$, model H10-11M-TC, modified, Coulbourn Instruments) was combined either with a stainless-steel shock floor or a grid floor. A custom-made cartridge (round or quadrangular) formed the walls of the chamber in order to create different local environmental contexts. Foot shocks were delivered via an external shocker (Precision Animal shocker, Coulbourn Instruments). Sounds were played from a L-22 soundcard with a maximal sampling frequency of $192\,\mathrm{kHz}$ (Lynx Studio Technology, Costa Mesa, CA, USA) and delivered via an amplifier (Model SLA-1, Applied Research and Technology, TEAC Europe GmbH, TASCAM Division, Wiesbaden, Germany), a modified equalizer (Model #351, Applied Research and Technology, TEAC Europe GmbH, TASCAM Division, Wiesbaden, Germany) and a custom-made speaker for free field delivery of sounds. The sound stimuli used were from the stimulus set used for two-photon calcium imaging of sound representations in auditory cortex (see below). $70\,\mathrm{ms}$ stimuli were repeated 15 times with a one-second-interval, resulting in a total duration of $15\,\mathrm{s}$. On- and offsets of stimuli were smoothed with a $10\,\mathrm{ms}$ long half-period cosine function. Sound levels for all stimuli used were normalized to a mean power of $78\,\mathrm{dB}$ sound pressure level (SPL). Peak sound levels ranged from 83 to $89\,\mathrm{dB}$ SPL.

5.4.10 *Conditioning session*

In the conditioning environment, lights were turned on ($20 - 30\,\mathrm{lx}$), and the roundish cartridges were used as walls of the chamber. A mild residual ethanol odor was present from previous cleaning of the chamber. Mice were placed in the chamber directly before the start of each session. After at least $1\,\mathrm{min}$, baseline ($60 - 90\,\mathrm{s}$) five sound-shock pairings ($0.75\,\mathrm{mA}$, one second, immediately following the sound) were given with a randomized inter-stimulus-interval ranging from 50 to $75\,\mathrm{s}$ (paired).

5.4.11 *Memory test session*

Four days after auditory fear conditioning (i.e. one day after the two-photon imaging paradigm was completed), mice were tested for freezing responses. In order to create a different environmental context, the quadrangular cartridges were used as chamber walls, lights were turned off, and the home cage embedding was placed underneath the metal grid to provide a familiar odor to the animals. After at least 1 min of baseline (60 – 90 s), the conditioned stimulus and one unconditioned sound stimulus were presented in five randomized presentation blocks with an inter-stimulus-interval of two seconds.

5.4.12 *Quantitative analysis of freezing behavior*

During conditioning and memory testing, movies were recorded at a frame rate of 2.8 frames per second. Movies were analyzed offline based on a similar approach as described previously (Kopec et al., 2007), which provides a rapid and unbiased analysis of animal behavior. In short, the number of 'significant motion pixels' (SMPs), i.e. pixels which varied by more than 20 gray values, was calculated for all pairs of consecutive frames using a custom Matlab R2007a script (MathWorks). For each movie, the size of the mouse was estimated by the median SMP value of the 25% highest SMPs calculated from pairs of frames at least two minutes apart, thus capturing the mouse likely at different positions in the chamber. The threshold for freezing was defined as fewer SMPs than corresponding to 0.3% of the mouse size, which separates SMP values during freezing and movement periods. Baseline freezing was assessed during 60 – 90 s baseline period of each protocol run.

5.4.13 *Confocal imaging*

Mice were deeply anaesthetized and perfused with a PBS/Heparin solution and subsequently with a 4% PFA solution following standard procedures. Brain sections of 70 µm thickness were cut on a vibratome (Leica Biosystems, Germany; VT-1000). Next, they were incubated for 30 min in a 5 $\mathrm{mg\,L^{-1}}$ 4′, 6-diamidino-2-phenylindole (DAPI) solution, and mounted on cover slips. Confocal images were acquired on a LSM780 microscope (Carl Zeiss, Germany) using a 40× immersion objective (Objective Plan-Apochromat 40×/1.4 Oil DIC M27, Carl Zeiss, Germany).

5.4.14 *Image processing of chronic two-photon data*

In order to track cells across days, the optimal affine transformation was identified to register regions of interest (ROI), encompassing the soma of individual neurons, onto each frame of the time series recorded from the same FOV across several days. ROIs were selected independently by two human experts and can be described by a set of several hundred points marking the centers of the mostly spherical neuronal somata. This set of points was transformed for each frame by an affine transformation consisting of rotation, scaling and shifting. The objective function value for the optimization of this transformation is the pixel-wise overlap between a band-pass filtered and binarized image of each frame and a mask generated from the transformed ROIs by drawing a circle with a three-pixel (4.30 µm) radius around the center of each ROI. This six-dimensional optimization problem (rotation angle, scale in x, scale in y, off-diagonal of scaling matrix, shift in x, shift in y) was solved numerically using Matlab's implementation of the Nelder-Mead-Simplex algorithm (fminsearch). This was done in two iterations, first for the entire frame, then for four equally sized horizontal segments to correct for full frame movements during the two-photon microscope scanning. In a third iteration individual ROIs were moved to the maximum in a two-pixel (2.87 µm) surrounding of a low-pass filtered image to allow for slight local distortions.

5.4.15 *ROI inclusion criteria*

Four quality criteria were defined in order to only include cells in the analysis that had a reliably present signal in the H2B::mCherry channel marking the neuronal somata. This was done on a frame-by-frame basis, so that at each given time point a cell was either reliably present or excluded.

Nearest Neighbor Distance (NND): Strongly overlapping cells in a given frame, i.e. cells with a center-to-center distance below three pixels (4.30 µm), were defined as unreliable in that respective frame. Thus, the chance to wrongly label individual cells was minimized.

Normalized Soma Signal Intensity (NSSI): For each cell at each time point, the difference between the mean signal intensity in the soma (two-pixel radius; 2.87 µm) and the mode of the intensity of the surrounding (ten-pixel radius; 14.34 µm) was computed and normalized by the 95-percentile of this difference. Cells with an intensity close to the background, an NSSI below the value of 0.2 were excluded.

Objective Function Value (OFV): The optimization described above resulted in the alignment and an objective function value, which describes the pixel wise overlap of the frame and the template. In order to rule out movement artifacts, individual frames in which the OFV was less than three standard deviations below the mode of the OFV for a given FOV were rejected.

Soma Signal to Noise ratio (SSN): The difference of the mean intensity of the soma (two-pixel radius; 2.87 µm) and the mode of the intensity of the surrounding (ten-pixel radius; 14.34 µm) was defined as signal. The standard deviation of a jittered version of the signal (same radii, but pseudo-random location of the "soma" in the ten-pixel radius) was defined as noise. In order to be included in the analysis, cells had to have a SSN value above one.

All quality criteria were tested and cells were excluded on a frame-by-frame basis. Excluded time points were treated as missing entries in the data. Cells that were not reliably detected on at least ten trials for each stimulus on a given day were completely excluded from the analysis.

5.4.16 *Calculation of* $\Delta F/F_0$ *and deconvolution*

The baseline F_0 used to compute $\Delta F/F_0$ was defined as a moving rank order filter, the 30th percentile of the 200 surrounding frames (100 before and 100 after). This $\Delta F/F_0$ was then deconvolved using the algorithm by Vogelstein et al., 2010.

5.4.17 *Stimulus-evoked sound responsiveness of single cells*

To classify single cells as sound responsive or not, all trials from a given stimulus were compared in a rank-sum test against twenty randomly picked patterns of spontaneous activity (from periods without sound presentation). A cell was classified as significantly responsive, if the p-value was below 0.01 after a Benjamini-Hochberg correction for multiple comparisons against number of days (4), number of stimuli (34), and number of cells (21,506) for at least one stimulus (Benjamini and Hochberg, 1995).

5.4.18 *Sound response profiles of single cells*

For each significantly sound responsive cell, sound response vectors to pure-tone frequencies and complex sound stimuli were max-normalized to the stimulus with highest response amplitude on the given day. The selection of cells for each day and sorting of cells on the y-axis was performed either on significantly sound responsive cells and their tuning from day one (Figure 5.6a), the given day (Figure 5.6b), or from day seven (Figure 5.6c). For the analysis to control for a sampling bias, the sorting was performed with only the first half of trials and the sound response vectors were plotted with the second half of trials (5.8).

5.4.19 *FOV inclusion criteria*

We included FOVs in our analysis that satisfied the following three criteria: (A) FOVs needed to contain at least 100 ROIs (i.e. neurons) which fulfilled the quality criteria described above, (B) FOVs needed more than ten significantly sound responsive neurons on each day and (C) neurons in the FOVs needed to respond to at least four stimuli on at least one day.

5.4.20 *Definition of response modes*

Response modes were defined for a given FOV. For each trial i, the population response of n simultaneously recorded neurons was characterized by an n-dimensional vector $\mathbf{v}$. Each entry of $\mathbf{v}$ was the mean deconvolved activity recorded in a time bin of $400\,\mathrm{ms}$ after stimulus onset. The response similarity between two stimuli p and q was then determined by

$$S(p,q) = \frac{1}{M_p M_q} \sum_{i=1}^{M_p} \sum_{j=1}^{M_q} \rho\left(\mathbf{v}_{p,i}, \mathbf{v}_{q,j}\right) \tag{5.1}$$

with trial numbers M_p, M_q, and Pearson's correlation coefficient $\rho\left(\mathbf{v}_1, \mathbf{v}_2\right)$ (Galton, 1886, Pearson, 1895). Note that stimuli p and q may refer to stimulus pairs presented at the same day or at two different days, in which case the same stimulus presented at different days is formally treated as two different stimuli. Response reliability for a given stimulus, at a given day was assessed by $S(p,p)$ (the mean correlation over all pairs of trials, excluding pairs with $i = j$, i.e. same trials). The response to a stimulus, for which $S(p,p) > 0.4$, was deemed reliable. Response modes were then estimated by hierarchical clustering of response similarity $S(p,q)$, restricted to (day-specific) stimuli p with reliable response, with $1 - S(p,q)$ as metric and unweighted average linkage clustering as linkage criterion. All responses of stimuli with non-reliable response ($S(p,p) < 0.4$) were assigned to the "null-mode" (0-mode). Choosing this threshold we found that most prominent clusters were well captured, as we verified by visual inspection of all FOVs. Importantly, our overall results are qualitatively similar when the threshold for response reliability was set to 0.3 or 0.5 (data not shown). The hierarchical clustering algorithm provided by Matlab (functions linkage and dendrogram) was used to sort stimulus responses. To estimate the number of relevant clusters objectively, the resulting cluster tree was cut at every possible cluster number and a Hubert's Γ (Hubert and Baker, 1977) was calculated as

$$\Gamma = \frac{2}{O(O-1)} \sum_{i=1}^{O} \sum_{j=i+1}^{O} \left(S_{ij} - c\right) T_{ij} \tag{5.2}$$

where O is the size of similarity matrix $\mathbf{S}$, c is a threshold, and $\mathbf{T}$ is a binary matrix of equal size with entries

$$T_{ij} = \begin{cases} 1, \text{if } i \text{ and } j \text{ are clustered together,} \\ 0, \text{otherwise.} \end{cases} \tag{5.3}$$

The threshold was set to $c = 0.4$, as for response reliability, ensuring that only 'reliably' correlated sound responses are considered to participate in the same cluster. Again, the overall results were qualitatively similar when using a slightly different threshold, e.g. 0.3 or 0.5 (data not shown). The response modes in a given FOV were then defined as the clusters obtained for the maximal value. This clustering was highly significant ($p \leqslant 0.001$) for all neuronal populations in a FOV compared to three surrogate data sets generated by (a) shuffling the stimulus identity across trials, (b) shuffling the stimulus identity for each cell individually and (c) shuffling the cell identity for each trial (5.13).

5.4.21 *Mode-associated responsiveness of single cells*

Similar to stimulus evoked sound responsiveness, we estimated whether a given cell is significantly responsive in a given mode. For each cell, we determined the rank-sum between activities from all trials associated with the mode to the same amount of spontaneous activity patterns drawn randomly from periods without sound presentation. A cell was significantly responsive in a given mode, if the p-value was below 0.01 after a Benjamini-Hochberg correction against number of days (4), number of modes (varying), and number of cells ($21,506$) (Benjamini and Hochberg, 1995).

5.4.22 *Sound decoding based on full response vectors*

A linear classifier (Matlab function lassoglm with L1 regularization) was trained to discriminate between responses to two different stimuli. For the analyses where training and testing were done on the same day, cross-validation was performed by leaving out one trial. Where training and testing were done on different days, training was done with all trials of a given day and the performance of the classifier was tested on each trial of a different day. The pairwise decoding performance was then defined as the percentage of correctly classified trials, and FOV decoding performance was defined as the mean pairwise decoding performance over all pairs of stimuli.

5.4.23 *Sound decoding based on a 34-fold classifier*

A support vector machine was trained to discriminate between all 34 stimuli using MATLAB's built in cecoc function. When training and testing was done on the same day, cross-validation was performed by leaving out one trial. When done on different days, the classifier was trained using all vectors of one day and tested with all vectors of the other day. The decoding performance was defined as the percentage of correctly classified trials.

5.4.24 *Dependence of decoding on number of modes and stimuli per mode*

We used a minimalistic model to study the dependence of sound discriminability on the number of modes and number of stimuli per mode in a FOV. For simplicity, we assume all trials of a given stimulus evoke responses in the same mode (which can be the 0-mode). This model suggests that in a regime with a 0-mode as large as observed in our experimental data, an increase in both number of modes and stimuli per mode improves decoding performance.

In the model, we assume two stimuli that are mapped to the same mode are indistinguishable, but can be distinguished from stimuli mapped to a different mode (including the 0-mode). The contribution of a mode to the overall decoding performance in a FOV is proportional to its size and to a mode specific decoding factor, given by the average discriminability associated with a stimulus in this mode. This factor is large if only few stimuli are mapped to this mode, as these are distinguishable from all other stimuli. Reversely, this factor is small for modes with many stimuli, since these can only be distinguished from few other stimuli.

We cast these considerations into mathematical form to reveal how the average discriminability depends on number of modes and stimuli per mode. The decoding performance of a full FOV P_{FOV} is given by

$$P_{FOV} = \frac{1}{N_{Stim}} \left(N_0 P_0 + \sum_{i=1}^{N_m} P_i L_i \right) \tag{5.4}$$

with the total number of stimuli N_{Stim}, the number of stimuli evoking no response N_0, the decoding factor P_0 specific to the 0-mode, the number of different response modes N_m, the decoding factor P_i specific to response mode and L_i, the number of stimuli mapped to response mode i. The mode specific decoding factor P_i was determined by the size of the mode L_i the stimulus is mapped to and the probabilities of its correct classification when compared to different stimuli within (P_{same}) and outside (P_{diff}) the mode:

$$P_i = \frac{(L_i - 1)P_{same} + (N_{Stim} - L_i)P_{diff}}{N_{Stim} - 1} \tag{5.5}$$

For simplicity, we considered the case of equally distributing the number of stimuli per mode. This results in an upper bound on the decoding performance, as can be seen by inserting P_i into the expression of P_{FOV} above. A numerical analysis revealed that this provided a reasonable approximation within the experimentally observed regime of number of modes and stimuli per mode. Thus, the expression for the decoding performance simplifies to:

$$P_{FOV} = \frac{1}{N_{Stim}} (N_0 P_0 + N_m L_m P_m) \tag{5.6}$$

with the mean number of stimuli per mode L_m, the number of modes N_m. For the mode specific decoding factors, we obtain

$$P_m = \frac{(L_m - 1)P_{same} + (N_{Stim} - L_m)P_{diff}}{N_{Stim} - 1} \tag{5.7}$$

and

$$P_0 = \frac{(N_{Stim} - L_m N_m - 1)P_{same} + L_m N_m P_{diff}}{N_{Stim} - 1}. \tag{5.8}$$

Assuming chance decoding within a given mode ($P_{same} = 0.5$) and perfect decoding between modes ($P_{diff} = 1$) and a total of 34 stimuli yields Figure 5.21a. In general, the decoding performance increases with number of modes, while the optimal number of stimuli per mode for a given number of modes is obtained for a uniform distribution of stimuli per mode including the 0-mode, which also follows from Chebyshev's sum inequality (Hardy et al., 1988).

5.4.25 *Statistics*

All statistical analyses were performed using MATLAB (Mathworks, Natick, MA, USA).

5.4.25.1 *Freezing behavior*

To compare the freezing behavior during the different time windows in the memory test session, a Kruskal-Wallis test (Kruskal and Wallis, 1952) was performed followed by a correction for multiple comparisons.

5.4.25.2 *Decoding comparison*

To compare decoding performance of a linear classifier trained with response vectors, or with maximum pooling from the best FOV of one mouse, a Kruskal-Wallis test (Kruskal and Wallis, 1952) was performed followed by a correction for multiple comparisons to reject the null-hypothesis that the decoding performances of all FOVs come from the same distribution.

5.4.25.3 *Comparison of population dynamics of mice undergoing ACFC to baseline cohort*

Bootstrapping was performed to test whether two mean values could origin from the same distribution. To do so, $10,000$ surrogate data sets were sampled with replacement from the original data sets. For each of the two respective values the means of the surrogate data sets were computed and compared. Next, the p-value was calculated as the probability that both the surrogate from the distribution of the smaller original mean was larger than the original mean, and vice versa. The p-values were corrected for multiple comparisons with a Bonferroni correction (Bonferroni, 1936). Stars denote significance levels: * $p < 0.05$.

6

A BASIS SET OF ELEMENTARY OPERATIONS CAPTURES RECOMBINATION OF NEOCORTICAL CELL ASSEMBLIES DURING BASAL CONDITIONS AND LEARNING.

In this chapter we will utilize the response mode picture, we developed in Chapter 5, to define a set of elementary operations and thus describe neuronal population dynamics in mouse auditory cortex. The chapter starts with an introduction to the concept of cell assemblies in Section 6.1, before we show some results quantifying the drift of cell assemblies in our data (Section 6.2.1). This drift can be captured by a set of elementary operations, which we introduce in Section 6.2.2. Then we use these newly defined operations to characterize drift under basal conditions (Section 6.2.3) and during auditory cued fear conditioning (ACFC, Section 6.2.4). Our results are discussed in Section 6.3, before we give a more detailed description of our methods (Section 6.4).

We use the same dataset that was previously used in Chapter 5, and build our analyses on the analyses performed there. Consequently, we will focus on new concepts and analyses in this chapter.

6.1 INTRODUCTION

In Chapter 5 we found that in mouse auditory cortex stimulus responses typically fall into a small set of possible response modes (Figure 5.11) and we looked at the changes these modes undergo under basal conditions and during learning. We found pronounced ongoing dynamics under basal conditions that were biased towards a differential generalization during fear conditioning. Now, we want to go further and deconstruct the representational dynamics into their most basic underlying operations to gain insight into their functional relevance. In contrast to our operational definition of response modes in the previous chapter, i.e. a local subgroup of neurons responding to multiple stimuli in a field of view (FOV), cell assemblies are defined on a functional level as a group of cells underlying a cognitive or behavioral function (Hebb, 1949). Adding a functional component to the observed response modes will lead to insight into functional properties and thus provide a link to the understanding of cell assemblies and their dynamics.

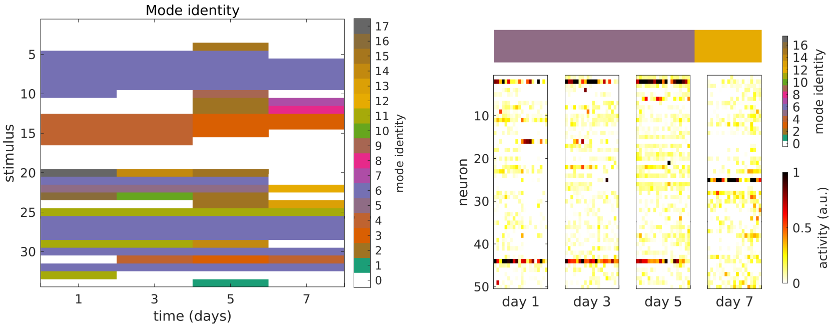
Cell assemblies (or rather the strong connections between neurons in a cell assembly) have been shown to have a number of interesting properties in theoretical studies, such as associative memory (Hopfield, 1982), multistability (Beer, 1995, Stern et al., 2014, Fasoli et al., 2016), a transition from winner-takes-all like dynamics to to multiple attractors (Miller, 2016, Chen and Miller, 2020), or correction of synaptic drift (Acker et al., 2019, Kossio et al., 2021). Their study in biological neural circuits remained a challenge due to their distributed nature until the recent development of techniques that allow the simultaneous recording of neuronal activity of several hundreds of neurons, resulting in a growing body of experimental evidence for cell assemblies in a number of different cortical systems (Harris, 2005, Buzsaki, 2010, Yuste, 2015, Holtmaat and Caroni, 2016). They have also been linked to functional properties, e.g. memory recall (Tonegawa et al., 2015).

Here, we further analyze the data described in Chapter 5 through the more functional lens of cell assemblies and their dynamics. As response modes undergo constant remodeling, it comes as no surprise that also cell assemblies are subject to ongoing remodeling, even under basal conditions, i.e. after behavioral habituation and without any apparent learning paradigm. We decompose these dynamics by means of a basis set of ten elementary operations, which provide a framework for the analysis of cell assembly remodeling during learning. We observed that the remodeling cannot be explained by a mere remapping of stimulus responses on an existing set of response modes, but coincides with a remodeling of response modes themselves. Furthermore, we found learning induced biases in the frequency of some operations leading to an increase of associations of sensory representations, via both the formation of new associations and the stabilization of old ones.

6.2 RESULTS

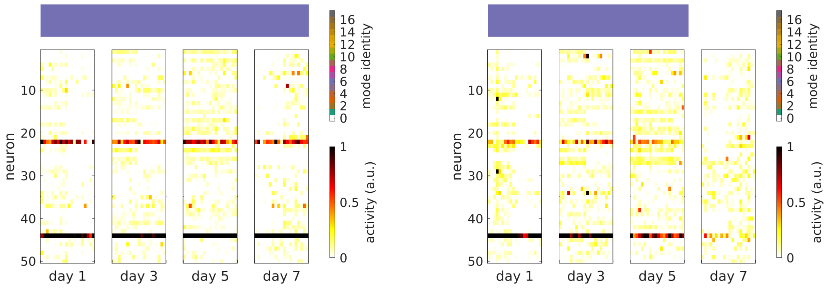
6.2.1 *Transitions between response modes*

The clustering of stimulus responses in local populations of neurons in mouse auditory cortex into these response modes can be used to track population responses over time. With this formalism being established, each stimulus response fell into a certain response mode (or evoked no response) at any imaging time point. This is visualized in Figure 6.1a by assigning a different color to each response mode and then tracking the response mode of each stimulus across all imaging days. Stimuli that did not evoke a response were grouped into the so-called $0-\mathrm{mode}$, colored white. While the response to some stimuli remained stable, other stimuli lost or gained a representation, and still others evoked a different response on consecutive days in this



(a) Response modes across time.

(b) Example response (Stim. 22).

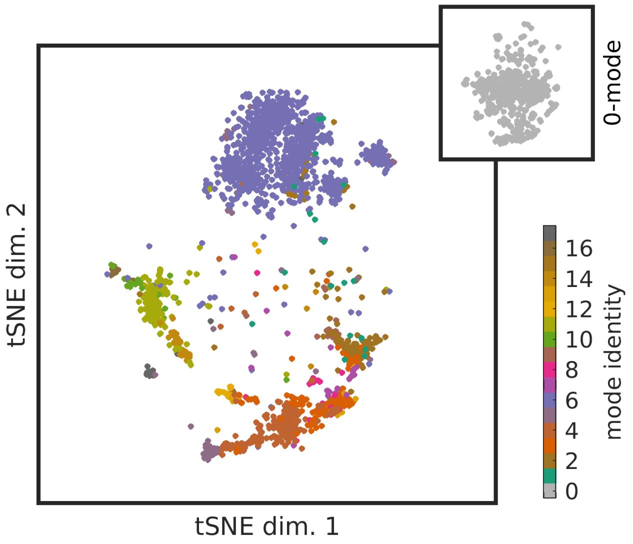


(c) Example response (Stim. 8).

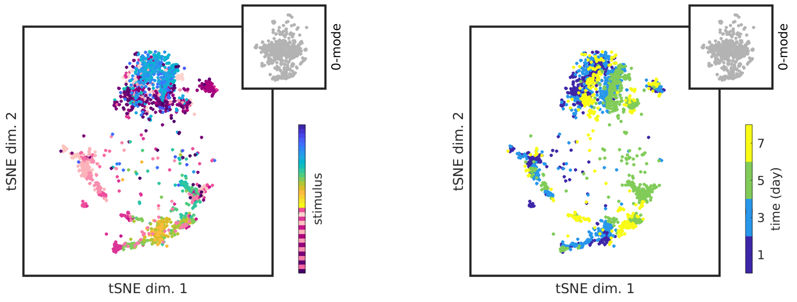
(d) Example response (Stim. 21).

Figure 6.1: Response modes change across time. (a) Response mode identity for each stimulus at each imaging time point in an example FOV. While the response mode identity for some stimuli remains stable throughout the imaging period, other stimuli gain or lose a response or change from one response mode to another. (b) Example for a stable population response on first three imaging days but a different response on day 7. Stimulus 22 from (a). (c) Example for a stable population response throughout the imaging period. Stimulus 8 from (a). (d) Example for a stimulus losing its response on the last imaging day. Stimulus 21 from (a). Note that this response mode is not lost on imaging day 7 (as can be seen in (c)), but rather this specific stimulus is not mapped to it any more.

FOV. Some examples of population response vectors from Figure 6.1a are shown in Figure 6.1b to Figure 6.1d. Some stimuli reliably evoked the same population response throughout the imaging period (CCCC, Figure 6.1c). Other stimuli gained or lost a response in a given FOV over time. Figure 6.1d shows a stimulus being mapped to the same response mode as the stimulus in Figure 6.1c on the first three imaging time points, but not any more on the last imaging day (CCC0). And then, the response to some stimuli changed with time. An example is shown in Figure 6.1b, where a stimulus evoked one response mode through three imaging time points and a different response mode on the last time point (AAAB).



(a) tSNE plot (colored by response mode).



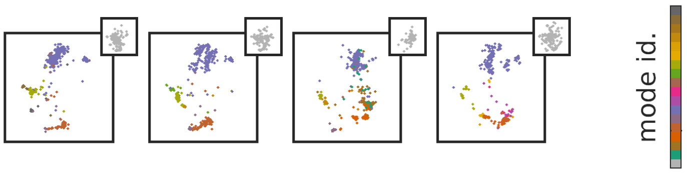
(b) tSNE plot (colored by stimulus). (c) tSNE plot (colored by day).

Figure 6.2: tSNE plots are used to visualize response modes in 2D. (a) tSNE plot of stimulus responses in example population from Figure 6.1. Each dot represents a trial of a stimulus on a day. All trials of all stimuli from all four imaging days are plotted in one plot, colored by response mode identity. Inlay: 0-mode. (b) Same as (a), but colored by stimulus identity. (c) Same as (a), but colored by imaging day.

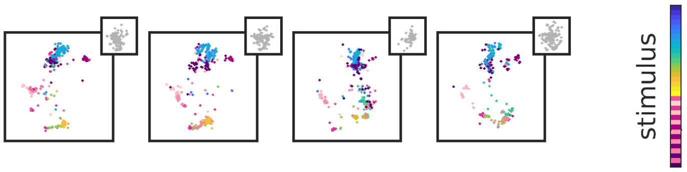
To illustrate response mode dynamics on a single trial basis we applied a dimensionality reduction technique called t-distributed stochastic neighbor embedding (tSNE, Maaten and Hinton, 2008). tSNE takes pairwise distances between high- and low-dimensional data points to compute their respective probability distributions. It then minimizes the distance between the two distributions using gradient descent (for details see Section 6.4). The result of this dimensionality reduction for response vectors to all stimuli on all days in the same example population as in Figure 6.1 can be seen in Figure 6.2. Here, every point depicts a single trial population response vector on a single day (in total 34 stimuli $\times$ 20 trials $\times$ 4 days $\approx 2,720$ points; few trials were removed due to movement artifacts, see Section 5.4). Stimuli that evoked no reliable response, i.e. fall into the 0-mode, are plotted as an inlay in the top right corner.

Coloring by response mode (Figure 6.2a) reveals that dimensionality reduction via tSNE is well suited to illustrate the response mode structure in the data. Apart from a small amount of exceptions, which are also visible in single trial response vectors in Figure 6.1, clusters of points are colored in the same color. This apparent agreement between color and the shape of the point clouds allows us to utilize tSNE for the purpose of visualization of respones modes. Coloring by stimulus (Figure 6.2b) reveals the mapping of stimuli onto response modes. Response modes consist of responses to various stimuli, visualized by different colors within each cluster of points. Pure tone stimuli of similar frequencies are often mapped onto the same response, but typically modes are evoked by both pure tones and complex sounds. Coloring by imaging day (Figure 6.2c) allows us to assess the temporal structure of response modes. While most response modes in this FOV are present at all four imaging time points (indicated by four colors inside point clusters) others are only present on some days (clusters missing one or multiple colors).

In order to track individual stimuli across time, we computed the dimesnionality reduction on all time points, but plotted only the subsets of points recorded on individual days (Figure 6.3a). These daily tSNE plots colored by response mode reveal changes across days. In this example most modes stayed stable across time, but some appeared newly or disappeared. The same plots colored by stimulus identity (Figure 6.3b) theoretically enable us to track the responses to single stimuli across time. They are, however too crowded, so in order to see what is happening, we had to focus on individual stimuli instead. In Figure 6.4 we colored all trials of the one stimulus we want to track in magenta and all other stimuli in gray. When following the same stimuli from Figure 6.1 across time in a tSNE plot we



(a) Daily tSNE plots (colored by response mode).



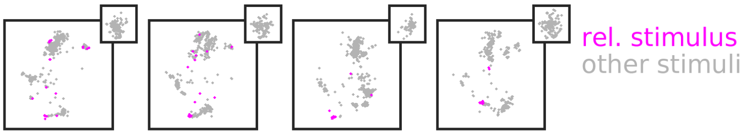
(b) Daily tSNE plots (colored by stimulus).

Figure 6.3: Single day tSNE plots. (a) Similar as Figure 6.2, tSNE was used to reduce the dimension of stimulus responses to 2, however the trials were split into days. Colored by response mode identity, Inlay: 0-mode. (b) Same as (a), but colored by stimulus identity.

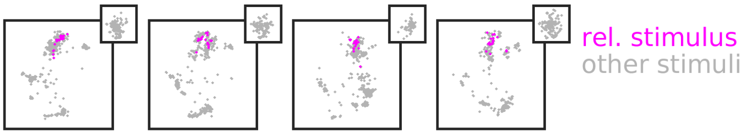
can again see that some stimuli move from evoking one response to evoking another response (in this case modes AAAB, Figure 6.4a), while other stimuli stay in the same response mode across all imaging days (mode CCCC, Figure 6.4b) or stay in the same response mode for some time, before not evoking a response anymore (CCC0, Figure 6.4c). We observed similar dynamics in all imaged FOVs with a varying degree of stability (see Chapter 5). Two principal forms of plasticity occurred: First, changes in the tuning of cell assemblies, i.e. the mapping of sounds onto cell assemblies could be different from day to day. Second, changes in the structure of cell assemblies, involving the de-novo formation of a cell assembly, loss of a cell assembly or substantial remodeling of a cell assembly by gaining or losing member neurons.

6.2.2 *A set of operations to describe response mode dynamics*

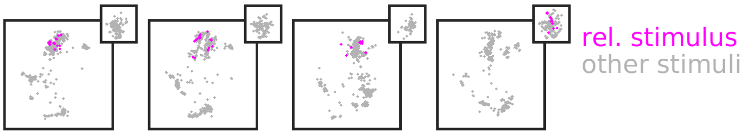
We wanted to systematically describe the dynamics of sound representations and response modes. To do so we defined a set of elementary operations that capture at the same time stimulus response changes and the response modes involved with this stimulus. We can approach this set of operations from two sides by starting with response modes and then looking at the involved stimulus responses or we can start with a stimulus response and then turn to the involved response modes. Both ways result in the same basis set of elementary operations. Here, for simplicity reasons, we only explain the approach starting with a stimulus response.



(a) Example tSNE plot (Stim. 22).



(b) Example tSNE plot (Stim. 8).



(c) Example tSNE plot (Stim. 21).

Figure 6.4: Single stimulus tSNE plots for example stimuli from Figure 6.1. tSNE was computed for all data, but plotted split by day. Inlay: 0-mode. (a) Stimulus response is stable for three imaging time points and changes to a different response mode on day 7. (b) Stimulus response is stable throughout the imaging period. (c) Stimulus response is stable at first until it disappears on the last day.

In Figure 6.5 we start with the stimulus response. We describe this schematic from left to right, from individual stimuli on the left hand side, via the response mode dynamics, ending on the definition of ten unique operations on the right hand side. We asked for each individual stimulus: Does it evoke a response at time point i? Does it evoke a response at time point $i+1$? And – if it evokes a response at both time points – does it fall into the same response mode or does it evoke different response modes? This already gives us five exclusive categories of transitions to start with: no response at both time points ($0 \rightarrow 0$), the same response mode at both time points ($A \rightarrow A$), no response at time point i and a response at time point $i+1$ ($0 \rightarrow A$), a response at time point i and no response at time point $i+1$ ($A \rightarrow 0$), or different response modes at both time points ($A \rightarrow B$).

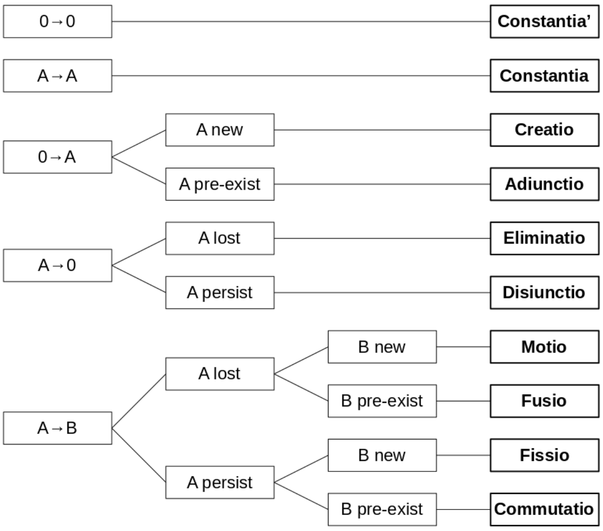


Figure 6.5: Schematic defining all possible operations of a stimulus response. Starting from the transitions on the left, describing the response change of a single stimulus, we find 10 distinct operations by asking, what is happening to the response modes, affected by this stimulus. E.g. when a stimulus evokes a response A at time i and no response at time $i+1$, is the response mode A lost (*eliminatio*) or are other stimuli still evoking this response mode A (*disiunctio*)?

As we were not only interested in whether a stimulus evokes a response or not, but also in the response mode dynamics of the involved response modes, we next asked, what is happening with the involved response modes: Are they only present at one of the two time points, i.e. newly emerging or disappearing? Or are they present at both time points regardless of the stimulus not evoking this response mode at the other time point? This is irrelavant for the trivial transitions $0 \rightarrow 0$ and $A \rightarrow A$, but it gives two options each to the $0 \rightarrow A$ and $A \rightarrow 0$ cases, as each of them involves one response mode, and four options to the $A \rightarrow B$ case, as it involves two response modes and each of them can be only present at one time point or at both.

This analysis allows us to classify any possible transition as one of the operations, described in Figure 6.5 and in a bit more detail below. We used latin names for the set of operations to distinguish the terminology from a mere description of the various forms of plasticity.

- *Constantia′* is the operation, when a stimulus evokes no response at both time points.
- *Constantia* describes the case, when a stimulus evokes the same response mode at both time points.
- *Creatio* we use for the operation, when a stimulus evokes no response at time i and a newly emerging response mode (that has not been present at the previous time point for any of the presented stimuli) at time $i+1$.
- *Eliminatio* is the inverse operation of *creatio*, where both the stimulus does not elicit a response any more at time $i+1$ and the response mode disappears, too (i.e. is not elicited anymore by any of the presented stimuli).
- *Adiunctio* describes a stimulus that does not evoke a response at time i and evokes a response mode at time $i+1$, that already existed previously.
- *Disiunctio* is the inverse to *adiunctio*, where a response mode is present at both time points, but a given stimulus does only elicit it at time i and not at time $i+1$.
- *Fusio* is the operation, where a stimulus evokes different response modes at both time points and the response mode it evokes at time i does not exist any more at time $i+1$, but the response mode it evokes at time $i+1$ is present at both time points. So, in other words, the stimulus joins a pre-existing response mode.
- *Fissio* is the inverse of *fusio*: a stimulus evokes different response modes at both time points and the response mode it evokes at time i is present at both time points, but the response mode it elicits at time $i+1$ is new. Or, in other words, the stimulus splits away from an existing response mode to form a new one.
- *Motio* describes, when a stimulus evokes different response modes on both time points and the response mode at time i

disappears and the response mode at time $i + 1$ is not present previously.

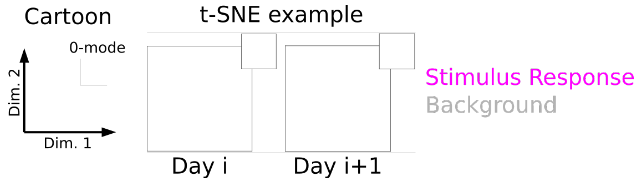
- *Commutatio* is the operation, when a stimulus switches between response modes and both response modes are present at both time points.

To make sure, the operations are not merely caused by fluctuations around some correlation threshold, we provide some supplementary statistics in Chapter 9: In Sup. Fig. 9.4 we plotted histograms of the mean correlation of the response to a stimulus (undergoing a certain operation) at time i (i.e. prior to the operation) and of the response to the same stimulus at time $i + 1$ (i.e. after the operation). For all operations (except *constantia*) this correlation is expected to be low and it indeed is. If the transitions were only caused by fluctuations around a threshold, this correlation would be close to this threshold. Additionally, we measured and plotted the histograms of the within mode correlations, in other words the correlation of a stimulus response to the responses of all the other stimuli evoking the same mode (Sup. Fig. 9.5). Those correlations are independent of operation and only depend on, whether a stimulus evokes a response at a given time point or not. And we show the distributions of the mode sizes prior to and after each operation (Sup. Fig. 9.6). Here, differences are visible for operations changing the response mode size.

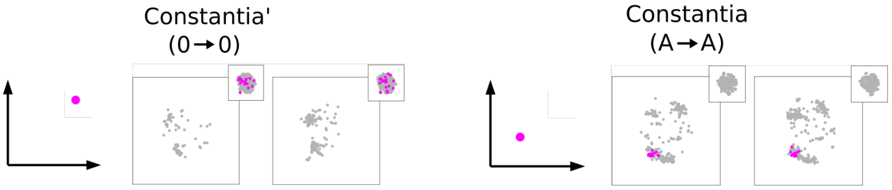
This approach, in principle, reflects the changes of information on sensory representations a putative readout neuron in a higher cortical area would receive via the projections originating from a sub-region of the auditory cortex. While it is highly unlikely that any readout neuron receives as input the exact neurons imaged in a given FOV, it is reasonable to assume that any subgroup of neurons in auditory cortex is not activated by all stimuli and shows overlapping response patters. Thus, the operations defined above for imaging FOVs are relevant from the point of view of readout neurons, too. Note that disappearing and emerging response modes are dependent on the set of used stimuli: We do not know, if different stimuli would still elicit a certain response mode. However, the same is true in general for readout neurons. It is impossible to play every existing stimulus in finite time, so some repsonse modes are bound to not be activated for a certain time span and thus "disappear".

6.2.3 *Ongoing recombination of cell assemblies*

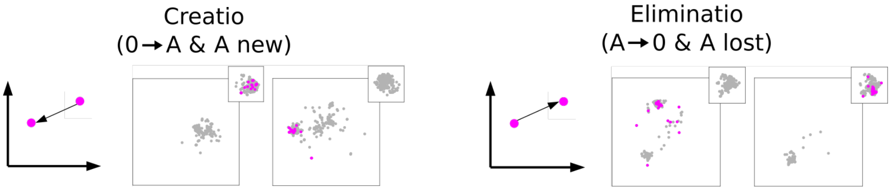
These ten operations can now be used to address several questions. As the operations constitute a complete set of all operations that theoretically exist, we first wondered, if all of them existed in our dataset. To that aim, we counted them and indeed, we found examples for



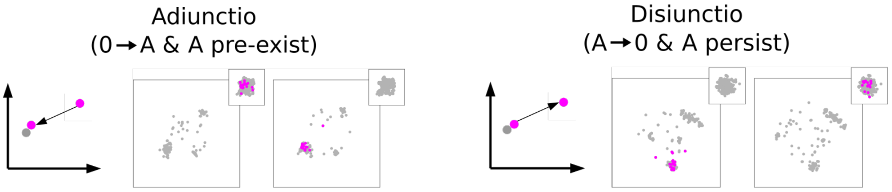
(a) Legend.



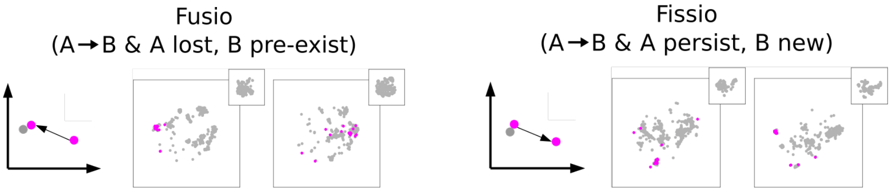
(b) *Constantia′*. (c) *Constantia*.



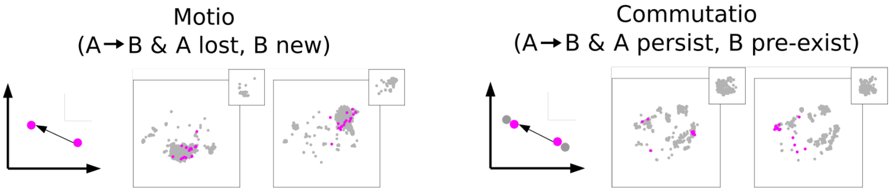
(d) *Creatio*. (e) *Eliminatio*.



(f) *Adiunctio*. (g) *Disiunctio*.



(h) *Fusio*. (i) *Fissio*.



(j) *Motio*. (k) *Commutatio*.

Figure 6.6: Example for every response mode operation as defined in Figure 6.5. (a) Legend. (b) to (k) Examples for each response mode operation, including schematic and tSNE plots before and after the operation.

all of them. Examples can be seen in Figure 6.6. There we show a schematic and an example tSNE representation for each of the ten operations described in the previous section. The frequencies of each

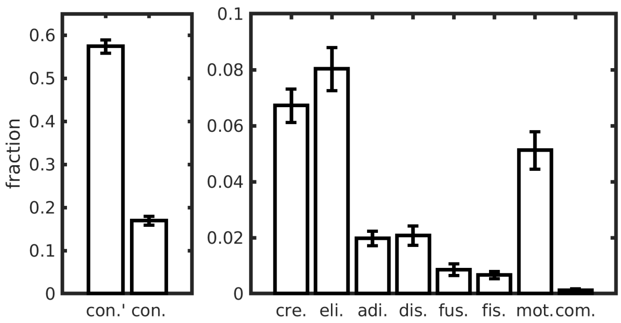


Figure 6.7: Response mode operation frequencies. Fraction of stimulus representations (mean ± SEM) undergoing each of the operations. Note that each of the 10 theoretically possible operations was found in our dataset.

of the operations in our dataset are given in Figure 6.7. While the majority of stimuli did not evoke a response on consecutive imaging time points (*constantia′*) and the majority of the remaining stimuli could be classified as *constantia*, we found a remarkable amount of remapping of stimulus responses between response modes and of disappearing/emerging response modes.

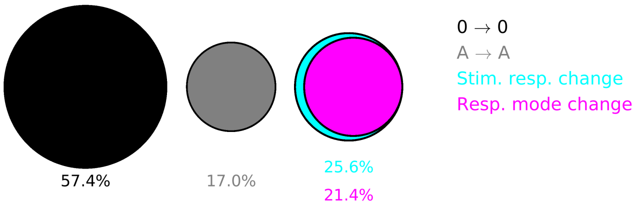


Figure 6.8: Venn diagram of stable and unstable response mode operations. While the majority of stimuli did not evoke any response (57.4%) or the same response on two consecutive imaging time points (17.0%), a large part of stimuli changed their response (25.6%). Of those changing their response most (21.4% of total) coincided with a newly appearing/disappearing response mode.

We wondered, if these changes were mostly a remapping of stimuli onto stable response modes or if the response modes themselves did undergo a remodeling. If mostly stimulus responses would be remapped between existing response modes, we would not find operations involving the gain or loss of a response mode. So, we grouped together all transitions that do not involve a gain or loss of a response mode (*adiunctio, disiunctio, commutatio*) and thus leave the response modes intact and compared them to all transitions that involve both a change in the stimulus response and a gain/loss of a response mode (*creatio, eliminatio, fusio, fissio, motio*). The result is plotted as a Venn diagram (Venn, 1880) alongside circles for *constantia′* and *constantia* for comparison. We found that from imaging time point to imaging time point (two days apart) 57.4% of stimuli did not evoke any response, 17.0% of all stimuli evoked the same response at both imaging days, 21.4% involved both a stimulus evoking a different response mode and a change of response modes and only 4.2% involved a stimulus change, while the involved response modes stayed the same. So, response modes are highly dynamic and most operations invovling the change of a stimulus response also involve a a remodeling of response modes. This might also explain the small amount of *commutatio* present in the data. As operations leaving the involved modes intact seem to be relatively rare, the operation leaving two modes intact, while the stimulus is remapped from one to the other, is even rarer. Note that in our picture of response mode transitions there are no operations that leave the stimulus response configuration intact while changing the response modes.

This set of response mode operations can capture ongoing dynamics under basal dynamics, showing the level of instability of cortical representations in mouse auditory cortex. In the next section we want to investigate the effects of fear conditioning on these dynamics making use of our set of response mode operations.

6.2.4 *The impact of learning on cell assembly dynamics*

In Chapter 5 we described the effect auditory cued fear conditioning has on the response modes typically found in auditory cortex, namely, that fear conditioning increased the amount of stimuli evoking a response (compared to the basal dataset Figure 5.20b), and that this was not due to an increase of the number of modes (Figure 5.22a), but rather to an increase of stimuli per mode (Figure 5.22b).

To validate the accuracy of our response mode operations, we wanted to know, if these finding were also reflected in the response mode operations. So, we combined operations changing the mode number (Figure 6.9) and operations changing the mode size (Figure 6.10). We could even further disentangle the response mode dynamics by splitting the operations into operations decreasing or increasing mode number or mode size respectively (for a full picture of single operation counts in both datasets see Figure 6.11 (for a stacked bar plot) or Sup. Fig. 9.7 (for line plots of individual operations)). We found no

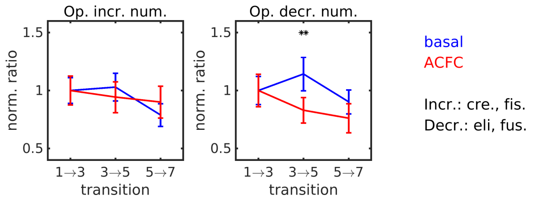


Figure 6.9: Operations changing the number of response modes. Increasing the number of response modes: *creatio*, *fissio*; decreasing the number of response modes: *eliminatio*, *fusio*. Significances: * $p < 0.5$, ** $p < 0.01$, *** $p < 0.001$, via bootstrapping.

significant difference between the basal and the ACFC dataset in the normalized ratio of operations increasing the mode number (*creatio*, *fissio*) and only a transient increase in the basal dataset as compared to an ongoing decline in the ACFC dataset for operations decreasing the mode number (*eliminatio*, *fusio*) (Figure 6.9). All other response mode operations are not explicitly associated with a change in the number of response modes. This finding is in line with our findings from Chapter 5.

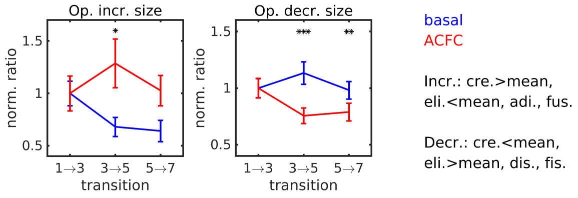
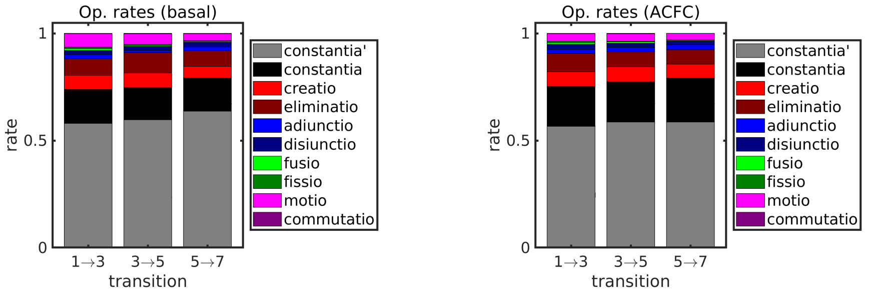


Figure 6.10: Operations changing the mode size (i.e. number of stimuli mapped to a response mode). Increasing mode size: *creatio* of mode with above mean mode size, *eliminatio* of mode with below mean mode size, *adiunctio*, *fusio*; decreasing mode size: *creatio* of mode with below mean mode size, *eliminatio* of mode with above mean mode size, *disiunctio*, *fissio*. Significances: * $p < 0.5$, ** $p < 0.01$, *** $p < 0.001$, via bootstrapping.

In contrast we found a significant difference in operations affecting the response mode size (Figure 6.10): While there was also only a transient increase (albeit significant) in operations increasing mode size (*creatio* of modes of larger size than the previous mean, *eliminatio* of modes of smaller size than the previous mean, *adiunctio*, *fusio*) in the ACFC dataset, fear conditioning seemed to have a longer lasting effect on operations decreasing mode size (*creatio* of modes of smaller size than the previous mean, *eliminatio* of modes of larger size than the previous mean, *disiunctio*, *fissio*). Those dropped in the

ACFC dataset, which would in turn lead to the increase in mode size, described in Chapter 5. So, this increase in mode size could mostly be attributed to a decrease of the amount of operations that decrease the response mode size.



(a) Basal dataset.

(b) ACFC dataset.

Figure 6.11: Fraction of each operation across time in (a) basal and (b) ACFC dataset.

We showed that our response mode operations were able to reliably reproduce (and even deepen our understanding of) findings made by more conventional methods, such as counting the number of modes or calculating the average number of stimuli mapped to a mode (i.e. mode size). We can, however, go further and answer some more questions about response mode dynamics. We could ask, for example, if

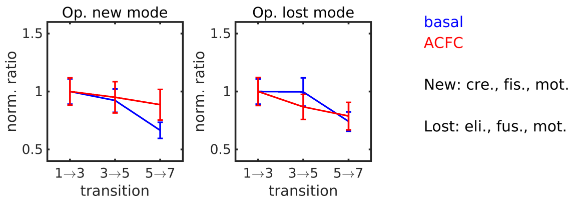


Figure 6.12: Operations changing response mode picture (i.e. operations including appearing or disappearing response patterns). Appearing response mode: *creatio*, *fissio*, *motio*; disappearing response mode: *eliminatio*, *fusio*, *motio*. Significances: * $p < 0.5$, ** $p < 0.01$, *** $p < 0.001$, via bootstrapping.

learning has more of an effect on response mode dynamics or if it is rather mostly affecting the stimulus responses and their mapping onto the modes. So, we first had a look at all operations that change the response mode picture, i.e. that include new or lost modes (Figure 6.12). We found no significant difference between the basal and the ACFC dataset for both operations including the formation of a new mode (*creatio*, *fissio*, *motio*) and operations during which a mode is lost (*eliminatio*, *fusio*, *motio*). Note that *motio* is included in both, as it consists of a stimulus response switching from one response to another, where the first response ceases to exist as well as the second mode did not exist before. So, we looked at stimuli changing

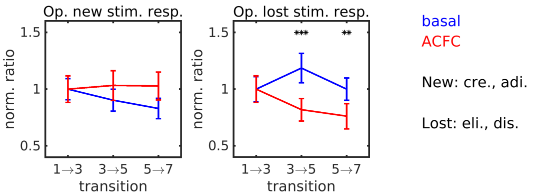


Figure 6.13: Operations changing stimulus responsiveness (i.e. from responsive to unresponsive or vice versa). Stimuli gaining response: *creatio*, *adiunctio*; stimuli losing response: *eliminatio*, *disunctio*. Significances: * $p < 0.5$, ** $p < 0.01$, *** $p < 0.001$, via bootstrapping.

their responsiveness, i.e. stimuli gaining or losing a response in FOV. We again compared the two datasets to each other (Figure 6.13) and while operations that involve a stimulus gaining a response (*creatio, adiunctio*) showed a – however, non-significant – decrease in the basal dataset that did not seem to be present in the ACFC dataset, operations involving the loss of a stimulus response (*eliminatio, disiunctio*) showed a significant decrease in the ACFC dataset compared to the basal one.

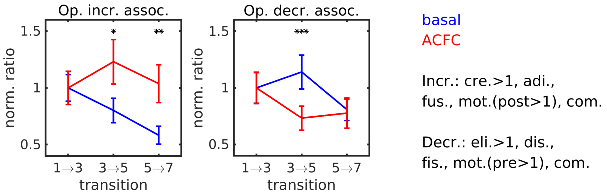


Figure 6.14: Operations changing stimulus associations (i.e. newly mapping stimuli to the same response mode or not mapping stimuli to the same response mode any more). Increasing associations: *creatio* of mode with more than one stimulus, *adiunctio*, *fusio*, *motio* to a mode with more than one stimulus, *commutatio*; decreasing associations: *eliminatio* of mode with more than one stimulus, *disiunctio*, *fissio*, *motio* from a mode with more than one stimulus, *commutatio*. Significances: * $p < 0.5$, ** $p < 0.01$, *** $p < 0.001$, via bootstrapping.

As the mapping of two different stimuli onto the same response mode could be - and has been (Besnard and Sahay, 2016, Cai et al., 2016, Grewe et al., 2017) - understood as an association of the two stimuli, and fear conditioning has been linked to generalization, we wondered, if fear conditioning had an effect on the associations of stimuli. We thought of associations on the cortical population level as two stimuli evoking the same response mode. Operations forming (breaking) associations were thus operations, where the stimulus evoked the same response as another stimulus at time $i + 1$ (i) and the stimulus response changed. We found a transient difference between the two datasets in operations decreasing associations (*eliminatio* of a mode with more than one entry, *disiunctio, fissio, motio* from a mode with more than one entry, *commutatio*) and a longer lasting difference in operations increasing associations (*creatio* of a mode with more than one entry, *adiunctio, fusio, motio* to a mode with more than one entry, *commutatio*). So, ACFC led to an increased formation of associations between responses to different stimuli.

6.3 DISCUSSION

We studied the dynamics of cell assemblies forming auditory representations in the cortex across several days. We identified a set of ten elementary operations to capture all possible transitions between cell assemblies for a given stimulus. These operational definitions are well suited to describe the data obtained in our experiments within a FOV of a particular size and using a fixed set of sensory stimuli. Regarding the entire auditory cortex and assuming responsive neurons for each sound in each mouse, all operations involving the 0-mode would become obsolete, so our set of operations would be reduced to five operations (*constantia, fusio, fissio, motio, commutatio*). As a potential readout neuron would only receive input from a subset of all neurons in auditory cortex, however, the complete set of ten operations seems necessary to capture all possible changes of feed forward inputs.

The driving forces underlying the drift of cell assemblies during basal conditions could most probably be found in the ongoing remodeling of the underlying synaptic connectivity. Such basal dynamcis in connectivity have been observed under behaviorally stable conditions and even during pharmacological blockade of neuronal activity (Yasumatsu et al., 2008,Loewenstein et al., 2011, Rubinski and Ziv, 2015, Dvorkin and Ziv, 2016, Nagaoka et al., 2016). Consistently, modelling work has shown that synaptic drift can lead to drift on the level of cell assemblies (Kappel et al., 2015, Kappel et al., 2018, Mongillo et al., 2018, Kossio et al., 2021).

When considering several cell assemblies, each representing a set of elements that send excitatory or inhibitory feed-forward projections to a higher-order readout cell (or cell assembly), computations could

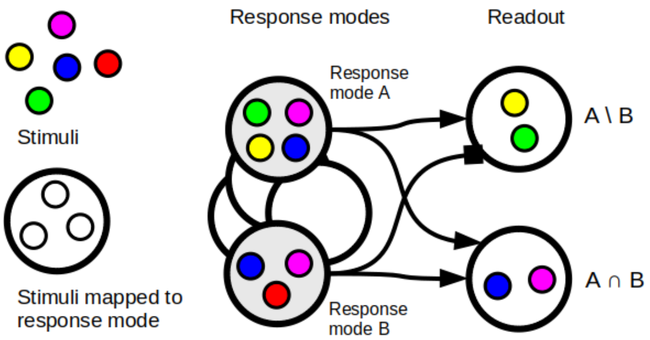


Figure 6.15: Response modes could be utilized to perform set operations in a readout layer, like $A \setminus B$ via excitatory input from A and inhibitory input from B or $A \cap B$ via excitatory input from both A and B.

be performed that are reminiscent of set-theory operations. For example, considering two cell assemblies each representing overlapping sets A and B of sensory elements, it could be possible to compute those elements in A that are unique to A and not overlapping with B (i.e. set difference: $A \setminus B$) by integrating excitatory input from cell assembly A and inhibitory input from cell assembly B. Cell assembly dynamics could then be seen as newly generating or combining these set-theory rules.

Our analysis of cell assembly operations indicates that auditory fear conditioning biases the frequency of operations that lead to an increase in the mapping of different stimuli onto a given cell assembly, possibly reflecting the formation of associations between sensory representations. This recombination of representations, i.e. their association as well as their dissociation, occurs to a substantial degree also during basal conditions and its functional relevance remains a matter of debate (Chambers and Rumpel, 2017). The ongoing recombination of cell assemblies could for example support cognitive processes that occur without explicit mental engagement, like the spontaneous creation of associations (Wallas, 1926) or their forgetting (Richards and Frankland, 2017). Our basis set of cell assembly operations provides a framework for these future studies.

6.4 METHODS

As this chapter builds upon the methods and results obtained in Chapter 5, we only describe further analyses, here. Please, have a look at Section 5.4 for further details on the datasets or response modes.

6.4.1 *t-distributed Stochastic Neighbor Embedding (tSNE)*

We used t-distributed Stochastic Neighbor Embedding (tSNE, Maaten and Hinton, 2008) to visualize in two dimensions the distribution of all single-trial response vectors recorded in a given FOV. tSNE utilizes pairwise distances between high-dimensional and low-dimensional data points to compute the probability distributions for both. The resulting low-dimensional data points are obtained by minimizing the Kullback-Leibler divergence D_{KL} (Kullback and Leibler, 1951) between both distributions using gradient descent from random initial conditions:

$$D_{KL}(P||Q) = \sum_{ij} p_{ij} \log \frac{p_{ij}}{q_{ij}}, \tag{6.1}$$

for pairwise distances p_{ij} in high dimensions and pairwise distances q_{ij} in low dimensions, which are defined as follows.

$$p_{ij} = (p_{j|i} + p_{i|j})/2N, \tag{6.2}$$

$$p_{j|i} = \frac{\exp(-d(\mathbf{x}_i, \mathbf{x}_j)^2/2\sigma_i^2)}{\sum_{k\neq 1} \exp(-d(\mathbf{x}_i, \mathbf{x}_k)^2/2\sigma_i^2)}, \quad p_{i|i} = 0, \tag{6.3}$$

$$q_{ij} = \frac{(1 + ||\mathbf{y}_j - \mathbf{y}_i||^2)^{-1}}{\sum_{k\neq l}(1 + ||\mathbf{y}_k - \mathbf{y}_l||^2)^{-1}}, \quad q_{i|i} = 0. \tag{6.4}$$

Here, $\mathbf{x}_i$ and $\mathbf{y}_i$ are high- and low-dimensional data points, respectively. For low-dimensional data points the Euclidean distance is used as a distance measure, for high-dimensional data points any measure d can be defined as a distance. We defined it as the correlation distance $1 - \rho$, where ρ is Pearson's correlation coefficient (Galton, 1886, Pearson, 1895):

$$\rho(a, b) = \frac{\sum_n (a_n - \bar{a})(b_n - \bar{b})}{\sqrt{\sum_n (a_n - \bar{a})^2} \sqrt{\sum_n (b_n - \bar{b})^2}}. \tag{6.5}$$

This optimization aims to preserve small distances in low-dimensional space. Errors in large distances are not penalized as strongly. So, the fine structure is preserved, whereas large distances are not. In our case, this would translate to: clusters are preserved, whereas the inter cluster distances and the distribution of clusters in two-dimensional space cannot be trusted.

We used correlation distance (1–ρ) as high-dimensional distance measure, instead of Euclidean distance used in the original work. We jointly mapped all single-trial response vectors to all stimuli, including responses across all days, and plot either all trials from a given day or trials from all days.

Part IV

A FIRING RATE MODEL OF POPULATION ACTIVITY IN MOUSE AUDITORY CORTEX

In the following part we investigate the neuronal population dynamics found in mouse auditory cortex (Part iii) using a firing rate model. The model is described in detail in Chapter 7. We find a regime with similar population activity patterns as in mouse auditory cortex for strong recurrent connections and strong inhibition. In Chapter 8 we apply synaptic drift to the network connectivity and find periods of stable response patterns interrupted by abrupt changes towards new response patterns. This suggests that, in the parameter regime similar to auditory cortex response changes are broadly distributed.

7

A FIRING RATE MODEL OF CLUSTERING IN MOUSE AUDITORY CORTEX

First, we give a short introduction to firing rate models that we used to reproduce clustering in mouse auditory cortex (Section 7.1), before discussing our model setup (Section 7.2). Next, we perform a parameter scan to understand the different dynamic regimes of our model (Section 7.3.1) and we have a closer look at the synaptic structure beneath this clustering (Section 7.3.2). We condclude this chapter with a brief discussion of the results (Section 7.4).

7.1 INTRODUCTION

In Part iii we analyzed experimental data from and showed evidence for response modes in mouse auditory cortex and their dynamics under basal conditions and during learning (Chapter 5). We extended the description of said dynamics towards a more functional understanding in Chapter 6. Here, we want to understand the link between the synaptic structure of a network and its activity. This is crucial to link synaptic drift to representational drift. Before we investigate this link in Chapter 8, however, we first need a model, capable of reproducing single time point dynamics of mouse auditory cortex.

We showed that in mouse auditory cortex responses to stimuli typically cluster into a near discrete set of responses. That means that different stimuli evoke the same population response as can be seen in Figure 5.11. Something similar has also been reported by Bathellier et al. (2012), See et al. (2018). These clusters typically consist of a subgroup of imaged neurons that are activated by multiple stimuli. This clustering might seem counterintuitive at first, however, the stimuli that are clustered together differ from FOV to FOV, in a way that allows for a unique global response, when taking together multiple FOVs throughout the auditory cortex.

Here, we want to address the question, what gives rise to this clustering. Which network connectivity statistics lead to a clustering of stimuli into a small set of possible response modes? We use a firing rate model of excitatory and inhibitory neurons with randomly drawn connections between each other. We vary different model parameters and find a regime similar to experimental data from mouse auditory cortex (for details see Chapter 5).

Already very small firing rate models consisting of two or three units display complex behavior such as simple and multiple hysteresis phenomena and unit cycle activity (Wilson and Cowan, 1972), different regimes of oscillations, fixed points, and transient responses (Wilson and Cowan, 1973), input and connectivity dependent bifurcations (Borisyuk and Kirillov, 1992, Beer, 1995), regimes of periodic, quasi-periodic and chaotic behavior (Pasemann, 2002), spontaneous symmetry breaking (Fasoli et al., 2016), or dependencies of dynamic distances on structural distances (Krauss et al., 2019b).

Attractors (i.e. states, the network converges to) in such networks can be learned (Hopfield, 1982) or arise in random networks for certain parameter settings. Apart from an attractor regime, where the system has multiple stable states, the literature often describes two more regimes, a uni-stable regime with only one attractor and a chaotic regime, depending on parameters such as synaptic gain (Wilson and Cowan, 1972, Sompolinsky et al., 1988, Zhang and Saggar, 2020), recurrent connection strength (Ostojic, 2014, Stern et al., 2014, Krauss et al., 2019a), strength of inhibition compared to excitation (Ostojic, 2014, Rost et al., 2018, Zhang and Saggar, 2020), randomness vs. structure (Kadmon and Sompolinsky, 2015, Mastrogiuseppe and Ostojic, 2018, Krauss et al., 2019a), stimulus strength (Wilson and Cowan, 1972, Rubin et al., 2015b), self excitation (Stern et al., 2014). While most of the computations in the brain resemble dynamics in the attractor regime, even chaotic networks can have many traits of cortical networks (Barak et al., 2013).

In this chapter we present evidence for two distinctive uni-stable regimes, an input driven regime, where each input evokes a unique response and a recurrence dominated regime, where all inputs evoke the same response. Moreover, we want to focus on another aspect of the multi-stable regime, namely that not only single stimuli evoke one out of a set of responses, depending on the initial condition, but also multiple stimuli can evoke the same response and thus form seemingly random response clusters, where different groups of stimuli evoke one out of a small set of possible responses. This seems similar to data recorded from mouse auditory cortex and can be found in a balanced regime (with strong excitation and inhibition). It seems to be mitigated by strong inhibition between competing groups of neurons.

We used a simple firing rate model as described in detail in Section 7.2. The found random clustering in a model regime governed by strong recurrent connections and strong inhibition presented in Section 7.3.1. These clusters could be characterized by weak inhibitory connections between neurons within a certain cluster and strong inhibitory connections between neurons from different clusters (Section 7.3.2).

7.2 MODEL SETUP

7.2.1 *Construction of connectivity matrices*

We used a firing rate model to reproduce the apparently random clustering of different stimuli onto response modes, as has been observed in mouse auditory cortex (Chapter 5). The model of a population consisting of N neurons can be described by a system of N coupled differential equations (Equation 7.1). The firing rate r_i of each neuron i is governed by:

$$\tau \frac{\partial}{\partial t} r_i = -r_i + f\left(\sum_{j=1}^{N} W_{ij} r_j + s_i(t)\right), \tag{7.1}$$

with the weight matrix of connections from neurons j to i, W_{ij}, time dependent input into each neuron i, $s_i(t)$, the time constant τ, and the nonlinearity $f(x)$. For simplicity, the nonlinearity function was chosen to be a rectification (Equation 7.2):

$$f(x) = [x]_+ = \begin{cases} x, & \text{if } x > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{7.2}$$

Without loss of generality we set $\tau = 1$. As we wanted to study a very general case, both the connectivity matrix $\mathbf{W}$ and the stimuli $\mathbf{s}$ were randomly generated. The fully connected connectivity matrix was drawn from two log-normal distributions, described by Equation 7.3.

$$w_k = \exp\left(\ln\left(\frac{\mu_k^2}{\sqrt{\mu_k^2 + \sigma_k^2}}\right) + Z\sqrt{\ln\left(1 + \frac{\sigma_k^2}{\mu_k^2}\right)}\right), \tag{7.3}$$

where w_k is a random synapse weight drawn from a log-normal distribution with mean μ_k and standard deviation σ_k for $k = \{E, I\}$ – excitatory or inhibitory – and Z is a standarad normally distributed random variable. We used two distributions – one for excitatory and one for inhibitory synapses – because cortical neurons have typically either only excitatory or only inhibitory outgoing synaptic connections (often referred to as Dale's principle, Dale, 1934). Log-normal distributions were used, because synaptic weights between cortical neurons have been experimentally measured to be distributed in a log-normal like manner (excitatory: e.g. Loewenstein et al., 2011, Buzsaki and Mizuseki, 2014, inhibitory: e.g. Minerbi et al., 2009, Rubinski and Ziv, 2015). As roughly 20% of neurons in cortex are inhibitory and 80% are excitatory (Sahara et al., 2012), we drew the first 20% of columns of $\mathbf{W}$ from a negative log-normal distribution (i.e. each weight was multiplied by -1) and the other 80% of columns from a positive log-normal distribution. The means μ_k of both distributions were set according to the position in parameter space, as will be discussed in detail in Section 7.3.1. The mean of the excitatory distribution μ_E was the connection strength μ divided by the number of neurons in the network N:

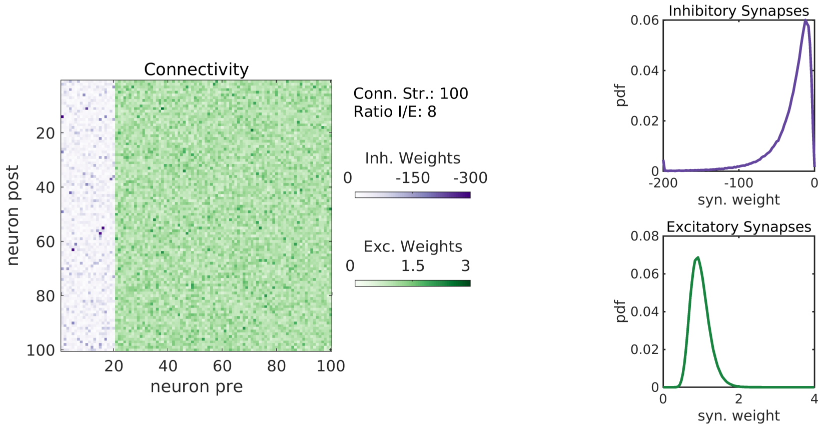
$$\mu_E = \frac{\mu}{N}. \tag{7.4}$$

The mean of the inhibitory distribution μ_I was the mean of the excitatory distribution multiplied by the ratio I/E ($R_{I/E}$) and the factor 4, to account for the 4 times more excitatory cells in the network:

$$\mu_I = \frac{N_E}{N_I} R_{I/E} \frac{\mu}{N} = 4 R_{I/E} \mu_E, \tag{7.5}$$

where N_E is the number of excitatory neurons and N_I is the number of inhibitory neurons. The standard deviations σ_k of the distributions were set relative to their means according to experimental measurements: $\sigma_E = 0.025\mu_E$ (Loewenstein et al., 2011) and $\sigma_I = 0.1\mu_I$ (Minerbi et al., 2009, Statman et al., 2014, Rubinski and Ziv, 2015), and divided by $\sqrt{N}$. The division of both means μ_k and variances σ_k^2 by the network size N normalizes the eigenvalue spectrum and makes it independent of the network size. Both connection strength μ and ratio of inhibition to excitation $R_{I/E}$ are thus independent of the network size. Note that it is not necessary to distinguish between excitatory and inhibitory neurons to achieve the results presented here, nor is it necessary for synapses to be drawn from log-normal distributions, as has been shown in earlier work for networks drawn from Gaussian distributions or binary networks drawn randomly from $\{0, 1\}$ (Eppler, 2015). We used log-normal distributions for two reasons: synaptic weights are distributed in a log-normal like way in cortical networks and more importantly, when applying synaptic drift (in Chapter 8) that has been fit to experimental recordings the log-normal distribution is its stationary distribution.

An example connectivity matrix can be seen in Figure 7.1a and the corresponding distributions from which inhibitory and excitatory connections are randomly drawn are displayed in Figure 7.1b. As stated in the figure, the network strength of this example is 100 and the ratio of inhibition to excitation is 8. As stated above, this means that the mean of the excitatory synapse distribution is given as $\mu_E = 100/N = 1$ and the mean of the inhibitory synapse distribution as $\mu_I = -1 \cdot \mu_E \cdot 4 \cdot 8 = -32$. The widths are set accordingly to be $\sigma_E = 0.025\mu_E = 0.025$ and $\sigma_I = 0.1\mu_I = 3.2$. An extensive description of the interpretation and effect of these parameters, as well as a parameter space scan can be found in Section 7.3.1.



(a) Example connectivity matrix.

(b) Histograms.

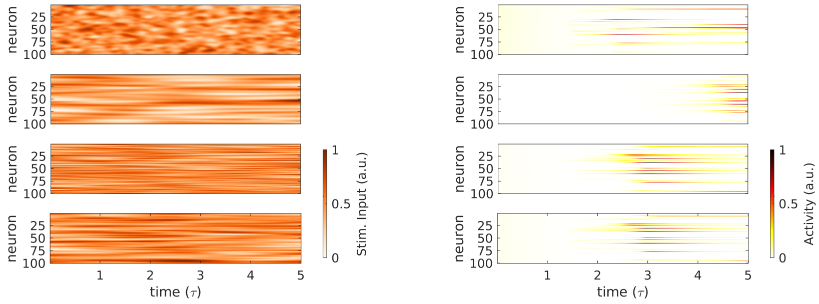
Figure 7.1: Connectivity matrices are drawn randomly from log normal distributions. (a) Example connectivity matrix for a random implementation of a network with recurrent connection strength of $\mu = 100$, and corrected ratio of mean inhibitory synapse strength to mean excitatory synapse strength of $R_{I/E} = 8$. (b) Histograms of inhibitory (top) and excitatory (bottom) synapse strengths accumulated over 100 implementations of the network (Inh.: $\mu_I = -32$, $\sigma_I = 3.2$; Exc.: $\mu_E = 1$, $\sigma_E = 0.025$).

7.2.2 *Construction of input*

Input to the network was assumed to be random with some correlations in time and across neurons to not make any too strong assumptions about the input a group of neurons in any cortical area is expected to receive. We did not model tonotopy, as we wanted to model a subpopulation of auditory cortex too small for an apparent tonotopic gradient. To this end, stimuli were generated by first sampling randomly from a uniform distribution in the interval $(0, 1)$, resulting in white noise matrices $g(x_i, t_i)$ of dimension $N \times T$, where T is the stimulation time in units of Δt, in our case 100×500. These matrices were convolved with Gaussians of random widths in both dimensions and then divided by N to keep the net input into the system constant with regard to the network size. The stimulus $s(x_i, t_i)$ into neuron x_i at time t_i is defined by the convolution of a uniform random matrix $g(x_i, t_i)$ with a two dimensional Gaussian kernel (Equation 7.6):

$$s(x_i, t_i) = \sum_{x_j} \sum_{t_j} g(x_j, t_j) \frac{1}{2\pi\sigma_x\sigma_t} e^{-\frac{(x_i - x_j)^2}{2\sigma_x^2} - \frac{(t_i - t_j)^2}{2\sigma_t^2}}, \tag{7.6}$$

where σ_x and σ_t are the width if the Gaussian in neuron space and time respectively, which were drawn randomly from $\sigma_x \in (0, 3)$ and $\sigma_t \in (0, 1/2\Delta t)$. These inputs were then divided by the number of neurons N to keep the mean input to the network stable against changes in network size. Four example stimuli of length 5τ and with



(a) Example stimuli.

(b) Example responses.

Figure 7.2: Stimuli & responses of an example network. (a) 4 example stimuli. (b) Network responses to the stimuli in (a) show sparse transient activations of subgroups of neurons. For an overview of all used stimuli and responses see Sup. Fig. 9.8 and Sup. Fig. 9.9.

temporal resolution $\Delta t = 0.01\tau$ can be seen in Figure 7.2a and an overview over all 40 stimuli that were used throughout this and the next chapter can be seen in Sup. Fig. 9.8. We used a relatively short stimulation length of 5τ in order to simulate the short sounds used in experimental data (see Chapter 5). The Δt had to be chosen sufficiently small for numerical solution of system of differential equations as detailed in the next section (Section 7.2.3)

7.2.3 *Clustered network responses*

The system was solved numerically using the Forward-Euler method (e.g. Atkinson, 1989), again, for a time span of 5τ and with a temporal resolution $\Delta t = 0.01\tau$, resulting in responses displayed in Figure 7.2b for the four stimuli shown in Figure 7.2a (or Sup. Fig. 9.9 for an overview of responses to all 40 stimuli respectively). In the network's initial state the activity of each neuron was set to 0. In the inhibition dominated regime this was not so important, as the strong inhibition would set almost any other activity to 0 at the second step anyway (for most initial conditions).

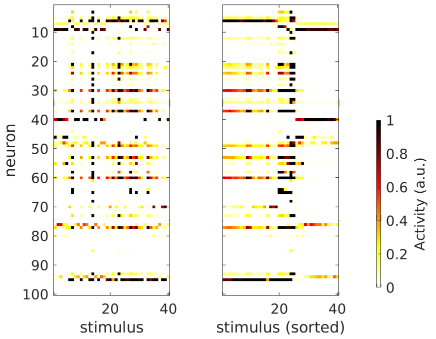
These time dependent responses were then used to calculate response vectors – as seen in Figure 7.3a – by taking the mean of the neuronal activities during the last $\tau = 100\Delta t$ of stimulation. This averaging was done both to simulate the effects of Calcium imaging, which implicitly averages neuronal activity on the time scale of tens of miliseconds, because the timescale of both Calcium dynamics and the dynamics of the Calcium indicator (GCaMP6m, Chen et al., 2013) are orders of magnitude slower than single action potentials ($\tau_{\mathrm{Ca}} \approx 1\,\mathrm{s}$ compared to $\tau_{\mathrm{AP}} \approx 1\,\mathrm{ms}$). This also facilitated further computations, which were done on 1×100 dimensional vectors instead of 500×100

dimensional matrices. So, it resulted in one response vector per stimulus. We then computed Pearson's correlation coefficient ρ (Galton, 1886, Pearson, 1895) between any two response vectors:

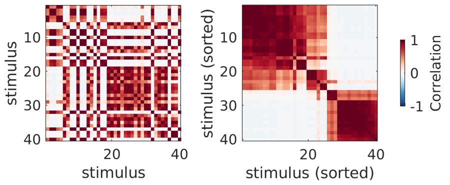
$$\rho(u, v) = \frac{\sum_i (u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\sum_i (u_i - \bar{u})^2}\sqrt{\sum_i (v_i - \bar{v})^2}} \tag{7.7}$$

resulting in a correlation matrix the size of the number of stimuli we used (in our examples 40×40). Here, u and v are response vectors, $\bar{u}$ and $\bar{v}$ are their respective means.

Response vectors are displayed before (left) and after (right) sorting with hierarchical clustering: Figure 7.3b shows the unsorted and sorted correlation matrix of the response vectors displayed in Figure 7.3a. The sorted correlation matrix shows a clustering of stimulus responses into response modes. In the depicted example, we can clearly see two large response clusters with high correlation (albeit some structure) within and low correlation between each other. Similar clustering was observed in most implementations of the network in the given parameter regime, although the number of found clusters and the stimuli whose responses were clustered together varied from network to network.



(a) Response vectors.



(b) Correlation matrix.

Figure 7.3: Response vectors & correlations fro the example from Figure 7.1. (a) Response vectors. Left: unsorted. Right: sorted via hierarchical clustering. (b) Corresponding correlation matrix.

To summarize, we devised a simple model capable of reproducing clustering of random inputs into what appears to be is well approximated by a discrete set of possible responses. This is reminiscent of what has been observed in experimental recordings from populations of neurons in mouse auditory cortex (Chapter 5). In order to understand this model better we next describe its different dynamic regimes.

7.2.4 *Normalizing the network's eigenvalue spectrum with respect to the network size*

We normalized the connectivity matrix $\mathbf{W}$ with respsect to the network size to control its eigenvalue spectrum. This normalization is based on Girko's circular law, which states that in the limit of large N all eigenvalues of a random matrix of size N drawn from a normal distribution are located within a cycle of radius $\sqrt{N}$ in the complex plane (Girko, 1984, Girko, 1990). This is of great importance for neural networks as the network behavior depends on some aspects of the eigenvalue spectrum of the underlying connectivity matrix. As the eigenvalue with the largest real part typically has a strong contribution the network dynamics (linear networks diverge, if it is > 1), the typical approach is to divide the standard deviation of network connections by $\sqrt{N}$. Later this has been shown for basically all uni-modal random distributions (e.g. Götze and Tikhomirov, 2007, Tao and Vu, 2010) and even for matrices observing Dale's law (Rajan and Abbott, 2006), which is especially helpful in our case.

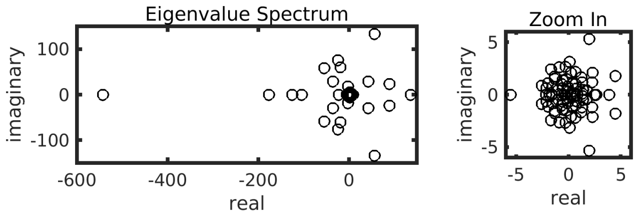
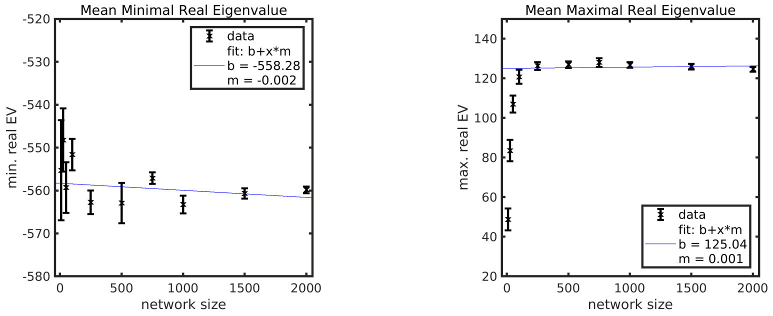


Figure 7.4: Eigenvalue spectrum of example network displayed in Figure 7.1. The dominant features of the network are determined by the minimal real eigenvalue and the radius of the distribution around the origin (i.e. the maximal real eigenvalue).

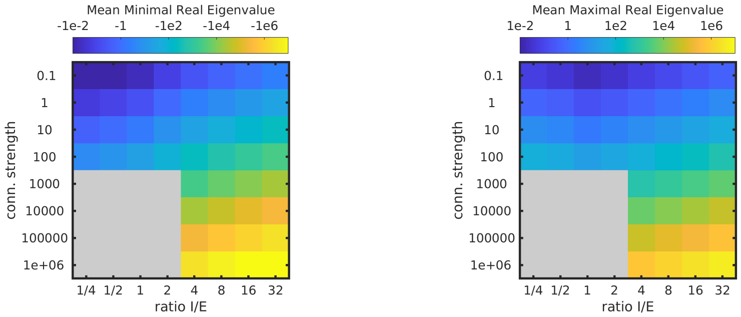
So, we divided the mean μ_k and the variance σ_k^2 of the distributions that were used to generate $\mathbf{W}$ (Equation 7.3) by the number of neurons N. This made sure that the eigenvalue spectrum is independent of the network size (see Rajan and Abbott, 2006, Wei, 2012). The eigenvalue spectrum for the example network from Figure 7.1 is displayed in Figure 7.4. The stability of the minimal and the maximal real eigenvalue, i.e. the in our case almost always negative mean of the matrix and the radius of the spectrum around the origin, with regard to changes in the network size is shown in Figure 7.5. This was done so the recurrent connections into each neuron are approximately independent of the network size. The recurrent connections into each neuron on average sum up to the fraction of excitatory or inhibitory neurons in the network multiplied by the parameters μ and $R_{I/E}$, which we varied in Section 7.3.1. Note that in our case, the system was working away from the linear regime and thus even eigenvalues with large positive real part were not enough to make



(a) Min. EV vs. Network Size (b) Max. EV vs. Network Size

Figure 7.5: Scalability of the eigenvalue spectrum of the connectivity matrix with network size. (a) Mean minimal real eigenvalue (errorbars denote SEM) vs. network size. Fit (performed on networks of size $N \geqslant 100$) is a straight line with slope $m \approx 0$. (b) Mean maximal real eigenvalue (errorbars denote SEM) vs. network size. Fit (performed on networks of size $N \geqslant 100$) is a straight line with slope $m \approx 0$. Connection strength and ratio I/E were set to 100 and 8 respectively. The choice of these parameters determines the layout of the eigenvalue spectrum independent of the network size. Small networks displayed statistical fluctuations due to small random samples and finite size effects and were thus excluded from the fit.

the network activity diverge.



(a) Min. EV (b) Max. EV

Figure 7.6: Minimal (a) and maximal (b) real eigenvalue averaged over 100 implementations of the network show proportionality to both connection strength and ratio of inhibition to excitation.

With this normalization in place we varied the mean recurrent connection strength μ and the fraction of inhibition to excitation $R_{I/E}$. This results in changes in the eigenvalue spectrum of the connectivity matrix and thus also changes the network dynamics. The smallest and the largest real eigenvalue as a function of μ and R_{I_E} can be seen in Figure 7.6. Eigenvalues with real part larger than one are typically associated with diverging network behavior in linear networks. Our network operates far away from the linear regime for strong inhibition. Also, for small networks this diverging behavior can be balanced by inhibition and finite size effects (Harish and Hansel, 2015).

Similarly, the stimulus input was divided by the network size N, to keep the average input to the network independent of the network size. Thus the mean network activity was independent of the network size.

7.3 RESULTS

7.3.1 *A scan of parameter space reveals a clustering regime for strong inhibition and strong recurrent connections*

To characterize the observed clustering of different stimuli onto a shared response mode, we wanted to understand in what parameter regime this clustering can be found. We kept the network size fixed at $N = 100$ and set the number of stimuli to 40 as before. We systematically varied the mean connection strength μ and the ratio between inhibition and excitation $R_{I/E}$ (as defined in Section 7.2.1. We first plotted the mean network activity over all neurons, networks and stimuli (Figure 7.7). We found a general dependency of this mean activity on both the ratio between inhibition and excitation $R_{I/E}$ and the mean recurrent connection strength μ. The stronger μ was the higher the network activity was and the stronger the inhibition was the weaker the network activity. This lead to a regime with divergent network activity (colored gray in Figure 7.7 and subsequent figures) for high recuerrent connectivity and weak inhibition.

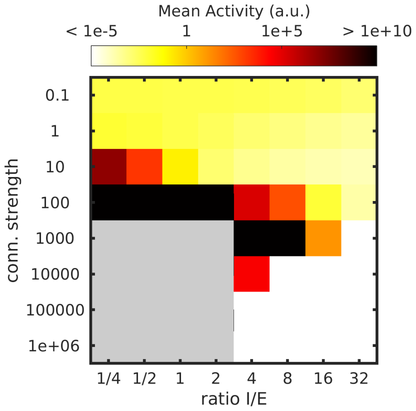


Figure 7.7: Parameter space scan for mean network activity reveals the effect of connection strength and ratio I/E on network activity. A regime of diverging network activity was found for high recurrent connectivity and low ratio of inhibition to excitation (colored gray).

We found different regimes of activity patterns in the network. These are shwon in Figure 7.8 and Figure 7.9 for our example connectivity from Section 7.2. In Figure 7.8 the response vectors of the network are displayed as a function of the mean recurrent connection strength μ and the ratio of inhibition to excitation $R_{I/E}$ sorted by hierarchical clustering (compare Figure 7.3a). For better visibility the colormap is adjusted individually for each pair of parameters. The change in mean network activity across parameter space is displayed in Figure 7.7. In Figure 7.9 one can see the corresponding correlation matrices (compare Figure 7.3b), also sorted by hierarchical clustering.

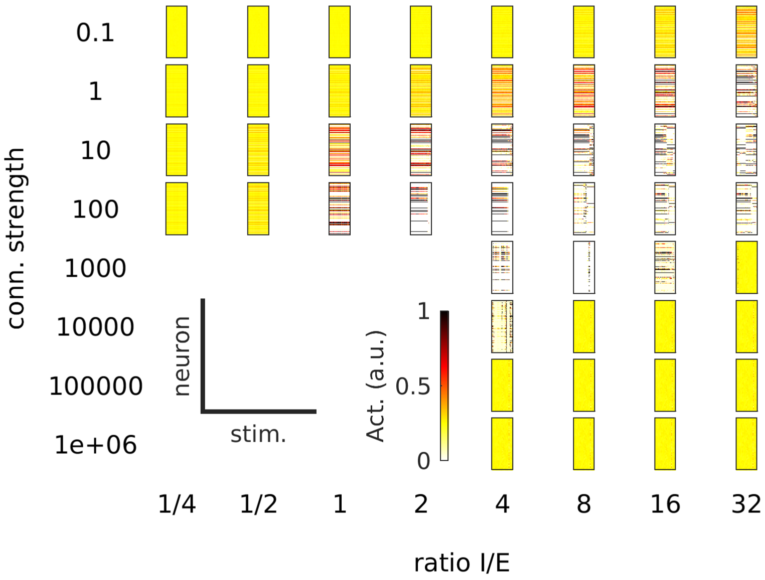


Figure 7.8: Response vectors (sorted via hierarchical clustering) for a systematic scan of mean recurrent connection strength and ratio of inhibition to excitation. Varying these parameters reveals different regimes of parameter space. Note: The lower left is empty due to divergence of network responses.

The lower left (i.e. strong recurrent connections and little inhibition) is left empty, because in this regime the network activity diverged. In the non-diverging part of parameter space we found four regimes. In the top left, for low connectivity strength and low relative inhibition, the network activity is dominated by the input, the correlations between responses to different stimuli are close to 0, as expected in an input dominated regime with all random inputs. As we increased the connection strength or the ratio of inhibition to excitation, we found a second trivial regime. Here, the network response to all the different stimuli is basically the same, the correlation between responses to any two stimuli is close to 1. Another regime can be seen in the lower right, i.e. for strong connections and for very high inhibition.

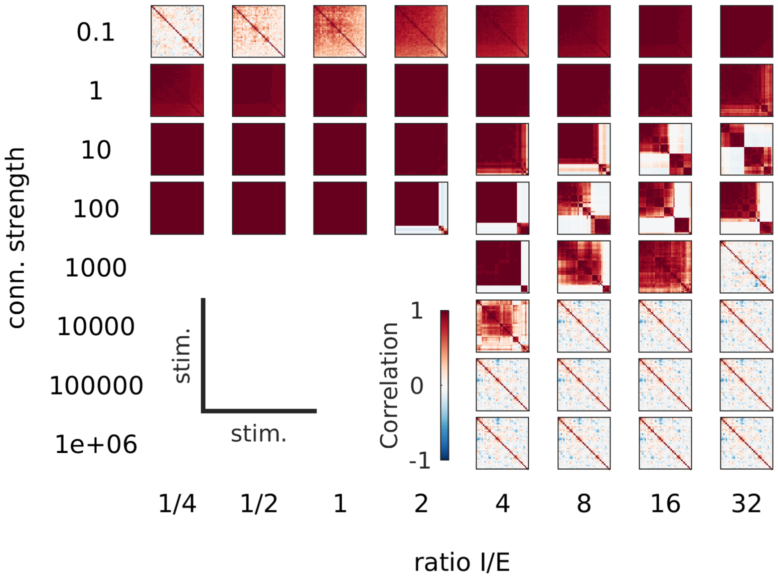


Figure 7.9: Correlation matrices of response vectors in Figure 7.8 (sorted via hierarchical clustering) for a systematic scan of mean recurrent connection strength and ratio of inhibition to excitation. Varying these parameters reveals different regimes of parameter space. Note: The lower left is empty due to divergence of network responses.

Here the responses seemed to be input dominated again, as each input evoked its specific network response. This can be understood by dampening: the network receives input at the very first time step of the stimulation. This is independent of the recurrent connections, as they only start to contribute, once the network becomes active. In this regime inhibition is so strong, it just inhibits all activity (i.e. the input into every neuron is 0 after the half-wave rectification), and thus the activity decays exponentially. The resulting correlations are a result of the not yet completely dampened activity of the very first time point of stimulation. And we found a fourth regime between the single response regime and the dampening regime. In this fourth regime, responses to random stimuli typically fall into one of a small – near discrete – set of possible responses. We called this regime the *clustering regime*, because these stereotypical responses are captured by a set of response clusters.

In order to get a better understanding of these regimes we turned to multiple statistical measures (to be defined below) while varying both the connectivity strength and the ratio of inhibition to excitation (Figure 7.10): (a) the mean activity across stimuli, neurons and network implementations (as we already did in Figure 7.7), (b) the Gini-coefficient (Gini, 1936) of the response activity distribution, (c) the mean correlation of responses to different stimuli in the same realization of the network, (d) the mean correlation of responses to the same stimulus across different implementations of the network, (e) the cluster number and (f) the dimensionality (computed as in Abbott et al., 2011). Together these six measures help us to understand the five different dynamic regimes.

(a) The *mean activity* (Figure 7.10a) is the mean activity across stimuli, neurons and random implementations of the network for each combination of the parameters connectivity strength and ratio of inhibition to excitation. As expected the mean activity diverged in networks with strong connectivity and little inhibition. The region of divergence beyond machine precision (1.79×10^{308}) is colored gray.

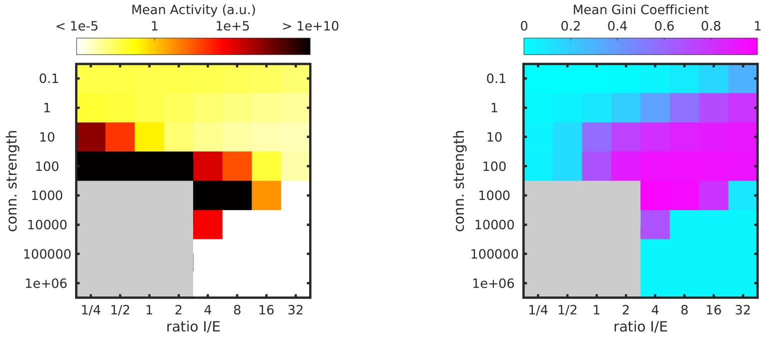
(b) The *Gini coeffecient* (Gini, 1936, Figure 7.10b) is a measure for the inequality of a distribution. It is probably best known from economics as a measurure for the inequality of the distribution of wealth in a society. It is defined as half of the relative mean absolute difference between all pairs of values from a distribution (Equation 7.8):

$$G = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} |x_i - x_j|}{2N^2\mu}, \tag{7.8}$$

with N values x_i and their mean $\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$. It can range from 0 to 1; a Gini coefficient of 0 indicates a perfectly equal distribution, whereas a Gini coefficient of 1 expresses maximal inequality.
We used this coefficient as an indicator of inequality in the distribution of neuronal responses, as the distribution of experimentally observed neural activity is typically log-normal like (e.g. Buzsaki and Mizuseki, 2014) with a high Gini coefficient, meaning most neurons are inactive most of the times, but some are very active at some times.

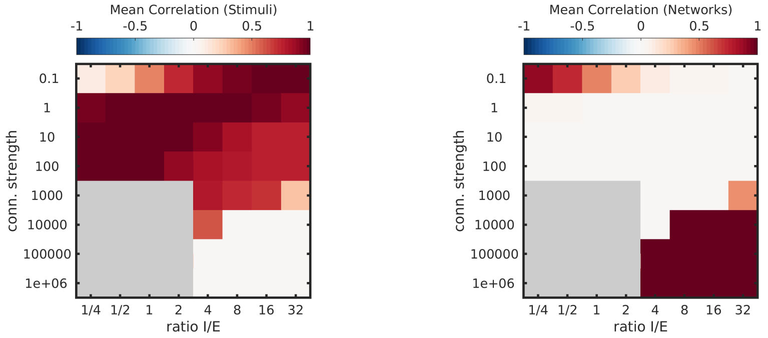
(c) The *mean correlation* of response vectors to different stimuli *within* a single implementation of the network (Figure 7.10c) is a measure of response structure. It was computed as the Pearson correlation coefficient (Galton, 1886, Pearson, 1895) between all responses of a single implementation of the network and then averaged across stimuli. It is 0 if there is no correlation at all between the stimulus responses and it is 1 if all stimuli evoke the exact same response.

(d) The *mean correlation* of response vectors to the same stimulus *across* implementations of the network (Figure 7.10d) describes the influence the afferent input has as compared to the recurrent



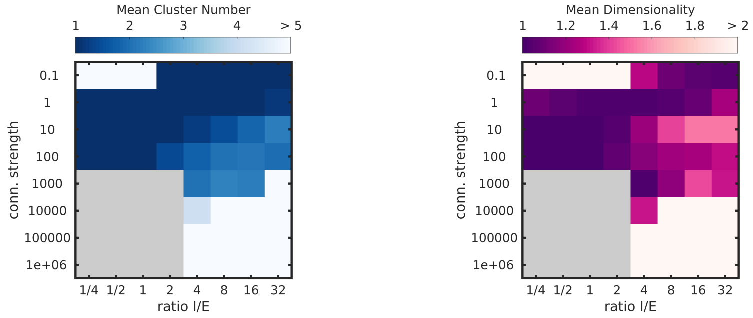
(a) Mean activity.

(b) Mean Gini coefficient.



(c) Mean correlation (stimuli).

(d) Mean correlation (networks).



(e) Mean cluster number.

(f) Mean dimensionality

Figure 7.10: Parameter space scan for multiple measures reveals qualitatively different regimes of our network model. Note: The gray area indicates divergence of network responses. Experimental distributions (where applicable) can be found in Figure 7.11. Experimentally observed values are best apporached for a connection strength between $\mu = \{10, 100, 1000\}$ and a ratio of inhibition to excitation between $R_{I/E} = 4, 8, 16, 32$. (a) Mean neuronal activity. (b) Mean Gini coefficient of activity distribution across 100 random implementations of the network. Experiment: 0.908 ± 0.004 (mean ± SEM). (c) Mean correlation across response vectors to different stimuli within a network. Experiment: 0.277 ± 0.006 (mean ± SEM). (d) Mean correlation of response vectors to the same stimulus across networks. (e) Mean cluster number. Experiment: 5.0 ± 0.2 (mean ± SEM). (f) Mean dimensionality. Experiment: 3.16 ± 0.06 (mean ± SEM).

connectivities. If the input dominates and the recurrent connections do not matter, the response of different networks to the same stimulus is the same no matter the recurrent connections and thus the correlation of response vectors from different networks to the same stimulus is close to 1. If recurrent connections dominate, the responses of random implementations of the network to the same stimulus are different and thus the correlation of response vectors from different networks to the same stimulus is close to 0.

(e) The *cluster number* (Figure 7.10e) is computed by application of Hubert's Γ statistics (Hubert and Baker, 1977) to infer clustering power. We sorted the data by hierarchical clustering and, in order to find the suitable number of clusters, cut the cluster tree at every possible level and computed a Γ value as the distance between the perfect clusters described by this clustering and the actual correlation matrix, sorted by this clustering. The cluster number was then chosen to be the number with maximum Γ value, which we defined as:

$$\Gamma = \frac{2}{O(O-1)} \sum_{i=1}^{O} \sum_{j=i+1}^{O} (S_{ij} - c) T_{ij}, \tag{7.9}$$

where S is the original correlation matrix of size O, c is a threshold and T is a binary matrix of size O with entries

$$T_{ij} = \begin{cases} 1, & \text{if } i \text{ and } j \text{ are clustered together,} \\ 0 & \text{otherwise.} \end{cases} \tag{7.10}$$

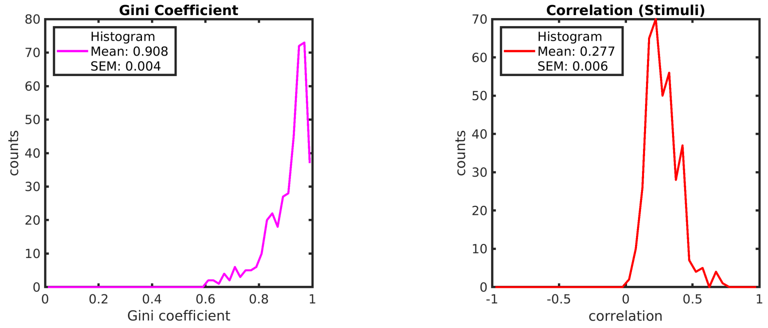
The threshold was set to $c = 0.4$. This Γ value becomes maximal for the clustering that describes the data best. This method was applied to experimental data, too (Section 5.4.20).

(f) The *dimensionality* (Figure 7.10f) is defined via the eigenvalues λ_i of the covariance matrix of the response vectors (Abbott et al., 2011) as:

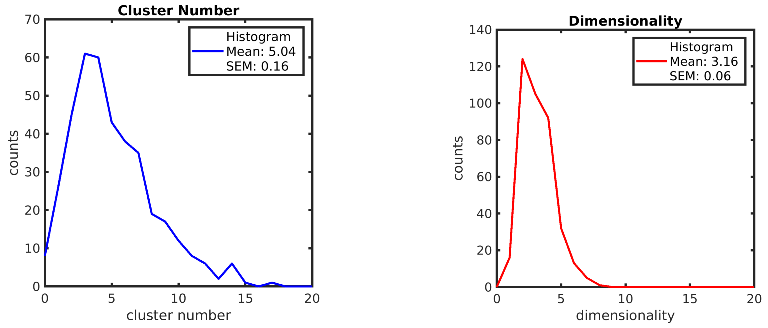
$$d = \frac{(\sum_i \lambda_i)^2}{(\sum_i \lambda_i^2)}. \tag{7.11}$$

It is a measure of the number of eigenvalues that are larger than the rest of the eigenvalues, so the number of dominant orthogonal dimensions.

As discussed in the following, based on these measures the different regimes can be understood as (1) a *linear regime* in the top left of the parameter space plot for low recurrent strength and little inhibition, (2) a *single-point-attractor regime* for stronger recurrent connections



(a) Gini coefficient (b) Correlation (stimuli)



(c) Cluster number (d) Dimensionality

Figure 7.11: Histograms of experimentally measured values for comparison to model data. Data taken from Chapter 5. (a) Gini coefficient. (b) Mean correlation between stimulus responses within a FOV (of all stimuli evoking a response). (c) Cluster number. (d) Dimensionality.

and/or stronger inhibition, (3) a *clustering regime* for even stronger recurrent connections and stronger inhibition, (4) a *diverging regime* for very strong recurrent connections and little inhibition (bottom left) and (5) a *dampening regime* for very strong recurrent connections and a lot of inhibition (bottom right).

The *linear regime* (1) is input dominated. Each stimulus evoked a different response (Figure 7.8 and Figure 7.9). Dimensionality (Figure 7.10f) and cluster number (Figure 7.10e) were high. However, there was a high similarity between responses of different implementations of the network to the same stimulus (Figure 7.10c). So, the recurrent connections played a minor roll, the input dominated. We called this regime *linear* as each stimulus was mapped to its own unique response.

In the *point attractor regime* (2) the recurrent connections dominated the activity patterns. Irrespective of the input (and even though the input is changing), each implementation of the network had its specific response pattern it converged to. Mean correlation between reponses within a network was 1 (Figure 7.10d), so were cluster number (Figure 7.10e) and dimensionality (Figure 7.10e).

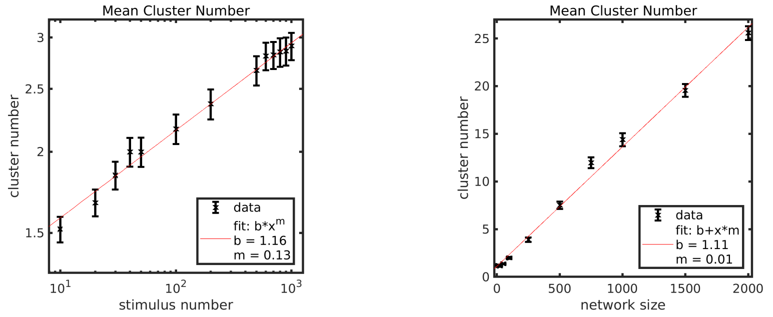
Classically, these two regimes are often not distinguished from each other and simply referred to as *uni-stable*, as in both cases each stimulus evokes a single clearly defined response. Rather than presenting multiple stimuli and defining both a stimulus dominated regime (1) and a network dominated regime (2), the distinction is traditionally only made between a *uni-stable* regime, a *bi-stable* (or a *multi-stable*) regime, where a single stimulus can ellicit two (or more) responses, depending on the initial condition of the netwok, and a chaotic regime (e.g. Wilson and Cowan, 1972, Sompolinsky et al., 1988, Fasoli et al., 2016, Zhang and Saggar, 2020).

When we furhter increased both connection strength and ratio I/E, we found a *clustering regime* (3). This is reminiscent of a classical *multi-stable* regime (Sompolinsky et al., 1988, Zhang and Saggar, 2020). However, apart from one stimulus being able to evoke multiple responses, multiple stimuli could also evoke the same response. So, stimuli were mapped seemingly randomly to one of the response modes. There were typically $2-3$ response modes (best to be seen in Figure 7.9). Taking together all the measures described above, this regime is both the most relevant regime as it is the one closest to experiment. While activity was at an intermediate level (Figure 7.10a), in this regime activity was also skewed such that most neurons were inactive, while some were highly active as captured by a Gini coefficient close to 1 (Figure 7.10b). Correlations between responses of the same network to different stimuli were at an intermediate level (Figure 7.10c), meaning there was more than one possible response (in which case corr. $\rho = 1$), but not a unique response to each stimulus (corr. $\rho = 0$). Correlations between networks were found to be 0 (Figure 7.10d), meaning that the observed clustering occurred to group stimuli together randomly. Additionally, the mean cluster number (Figure 7.10e) as well as the mean dimensionality (Figure 7.10f) were at an intermediate level. They both were between 1 in the *single-point-attractor* regime, where only one response is possible per network and the *linear* regime, where each stimulus evokes its specific response.

The *diverging regime* (4) is characterized by very strong recurrent connections and not enough inhibition present to possibly counter balance the strong excitation. The network activity is diverging.

In the *dampening regime* (5) the inhibition grows so strong that it supresses every activity exponentially. As we cut off the network activity after 5τ, however, the network activity did not reach 0, yet, and the activity was basically a dampened version of the input. Thus, we found, similar to the linear regime that the mean correlation of responses of the same network to different stimuli (Figure 7.10c) was close to 0, whereas the correlation of responses of different networks to the same stimulus (Figure 7.10d) was close to 1.

A regime matching to experimental data best could be found within the *clustering regime*. Distributions of experimental measurements (where applicable) are shown in Figure 7.11. In the experimental data (taken from all FOVs and all imaging days of the basal dataset from Chapter 5) the Gini coefficient was found to be 0.908 ± 0.004 (Figure 7.11a, mean ± SEM), the correlation across response vectors within a FOV was 0.277 ± 0.006 (Figure 7.11b), the mean cluster number was 5.0 ± 0.2 (Figure 7.11c), and the dimensionality was 3.16 ± 0.06 (Figure 7.11d). The absolute value of activity was of no interest as the model network activity scales with the input and the correlation of response vectors to the same stimulus across FOVs could not be computed in the experiment, as the FOVs consisted of different neurons and did not receive the exact same stimulation. The similarity between experiment and model appeared without any fine tuning whatsoever and for all of the above measures in in the same region and only in this region of parameter space. We just varied the mean connectivity and the ratio of inhibition to excitation and found a region corresponding to experimental data for strong recurrent connections and strong inhibition. This means that the observed experimental clustering of stimulus responses into response modes can be explained by random connectivity in a strongly recurrent network with relatively strong inhibitory connections.



(a) Cl. Num. vs. Stim. Num. (b) Cl. Num. vs. Network Size

Figure 7.12: Model scalability. (a) The growth of cluster number with the number of stimuli can be fit by a power law with exponent 0.13. Note: Axes are loglog scale. (b) For a fixed number of stimuli ($n_{\text{Stim}} = 40$) the growth of cluster number with the network size can be fit by a line of slope 0.01. Cluster number is computed via Hubert's Γ statistics (Hubert and Baker, 1977) and averaged over 100 random implementations of the network.

The clustering in the *clustering regime* was found independent of network size and stimulus number. However, when we increased the stimulus number, we found a power law dependency on the cluster number (Figure 7.12a, $n_{Cl} = 1.16x^{n_{Stim}}$). So, if the stimulus number is increased, there tends to be a stimulus evoking a new response mode. When we increased the network size, we found a linear dependency between the cluster number and the network size (Figure 7.12b, $n_{Cl} = 1.11 + 0.01N$), at least for network sizes of up to $N = 2000$ neurons. For a finite number of stimuli this has to be bounded, however. We also found an increase of cluster number with network size in auditory cortex, where on a global scale each stimulus evoked a unique response through combinations of clustered responses in single FOVs (Figure 5.10b).

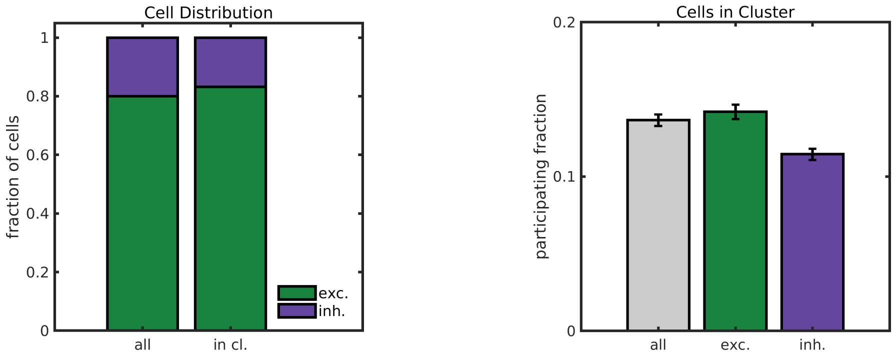
Investigating the eigenvalue spectra or specifically the minimal and maximal real eigenvalues in parameter space, we found a dependency of both connection strength and the ratio between inhibition and excitation. The minimal real eigenvalue was either determined by inhibition, if inhibition was stronger than excitation or by the width of the distribution, if excitation and inhibition were of similar strength (Figure 7.6a). The maximal real eigenvalue was (as a proxy for the radius around the origin of the complex plane) determined by the larger of both inhibition or excitation, so it was mirrored at a ratio of 1 and it was growing with the connection strength (Figure 7.6b). In the relevant (i.e. the clustering) regime the eigenvalue spectrum has one large negative eigenvalue and the radius of the circular distribution of eigenvalues around the origin is well beyond 1. So, the system works far away from the linear regime (where activity would diverge). The large negative eigenvalue is determined by the overall strength of inhibition, the large positive eigenvalue accounts for strong network activations. Together this leads to highly non-linear dynamics via the static non-linearity and high activity, especially of inhibitory neurons.

To sum up, we found four intuitive regimes, a *linear* regime, a *single attractor* regime (both *uni-stable*), a *diverging* regime, and a *dampening* regime – plus a *clustering* regime (which is reminiscent of the *multi-stable* regime in the literature) – by varying the strength of recurrent connections and the ratio of inhibition to excitation. The *clustering* regime that is consistent with experimental data from mouse auditory cortex (Chapter 5), was found for relatively strong recurrent connections and strong inhibition.

7.3.2 *Clusters are formed by weaker than average inhibitory synapses between participating neurons*

This section describes joint work with Lusie Schulte as part of her B.Sc. project (Schulte, 2017).

We next wondered what is the underlying neuronal structure of the clusters of stimulus responses. With this aim we looked at the neurons that were active when one of these response modes was evoked in networks in the *clustering* regime described in (Section 7.3.1). We set the mean connectivity to 100 and the ratio of inhibition to excitation to be 8. One obvious hypothesis would be that the activity patterns associated with these clusters mostly consist of excitatory neurons that are interconnected and thus form a positive feedback loop. So, we counted the neurons that were active during population responses. Neurons were considered active, when their average activity to all stimuli within a cluster was larger than the average activity of all neurons to these stimuli. And indeed we found a slight excess of excitatory neurons associated with response modes as compared to the overall distribution of neurons (Figure 7.13a). While the overall fraction of excitatory cells was set to be 80% (and thus the fraction of inhibitory cells was 20%), these fractions were shifted to 83.23 ± 2.74% of excitatory and only 16.77 ± 0.53% of inhibitory cells within responses to clustered stimuli.



(a) Distribution of cells.

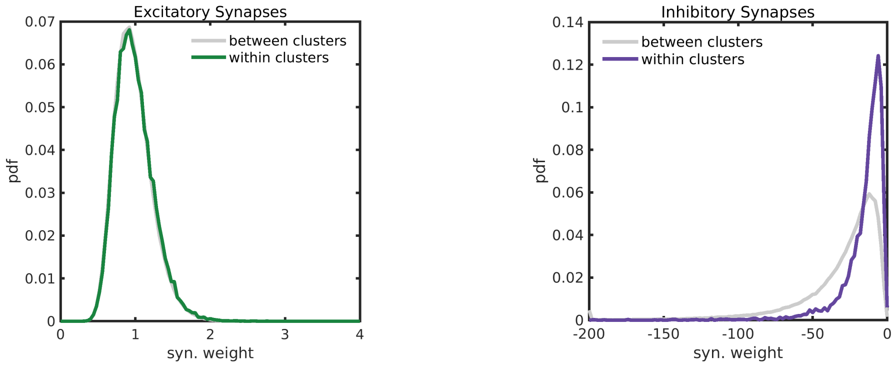
(b) Cells in cluster.

Figure 7.13: Activity patterns underlying stimulus response clusters are characterized by slightly more excitatory cells. (a) Distribution of inhibitory and excitatory cells in clustered response patterns as compared to the entire dataset. (b) Fraction of cells participating in at least one clustered response pattern. From left to right: fraction of all cells, excitatory cells, inhibitory cells.

This slight imbalance could also be seen, when we looked at the fraction of cells that participated in the activity pattern for at least one cluster (Figure 7.13b). 13.66 ± 0.38% of all cells were active, 14.21 ±

0.47% of excitatory cells, and 11.45 ± 0.36% of inhibitory cells.

To gain more insight, we also looked at the distribution of synapses between neurons activated by stimuli within clusters and compared them to synapses between neurons that were not active together. Figure 7.14 shows the distributions of excitatory (Figure 7.14a) and inhibitory synapses (Figure 7.14b) both within and between (i.e. outside) clustered activity patterns. The distribution of excitatory synapses is basically the same within and between those patterns. For inhibitory synapses there is a difference between these distributions: Inhibitory synapses between neurons inside a clustered response pattern were typically weaker than inhibitory synapses outside of clustered response patterns.



(a) Excitatory synapses.

(b) Inhibitory synapses.

Figure 7.14: Activity patterns of stimulus response clusters are characterized by weak inhibitory connections between participating neurons. (a) Distribution of excitatory synaptic connection strengths is the same for neurons within and between clustered activity patterns. (b) Distribution of inhibitory synaptic connection strengths reveals a difference between synapses between neurons within a clustered activity pattern and synapses between neurons outside clustered activity patterns.

In conclusion, while we did find a slight excess of excitatory neurons inside population responses to clustered stimuli compared to the overall distribution, the clustered activity patterns could be described best by a lack of strong inhibitory connections within. Intuitively this makes sense as well, as a strong inhibitory synapse from one neuron to another would inhibit the second neuron and thus remove it from the pattern. So, a *clustering* regime, reminiscent of a *multi-stable* regime, as has been described in literature, can arise in random networks via strong inhibitory connections. Those lead to exclusive activation of response patterns depending to some extent on the stimulation, different from the uni-modal network dominated regime, where any activation will lead to only one response, as the differentiation via structured inhibition is impossible due to only weak inhibitory synapses.

7.4 DISCUSSION

We found that a random firing rate model was able to capture key features of stimulus response dynamics in mouse auditory cortex. Those key features include sparse neuronal activity, best described by a skewed distribution towards high rates, and a clustering of stimulus responses into response modes. In conclusion this makes the presented firing rate model a liable model of neuronal population responses in mouse auditory cortex, especially as it works on few assumptions and a working regime is found already in random (i.e. non-structured) connectivity matrices.

A clustering of stimulus responses was found in a regime of parameter space with strong excitatory and even stronger inhibitory recurrent connections. This clustering means that different stimuli evoked the same population response. The cell assemblies behind the clusters (i.e. the network units activated by the stimuli) were characterized by weak inhibitory synapses among them and strong inhibitory synapses between.

The observed attractor-like dynamics are reminiscent of so-called *winner-takes-all* networks (e.g. Wilson and Cowan, 1973). Such networks typically consist of multiple units (or groups of units) and the activation of one unit (or one group of units) inhibits all other activity. In these networks the structure is usually engineered to work like this. We find that strong inhibitory synapses between neurons underlying one response mode and other neurons not activated by this response mode are able to produce similar dynamics in random networks.

This clustering of different stimuli onto a response mode seems to be characteristic of certain regions in the cortex, like the auditory cortex (Bathellier et al., 2012, Atencio and Schreiner, 2013, See et al., 2018), while it has not been reported in other regions such as e.g. visual cortex. Our model suggests that these different cortical areas might work in different regimes and that clustering is associated with strong inhibition. Interestingly, auditory cortex is thought to be more inhibition dominated than other areas, e.g. visual cortex (Liang et al., 2019). A similar and potentially related effect (although termed differently) has also been reported by Chen and Miller, 2020, who found that in small chaotic networks the number of states decreased by increasing both recurrent connections and inhibition.

Apart from the clustering regime we were able to identify four more regimes. While two of them were rather trivial (diverging responses for too strong excitation and exponentially decaying responses for too strong inhibition), we found two uni-stable regimes. We were able to make the distinction into an input dominated regime, where each input evoked a unique response (regardless even of the network configuration), and a network dominated regime, where each stimulus evoked the same response (only depending on the network configuration). This distinction is not always made in previous work, probably because networks are typically probed with a single stimulus, especially for parameter scans, to save time.

Similar studies that performed parameter scans of different parameters, typically found three regimes: a uni-stable regime, a bi- or multi-stable regime, and a chaotic regime (Sompolinsky et al., 1988, Stern et al., 2014, Kadmon and Sompolinsky, 2015, Zhang and Saggar, 2020). As written above, we identified two distinct uni-stable regimes. A bi- or multi-stable regime produced clustering into response modes in our model. We did not find chaotic behavior, probably due to two reasons: (1) Our non-linearity was not bounded, allowing activity to diverge for strong recurrent connections and no strong inhibition. (2) Inhibition (together with finite size effects) is able to prevent chaotic behavior, even for connectivity matrices with large positive real eigenvalues (Ostojic, 2014, Harish and Hansel, 2015, Fasoli et al., 2016).

8

ABRUPT TRANSITIONS OF RESPONSE PATTERNS EMERGING FROM GRADUAL CHANGES OF NETWORK CONNECTIVITY

In this chapter we discuss the occurence of abrupt transitions of response patterns that follow gradual changes of network connectivity. Section 8.1 gives an introduction to considerations on the effect of changes in the network's structure on the network response. Then, we will explain the model setup (Section 8.2), before we detail our findings in Section 8.3, namely, that abrupt transitions of network response patterns are linked to qualitative changes in the fixed points picture of the network. We end this chapter with a discussion (Section 8.4).

8.1 INTRODUCTION

In the previous chapter (Chapter 7) we devised a model capable of reproducing single time point dynamics of neuronal population responses in mouse auditory cortex (as described in Chapter 5), most notably a clustering of responses to stimuli into a near discrete set of response modes. This clustering into response modes was found in random networks with strong recurrent connections and a high ratio of inhibition to excitation. In Chapter 5 we also saw considerable remodeling of said response modes even under basal conditions. In this chapter we want to understand, how synaptic drift can lead to such remodeling of representations. How stable are network responses towards synaptic drift? And is there a way to link representational drift to the underlying synaptic drift?

Recent studies in various regions of mouse neocortex found that both neuronal population activity and the underlying synaptic connectivity change over time. Synapses change in strength, they emerge and disappear (Rumpel and Triesch, 2016). While learning induced changes of synaptic connections have been studied extensively in the past (brought to attention by Hebb, 1949, for an overview of recent work, see Humeau and Choquet, 2019), the focus has partly shifted in recent years towards synaptic changes in the absence of an explicit learning paradigm. This so-called synaptic drift, i.e. seemingly random changes in synapse strength, including disappearing and emerging synapses, was found in both excitatory (Yasumatsu et al., 2008, Loewenstein et al., 2011, Loewenstein et al., 2015, Berry and Nedivi, 2017, Ziv and Brenner, 2018) and inhibitory synapses (Rubinski and Ziv, 2015, Dvorkin and Ziv, 2016, Villa et al., 2016). But not only the connectivity of neural networks changes, also neuronal population activity patterns change across time as shown by us (see Chapter 5 and Chapter 6) and others (e.g. Clopath et al., 2017, Rule et al., 2019).

A lot of work has been focused on the question, how stability on the level of cortical representations can be maintained in the presence of synaptic drift (Vogels et al., 2011, Litwin-Kumar and Doiron, 2014, Mongillo et al., 2017, Mongillo et al., 2018, Fauth and Rossum, 2019, Montangie et al., 2020). In these studies, this is mostly achieved by homeostatic plasticity rules and/or by leaving relevant parts of the network stable. Homeostatic plasticity rules can also be applied to the readout units of networks with changing activity patterns, at least for not too fast representational drift (Acker et al., 2019, Kossio et al., 2021).

Here, however, we do not want to explain stability, but want to instead study the influence of changes in network connectivity on network activity. As both synaptic connectivity and neuronal population activity have been shown to be not as stable as previously assumed in various regions of mouse auditory cortex – even under basal conditions – we investigate the link between these two. This drift seems to be ever present, so it probably plays a role for different processes and has to be accounted for, before external influences – like learning – can be addressed. To our knowledge this has not been investigated before.

We used the firing rate model as described in Chapter 7 in the regime, we found, similar to experimental recordings of mouse auditory cortex (Chapter 5) and apply synaptic drift as fit to experimental data by Loewenstein et al. (2011). We observed that gradual synaptic drift leads to periods of stable responses to stimulation, that are interrupted by abrupt transitions towards new network response patterns. We studied the underlying mechanism of these changes by analyzing the fixed point topology of the network model, extending a method from Sussillo and Barak (2013). We found that abrupt transitions cannot be explained by a displacement of existing fixed points, but typically coincide with qualitative changes in the fixed point structure.

8.2 MODEL SETUP

For the analyses in this chapter we used the circuit model introduced and discussed in detail in Chapter 7.

8.2.1 *Modeling drift of excitatory synapses*

Synaptic drift of excitatory synapses was modelled as described by Loewenstein et al. (2011). Here, we want to briefly summarize their model of synaptic drift. In experiments, dendritic spine size is taken as a proxy for postsynaptic efficacy of *excitatory* synapses. We use all-to-all connectivity matrices in our model and thus neglected the removal and (re-)appearance of synapses.

As the steady state distribution of dendritic spine sizes is log-normal distributed, the spine size dynamics were modelled by a multiplicative random process. The starting point are two independent Ornstein-Uhlenbeck processes (Equation 8.1, Uhlenbeck and Ornstein, 1930). Each of the two processes ($i = 1, 2$) can be described by:

$$\tau_i \dot{X}_i = -X_i + \xi_i, \tag{8.1}$$

where τ_i determines the time scale of the process, X_i are the dynamic variables and ξ_i is a white noise term with mean $\langle \xi_i \rangle = 0$ and covariance $\langle \xi_i(t) \xi_j(t') \rangle = 2\tau_i \sigma_i^2 \delta_{ij} \delta(t - t')$. σ_i^2 are the stationary variances of the respective processes. Now, the logarithm of the spine size S was fit to the sum of two such dynamic variables (Equation 8.2):

$$\log(S) = X_1 + X_2 + \mu, \tag{8.2}$$

where the constant μ is the mean of the logarithm of all synapse sizes. Thus, the system converges to a steady state, where synapse sizes are log-normal distributed, and we can solve for the probability of a synapse size O_t at time t dependent on its size O_{t-1} at time $t - 1$ (Equation 8.3):

$$P(O_t | O_{t-1}) = \frac{1}{O_t} \frac{\exp\left(\frac{\log O_t - (\beta \log O_{t-1} + (1-\beta)\mu)^2}{2\sigma^2_{O_t O_{t-1}}} \right)}{\sqrt{2\pi\sigma^2_{O_t O_{t-1}}}} \tag{8.3}$$

with $\beta = \frac{\sigma_1 \exp(-t/\tau_1) + \sigma_2 \exp(-t/\tau_2)}{\sigma^2_{O_{t-1}}}$, $\sigma_{O_t O_{t-1}} = \sigma_{O_{t-1}}(1 - \beta^2)$, where $\sigma^2_{O_{t-1}} = \sigma_1^2 + \sigma_2^2$ is the variance of the synapse size distribution. σ_1 and σ_2 were experimentally determined by Loewenstein et al. (2011) to be $\sigma_1 = 0.0683$ and $\sigma_2 = 0.0292$. As we handed the variance of synapse sizes $\sigma^2_{O_{t-1}}$ as input to our model (see Section 7.2), we computed σ_1 and σ_2 as $\sigma^2_{1,\,\text{model}} = \sigma^2_{O_{t-1}} \sigma_1^2 / (\sigma_1^2 + \sigma_2^2)$ and $\sigma^2_{2,\,\text{model}} = \sigma^2_{O_{t-1}} \sigma_2^2 / (\sigma_1^2 + \sigma_2^2)$ respectively. This process converges to a log-normal distribution or – if we already start from a log-normal distribution with the correct mean $\exp(\mu)$ and variance $\sigma^2_{O_{t-1}}$ – keeps this distribution of synapse sizes in a steady state. Note that the original model fitted in Loewenstein et al. (2011) included a term to account for experimental noise, which we omitted in the model. The time constants for the drift of excitatory synapses were obtained in the original publication and the best fit to experimental data was found to be $\tau_1 = 2.87\,\mathrm{d}$ and $\tau_2 = 212\,\mathrm{d}$. The mean logarithm of synapse sizes μ (and thus also the variance) was set in the model according to the regime in parameter space (as described in detail in Chapter 7). The mean connectivity strength was chosen to be $\mu = 100$ and the ratio $R_{I/E} = 8$.

8.2.2 *Implementation of synaptic drift*

To efficiently sample from the distribution of possible synapse sizes in Equation 8.3, we followed the standard procedure and projected it onto a distribution, we can sample from numerically, e.g. the uniform distribution. This was done for every synaptic weight at every time step. Mathematically this projection onto a uniform distribution is described by

$$u = \int_0^x P(O_t|O_{t\text{-}1})dO_t, \tag{8.4}$$

where u is a random number drawn from a uniform distribution and x is a random number drawn from $P(O_t|O_{t\text{-}1})$. We inverted this equation and solved it for x to get the corresponding synapse size. To achieve this we simplify Equation 8.3 by substituting $a = 2\sigma^2_{O_t|O_{t\text{-}1}}$, $b = (\beta\log(O_{t\text{-}1}) + (1-\beta)\mu)$, leading to:

$$P(x|O_{t\text{-}1}) = \frac{1}{\sqrt{\pi a}}\exp\left(\frac{-b^2}{a}\right)x^{\frac{2b}{a}-1}\exp\left(\frac{\log^2 x}{a}\right), \tag{8.5}$$

which can be further simplified by substituting $g = \frac{2b}{a} - 1$ and $h = \frac{1}{\sqrt{\pi a}}\exp\frac{-b^2}{a}$ to:

$$P(x|O_{t\text{-}1}) = hx^g\exp\left(-\frac{\log^2 x}{a}\right). \tag{8.6}$$

The cumulative sum of this expression up to a value x is then:

$$u = \int_0^x hy^g\exp\left(-\frac{\log^2 y}{a}\right)dy \tag{8.7}$$

$$= -\frac{1}{2}\left[\operatorname{erf}\left(\frac{b-\log x}{\sqrt{a}}\right) - 1\right]. \tag{8.8}$$

Now, this can be inverted and reads:

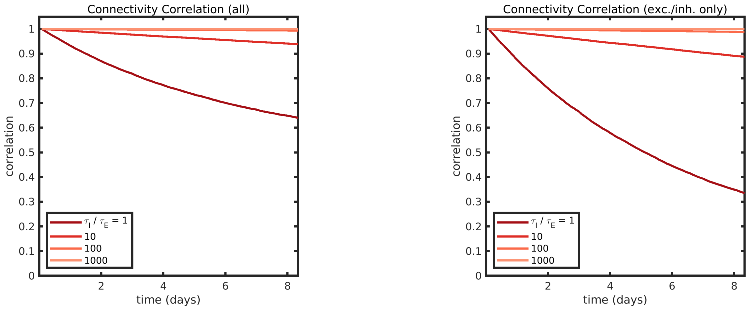
$$x = \exp\left[b - \sqrt{a}\operatorname{erf}^{-1}(1-2u)\right] \tag{8.9}$$

$$= \exp\left[(\beta\log(O_{t\text{-}1}) + (1-\beta)\log(\mu)) - \sqrt{2}\sigma_{O_t|O_{t\text{-}1}}\operatorname{erf}^{-1}(1-2u)\right]. \tag{8.10}$$

With this expression, we were able to sample u from a uniform distribution and efficiently compute the corresponding value drawn from $P(O_t|O_{t\text{-}1})$.

8.2.3 *Modeling drift of inhibitory synapses*

As there are – up to now – no direct experimental measurements of the drift of inhibitory synapses, we assumed that inhibitory synapses changed in the same way excitatory synapses do, but not necessarily on the same time scale as excitatory synapses. So, we varied the inhibitory time scale (i.e. τ_1 and τ_2) by factors of $\tau_I/\tau_E = \{1, 10, 100, 1000\}$. Figure 8.1a shows the correlation of connectivity weights at any time point to the intitial connectivity weigths, averaged over 100 random initializations of the network, starting from steady state. The larger the time scale τ_I is, the slower is the decay in this correlation.



(a) Conn. corr. all.

(b) Conn. corr. exh./inh. only

Figure 8.1: Correlation of network connectivity to connectivity at $t = 0$ (starting from steady state). (a) Correlation of connectivity weights to connectivity weights at $t = 0$ for the entire connectivity matrix (for different ratios $\tau_I/\tau_E = \{1; 10; 100; 1,000\}$). Mean over 100 implementations of the network. Shaded area is SEM. (b) Correlation of connectivity weigths to connectivity weights at $t = 0$ for connectivity matrices draawn from one log-normal distribution (only excitatory or only inhibitory; for different ratios τ/τ_E). Mean over 100 implementations of the network. Shaded area is SEM.

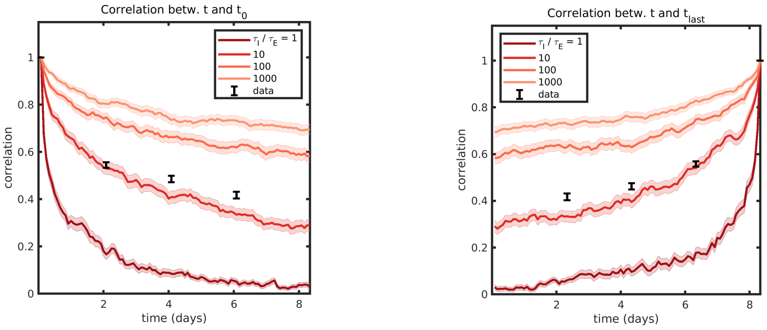
Overall, this plot shows that the correlation of connectivity weights is decaying rather slowly and as inhibitory weights are larger, increasing the time scale of drift of inhibitory synapses adds further to the slowness of this decay. To check whether this slow decay was due to the conservation of the overall structure of the network – i.e. inhibitory synapses could never become excitatory synapses and vice versa – we constructed networks of only excitatory (or inhibitory) synapses drawn from the same distributions as the original networks. We applied drift and found that the correlation to the initial connectivity weights decayed faster, but not by much (Figure 8.1b). The correlation decay was independent of mean and standard deviation of the distribution (as is expected for Pearson's correlation coefficent) and thus, only one line is shown for each time scale.

So, while the synaptic drift has quite a dramatic effect on individual synapses, as described in Loewenstein et al. (2011), the overall synaptic connectivity changes on a rather slow time scale. Note that in Figure 8.1 the standard deviation is small, too, in both cases. This indicates that individual implementations of the network all change with approximately the same rate.

8.3 RESULTS

8.3.1 *Drift of inhibitory synapses is predicted to be an order of magnitude slower than drift of excitatory synapses*

In order to investigate response changes due to the synaptic drift described above, we computed the responses to the same 40 stimuli as described in Section 7.2.2 using the connectivity matrix (Section 7.2.1) with $\mu = 100$ and $R_{I/E} = 8$ at each time step (starting from steady state with synaptic drift as described in Section 8.2). The response vectors were then correlated to the response vector for the same stimulus at the initial time point. The mean correlations to the first (and last time point) over 40 stimuli and 100 implementations of the network are shown in Figure 8.2. The correlation decay of response vectors is faster than the correlation decay of network connectivity. This means that relatively small changes to the network connectivity have a more pronounced effect on the network activity.



(a) Resp. corr. t_0

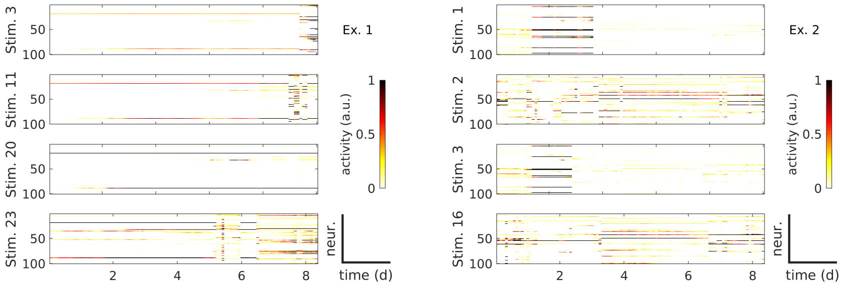
(b) Resp. corr. t_{last}

Figure 8.2: Correlation of network response to response at $t = 0$. (a) Correlation of response to response at $t = 0$ (for different ratios $\tau_I/\tau_E = \{1; 10; 100; 1,000\}$). Mean over the same 100 implementations of the network as in Figure 8.1 and 40 stimuli. Shaded area is SEM over networks. (b) Correlation of response to response at $t = t_{last}$ (for different ratios τ_I/τ_E). Mean over 100 implementations of the network and 40 stimuli. Shaded area is SEM over networks. An inhibitory synaptic turnover 10 times slower than excitatory turnover reflects experimental data best.

To infer the biologically plausible rate of drift for inhibitory synapses, we compared the correlation decay of network responses for different rates in the model to experimental data (taken from Chapter 5, i.e. Aschauer et al., 2022). For better comparison we computed the median response vectors for each stimulus on each day (to account for experimental noise), correlated them to the median response vector at the first day (day 0) or at the last day (day 6) of imaging, and plotted the mean over imaging fields of view and stimuli. We found that in order to achieve a similar level of change as in the experiment, inhibitory synapses had to be at least a factor 10 times more stable than excitatory synapses. This might be due to the fact that the model works in an inhibition dominated regime. This decay in correlation was mirrored, when we computed the correlation to the last time point instead of the first (Figure 8.2b). Note that the shape of the model decay and the shape of the correlation decay in the experiment did not match exactly, hinting at dynamics not included in the random drift of the model. While the decay from the intial time point to the next matched, the experimental data seems to be more stable on consecutive time points.

8.3.2 *Abrupt transitions of response patterns are caused by ongoing synaptic drift*

We next wondered, if the observed monotonous decay of mean response correlation – averaged over 40 stimuli and 100 implementations of the network (Figure 8.2) – would also be present in stimulus responses of individual networks. Stimulus responses of individual networks, however, showed a qualitatively different behavior. While the connectivity was changing slowly and steadily, also on the level of individual networks, most of the time responses to stimuli appeared

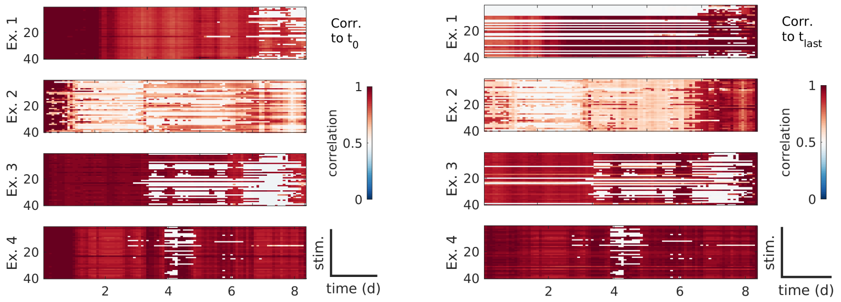


(a) Example Network 1.

(b) Example Network 2.

Figure 8.3: Synaptic turnover leads to phases of stable response to stimuli interrupted by abrupt changes. (a), (b) Responses to example stimuli of networks changing according to Equation 8.3.

to be affected only little by these changes, except for some time points, where abrupt transitions occurred. Example stimulus responses for four stimuli in two example networks can be seen in Figure 8.3. Periods of stable responses were interspersed by abrupt transitions, when neurons activated by a given stimulus changed from one time step to another (e.g. network 1, stimulus 3, around day 8 or network 1, stimulus 23, around day 6). Sometimes there was also an abrupt transition from no response to a response or vice versa (e.g. network 2, stimulus 1 or stimulus 3).



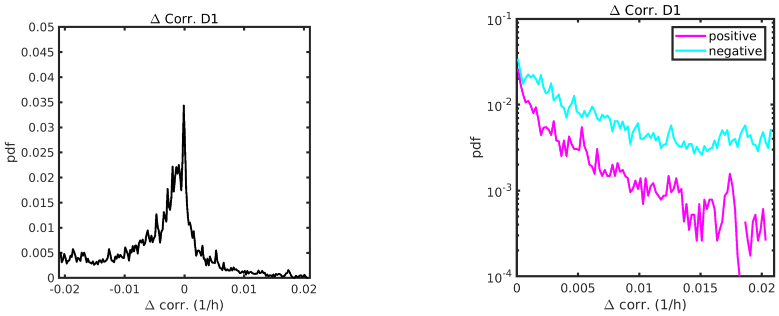
(a) Example corr. to t_0.

(b) Example corr. to t_{last}.

Figure 8.4: Response correlations show stable periods and abrupt transitions. (a) Correlation of responses at time t to t_0 (Examples 1 and 2 from Figure 8.3 shown along with two more examples shown in Sup. Fig. 9.10). (b) Correlation of responses at time t to t_{last} (Examples 1 and 2 from Figure 8.3 and two more examples shown in Sup. Fig. 9.10).

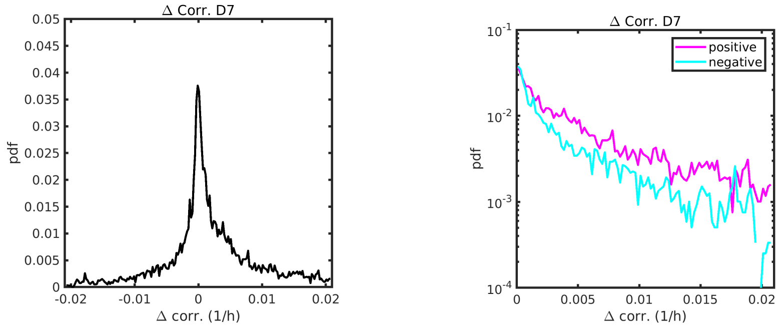
We visualized these transitions by computing for each stimulus the correlation of the response at any time point to the response at the first or last time point (Figure 8.4a and Figure 8.4b). We computed this correlation to a unique time point (and not between consecutive time points), because we wanted to have a fixed reference point. This computation was done to both the first and the last time point to check for temporal symmetry.

At any given time point, we either see little to no change in the correlation or abrupt transitions. Transitions occurred both away from the initial response, but sometimes also toward a (partial) recovery of an earlier response that had been transiently lost at an intermediate time point. Periods of little to no change are visualized by a red area (or horizontal line) meaning the response has an unchanging correlation to the response at the initial or last time point. Slow, gradual changes would be signified by a gradual change in color across an extended time, however, those are not present. White areas indicate near 0 correlation to the initial or last time point. For near 0 periods we cannot say much about the changes in the system, because the responses could change. As long as they do not become more or less similar to the initial or last response this would be difficult to notice based on this figure alone. This is shown in Figure 8.4 for four example implementations of the network, the two examples from Figure 8.3 and two more examples, shown in Sup. Fig. 9.10. Note that periods of stability can be of very different extent in time. Sometimes these periods last almost across the entire simulation length (e.g. the red region in example 1), whereas some of these periods only last for a very brief timespan (e.g. the white region in example 4).



(a) Changes corr. day 1.

(b) Changes corr. day 1. (semi log).

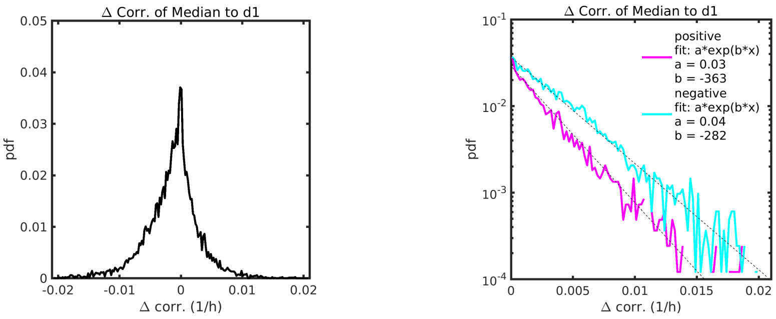


(c) Changes corr. day 7.

(d) Changes corr. day 7 (semi log).

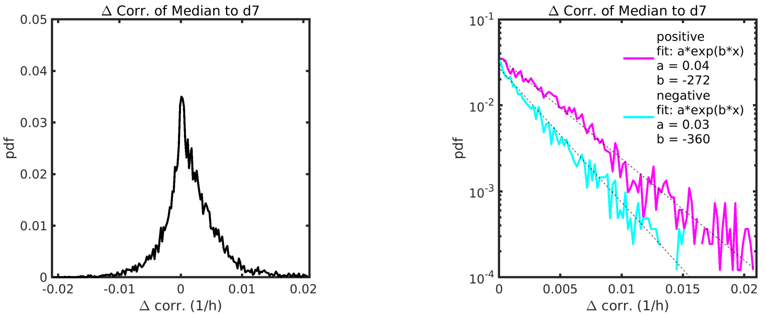
Figure 8.5: Correlation changes are broadly distributed in the *model*. (a), (b) Histogram of changes in correlation to day 1 on normal (a) and semi log scale axes (b). (c), (d) Histogram of changes in correlation to day 7 on normal (c) and semi log scale axes (d). For better comparability to experimental data, responses were taken from four days at two day intervals (days 1, 3, 5, 7).

To quantify the dynamics of the changes in these correlations we computed the difference in correlation (to the state at $t = 0$) between consecutive time points. The time points were chosen at a two day interval for comparison to the experimental data and the changes were converted to be of unit $1\,h^{-1}$. Figure 8.5 shows the distribution of these changes in correlation to the first time point in normal scale (Figure 8.5a) and semi-log scale (Figure 8.5b) and the changes in correlation to the last time point in normal scale (Figure 8.5c) and semi-log scale (Figure 8.5d). We found a broad distribution of correlation changes both away from the initial response and towards the initial response. This means that responses typically changed very little between time points, but sometimes big changes occured, interestingly, both negative and positive. Negative changes in the correlation to the initial response mean that the response became less similar to the initial response, i.e. following the global trend; positive changes imply a (partial) recovery of the initial response. Interestingly, we also observed a considerable amount of changes in the opposite direction, i.e. response changes reversing the global trend towards a response more similar to the initial response. This distribution of correlation changes was largely mirrored in time, when regarding the correlation of a response at a given time point to the response at the last time point (i.e. Figure 8.5a looks like a mirrored version of Figure 8.5c), indicating a temporal symmetry as expected of a random stationary process.



(a) Changes corr. day 1.

(b) Changes corr. day 1 (semi log).



(c) Changes corr. day 7.

(d) Changes corr. day 7 (semi log).

Figure 8.6: Correlation changes are broadly distributed in the *experiment* (data taken from Chapter 5). (a), (b) Histogram of changes in correlation to day 1 on normal (a) and semi log scale axes (b). (c), (d) Histogram of changes in correlation to day 7 on normal (c) and semi log scale axes (d).

Remarkably, the corresponding distributions in experimental data of the very system we based our model on (same as in Chapter 5, Aschauer et al., 2022) were broadly distributed, too. This would not be expected for a system with slow representational drift without abrupt transitions, where this distribution would be narrower. To ease comparison, we calculated the difference in correlation of the median response over trials on a given day to the median response on the first and last imaging day. This is comparable to the model analysis, as imaging days were two days apart. Furthermore the model is deterministic, so we only have one response per stimulus. The median response was used for experimental data to have a robust estimate of the response and one stereotypical response per stimulus and imaging time point. The resulting distributions are shown in Figure 8.6. As in the model, this change in correlation was peaked around 0 with tails in both directions. While change was mostly close to 0, there were also time points with substantial response changes. Again, the distributions were biased towards changes away from the initial response (towards the last response), but also in the experiment, we could see evidence for (partial) recovery of previous response states. Moreover, the correlation changes in experimental data seem to be exponentially distributed. This exponential distribution is not captured by the model, but might hint at some underlying mechansism not captured by random synaptic drift, merely fit to experimentally measured synaptic drift.

In summary, we found that steady changes of synaptic connections in our firing rate model lead to periods of relatively stable stimulus responses, interrupted by more abrupt transitions toward different response patterns. These changes were broadly distributed and sometimes led to a recovery of a previously lost response, all in accordance with experimental data.

8.3.3 *The fixed point topology of a neural network*

Next, we asked whether these abrupt changes of responses can be linked to changes in the network structure. The network structure – in our case the connectivity matrix – together with the input can be understood to shape the high-dimensional energy landscape governing the network dynamics (e.g. Hopfield, 1982, Sompolinsky et al., 1988). The network activity would then always find its way to a local minimum of this energy landscape (see Figure 8.7 for a one-dimensional example). A two dimensional example could be thought of as an actual landscape with mountains and valleys and the energy would be the potential energy, which causes everything that is not fixed to move to a local minimum. As a high-dimensional energy landscape is hard to imagine and thus changes in this high-dimensional landscape are hard to comprehend, we investigated the fixed points of

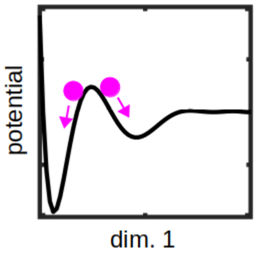


Figure 8.7: Schematic of a one-dimensional energy landscape. A (one-dimensional) function can be understood as an energy landscape. Like e.g. water in an actual landscape, anything would be drawn to a local (or the global) minimum of the function.

the system. We want to utilize the fixed points of the system in order to investigate how change in the network is different during abrupt response transitions from change in the network during periods of near stable response. Are response transitions really abrupt? Or are they continuous on a faster time scale and we merely lack the temporal resolution? Abrupt transitions would be linked to bifurcations, i.e. qualitative changes in the fixed point topology; fixed points appear or disappear. Quantitative fixed point changes, i.e. moving fixed points, could still be able to produce abrupt response transitions, but there would be no transition between network states.

For a given (static) stimulus input and initial condition the high dimensional network activity typically evolves along some trajectory that approaches a sequence of *unstable* fixed points before it converges to one of the *stable* fixed points. We illustrate this in a 2D energy landscape, i.e. the energy is a function of 2 dimensions $f(x, y)$. The requirement for a fixed point (x_0, y_0) is then, that the derivative at this point in both dimensions is $\frac{\partial}{\partial x} f(x_0, y_0) = \frac{\partial}{\partial y} f(x_0, y_0) = 0$. This is of course the case for all (local) extrema, i.e. for minima (*stable* fixed point) and maxima (*unstable* fixed point) of the function. However, there is another possibility for an *unstable* fixed point, a saddle point. Here, e.g. the derivative of f at (x_0, y_0) is 0 aswell, but $f(x_0, y_0)$ has a maximum in x and a minimum in y (or vice versa). The function could of course also be rotated and then the minimum and maximum would occur along mixed dimensions.

In higher dimensions saddle points typically occur more often than in 2D (relative to the total number of fixed points), mostly because any given fixed point would need to be a minimum in every dimension to be a stable fixed point. A maximum in only one of these dimensions would make it a saddle point by definition. The increase of the number of saddle points with dimension has been shown theoretically for random matrices drawn from a Gaussian distribution by Bray and

Dean (2007) and heuristically for more realistic neural networks by Dauphin et al. (2014). Saddle points can be attractive in any dimension but one and thus activity trajectories are heavily influenced by (and quite often pass through the basin of attraction of) saddle points.

To determine whether a fixed point is stable or unstable the system can be linearized around this fixed point. To do so the entire expression in Equation 7.1 – including stimulus and non-linearity – has to be linearized. This can be done by computing the Taylor expansion (Taylor, 1715) of the system around a fixed point. Ignoring any higher order terms leaves us with the first derivative of the right hand side of Equation 7.1 with respect to each neuron's rate $\frac{\partial}{\partial t} r_i$. This matrix **J** – called the Jacobian matrix – can then be used to investigate the stability of a fixed point via Equation 8.11:

$$\mathbf{J}\mathbf{v}_i = \lambda_i \mathbf{v}_i, \tag{8.11}$$

where $\mathbf{v}_i$ is the eigenvector and λ_i the corresponding eigenvalue. The eigenvalues of the linearized system around the fixed point determine its stability. If all of these eigenvalues are negative the system will always converge in this fixed point when it is disturbed in any direction and thus the fixed point is a *stable* fixed point. If any of these eigenvalues has a positive real part, a small perturbance in the associated direction will lead to divergence and thus the fixed point is an *unstable* fixed point. A non-zero imaginary part of an eigenvalue leads to oscillatory behavior, whereas eigenvalues without an imaginary part lead to divergence or convergence without oscillations. To determine whether a fixed point is a local maximum, minimum or saddle point we can count the number of divergent dimensions: if all eigenvalues are positive, the fixed point is a maximum; if all eigenvalues are negative the fixed point is a minimum; if some eigenvalues are positive and others are negative, the fixed point is a saddle point with as many divergent dimensions as it has positive eigenvalues. The computation of the Jacobian and of its eigenvalues can be done numerically.

A complex system can have no fixed points, a single fixed point, or multiple fixed points. Typically random high dimensional systems have multiple fixed points. This means that a single stimulus can evoke different responses, i.e. the network activity can converge to different stable fixed points, depending on the initial condition of the system. To find them all a network has to be run several times with different initial conditions. However, this only results in stable fixed points, as unstable fixed points are only found, if the system is initialized at the exact location of an unstable fixed point, which is highly unlikely. We approached this employing a method from Sussillo and Barak (2013), described in detail in Section 8.3.4.

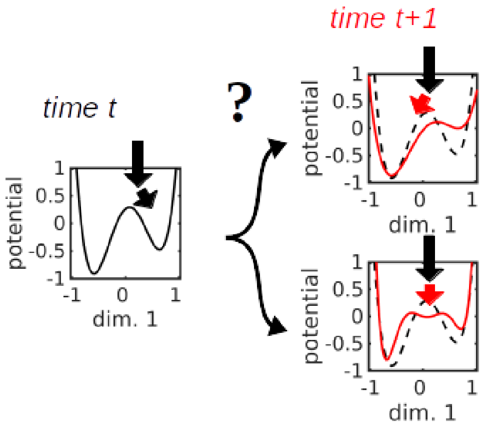


Figure 8.8: Schematic of possible changes to a one-dimensional energy landscape. Abrupt transitions could be caused by minor changes in the fixed point topology, e.g. a rerouting of the activity trajectory due to a translation of an unstable fixed point (top right) or by qualitative changes of the fixed point topology, e.g. the emergence of a new stable fixed point (bottom right).

We want to use the fixed points of our network model to gain an understanding of its energy landscape and how it is altered by synaptic drift. When we have the fixed points of the network (for a given stimulus) at two consecutive time points, we can investigate changes in the fixed point structure of the system, comparing cases, when there is minor change in the network activity, with those showing an abrupt transition. Different changes in the fixed point structure of a network might affect the activity outcome to a different extent. Are abrupt transitions of network responses caused by shifts, i.e. quantitative changes of the network? Already small changes might lead to a rerouting of the activity trajectory. Or are these abrupt transitions caused by qualitative changes in the fixed point topology , e.g. newly emerging or disappearing fixed points? Both of these possible outcomes are illustrated in Figure 8.8).

8.3.4 *Finding fixed points numerically*

To find all the fixed points of a given network, we employed a method adapted from Sussillo and Barak (2013). A fixed point in a firing rate network is defined as a point, where the activity does not change any more and thus the right hand side of Equation 7.1 is 0 for all neurons (Equation 8.12):

$$\tau \frac{\partial r_i}{\partial t} = -r_i + f\left(\sum_{j=1}^{N} W_{ij} r_j + s_i\right) \overset{!}{=} 0 \qquad \forall i \in [1, N]. \tag{8.12}$$

Here (as above), r_i is the firing rate of neuron i, W_{ij} is the connection between neurons i and j, s_i is the stimulus input into neuron i, τ is the characteristic time constant and f the nonlinearity. Note that the stimulus can not be time dependent any more, as the system would never converge for an ever changing stimulus.

The trick used by Sussillo and Barak (2013) to find the roots of this expression was to construct another expression where all the roots of Equation 8.12 would become minima. This was done by taking the sum of the squares of the left hand side of Equation 8.12 over all neurons and dividing it by 2 for normalization, leaving the optimization problem (Equation 8.13):

$$\underset{\mathbf{r}}{\operatorname{argmin}} \left(0.5 \sum_i \left(-r_i + f \left(\sum_j W_{ij} r_j + s_i \right) \right)^2 \right), \tag{8.13}$$

that could consecutively be solved numerically. This way, we are able to find all the fixed points of the system and not only stable fixed points, as both stable and unstable fixed points are minima of this new expression.

In order to make sure the fixed point finder finds all fixed points in a reasonable amount of time, we adjusted our network model (compared to Chapter 7):

- The number of neurons was decreased from $N = 100$ to $N = 10$. This left the network complex enough for all the interesting dynamics to appear, but drastically reduced the amount of run time. Running the fixed point finder for a single network with $N = 100$, a single stimulus, and a single time point took about $30\,\mathrm{min}$, i.e. $> 8000\,\mathrm{d}$ of computation for 100 networks , 40 stimuli and 100 time points (exact timing depended on the number of steps prior to convergence of the optimizer for each initial condition). Reducing the network size to $N = 10$, we were able to run it for 100 implementations of the network, 40 stimuli, and 100 time steps in several hours (again the exact timing depended on the number of steps prior to convergence for each initial condition).
- For the optimization to converge in a minimum of Equation 8.13, the stimuli had to be independent of time, so we used random static stimuli as input. Otherwise, the randomness would lead to non-zero changes of the network's firing rates.
- As non-linearity we used

$$f(x) = \ln\left(1 + e^x\right), \tag{8.14}$$

which is a smoothed version of the half wave rectification used above, because the rectification together with the strong inhibition leads to large regions of activity space, where the strong inhibition sets all activity to 0 and the fixed point finder cannot further converge.

- To then be in a regime, where the exact shape of the non-linearity does not matter to the network activity, we also increased the stimulus strength, so the network activity was never close to 0, where the smoothed rectification differs from the standard rectification. Thus, the network activity is not affected by the change to this new non-linearity, but the fixed point finder is able to find fixed points.

Apart from these changes, we kept everything as before, so synaptic changes were implemented as described in detail above (Section 8.3. In order to find all fixed points, i.e. all minima of Equation 8.13, we had to make sure to run the optimizer with enough random initial conditions. We found that for networks of ten neurons the number of identified fixed points started to converge at $1,000$ initial conditions, so we ran the fixed point finder with $10,000$ initial conditions.

To check, whether the resulting fixed points were stable or unstable, we computed the eigenvalues of the Jacobian matrix. When all of them are positive the fixed point is at a local maximum, when all of them are negative it is at a local minimum and if some are positive and others are negative the found fixed point is a saddle point. Minima are stable fixed points, maxima and saddle points are unstable fixed points.

8.3.5 *Abrupt transitions coincide with qualitative changes of the fixed point topology*

Next, we wanted to use the fixed point finder to investigate the fixed point dynamics during abrupt transitions and compare them to fixed point dynamics during small response changes. To this end, we implemented synaptic drift in 100 random implemetations of the network for 100 time steps at a time interval of two hours. Thus, we covered slightly more then a week. Note that the size of the time steps had little influence on the presenence of abrupt transitions. The probability for abrupt transitions to occur during a given time interval remained constant for small enough time steps (avoiding undersampling). Smaller time steps did not lead to more continuous transitions, but rather to a lower probability for abrupt transitions per time step (data not shown).

At each time step we took the connectivity matrix as our network and computed the network responses to stimuli (for details see Chapter 7). To measure the abruptness of a response change we correlated the response of the network to a stimulus at time $t-1$ to its response to the same stimulus at time t using Pearson's correlation coefficient. This correlation of network responses at consecutive time points is a direct measure of absolute change between those time points, unlike the previously used difference in correlation to the first (last) time point. Here, we were not interested in a fixed reference, but we wanted to investigate the connection between the amount of response change and changes of the fixed point topology.

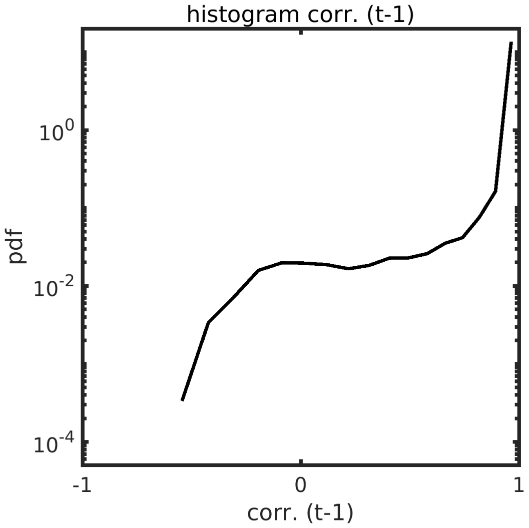


Figure 8.9: Histogram of correlations between response vectors on consecutive time points (2 h apart) shows a broad distribution.

The distribution of these correlations can be seen in Figure 8.9. It is clearly peaked in the highest bin (i.e. close to 1), so for most time steps the network responses change little. Nevertheless, the histogram has a shoulder at small correlations; it is almost bimodal, although the logarithmic y-axis might bias this impression. This is in line with findings from Section 8.3.2, where periods of stability towards synaptic drift were found to be interspersed by abrupt transitions.

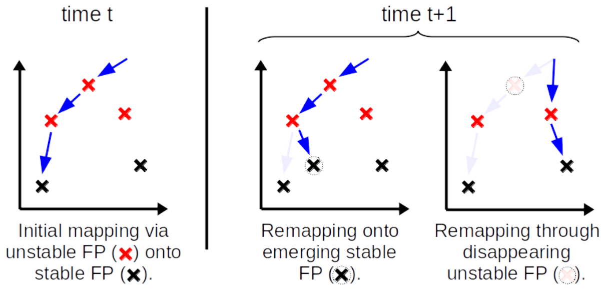
Using the fixed point finder described in Section 8.3.4, we were able to detect both stable and unstable fixed points for each stimulus at each time point in every implementation of the network. We found 2.8 ± 5.8 (mean $\pm$ standard deviation (SD)) fixed points per stimulus and network at a given time point. Of those, 1.8 ± 3.6 were stable fixed points and 1.0 ± 3.1 were unstable. These numbers were independent of stimulus, implementation of the network and time – as could be expected from a stationary system. However, they were highly variable, as shown by the large standard deviations. Minimum counts were 0 fixed points and maximum counts were 128 (total), 77 (stable) and 94 (unstable). 95% confidence intervals were from 0 to 13 (total), 9 (stable), 6 (unstable).

Next, we wondered, what was happening to the fixed points during response changes. If a fixed point was identified at the same location on consecutive time points, we called it a non-moving fixed point. For all other fixed points we ideally want to distinguish between the two possible outcomes sketched in Figure 8.8. Are fixed points drifting? Or are bifurcations leading to emerging/disappearing fixed points? We identified multiple fixed points per time point, however, and they can appear or disappear through various bifurcations (some of which simultaneously effecting multiple fixed points in different locations). So, we are not able to make a clear distinction between drifting fixed points and bifurcations. We instead chose to distinguish between fixed points emerging/disappearing *near* other fixed points and fixed points emerging/disappearing *far* from other fixed points. In this way, drifting fixed points were always classified as near other fixed points and we could be sure that emerging/disappearing fixed points far from others were caused by a bifurcation. Fixed points emerging/disappearing near others, however, could aso be caused by bifurcations and were not distingusihable from drifting fixed points.

FIXED POINT TRANSITION	HISTOGRAM	RATIO	BAR PLOT
stable → stable unstable → stable stable → unstable unstable → unstable	Figure 8.11	Figure 8.12	Figure 8.15
new stable (near) new stable (far) lost stable (near) lost stable (far)	Sup. Fig. 9.11	Figure 8.13	Figure 8.16
new unstable (near) new unstable (far) lost unstable (near) lost unstable (far)	Sup. Fig. 9.12	Figure 8.14	Figure 8.17

Table 8.1: A table of fixed point changes listing all possible changes and where to find the different plots. *near* means in the vicinity of another fixed point, *far* means not in the vicinity of another fixed point. The four subplots of the respective figures are always ordered from top left to bottom right.

So, we counted non-moving fixed points, as well as appearing and disappearing fixed points. The non-moving fixed points were classified into four groups, depending on their stability (*stable* → *stable*, *unstable* → *unstable*, *stable* → *unstable*, *unstable* → *stable*). Newly emerging and disappearing fixed points were *stable* or *unstable* and further grouped into fixed points appearing or disappearing near ($< 1/10$ SD of nearest neighbor distances across all time points, networks, and stimuli) or far from ($> 1/10$ SD of nearest neighbor distances) another fixed point. An overview of these transitions and where to find the corresponding plots, which we discuss in detail in the following, is given in Table 8.1. With these twelve categories we were able to better understand the network dynamics during abrupt response changes. We first want to compare the histogram of correlations from all time steps (Figure 8.9) to the histograms of correlations from time steps including each of these twelve fixed point changes. Then we make a distinction into weak and strong changes of network responses and find out which fixed point changes are more often associated to them.



Initial mapping via unstable FP (✖) onto stable FP (✖).

Remapping onto emerging stable FP (✖).

Remapping through disappearing unstable FP (◌).

Figure 8.10: Schematic fixed point changes. Left: At time t a stimulus response is mapped via multiple *unstable* fixed points onto a *stable* fixed point. Right: If an abrupt response transitions occurs between times t and $t+1$, this typically coincides with either a remapping to an emerging (or away from a disappearing) *stable* fixed point in the vicinity of an already existing fixed point (middle) or with a rerouting due to an appearing (or disappearing) *unstable* fixed point (right).

If the abrupt transitions of responses were mediated by quantitative changes of fixed points, we would expect no heightened association of bifurcations with these respective time points. Emerging/disappearing fixed points far from other fixed points would either not be present at all or they would be equally distributed among all time points. However, if abrupt transitions were linked to qualitative changes of fixed points (i.e. bifurcations), we would find an excess overlap of time points with low correlation and time points with bifurcations. Indeed, we do find that low correlation time points coincide with certain qualitative fixed point changes, e.g. the two shown in Figure 8.10, namely the emergence of a stable fixed point in the vicinity of another fixed point or the disappearing of an unstable fixed point far from other fixed points. While the first could also be linked to quantitative changes, the latter cannot. Thus, we can link some strong response changes to abrupt transitions mitigated by bifurcations.

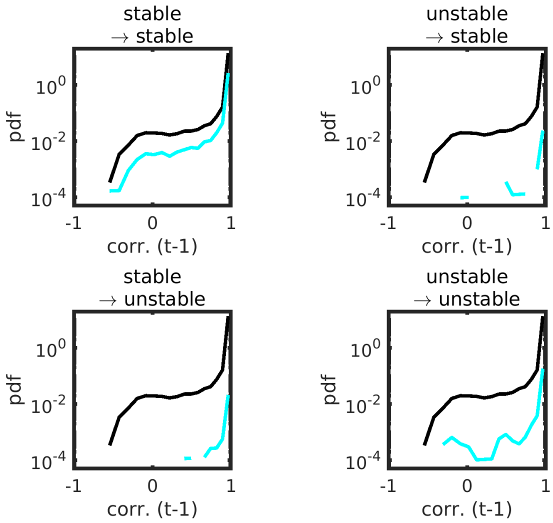


Figure 8.11: Histograms of correlations between response vectors on consecutive time points for all time points that include non-moving fixed points (cyan). Non-moving fixed points can either not change and remain *stable* or *unstable* or change from *stable* to *unstable* or vice versa. For comparison the entire distribution (Figure 8.9) is plotted (black). Partial disappearance of lines is caused by 0 counts and logarithmic y-axes.

We started by plotting, along with the histogram of all response change correlations, the histograms of only those time points associated with the corresponding change in the fixed points structure. As an example, we show the non-moving fixed points in Figure 8.11 (for *stable* and *unstable* fixed points, see Sup. Fig. 9.11 and Sup. Fig. 9.12). The black line is the complete data histogram (Figure 8.9) and the colored line is the histogram of all time points during which the specified change in the fixed points structure occurred. In Figure 8.11 a *stable* fixed point stayed at the same position (top left), a fixed point stayed at the same position, but changed from being *unstable* to being *stable* (top right), a fixed point changed from being *stable* to *unstable* (bottom left) and an *unstable* fixed point stayed at the same position.

As these histograms span orders of magnitudes on the y-axis, it is difficult to assess, what fraction of the total (black) is associated with any of the changes in the fixed points picture (colored). Therefore, we computed for each bin the fraction of the total response changes

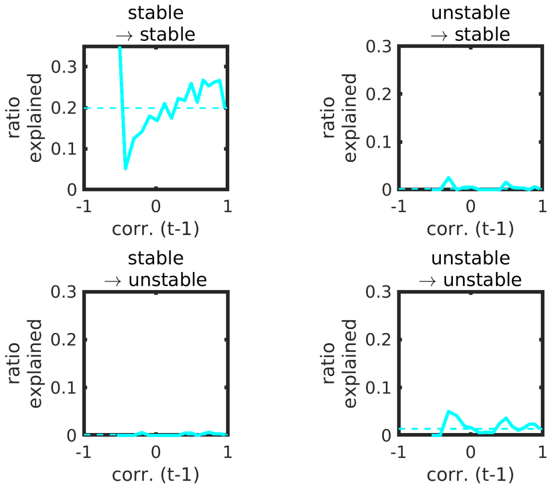


Figure 8.12: Fraction of time points with non-moving fixed points as a function of the correlation between response vectors on consecutive time points (solid line). Fraction independent of correlation (dashed line). The amount of non-moving *stable* fixed points seems to be correlated to the response correlation; other non-mocing fixed points show no such dependency.

in the correlation bin that was associated with the specific change. Non-moving *stable* fixed points (Figure 8.12 top left) appear to be correlated to the correlation of response vectors at consecutive time points. The more *stable* fixed points remain stable, the less change was found between responses (except for the lowest bin, but there were only two entries and one of them was associated with a non-moving *stable* fixed point). No such trend was observed for the other non-moving fixed points (Figure 8.12 top right, bottom left and bottom right).

We followed the same procedure (computing the ratio of histograms) for newly appearing or disappearing *stable* fixed points (Figure 8.13). We found that both appearing and disappearing *stable* fixed points in the vicinity of other fixed points were associated with response changes. Except for the top bin (almost no response change), a response change was associated with both a new and a lost *stable* fixed point almost 70% of the time. This was not the case, however, for appearing or disappearing *stable* fixed points far from other fixed points. Intuitively, this can be understood as a *stable* fixed point in the vicinity of the fixed point the stimulus response converged to deflecting the stimulus response trajectory onto itself (Figure 8.10 center). Similarly, a disappearing *stable* fixed point close to another one, might just lead to a remapping to this other fixed point. A *stable* fixed point far away has little to no effect on this trajectory. The magnitude of this response change (Figure 8.13 top left and bottom left) is broadly dis-

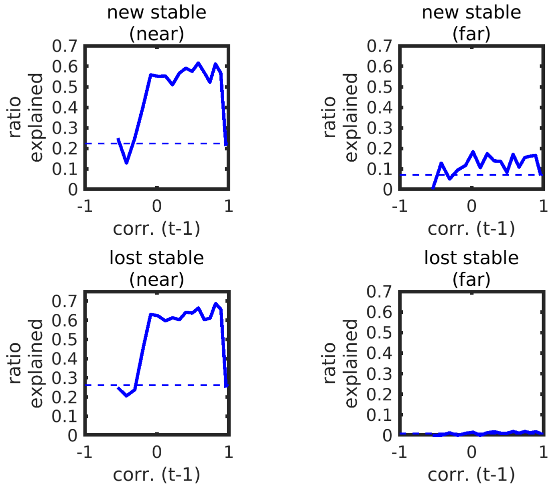


Figure 8.13: Fraction of time points with changes in *stable* fixed points as a function of the correlation between response vectors on consecutive time points (solid line). Fraction independent of correlation (dashed line). Emerging an d disappearing *stable* fixed points in the vicinity of other fixed points are associated with lower correlations.

tributed, ranging from close to 0 to almost 1. This might be due to the fact, that these new or lost *stable* fixed points can be close or farther away from the *stable* fixed point, the network activity converged in. In summary, *stable* fixed points emerging or disappearing in the vicinity of other fixed points coincided with strong changes in the network response. This was not the case for emerging or disappearing *stable* fixed points far from other fixed points.

But not only movement of *stable* fixed points affected the network response, also changing *unstable* fixed points are associated with large response changes (Figure 8.14). The figure shows trends of an association with low correlations for appearing and disappearing *unstable* fixed points in the vicinity of already existing fixed points. However, contrary to *stable* fixed points this association was found also for new *unstable* fixed points far away from any fixed points. These three cases (*unstable* fixed points emerging and disappearing in the vicinity of other fixed points and *unstable* fixed points emerging far from other fixed points, Figure 8.14 top left, bottom left and top right) were more often associated to larger changes in the network response than to smaller changes. And even the fourth case (disappearing *unstable* fixed points, Figure 8.14 bottom right) shows the same dependency, albeit on a lower level. To sum up changing *unstable* fixed points conincided with strong response changes, when the emerged or disappeared in the vicinity of other fixed points, but also when they appeared far from other fixed points (compare Figure 8.10 right). The

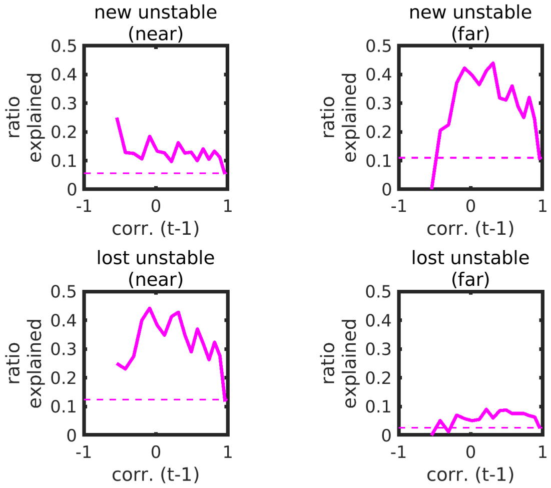


Figure 8.14: Fraction of time points with changes in *unstable* fixed points as a function of the correlation between response vectors on consecutive time points (solid line). Fraction independent of correlation (dashed line). Changes in *unstable* fixed points are lost or new *unstable* fixed points close to or far away from already existing fixed points.

association between disappearing *unstable* fixed points far from other fixed points and response changes was weaker.

Another way to investigate the dependencies of response changes on the twelve categories of fixed point changes, however, with less noise caused by small bin sizes, was computed by grouping network response changes into two groups, one with little change, i.e. a high correlation of stimulus responses at consecutive time points ($\rho > 0.5$) and one with large change, i.e. a low correlation of stimulus responses at consecutive time points ($\rho < 0.5$). The threshold of 0.5 was chosen as the distribution had a plateau at this point and thus changing it to some extent did not affect the outcome, especially, as the distribution drops off rapidly when moving away from $\rho \approx 1$ towards slightly lower values (Figure 8.9).

For Figure 8.15 we computed the fraction of time points each of the non-moving fixed points changes appeared (top left: *stable* → *stable*, top right: *stable* → *unstable*, bottom left: *unstable* → *stable*, bottom right: *unstable* → *unstable*). This is displayed for all time points regardless of response correlation between consecutive time points as a black bar and for high correlations and low correlations as a filled or empty colored bar, respectively. All four panels show no significantly different fraction in either direction for both high and low correlations. Response changes are thus not correlated with non-moving fixed points.

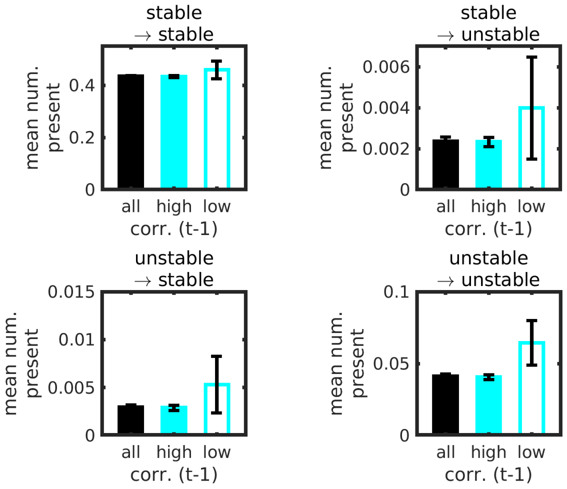


Figure 8.15: Fixed points changes associated to high (> 0.5) and low (< 0.5) correlations between consecutive time points for non-moving fixed points. On the y-axis is the mean number of the specific fixed point change per time step for each group. There was no significant difference between the high correlation group and the low correlation group for any of the non-moving fixed points.

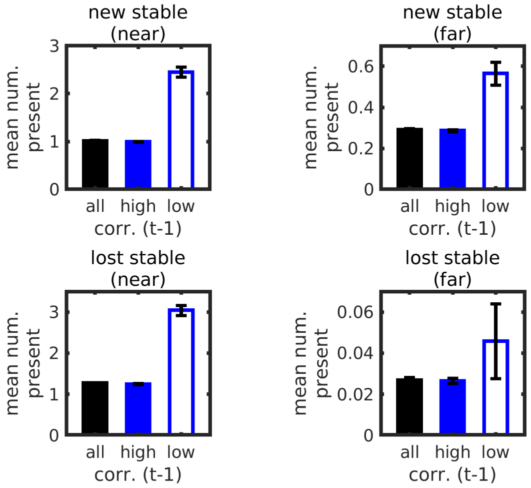


Figure 8.16: Fixed points changes associated to high (> 0.5) and low (< 0.5) correlations between consecutive time points for changes in *stable* fixed points. On the y-axis is the mean number of the specific fixed point change per time step for each group. Here, for all fixed point changes the low correlation group shows a significant increase compared to all time steps.

Figure 8.16 shows the corresponding plots for emerging and disappearing *stable* fixed points near to or far from other fixed points (top left: new *stable* fixed points close to another fixed point, top right: new *stable* fixed points far from another fixed point, bottom left: lost *stable* fixed point close to another fixed point, bottom right: lost *stable* fixed point far from another fixed point). All four panels show a similar picture, namely a significant increase of the number of each specific fixed point change for time points with low response correlation compared to all time points, and no such increase for time points with high response correlation. For each of those fixed point changes the increase was roughly by a factor of 2.

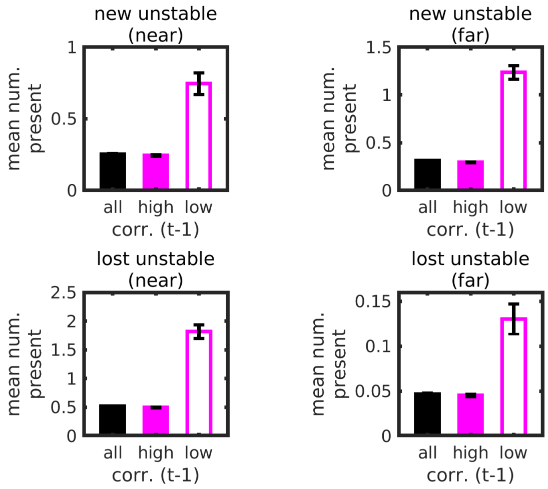


Figure 8.17: Fixed points changes associated to high (> 0.5) and low (< 0.5) correlations between consecutive time points for changes in *unstable* fixed points. On the y-axis is the mean number of the specific fixed point change per time step for each group. Here, for all fixed point changes the low correlation group shows a significant increase compared to all time steps.

Similarly, we investigated time points involving changes of *unstable* fixed points in Figure 8.17 (top left: new *unstable* fixed points close to another fixed point, top right: new *unstable* fixed points far from another fixed point, bottom left: lost *unstable* fixed point close to another fixed point, bottom right: lost *unstable* fixed point far from another fixed point). Here, comparable to *stable* fixed point changes, we again found for all four cases a significant increase of the number of fixed point changes for time points with low response correlation and no such increase for time points with high response correlation. For *unstable* fixed point changes this increase was roughly by a factor of 3.

With this analysis the effect of non-moving fixed points seemed to be negligible (Figure 8.15). For all changes involving changes to the *stable* (Figure 8.16) and *unstable* (Figure 8.17) fixed points, however, this analysis revealed an increase of the respective change for low correlations and thus larger changes of the network responses, by a factor of roughly 2 for *stable* fixed points and even by a factor of roughly 3 for *unstable* fixed points.

So, in brief, we found that the more pronounced response transitions we observed in network responses typically coincided with either a remapping including an emerging or a disappearing *stable* fixed point in the vicinity of another fixed point, or a rerouting through an appearing (or disappearing) *unstable* fixed point (as illustrated by Figure 8.10). Abrupt network response transitions were not predominantly caused by minor changes of single fixed points, but rather by qualitative changes of the fixed point topology.

8.4 DISCUSSION

We found that ongoing synaptic change in our firing rate model led to periods of stable stimulus responses that were interupted by abrupt transitions toward different responses. These abrupt transitions were associated with qualitative changes in the fixed point topology of the network model.

Abrupt transitions are difficult to detect in experimental data, because of two reasons: (1) There is typically a lot of time between experimental sessions, which are typically rather short. So, the probability to record an abrupt transition is small. (2) There is noise in experimental recordings, which means that even if you record an abrupt tranition during the comparatively small time span of experimental observation it might be hardly distinguishable from recording noise. This can be exemplified by our data from Chapter 5. On the one hand, it is recorded at two day intervals and we recorded clear changes between sessions. Did they occur abruptly? Or was it rather a continuous representational drift over two days leading to clearly distinct activity patterns? As some of the patterns were stable across the entire time of the recording, however, we can make the point that the transitions occur on a fast time scale compared to these apparently present periods of stability. On the other hand, it is hard to point at transitions during recording sessions: We defined a response reliability using multiple trials of a stimulus presentation. But what does it mean, when a stimulus evokes a clear response in almost all trials, but no (or a different) response during the last three trials? Is this an abrupt transition? Or is this noise in the biological system? Or experimental noise? Maybe the animal moved and the imaging plane was slightly tilted, leading to individual cells – and their potential activity – being lost. To pinpoint the exact moment of such an abrupt transition is extremely difficult and to our knowledge, the ideal experiment has not been designed, yet. Especially, keeping in mind the broad distribution of transitions that would need long lasting experiments with a high temporal resolution.

Although abrupt transitions are notoriously difficult to measure, there have still been reports of some in different systems as hippocampus (Rubin et al., 2015a, Sheintuch et al., 2020), parietal cortex (Driscoll et al., 2017), auditory cortex (Kobak et al., 2019), entorhinal cortex (Low et al., 2021) and visual cortex (Deitch et al., 2021). All of those report changes that are fast compared to periods of stability.

We speculate that we might even know such transitions from our daily experience. There seem to be two rather obvious types of learning, we all know: It can be a tedious process of becoming better at a task, which is rather slow and continuous, but there also is this other type, when a learning process (maybe as tedious and long) leads to a sudden insight. This could be linked to an abrupt transition in neuronal activity, especially as the underlying mechanism, we describe in our model, is generic enough to be present in any changing random network with strong recurrent connections and strong inhibtion.

As the evidence for representational unstability is growing, the question remains, how we form seemingly stable perceptions, lasting for days, months, or even years, when the underlying substrate is not stable. There might be different mechanisms, such as representational redundancy, stability only at higher processing levels, or a continuous retuning of the readout, discussed in detail in e.g. Chambers and Rumpel, 2017 and Susman et al., 2019. A continuous retuning has recently been shown to work in a model using a Hebbian like plasticity rule on the connections to the readout neurons (Kossio et al., 2021).

Part V

DISCUSSION

In this thesis we asked how sensory representations of stimuli in mouse auditory cortex change across time, both under basal conditions and during learning. We further wanted to know how the observed representational drift is linked to synaptic drift in the underlying network structure. We addressed these questions by analyzing both empirical data and a computational model. Experimental data was recorded via two-photon imaging from local neuronal populations in mouse auditory cortex. Computational studies were performed using a circuit model of such a local population. We found that population responses in mouse auditory cortex are clustered into a small set of response modes. These response modes change across time even in the absence of learning and this so-called representational drift is biased towards the formation of associations during learning (Chapter 5). We then defined a set of elementary operations to deconstruct and describe response mode transitions under basal conditions and during auditory cued fear conditioning (ACFC), where a sound was associated with an aversive stimulus. The elementary operations helped to further disentangle the dynamics and show that the formation of associations during learning was due to both more operations increasing the formation of new associations and less operations breaking up existing associations (Chapter 6). Next, we devised a model to investigate the neural population dynamics in auditory cortex. Apart from previously described regimes of random network models (e.g. Wilson and Cowan, 1972, Sompolinsky et al., 1988, Zhang and Saggar, 2020) we were able to identify a regime, where stimuli are clustered into response modes similar to experimental findings, for strong recurrent connections and strong inhibition (Chapter 7). This model was then used to show that ongoing synaptic drift (in the relevant regime) leads to periods with stable stimulus responses, interupted by abrupt transitions toward new responses (Chapter 8). These abrupt transitions coincide with qualitative changes in the fixed point topology of the network.

In the empirical part we were able to reproduce earlier findings corroborating that different stimuli evoke the same response in a group of neurons in auditory cortex (Bathellier et al., 2012, Atencio and Schreiner, 2013, See et al., 2018). This seems counter-intuitive at first, because based on this group of neurons we are not able to discriminate between the stimuli that evoke the same response. The fact, however, that different stimuli are grouped together in different fields of view in the same animal (i.e. different parts of primary auditory cortex) leads to a unique cortex-wide response per stimulus and thus discriminability between different stimuli (see 5.10b). It is nevertheless valid to consider response modes recorded from a subpopulation of neurons in sensory cortex, as a potential readout neuron higher up in cortical hierarchy would not receive input from the entire auditory cortex, either, but rather from a subgroup of all cortical neurons.

Recently, neuronal representations have been shown to be subject to change in many different cortical (and non-cortical) areas, e.g. in mouse hippocampus, barrel, olfactory, visual, motor and posterior perietal cortex (Rokni et al., 2007, Huber et al., 2012, Mankin et al., 2012, Margolis et al., 2012, Ziv et al., 2013, Clopath et al., 2017, Driscoll et al., 2017, Hainmueller and Bartos, 2018, Rule et al., 2019, Deitch et al., 2021, Schoonover et al., 2021). Here, we showed that this representational drift is also present under behaviorally stable conditions, i.e. without any explicit learning paradigm. On the one hand this might not be too surprising, as the same has been shown for the underlying substrate: synapses appear and disappear on the time scale of days, and those remaining change their sizes (Loewenstein et al., 2011, Loewenstein et al., 2015, Villa et al., 2016, Berry and Nedivi, 2017). While often attributed to learning (e.g. Hebb, 1949) this synaptic drift is not only present in the absence of an explicit learning paradigm, but even, when neuronal activity has been silenced (Yasumatsu et al., 2008, Rubinski and Ziv, 2015, Dvorkin and Ziv, 2016, Nagaoka et al., 2016). On the other hand this representational drift is far from trivial, as the question remains, at what level the perceived robustness is achieved.

We showed that auditory cued fear conditioning led to a bias in the ongoing recombination of response modes. While leaving the overall response mode statistics largely intact, it led to an increased co-mapping of stimuli onto the same mode, i.e. after fear conditioning more stimuli elicited the same response mode. Two stimuli evoking the same response has been understood as an association (Grewe et al., 2017), so we find that learning leads to more associations, possibly in line with a generalization, as also witnessed in post-traumatic stress disorder (Besnard and Sahay, 2016). There, traumatized patients often generalize between the traumatic stimulus (e.g. explosions) and other similar, but harmless stimuli (e.g. the banging of a door). The observed increased co-mapping of stimuli similar to the conditioned stimulus onto the same response mode with th econditioned stimulus might be a potential mechanism.

We defined a set of ten elementary operations (*constantia, constantia′, creatio, eliminatio, adiunctio, disiunctio, fusio, fissio, commutatio, motio*) and were able to deconstruct response mode dynamics even further. These operations allow for a more detailed analysis and revealed that associations among stimuli representations are due to both an increase in operations forming associations and a decrease in operations breaking associations. Thus, ACFC has both a stabilizing effect on existing operations and leads to the formation of new associations.

As we can safely assume that any stimulus (inside the hearing range) has a representation in auditory cortex, some of our response mode operations – those involving the 0-mode – are obsolete on the level of the entire auditory cortex, and the number of operations is reduced to five relevant ones (*constantia, fusio, fissio, commutatio, motio*). It would be interesting to know, if on the global level those five can all be observed. However, on a local scale all ten operations have to be taken into account, as some stimuli don't evoke a local population response in a given field of view. The same is probably true for putative readout neurons. They can hardly receive input from all neurons at once, so for some stimuli they are deemed to receive no input. Thus, the entire set of ten operations is worth considering, when describing cortical population dynamics.

A firing rate model was able to reproduce key characteristics of population activity in mouse auditory cortex, including the clustering of stimulus responses into response modes. Varying the overall recurrent connection strength and the ratio of inhibition to excitation, we found five dynamic regimes of the model: The network produced two *uni-stable* regimes. One was input dominated, i.e. every stimulus evoked its own response, largely independent of the recurrent connections. The other was network dominated, so every stimulus evoked the same response, but different for each network. Classically, these regimes are often considered the same, as the previous work was rarely focused on responses to multiple stimuli. Apart from two more rather trivial regimes, a dampening and a diverging regime, we found clustering of stimuli into response modes in a *multi-stable* regime. In this regime, a stimulus is able to produce multiple responses, depending on the initial condition, and this behavior is well described in the literature – as are all the other regimes, at least from the point of view of a single stimulus (e.g. Wilson and Cowan, 1972, Sompolinsky et al., 1988, Zhang and Saggar, 2020). In our case, however, we found this multi-stable regime in a random network for strong recurrent connections and strong inhibition. Additionally, in this multi-stable regime stimulus responses are clustered.

Applying synaptic drift (modeled on Loewenstein et al., 2011) we find periods of stable network responses interrupted by abrupt transitions to new responses. This finding is in line with experimental data, as far as the two can be compared despite the lower temporal resolution of the experimental data. These coincide with qualitative changes in the fixed point topology and are not just a rerouting through minor displacement of fixed points. The qualitative changes associated with these transitions are typically a rerouting via the emergence (or disappearing) of an *unstable* fixed point somewhere along the activity trajectory or the rerouting onto an emerging (or away from a disappearing) *stable* fixed point that lies close to the former final state.

Taking all these findings together, we can conclude that the brain is not stable. Synaptic drift leads to representational drift, yet somehow, perceptions seem to be stable. The big question remaining is, how is this perceptional stability achieved? Or, is our percept of stability really true stability? Our memories change all the time, so can we be sure about this? This will require future investigation.

In a not so distant future there are a lot of interesting directions: Can we find the different regimes described by the model in real experimental data? One relevant parameter – the strength of the inhibition – could be adjusted pharmacologically. A decrease should drive the system into a regime without clustering, more similar to maybe visual cortex. Or vice versa, would we find clustering, if we increase the inhibition in visual cortex?

Another interesting question, more on the modeling side would be, if we find different types of drift in different regimes of the model. Can we find a regime without abrupt transitions? We could then describe the two modes of operations, we typically observe in our day to day lives: incremental changes in contrast to abrupt insights. If we find a mechansim to switch between those, already in simple network models, there is a high likelihood, these are also present in any drifting network. A step in a similar direction would be to observe a learning network during its epochs of learning. Could we identify different regimes in machine learning, too? Do machines learn incrementally or abruptly? Or both? There have been studies investigating the fixed points of networks that are fully trained on simple tasks (Sussillo and Barak, 2013), but what happens prior to the network reaching this static state?

A bit more on the experimental side, it would be interesting to perform fixed point analysis on functional or effective networks reconstructed from actual experimental recordings. It would give a deeper understanding of, whether the transitions that can be observed in recordings are really as abrupt as predicted by the model or if the transitions are qualitatively different. Of course, the temporal resolution in experimental data can never be as fine as in a model, however, we find stable response modes on the time scale of days (both in the experiment and in the model), so we might be able to observe and track changes in the fixed point topology. Along the same lines, it might be interesting to see what is happening during different cognitive processes. Starting from simple tweaks to the task (i.e. what do we see, when the mice do not undergo fear conditioning, but rather learn to perform a discrimination?) to more complex behavior like maternal care for pups or even abstract ideas as forgetting or insight. The set of elementary operations we defined could be useful to discrimiate what is happening there.

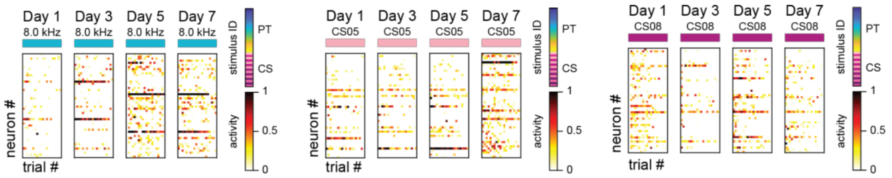
This leaves the questions: Why is there drift in the cortex? Is synaptic drift good for anything? And the same for representational drift? This is a relatively new question. It has often been thought to be a shortcoming of the brain and one of the main questions has been, how to counteract this drift (Chambers and Rumpel, 2017, Susman et al., 2019). There are many ideas, how stability can be achieved from dynamic components (e.g. Kossio et al., 2021). However, recently ideas have been brought to attention regarding the benefits of such a dynamic system. What if the brain has not evolved as a memory storage capacity but rather as a tool for fast adaptation? Maybe forgetting is as important for a good performance as is learning. This has been pushed from a psychological point of view, as well as from a machine learning point of view. Psychologically it makes a lot of sense to forget traumata, for example. Also, it typically does not help to remember every detail, as we are quite often in similar situations, but hardly ever in the exact same situation twice and we still need to know what to do (Richards and Frankland, 2017). Arguments from the machine learning side go In the same direction: random drift can be used as a tool to prevent overfitting. In neural networks there are different tools for such tasks, but a surprisingly good one is *dropout*, where a percentage of connections is removed randomly at each step, which prevents overfitting and thus increases the performance. A similar mechanism is potentially helpful in the brain, too. And drift leads to similar results as dropout (Aitken et al., 2021).

Part VI

APPENDIX

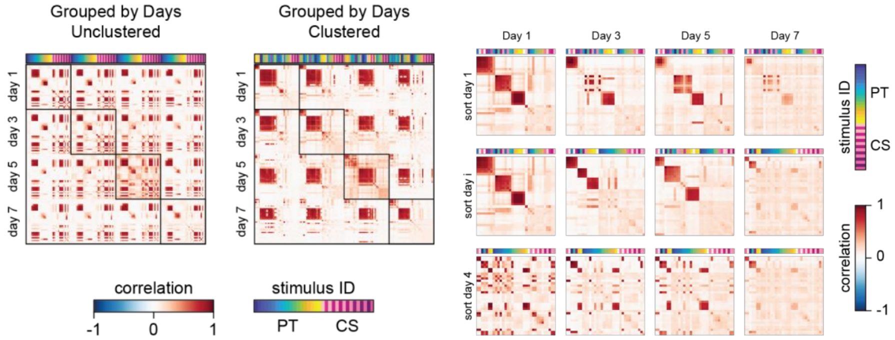
9

SUPPLEMENTARY FIGURES



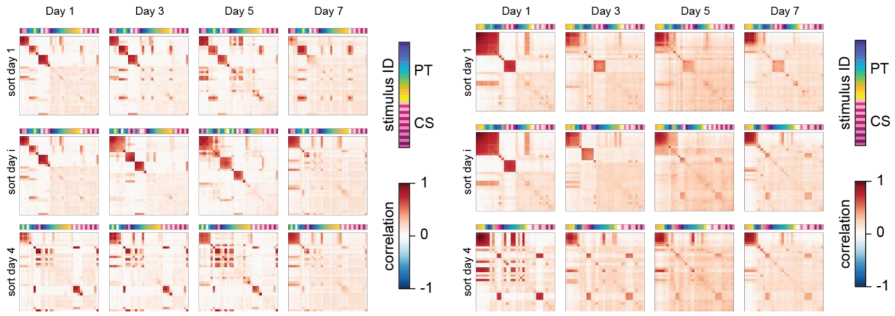
(a) Example response C.
(b) Example response D.
(c) Example response E.

Sup. Fig. 9.1: Further examples of single trial population response vectors for two example stimuli over time. For illustrative purposes, only fifty most active cells are shown, and trials are sorted by descending mean activity (PT: pure tones, CS: complex sounds)..



(a) Clustering procedure.

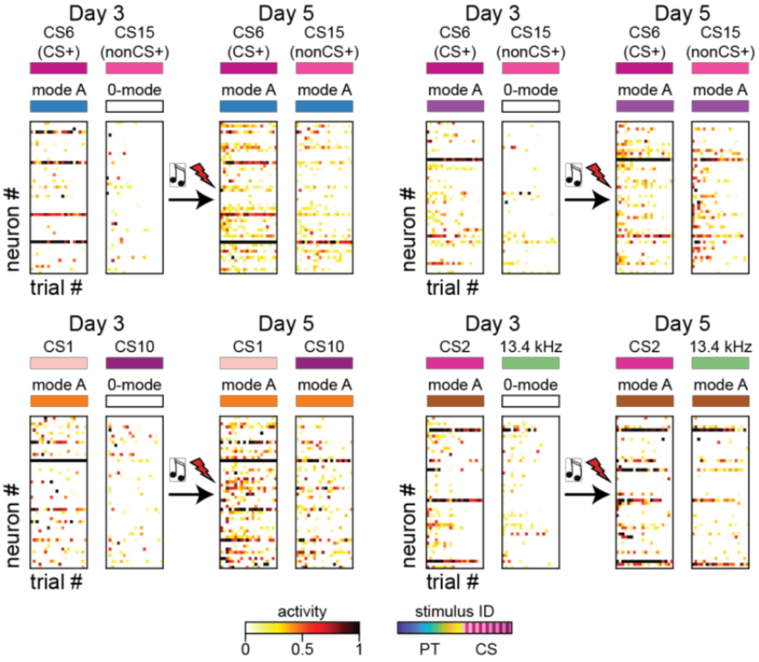
(b) Mode changes example A.



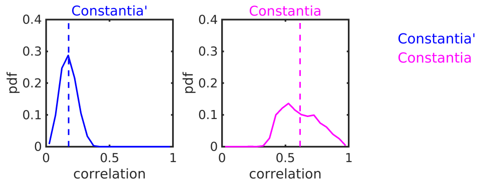
(c) Mode changes example B.

(d) Mode changes example C.

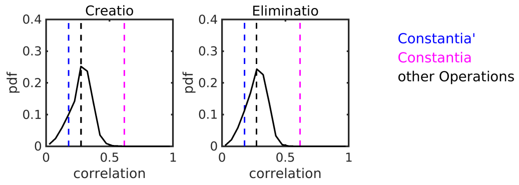
Sup. Fig. 9.2: Dynamics of response modes across days. (a) Left: Unsorted Similarity matrix for all sound-evoked responses from all four imaging days recorded from the example FOV shown in Figure 5.11 (PT: pure tones, CS: complex sounds). Right: Same as left, but sorted by hierarchical clustering of stimulus responses with high self-reliability and grouped by day (PT: pure tones, CS: complex sounds). (b)-(d), Single day similarity matrices of sound evoked responses sorted by hierarchical clustering for three additional example FOVs. Top: Sorting from day 1 is applied to the subsequent days. Middle: Sorting on each day individually. Bottom: Sorting from day 7 is applied to the previous days (PT: pure tones, CS: complex sounds).



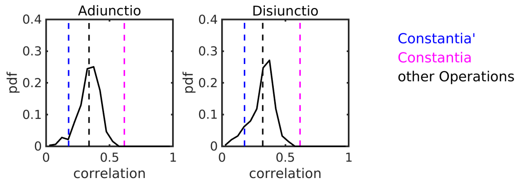
Sup. Fig. 9.3: Further examples of co-mapping. Top row: Further examples of single trial population response vectors from two FOVs showing the responses of the conditioned stimulus (CS+) and the non-conditioned stimulus (non-CS+). Prior to fear conditioning the non-CS+ did not elicit a significant response (0-mode), whereas after fear conditioning its response became similar to that of the CS+ (mode A). Top: Stimulus identity; Middle: Mode identity; Bottom: Single trial population response vectors. For illustrative purposes, only fifty most active cells are shown in random order, and trials are sorted by descending mean activity (PT: pure tones, CS: complex sounds, compare to Figure 5.27). Bottom row: Similar to top row, but showing the gain of a response mode representation for two other stimuli which were not presented during conditioning and gain a neuronal association to the response mode of another stimulus after fear conditioning..



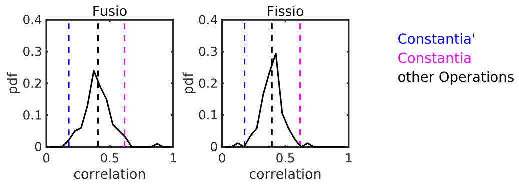
(a) *Constantia′* and *constantia*.



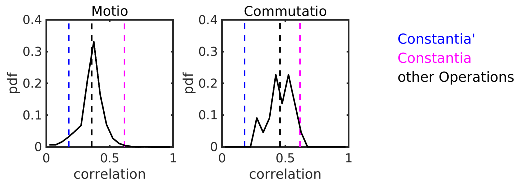
(b) *Creatio* and *eliminatio*.



(c) *Adiunctio* and *disiunctio*.

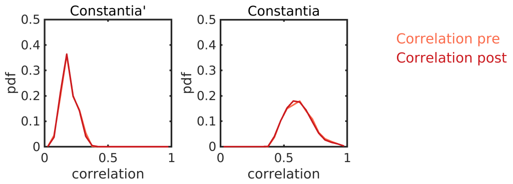


(d) *Fusio* and *fissio*.

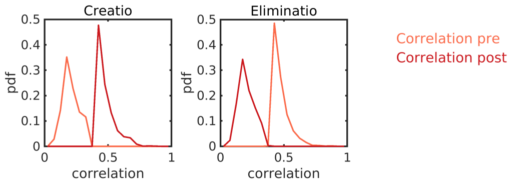


(e) *Motio* and *commutatio*.

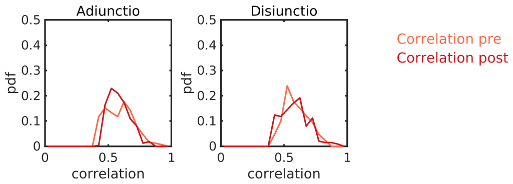
Sup. Fig. 9.4: Mean response correlation between all trials of a stimulus prior to a response mode operation and all trials of the same stimulus after the operation. This correlation is lowest for stimuli not evoking any response at all (*constantia′*) and highest for stimuli evoking the same response (*constantia*). All other operations fall in between.



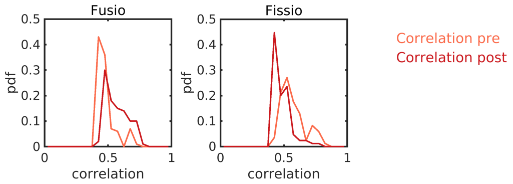
(a) *Constantia′* and *constantia*.



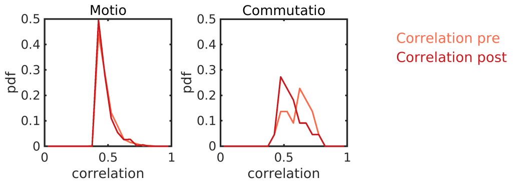
(b) *Creatio* and *eliminatio*.



(c) *Adiunctio* and *disiunctio*.

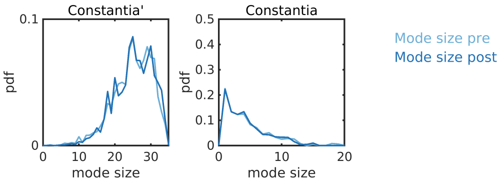


(d) *Fusio* and *fissio*.

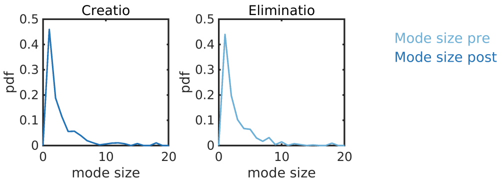


(e) *Motio* and *commutatio*.

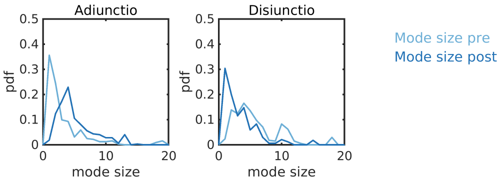
Sup. Fig. 9.5: Mean within response mode correlations before and after operations. Within mode correlations are roughly the same for all operations, depending solely on whether the stimulus is evoking the 0-mode or any other response mode.



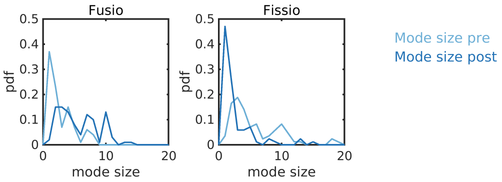
(a) *Constantia′* and *constantia*.



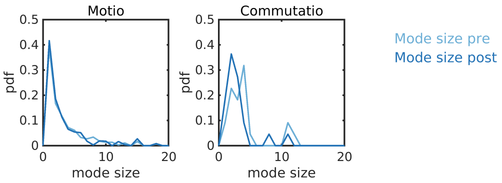
(b) *Creatio* and *eliminatio*.



(c) *Adiunctio* and *disiunctio*.

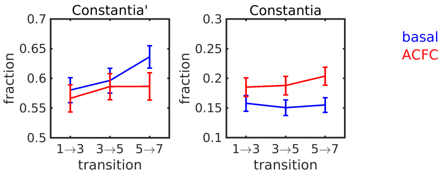


(d) *Fusio* and *fissio*.

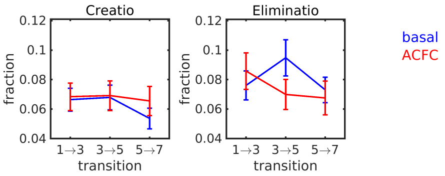


(e) *Motio* and *commutatio*.

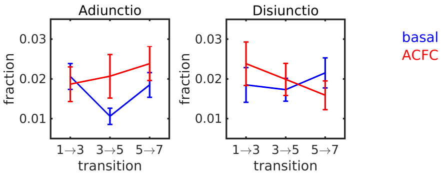
Sup. Fig. 9.6: Response mode size (i.e. number of stimuli mapped to a mode) distributions before and after operations. For *creatio* (*eliminatio*) only the response size after (before) the operation is shown.



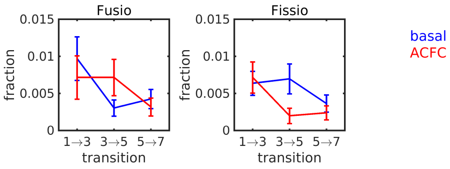
(a) *Constantia′* and *constantia*.



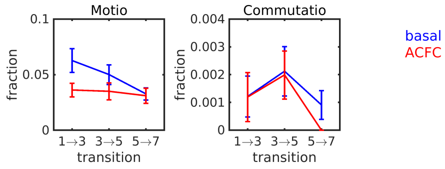
(b) *Creatio* and *eliminatio*.



(c) *Adiunctio* and *disiunctio*.

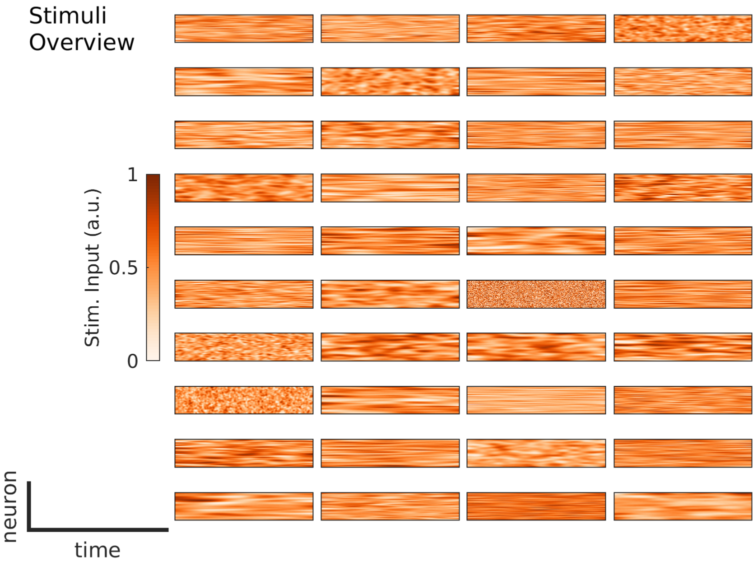


(d) *Fusio* and *fissio*.

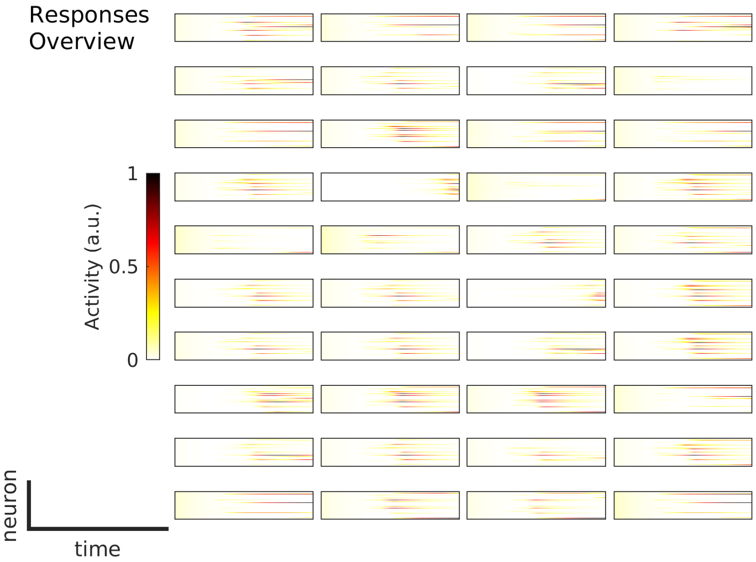


(e) *Motio* and *commutatio*.

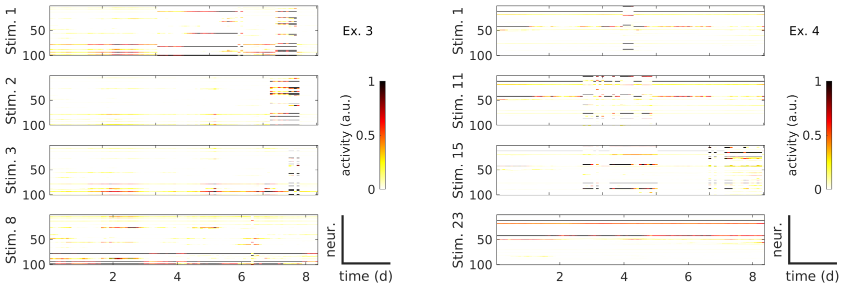
Sup. Fig. 9.7: Changes of response mode operations across time in both datasets.



Sup. Fig. 9.8: Overview of all 40 stimuli used for analyses in networks with 100 neurons. Stimuli were generated randomly by drawing from a uniform distribution in the interval $(0, 1)$ and filtering with normalized Gaussian filters of random mean and variance.



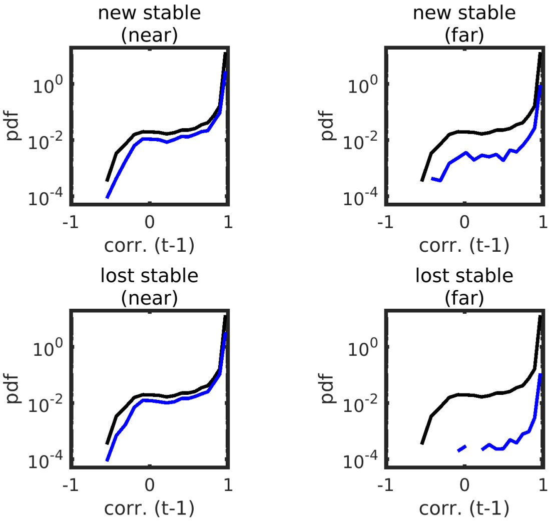
Sup. Fig. 9.9: Responses of example network displayed in Figure 7.1 to all stimuli shown in Sup. Fig. 9.8.



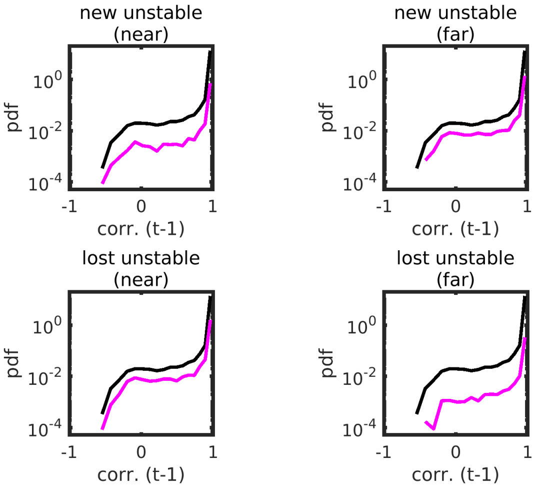
(a) Example Network 3.

(b) Example Network 4.

Sup. Fig. 9.10: Further examples of synaptic turnover. (a), (b) Responses to example stimuli of networks changing according to Equation 8.3 for example networks from Figure 8.4.



Sup. Fig. 9.11: Histograms of correlations between response vectors on consecutive time points for all time points where *stable* fixed points are changed (blue). For comparison the entire distribution (Figure 8.9) is plotted (black). These changes in *stable* fixed points are lost or new *stable* fixed points close to or far away from already existing fixed points. Partial disappearance of lines is caused by 0 counts and logarithmic y-axes.



Sup. Fig. 9.12: Histograms of correlations between response vectors on consecutive time points for all time points where *unstable* fixed points are changed (magenta). For comparison the entire distribution (Figure 8.9) is plotted (black). These changes in *unstable* fixed points are lost or new *unstable* fixed points close to or far away from already existing fixed points. Partial disappearance of lines is caused by 0 counts and logarithmic y-axes.

BIBLIOGRAPHY

Abbott, L. F., Rajan, K., and Sompolinsky, H. (2011). "Interactions between Intrinsic and Stimulus-Evoked Activity in Recurrent Neural Networks." In: *The Dynamic Brain: An Exploration of Neuronal Variability and Its Functional Significance*. Ed. by M. Ding and D. Glanzman. Oxford University Press.

Acker, D., Paradis, S., and Miller, P. (2019). "Stable memory and computation in randomly rewiring neural networks." In: *J Neurophysiol* 122, pp. 66–80.

Ades, H. W. and Engström, H. (1974). "Anatomy of the Inner Ear." In: *Auditory System: Anatomy Physiology (Ear)*. Ed. by W. D. Keidel and W. D. Neff. Springer Berlin Heidelberg, pp. 125–158.

Ahmadian, Y., Rubin, D. B., and Miller, K. D. (2013). "Analysis of the stabilized supralinear network." In: *Neural Comput* 25, pp. 1994–2037.

Aitken, K., Garrett, M., Olsen, S., and Mihalas, S. (2021). "The geometry of representational drift in natural and artificial neural networks." In: *bioRxiv*.

Aizenberg, M. and Geffen, M. N. (2013). "Bidirectional effects of aversive learning on perceptual acuity are mediated by the sensory cortex." In: *Nat Neurosci* 16, pp. 994–996.

Amit, D. J. and Brunel, N. J.-B. (1997). "Dynamics of a recurrent network of spiking neurons before and following learning." In: *Netw Comput Neural Syst* 8, pp. 373–404.

Armony, J. L., Servan-Schreiber, D., Romanski, L. M., Cohen, J. D., and Le-Doux, J. E. (1997). "Stimulus generalization of fear responses: effects of auditory cortex lesions in a computational model and in rats." In: *Cereb Cortex* 7, pp. 157–165.

Aschauer, D. F., Eppler, J.-B., Ewig, L., Chambers, A., Pokorny, C., Kaschube, M., and Rumpel, S. (2022). "Learning-induced biases in the ongoing dynamics of sensory representations predict stimulus generalization." In: *Cell Rep*.

Aschauer, D. and Rumpel, S. (2018). "The Sensory Neocortex and Associative Memory." In: *Curr Top Behav Neurosci* 37, pp. 177–211.

Atencio, C. A. and Schreiner, C. E. (2013). "Auditory cortical local subnetworks are characterized by sharply synchronous activity." In: *J Neurosci* 33, pp. 18503–18514.

Atkinson, K. A. (1989). *An Introduction to Numerical Analysis (2nd ed.)* John Wiley & Sons.

Aviel, Y. and Gerstner, W. (2006). "From spiking neurons to rate models: A cascade model as an approximation to spiking neuron models with refractoriness." In: *Phys Rev E* 73.

Barak, O., Sussillo, D., Romo, R., Tsodyks, M., and Abbott, L. F. (2013). "From fixed points to chaos: Three models of delayed discrimination." In: *Prog Neurobiol* 103, pp. 214–222.

Bathellier, B., Ushakova, L., and Rumpel, S. (2012). "Discrete neocortical dynamics predict behavioral categorization of sounds." In: *Neuron* 76, pp. 435–449.

Beer, R. D. (1995). "On the Dynamics of Small Continuous-Time Recurrent Neural Networks." In: *Adapt Behav* 3, pp. 469–509.

Benjamini, Y. and Hochberg, Y. (1995). "Controlling the false discovery rate – a practical and powerful approach to multiple testing." In: *J R Stat Soc B* 57, pp. 289–300.

Berry, J. A. and Davis, R. L. (2014). "Chapter 2 – Active Forgetting of Olfactory Memories in Drosophila." In: *Odor Memory and Perception*. Ed. by E. Barkai and D. A. Wilson. Vol. 208. Progress in Brain Research. Elsevier, pp. 39–62.

Berry, K. P. and Nedivi, E. (2017). "Spine dynamics: Are they all the same?" In: *Neuron* 96, pp. 43–55.

Besnard, A. and Sahay, A. (2016). "Adult hippocampal neurogenesis, fear generalization, and stress." In: *Neuropsychopharmacology* 41, pp. 24–44.

Bliss, T. V. and Collingridge, G. L. (1993). "A synaptic model of memory: long-term potentiation in the hippocampus." In: *Nature* 361, pp. 31–39.

Boddez, Y., Finn, M., and De Houwer, J. (2021). "The (shared) features of fear: toward the source of human fear responding." In: *Curr Opin Psychol* 41, pp. 113–117.

Bonferroni, C. E. (1936). *Teoria statistica delle classi e calcolo delle probabilita*. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze.

Borisyuk, R. M. and Kirillov, A. B. (1992). "Bifurcation analysis of a neural network model." In: *Biol Cybern* 66, pp. 319–325.

Brainard, M. S. and Knudsen, E. I. (1993). "Experience-dependent plasticity in the inferior colliculus: a site for visual calibration of the neural representation of auditory space in the barn owl." In: *J Neurosci* 13, pp. 4589–4608.

Bray, A. J. and Dean, D. S. (2007). "Statistics of critical points of gaussian fields on large-dimensional spaces." In: *Phys Rev Lett* 98.

Brette, R. (2015). "Philosophy of the Spike: Rate-Based vs. Spike-Based Theories of the Brain." In: *Front Syst Neurosci* 9.

Brunel, N. (2000). "Dynamics of Sparsely Connected Networks of Excitatory and Inhibitory Spiking Neurons." In: *J Comput Neurosci* 8, pp. 183–208.

Brunel, N., Chance, F., Foucard, N., and Abbott, L. (2001). "Effects of synaptic noise and filtering on the frequency response of spiking neurons." In: *Phys Rev Lett* 86, pp. 2186–2189.

Brunel, N. and Vincent, H. (1999). "Fast Global Oscillations in Networks of Integrate-and-Fire Neurons with Low Firing Rates." In: *Neural Comput* 11, pp. 1621–1671.

Buzsaki, G. (2010). "Neural syntax: cell assemblies, synapsembles, and readers." In: *Neuron* 68, pp. 362–385.

Buzsaki, G. and Mizuseki, K. (2014). "The log-dynamic brain: how skewed distributions affect network operations." In: *Nature reviews. Neuroscience* 15, pp. 264–278.

Cai, D. J., Aharoni, D., Shuman, T., Shobe, J., Biane, J., Song, W., Wei, B., Veshkini, M., La-Vu, M., Lou, J., Flores, S. E., Kim, I., Sano, Y., Zhou, M., Baumgaertel, K., Lavi, A., Kamata, M., Tuszynski, M., Mayford, M., Golshani, P., and Silva, A. J. (2016). "A shared neural ensemble links distinct contextual memories encoded close in time." In: *Nature* 534, pp. 115–118.

Cajal, S. Ramon y (1911). *Histologie du Systeme Nerveux de l'Homme et des Vertebres*. Maloine Paris.

Cessac, B. (1995). "Increase in Complexity in Random Neural Networks." In: *J Phys I* 5, pp. 409–432.

Chambers, A. R. and Rumpel, S. (2017). "A stable brain from unstable components: Emerging concepts and implications for neural computation." In: *Neuroscience* 257, pp. 172–184.

Chaurasiya, H. (2020). "Time-Frequency Representations: Spectrogram, Cochleogram and Correlogram." In: *Procedia Comput Sci* 167, pp. 1901–1910.

Chen, B. and Miller, P. (2020). "Attractor-state itinerancy in neural circuits with synaptic depression." In: *J Math Neurosci* 10.

Chen, T. W., Wardill, T. J., Y. Sun, S. R. Pulver, Renninger, S. L., Baohan, A., Schreiter, E. R., Kerr, R. A., Orger, M. B., Jayaraman, V., Looger, L. L., Svoboda, K., and Kim, D. S. (2013). "Ultrasensitive fluorescent proteins for imaging neuronal activity." In: *Nature* 499, pp. 295–300.

Cho, J. H., Rendall, S. D., and Gray, J. M. (2017). "Brain-wide maps of Fos expression during fear learning and recall." In: *Learn Memory* 24, pp. 169–181.

Clopath, C., Bonhoeffer, T., Hubener, M., and Rose, T. (2017). "Variance and invariance of neuronal long-term representations." In: *Philos Trans R Soc Lond B Biol Sci* 372.

Cohen, L. D. and Ziv, N. E. (2019). "Neuronal and synaptic protein lifetimes." In: *Curr Opin Neurobiol* 57, pp. 9–16.

Crash Tokio (2006). *We are plastic*. We are plastic.

Curto, C. and Morrison, K. (2016). "Pattern Completion in Symmetric Threshold-Linear Networks." In: *Neural Comput* 28, pp. 2825–2852.

Dale, H. H. (1934). "Pharmacology and Nerve-endings (Walter Ernest Dixon Memorial Lecture)." In: *J R Soc Med* 28 (3), pp. 319–330.

Dalmay, T., Abs, E., Poorthuis, R. B., Hartung, J., Pu, D. L., Onasch, S., Lozano, Y. R., Signoret-Genest, J., Tovote, P., Gjorgjieva, J., and Letzkus, J. J. (2019). "A Critical Role for Neocortical Processing of Threat Memory." In: *Neuron* 104, 1180–1194 e1187.

Dauphin, Y., Pascanu, R., Gulcehre, C., Cho, K., Ganguli, S., and Bengio, Y. (2014). "Identifying and attacking the saddle point problem in high-dimensional non-convex optimization." In: *arxiv*. arXiv:1406.2572.

Deitch, D., Rubin, A., and Ziv, Y. (2021). "Representational drift in the mouse visual cortex." In: *Curr Biol* 31, pp. 4327–4339.

Denk, W., Strickler, J. H., and Webb, W. W. (1990). "Two-Photon Laser Scanning Fluorescence Microscopy." In: *Science* 248, pp. 73–76.

Desai, N. S., Rutherford, L. C., and Turrigiano, G. G. (1999). "Plasticity in the intrinsic excitability of cortical pyramidal neurons." In: *Nat Neurosci* 2, pp. 515–520.

Dichter, M. A. and Ayala, G. F. (1987). "Cellular mechanisms of epilepsy: A status report." In: *Science* 237, pp. 157–164.

Driscoll, L. N., Pettit, N. L., Minderer, M., Chettih, S. N., and Harvey, C. D. (2017). "Dynamic Reorganization of Neuronal Activity Patterns in Parietal Cortex." In: *Cell* 170, 986–999 e916.

Dudai, Y. and Carruthers, M. (2005). "The Janus face of Mnemosyne." In: *Nature* 434.

Dunsmoor, J. E. and Paz, R. (2015). "Fear generalization and anxiety: behavioral and neural mechanisms." In: *Biol Psychiatry* 78, pp. 336–343.

Dvorkin, R. and Ziv, N. E. (2016). "Relative Contributions of Specific Activity Histories and Spontaneous Processes to Size Remodeling of Glutamatergic Synapses." In: *PLoS Biol* 14.

Eppler, J.-B. (2015). "Spontaneous Categorization in Mouse Auditory Cortex." M.Sc. Thesis. Goethe University Frankfurt am Main.

Ermentrout, B. (1994). "Reduction of conductance-based models with slow synapses to neural nets." In: *Neural Comput* 6, pp. 679–695.

Fasoli, D., Cattani, A., and Panzeri, S. (2016). "The Complexity of Dynamics in Small Neural Circuits." In: *PLoS Comput Biol* 12.

Fauth, M. J. and Rossum, M. C. van (2019). "Self-organized reactivation maintains and reinforces memories despite synaptic turnover." In: *eLife* 8.

Frankland, P. W., Köhler, S., and Josselyn, S. A. (2013). "Hippocampal neurogenesis and forgetting." In: *Trends Neurosci* 36, pp. 497–503.

Frisina, R. D. (2001). "Subcortical neural coding mechanisms for auditory temporal processing." In: *Hear Res* 158, pp. 1–27.

Galton, F. (1886). "Regression towards mediocrity in hereditary stature." In: *J R Anthropol Inst* 15, pp. 246–263.

Gauer, J., Nagathil, A., Martinr, R., Thomas, J. P., and Völter, C. (2019). "Interactive Evaluation of a Music Preprocessing Scheme for Cochlear

Implants Based on Spectral Complexity Reduction." In: *Front Neurosci* 13.

Gentet, L. J., Stuart, G. J., and Clements, J. D. (2000). "Direct measurement of specific membrane capacitance in neurons." In: *Biophys J* 79.

Gerstner, W. (1995). "Time structure of the activity in neural network models." In: *Phys Rev E Stat Nonlin Soft Matter Phy* 51, pp. 738–758.

— (2000). "Population dynamics of spiking neurons: fast transients, asynchronous states, and locking." In: *Neural Comput* 12, pp. 43–89.

Gerstner, W. and Kistler, W. M. (2002). *Spiking Neuron Models*. Cambridge University Press.

Gerstner, W., Kistler, W. M., Naud, R., and Paninski, L. (2014). *Neuronal Dynamics: From single neurons to networks of cognition and beyond*. Cambridge University Press.

Gillet, S. N., Kato, H. K., Justen, M. A., and Isaacson, J. S. (2017). "Fear Learning Regulates Cortical Sensory Representations by Suppressing Habituation." In: *Front Neural Circuits* 11.

Gini, C. (1936). "On the Measure of Concentration with Special Reference to Income and Statistics." In: *Colorado College Publication* 208, pp. 73–79.

Girko, V. L. (1984). "Circular Law." In: *Theory Probab Appl* 29, pp. 694–706.

— (1990). *Theory of Random Determinants*. Boston, MA: Kluwer.

Gray, L. (1997). "Auditory system: pathway and reflexes." In: *Neuroscience Online: An Electronic Textbook for the Neurosciences*. Ed. by J. H. Byrne and N. Dafny. The University of Texas Medical School at Houston.

Grewe, B. F., Gründemann, J., Kitch, L. J., Lecoq, J. A., Parker, J. G., Marshall, J. D., Larkin, M. C., Jercog, P. E., Grenier, F., Li, J. Z., Lüthi, A., and Schnitzer, M. J. (2017). "Neural ensemble dynamics underlying a long-term associative memory." In: *Nature* 543, pp. 670–675.

Grimm, D., Kay, M. A., and Kleinschmidt, J. A. (2003). "Helper virus-free, optically controllable, and two-plasmid-based production of adeno-associated virus vectors of serotypes 1 to 6." In: *Mol Ther* 7, pp. 839–850.

Grossberg, S. (1980). "How does a brain build a cognitive code?" In: *Psychol Rev* 87, pp. 1–51.

Guillery, R. W. and Sherman, S. M. (2011). "Branched thalamic afferents: what are the messages that they relay to the cortex?" In: *Brain Res Rev* 66, pp. 205–219.

Göppert-Mayer, M. (1931). "Über Elementarakte mit zwei Quantensprüngen." In: *Ann. Phys.* 401, pp. 273–294.

Götze, F. and Tikhomirov, A. (2007). "On the circular law." In: *arxiv*. arXiv:math/0702386.

Hainmueller, T. and Bartos, M. (2018). "Parallel emergence of stable and dynamic memory engrams in the hippocampus." In: *Nature* 558, pp. 292–296.

Hansel, D., Mato, G., Meunier, C., and Neltner, L. (1998). "On numerical simulations of integrate-and-fire neural networks." In: *Neural Comput* 10, pp. 467–483.

Hansel, D. and Sompolinsky, H. (1992). "Synchronization and computation in a chaotic neural network." In: *Phys Rev Lett* 68, pp. 718–721.

— (1996). "Chaos and synchrony in a model of a hypercolumn in visual cortex." In: *J Comput Neurosci* 3, pp. 7–34.

Hansel, D. and Vreeswijk, C. van (2002). "How noise contributes to contrast invariance of orientation tuning in cat visual cortex." In: *J Neurosci* 22, pp. 5118–5128.

Hardt, O., Nader, K., and Nadel, L. (2013). "Decay happens: the role of active forgetting in memory." In: *Trends Cogn Sci* 17, pp. 111–120.

Hardy, G. H., Littlewood, J. E., and Polya, G. (1988). *Inequalities*. Cambridge University Press.

Harish, O. and Hansel, D. (2015). "Asynchronous Rate Chaos in Spiking Neuronal Circuits." In: *PLoS Comput Biol* 11.

Harris, K. D. (2005). "Neural signatures of cell assembly organization." In: *Nat Rev Neurosci* 6, pp. 399–407.

Harris, K. D. and Mrsic-Flogel, T. D. (2013). "Cortical connectivity and sensory coding." In: *Nature* 503, pp. 51–58.

Hebb, D. O. (1949). *The Organization of Behavior*. John Wiley & Sons.

Hendry, S. H., Schwark, H. D., Jones, E. G., and Yan, J. (1987). "Numbers and proportions of GABA-immunoreactive neurons in different areas of monkey cerebral cortex." In: *J Neurosci*.

Hennequin, G., Ahmadian, Y., Rubin, D. B., Lengyel, M., and Miller, K. D. (2018). "The Dynamical Regime of Sensory Cortex: Stable Dynamics around a Single Stimulus-Tuned Attractor Account for Patterns of Noise Variability." In: *Neuron* 98, pp. 846–860.

Herculano-Houzel, S. (2009). "The human brain in numbers: a linearly scaled-up primate brain." In: *Front Hum Neurosci* 3.

Herculano-Houzel, S., Watson, C., and Paxinos, G. (2013). "Distribution of neurons in functional areas of the mouse cerebral cortex reveals quantitatively different cortical zones." In: *Front Neuroanat* 7.

Hodgkin, A. L. and Huxley, A. F. (1952). "A Quantitative Description of Membrane Current and its Application to Conduction and Excitation in Nerve." In: *J Physiol* 117, pp. 500–544.

Holtmaat, A. and Caroni, P. (2016). “Functional and structural underpinnings of neuronal assembly formation in learning.” In: *Nat Neurosci* 19, pp. 1553–1562.

Hopfield, J. J. (1982). “Neural networks and physical systems with emergent collective computational abilities.” In: *Proc Natl Acad Sci USA* 79, pp. 2554–2558.

Hromádka, T., DeWeese, M. R., and Zador, A. M. (2008). “Sparse Representation of Sounds in the Unanesthetized Auditory Cortex.” In: *PLoS Biology* 6.

Hromádka, T. and Zador, A. M. (2009). “Representations in auditory cortex.” In: *Curr Opin Neurobiol* 19, pp. 430–433.

Huber, D., Gutnisky, D. A., Peron, S., O’Connor, D. H., Wiegert, J. S., Tian, L., Oertner, T. G., Looger, L. L., and Svoboda, K. (2012). “Multiple dynamic representations in the motor cortex during sensorimotor learning.” In: *Nature* 484, pp. 473–478.

Hubert, L. J. and Baker, F. B. (1977). “Analyzing Distinctive Features.” In: *J Educ Stat* 2, pp. 79–98.

Humble, J., Hiratsuka, K., Kasai, H., and Toyoizumi, T. (2019). “Intrinsic Spine Dynamics Are Critical for Recurrent Network Learning in Models With and Without Autism Spectrum Disorder.” In: *Front Comput Neurosci* 13.

Humeau, Y. and Choquet, D. (2019). “The next generation of approaches to investigate the link between synaptic plasticity and learning.” In: *Nat Neurosci* 22, pp. 1536–1543.

Jahnke, S., Memmesheimer, R., and Timme, M. (2009). “How chaotic is the balanced state?” In: *Front Comput Neurosci* 3.

Jones, E. G. (1975). “Varieties and distribution of non-pyramidal cells in the somatic sensory cortex of the squirrel monkey.” In: *J Comp Neurol* 160, pp. 205–267.

Kadmon, J. and Sompolinsky, H. (2015). “Transition to Chaos in Random Neuronal Networks.” In: *Phys Rev X* 5.

Kaiser, W. and Garrett, C. G. B: (1961). “Two-Photon Excitation in CaF_2: Eu^{2+}.” In: *Phys Rev Lett* 7, pp. 229–231.

Kandel, E. R. (2001). “The Molecular Biology of Memory Storage: A Dialogue Between Genes and Synapses.” In: *Science* 294, pp. 1030–1038.

— (2013). *Principles of neural science (5th ed.)* New York: McGraw-Hill.

Kappel, D., Habenschuss, S., Legenstein, R., and Maass, W. (2015). “Network Plasticity as Bayesian Inference.” In: *PLoS Comput Biol* 11.

Kappel, D., Legenstein, R., Habenschuss, S., Hsieh, M., and Maass, W. (2018). “A Dynamic Connectome Supports the Emergence of Stable Computational Function of Neural Circuits through Reward-Based Learning.” In: *ENeuro* 5.

Kato, H. K., Gillet, S. N., and Isaacson, J. S. (2015). "Flexible Sensory Representations in Auditory Cortex Driven by Behavioral Relevance." In: *Neuron* 88, pp. 1027–1039.

Kawaguchi, Y and Kubota, Y (1997). "GABAergic cell subtypes and their synaptic connections in rat frontal cortex." In: *Cereb Cortex* 7, pp. 476–486.

Keller, D., Erö, C., and Markram, H. (2018). "Cell Densities in the Mouse Brain: A Systematic Review." In: *Front Neuroanat* 12.

Kentridge, R. W., Heywood, C. A., and Weiskrantz, L. (1999). "Attention without awareness in blindsight." In: *Proc Biol Sci* 266, pp. 1805–1811.

Kettcar (2002). *Ich danke der Academy*. Du und wieviel von deinen Freunden.

Kobak, D., Pardo-Vazquez, J. L., Valente, M., Machens, C. K., and Renart, A. (2019). "State-dependent geometry of population activity in rat auditory cortex." In: *elife* 8.

Kopec, C. D., Kessels, H. W., Bush, D. E., Cain, C. K., LeDoux, J. E., and Malinow, R. (2007). "A robust automated method to analyze rodent motion during fear conditioning." In: *Neuropharmacology* 52, pp. 228–233.

Kossio, Y. F. K., Goedeke, S., Klos, C., and Memmesheimer, R. (2021). "Drifting assemblies for persistent memory: Neuron transitions and unsupervised compensation." In: *Proc Natl Acad Sci USA* 118.

Krauss, P., Schuster, M., Dietrich, V., Schilling, A., Schulze, H., and Metzner, C. (2019a). "Weight statistics controls dynamics in recurrent neural networks." In: *PLoS One* 14.

Krauss, P., Zanki, A., Schilling, A., Schulze, H., and Metzner, C. (2019b). "Analysis of Structure and Dynamics in Three-Neuron Motifs." In: *Front Comput Neurosci* 13.

Kriegeskorte, N., Mur, M., and Bandettini, P. (2008). "Representational similarity analysis – connecting the branches of systems neuroscience." In: *Front Syst Neurosci* 2.

Kruskal, W. H. and Wallis, W. A. (1952). "Use of ranks in one-criterion variance analysis." In: *J Am Stat Assoc* 47, pp. 583–621.

Kullback, S. and Leibler, R. A. (1951). "On information and sufficiency." In: *Ann Math Stat* 22, pp. 79–86.

König, K. (2018). *Multiphoton Microscopy and Fluorescence Lifetime Imaging: Applications in Biology and Medicine*. Walter de Gruyter GmbH & Co KG.

Lai, C. S. W., Adler, A., and Gan, W. B. (2018). "Fear extinction reverses dendritic spine formation induced by fear conditioning in the mouse auditory cortex." In: *Proc Natl Acad Sci USA* 115, pp. 9306–9311.

Lapicque, L. (1907). "Recherches quantitatives sur l'excitation électrique des nerfs traitée comme une polarization." In: *J Physiol Pathol* 9, pp. 620–635.

Lazar, A., Pipa, G., and Triesch, J. (2009). "SORN: A self-organizing recurrent neural network." In: *Front Comput Neurosci* 3.

Letzkus, J. J., Woldd, S. B., Meyer, E. M., Tovote, P., Courtin, J., Herry, C., and Luthi, A. (2011). "A disinhibitory microcircuit for associative fear learning in the auditory cortex." In: *Nature* 480.

Liang, F., Li, H., Chou, X., Zhou, M., Zhang, N. K., Xiao, Z., Zhang, K. K., Tao, H. W., and Zhang, L. I. (2019). "Sparse representation in awake auditory cortex: cell-type dependence, synaptic mechanisms, developmental emergence, and modulation." In: *Cereb Cortex* 29, pp. 3796–3812.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (1967). "Perception of the speech code." In: *Psychol Rev* 74, pp. 431–461.

Linss, V., Linss, W., Emmerich, E., and Richter, F. (2007). "The cochleogram of the guinea pig." In: *Eur Arch Otorhinolaryngol* 264, pp. 369–375.

Litwin-Kumar, A. and Doiron, B. (2014). "Formation and maintenance of neuronal assemblies through synaptic plasticity." In: *Nat Commun* 5.

Loewenstein, Y., Kuras, A., and Rumpel, S. (2011). "Multiplicative dynamics underlie the emergence of the log-normal distribution of spine sizes in the neocortex in vivo." In: *J Neurosci* 31, pp. 9481–9488.

Loewenstein, Y., Yanover, U., and Rumpel, S. (2015). "Predicting the Dynamics of Network Connectivity in the Neocortex." In: *J Neurosci* 35, pp. 12535–12544.

Low, I. I. C., Williams, A. H., Campbell, M. G., Lindeman, S. W., and Giocomo, L. M. (2021). "Dynamic and reversible remapping of network representations in an unchanging environment." In: *Neuron* 109, pp. 2967–2980.

Lubda, M. (2021). "Predicting neural responses to auditory stimuli in the primary auditory cortex." M.Sc. Thesis. Goethe University Frankfurt am Main.

Luria, A. R. (1968). *The Mind of a Mnemonist: A Little Book about a Vast Memory*. Harvard University Press.

Ma, X. and Suga, N. (2009). "Specific and nonspecific plasticity of the primary auditory cortex elicited by thalamic auditory neurons." In: *J Neurosci* 29, pp. 4888–4896.

Maaten, L. van der and Hinton, G. (2008). "Visualizing data using t-SNE." In: *J Mach Learn Res* 9, pp. 2579–2605.

Maczulska, K. E., Tinter-Thiede, J., Peter, M., Ushakova, L., Bathellier, B., and Rumpel, S. (2013). "Dynamics of dendritic spines in the mouse auditory cortex during memory formation and memory recall." In: *Proc Natl Acad Sci USA* 110, pp. 18315–18320.

Mankin, E. A., Sparks, F. T., Slayyeh, B., Sutherland, R. J., Leutgeb, S., and Leutgeb, J. K. (2012). "Neuronal code for extended time in the hippocampus." In: *Proc Natl Acad Sci USA* 109, pp. 19462–19467.

Margolis, D. J., Lutcke, H., Schulz, K., Haiss, F., Weber, B., Kugler, S., Hasan, M. T., and Helmchen, F. (2012). "Reorganization of cortical population activity imaged throughout long-term sensory deprivation." In: *Nat Neurosci* 15, pp. 1539–1546.

Mastrogiuseppe, F. and Ostojic, S. (2017). "Intrinsically-generated fluctuating activity in excitatory-inhibitory networks." In: *PLoS Comput Biol* 13.

— (2018). "Linking Connectivity, Dynamics, and Computations in Low-Rank Recurrent Neural Networks." In: *Neuron* 99, pp. 609–623.

Mattia, M. and Giudice, P. D. (2002). "Population dynamics of interacting spiking neurons." In: *Phys Rev E Stat Nonlin Soft Matter Phys* 66.

McCulloch, W. S. and Pitts, W. (1943). "A logical calculus of the ideas immanent in nervous activity." In: *Bull Math Biophys* 5, pp. 115–133.

McGaugh, J. L. (2000). "Memory – a Century of Consolidation." In: *Science* 287, pp. 248–251.

Mendez, M. F. and Geehan Jr, G. R. (1988). "Cortical auditory disorders: clinical and psychoacoustic features." In: *J Neurol Neurosurg Psychiatry* 51, pp. 1–9.

Metzger, C., Werf, Y. van der, and Walter, M. (2013). "Functional mapping of thalamic nuclei and their integration into cortico-striatal-thalamo-cortical loops via ultra-high resolution imaging – from animal anatomy to in vivo imaging in humans." In: *Front Neurosci* 7.

Middlebrooks, J. C. and Knudsen, E. I. (1984). "A neural code for auditory space in the cat's superior colliculus." In: *J Neurosci* 4, pp. 2621–2634.

Miller, K. D. and Fumarola, F. (2012). "Mathematical equivalence of two common forms of firing rate models of neural networks." In: *Neural Comput* 24, pp. 25–31.

Miller, P. (2016). "Itinerancy between attractor states in neural systems." In: *Curr Opin Neurobiol* 40.

Miner, D. and Triesch, J. (2016). "Plasticity-Driven Self-Organization under Topological Constraints Accounts for Non-random Features of Cortical Synaptic Wiring." In: *PLoS Comput Biol* 12.

Minerbi, A., Kahana, R., Goldfeld, L., Kaufman, M., Marom, S., and Ziv, N. E. (2009). "Long-term relationships between synaptic tenacity, synaptic remodeling, and network activity." In: *PLoS Biology* 7.

Mongillo, G., Rumpel, S., and Loewenstein, Y. (2017). "Intrinsic volatility of synaptic connections – A challenge to the synaptic trace theory of memory." In: *Curr Opin Neurobiol* 46, pp. 7–13.

— (2018). “Inhibitory connectivity defines the realm of excitatory plasticity.” In: *Nat Neurosci* 21, pp. 1463–1470.

Montangie, L., Miehle, C., and Gjorgjieva, J. (2020). “Autonomous emergence of connectivity assemblies via spike triplet interactions.” In: *PLoS Comput Biol* 16.

Morel, A. and Kaas, J. H. (1992). “Subdivisions and connections of auditory cortex in owl monkeys.” In: *J Comp Neurol* 318, pp. 27–63.

Murphy, B. K. and Miller, K. D. (2009). “Balanced amplification: a new mechanism of selective amplification of neural activity patterns.” In: *Neuron* 61, pp. 635–648.

Nagaoka, A., Takehara, H., Hayashi-Takagi, A., Noguchi, J., Ishii, K., Shirai, F., Yagishita, S., Akagi, T., Ichiki, T., and Kasai, H. (2016). “Abnormal intrinsic dynamics of dendritic spines in a fragile X syndrome mouse model in vivo.” In: *Sci Rep* 6.

Nathanson, J. L., Yanagawa, Y., Obata, K., and Callaway, E. M. (2009). “Preferential labeling of inhibitory and excitatory cortical neurons by endogenous tropism of adeno-associated virus and lentivirus vectors.” In: *Neuroscience* 161, pp. 441–450.

Ni, G., Elliott, S. J., Ayat, M., and Teal, P. D. (2014). “Modelling Cochlear Mechanics.” In: *Biomed Res Int* 2014.

Noback, C. R., Strominger, N. L., Demarest, R. J., and Ruggiero, D. A. (2005). *The Human Nervous System: Structure and Function (Sixth ed.)s*. Totowa, NJ: Humana Press.

Ostojic, S. (2014). “Two types of asynchronous activity in networks of excitatory and inhibitory spiking neurons.” In: *Nat Neurosci* 17, pp. 594–600.

Ostojic, S. and Brunel, N. (2011). “From Spiking Neuron Models to Linear-Nonlinear Models.” In: *PLoS Comput Biol* 7.

Palmer, A. R. and Kuwada, S. (2005). “Binaural and spatial coding in the inferior colliculus.” In: *The inferior colliculus*. Ed. by J. A. Winer and C. E. Schreiner. Springer New York, pp. 377–410.

Pan, S. (2018). “Cochlea Modelling and its Application to Speech Processing.” Ph.D. Thesis. University of Southampton.

Pannese, A., Herrmann, C. S., and Sussman, E. (2015). “Analyzing the Auditory Scene: Neurophysiologic Evidence of a Dissociation Between Detection of Regularity and Detection of Change.” In: *Brain Topogr* 28, pp. 411–422.

Pasemann, F. (2002). “Complex dynamics and the structure of small neural networks.” In: *Netw Comput Neural Syst* 13, pp. 195–216.

Pavlov, I. P. and Anrep, G. V. (1927). “Conditioned reflexes; an investigation of the physiological activity of the cerebral cortex.” In: *Ann Neurosci* 17, pp. 136–141.

Pearson, K. (1895). “Notes on regression and inheritance in the case of two parents.” In: *Proc R Soc Lond* 58, pp. 240–242.

Peter, M., Scheuch, H., Burkard, T. R., Tinter, J., Wernle, T., and Rumpel, S. (2012). "Induction of immediate early genes in the mouse auditory cortex after auditory cued fear conditioning to complex sounds." In: *Genes Brain Behav* 11, pp. 314–324.

Ponulak, F. and Kasinski, A. (2011). "Introduction to spiking neural networks: Information processing, learning and applications." In: *Acta Neurobiol Exp* 71, pp. 409–433.

Poo, M. M., Pignatelli, M., Ryan, T. J., Tonegawa, S., Bonhoeffer, T., Martin, K. C., Rudenko, A., Tsai, L. H., Tsien, R. W., Fishell, G., Mullins, C., Gonçalves, J. T., Shtrahman, M., Johnston, S. T., Gage, F. H., Dan, Y., Long, J., Buzsáki, G., and Stevens, C. (2016). "What is memory? The present state of the engram." In: *BMC Biol* 19, pp. 14–40.

Poort, J., Khan, A. G., Pachitariu, M., Nemri, A., Orsolic, I., Krupic, J., Bauza, M., Sahani, M., Keller, G. B., Mrsic-Flogel, T. D., and Hofer, S. B. (2015). "Learning Enhances Sensory and Multiple Non-sensory Representations in Primary Visual Cortex." In: *Neuron* 86, pp. 1478–1490.

Quirk, G. J., Armony, J. L., and LeDoux, J. E. (1997). "Fear conditioning enhances different temporal components of tone-evoked spike trains in auditory cortex and lateral amygdala." In: *Neuron* 19, pp. 613–624.

Rajan, K. and Abbott, L. F. (2006). "Eigenvalue spectra of random matrices for neural networks." In: *Phys Rev Lett* 97, pp. 2–5.

Rajan, K., Abbott, L. F., and Sompolinsky, H. (2010). "Stimulus-dependent suppression of chaos in recurrent neural networks." In: *Phys Rev E* 82.

Reale, R. A. and Imig, T. J. (1980). "Tonotopic organization in auditory cortex of the cat." In: *J Comp Neurol* 192, pp. 265–291.

Richards, B. A. and Frankland, P. W. (2017). "The persistence and transience of memory." In: *Neuron* 94, pp. 1071–1084.

Robles, L. and Ruggero, M. A. (2001). "Mechanics of the mammalian cochlea." In: *Physiol Rev* 81, pp. 1305–1352.

Rokni, U., Richardson, A. G., Bizzi, E., and Seung, H. S. (2007). "Motor learning with unstable neural representations." In: *Neuron* 54, pp. 653–666.

Romani, G. L., Williamson, S. J., and Kaufman, L. (1982). "Tonotopic Organization of the Human Auditory Cortex." In: *Science* 216, pp. 1339–1340.

Rosenblatt, F. (1958). "The perceptron – a probabilistic model for information storage and organization in the brain." In: *Psychol Rev* 65.

Rost, T., Deger, M., and Nawrot, M. P. (2018). "Winnerless competition in clustered balanced networks: inhibitory assemblies do the trick." In: *Biol Cybern* 112, pp. 81–98.

Rubin, A., Geva, N., Sheintuch, L., and Ziv, Y. (2015a). "Hippocampal ensemble dynamics timestamp events in long-term memory." In: *elife* 4.

Rubin, D. B., Hooser, S. D. van, and Miller, K. D. (2015b). "The Stabilized Supralinear Network: A Unifying Circuit Motif Underlying Multi-Input Integration in Sensory Cortex." In: *Neuron* 85, pp. 402–417.

Rubinski, A. and Ziv, N.E. (2015). "Remodeling and Tenacity of Inhibitory Synapses: Relationships with Network Activity and Neighboring Excitatory Synapses." In: *PLoS Comput Biol* 11.

Rudnicki, M., Schoppe, O., Isik, M., Völk, F., and Hemmert, W. (2015). "Modeling auditory coding: from sound to spikes." In: *Cell Tissue Res* 361, pp. 159–175.

Rule, M. E., O'Leary, T., and Harvey, C. D. (2019). "Causes and consequences of representational drift." In: *Curr Opin Neurobiol* 58, pp. 141–147.

Rumpel, S. and Triesch, J. (2016). "Thy dynamic connectome." In: *e-Neuroforum* 22, pp. 48–53.

Russo, M., Stella, M., Sikora, M., and Šarić, M. (2019). "Cochlea-inspired speech recognition interface." In: *Med Biol Eng Comput* 57, pp. 1393–1403.

Sahara, S., Yanagawa, Y., O'Leary, D. D. M., and Stevens, C. F. (2012). "The Fraction of Cortical GABAergic Neurons Is Constant from Near the Start of Cortical Neurogenesis to Adulthood." In: *J Neurosci* 31(14), pp. 4755–4761.

Saladin, K. (2011). *Human anatomy (3rd ed.)* McGraw-Hill.

Schacter, D., Addis, D., and Buckner, R. (2007). "Remembering the past to imagine the future: the prospective brain." In: *Nat Rev Neurosci* 8, pp. 657–661.

Schmitt, M., Mayerhöfer, T., Popp, J., Kleppe, I., and Weisshart, K. (2013). "Light–Matter Interaction." In: *Handbook of Biophotonics*. John Wiley & Sons, Ltd, pp. 87–261.

Schoonover, C. E., Ohashi, S. N., Axel, R., and Fink, A. J. P. (2021). "Representational drift in primary olfactory cortex." In: *Nature* 594, pp. 541–546.

Schrauwen, B., Wardermann, M., Verstraeten, D., Steil, J. J., and Stroobandt, D. (2008). "Improving reservoirs using intrinsic plasticity." In: *Neurocomputation* 71, pp. 1159–1171.

Schulte, L. (2017). "Effects of Synaptic Changes to Responses in a Network Model." B.Sc. Thesis. Goethe University Frankfurt am Main.

See, J. Z., Atencio, C. A., Sohal, V. S., and Schreiner, C. E. (2018). "Coordinated neuronal ensembles in primary auditory cortical columns." In: *elife* 7.

Shadlen, M. N. and Newsome, W. T. (1994). "Noise, neural codes and cortical organization." In: *Curr Opin Neurobiol* 4, pp. 569–579.

Shadlen, M. N. and Newsome, W. T. (1998). "The variable discharge of cortical neurons: Implications for connectivity, computation, and information coding." In: *J Neurosci* 18, pp. 3870–3896.

Sheintuch, L., Geva, N., Baumer, H., Rechavi, Y., Rubin, A., and Ziv, Y. (2020). "Multiple maps of the same spatial context can stably coexist in the mouse hippocampus." In: *Curr Biol* 30.

Sherman, S. M. (2007). "The thalamus is more than just a relay." In: *Curr Opin Neurobiol* 17, pp. 417–422.

— (2012). "Thalamocortical interactions." In: *Curr Opin Neurobiol* 22, pp. 575–579.

Shi, T., Chang, D. E., and Cirac, J. I. (2015). "Multiphoton-scattering theory and generalized master equations." In: *Phys Rev A* 92.

Sillito, A. M. (1975). "The contribution of inhibitory mechanisms to the receptive field properties of neurones in the striate cortex of the cat." In: *J Physiol* 250, pp. 305–329.

Singheiser, M., Gutfreund, Y., and Wagner, H. (2012). "The representation of sound localization cues in the barn owl's inferior colliculus." In: *Front Neur Circuits* 6.

Sompolinsky, H., Crisanti, A., and Sommers, H. J. (1988). "Chaos in Random Neural Networks." In: *Phys Rev Lett* 61, pp. 259–262.

Statman, A., Kaufman, M., Minerbi, A., Ziv, N. E., and Brenner, N. (2014). "Synaptic Size Dynamics as an Effectively Stochastic Process." In: *PLoS Computational Biology* 10.

Steil, J. J. (2007). "Online reservoir adaptation by intrinsic plasticity for backpropagation-decorrelation and echo state learning." In: *Neural Netw* 20, pp. 353–364.

Stern, M., Sompolinsky, H., and Abbott, L. F. (2014). "Dynamics of Random Neural Networks with Bistable Units." In: *Phys Rev E* 90.

Studer, F. and Barkat, T. R. (2022). "Inhibition in the auditory cortex." In: *Neurosci Biobehav Rev* 132, pp. 61–75.

Susman, L., Brenner, N., and Barak, O. (2019). "Stable memory with unstable synapses." In: *Nat Commun* 10.

Sussillo, D. and Barak, O. (2013). "Opening the Black Box: Low-Dimensional Dynamics in High-Dimensional Recurrent Neural Networks." In: *Neural Comput* 25, pp. 626–649.

Svoboda, K. and Yasuda, R. (2006). "Principles of Two-Photon Excitation Microscopy and Its Applications to Neuroscience." In: *Neuron* 50, pp. 823–839.

Tao, T. and Vu, V. (2010). "Random matrices: Universality of ESDs and the circular law." In: *Ann Probab* 38, pp. 2023–2065.

Taylor, B. (1715). *Methodus Incrementorum Directa et Inversa*. London.

Theunissen, F. E. and Elie, J. E. (2014). "Neural processing of natural sounds." In: *Nat Rev Neurosci* 15, pp. 355–266.

Thompson, R. F. (1962). "Role of the cerebral cortex in stimulus generalization." In: *J Comp Physiol Psychol* 55, pp. 279–287.

Tonegawa, S., Liu, X., Ramirez, S., and Redondo, R. (2015). "Memory engram cells have come of age." In: *Neuron* 87, pp. 918–931.

Toutounji, H. and Pipa, G. (2014). "Spatiotemporal Computations of an Excitable and Plastic Brain: Neuronal Plasticity Leads to Noise-Robust and Noise-Constructive Computations." In: *PLoS Comput Biol* 10.

Triesch, J., Vo, A. D., and Hafner, A. (2018). "Competition for synaptic building blocks shapes synaptic plasticity." In: *eLife* 7.

Tsukano, H., Horie, M., Ohga, S., Takahashi, K., Kubota, Y., Hishida, R., Takebayashi, H., and Shibuki, K. (2017). "Reconsidering Tonotopic Maps in the Auditory Cortex and Lemniscal Auditory Thalamus in Mice." In: *Front Neural Circuits* 11.

Turrigiano, G. G., Leslie, K. R., Desai, N. S., Rutherford, L. C., and Nelson, S. B. (1998). "Activity-dependent scaling of quantal amplitude in neocortical neurons." In: *Nature* 391, pp. 892–896.

Uhlenbeck, G. E. and Ornstein, L. S. (1930). "On the theory of brownian motion." In: *Phys Rev* 36, pp. 823–841.

Venn, J. (1880). "I. On the Diagrammatic and Mechanical Representation of Propositions and Reasonings." In: *Lond Edinb Dublin Philos Mag J Sci* 10, pp. 1–18.

Villa, K. L., Berry, K. P., Subramanian, J., Cha, J. W., Oh, W. C., Kwon, H., Kubota, Y., So, P. T. C., and Nedivi, E. (2016). "Inhibitory synapses are repeatedly assembled and removed at persistent sites in vivo." In: *Neuron* 89, pp. 756–769.

Vogels, T. P., Sprekeler, H., Zenke, F., Clopath, C., and Gerstner, W. (2011). "Inhibitory plasticity balances excitation and inhibition in sensory pathways and memory networks." In: *Science* 334, pp. 1569–1573.

Vogelstein, J. T., Packer, A. M., Machado, T. A., Sippy, T., Babadi, B., Yuste, R., and Paninski, L. (2010). "Fast nonnegative deconvolution for spike train inference from population calcium imaging." In: *J Neurophysiol* 104, pp. 3691–3704.

Vreeswijk, C. van and Sompolinsky, H. (1996). "Chaos in Neuronal Networks with Balanced Excitatory and Inhibitory Activity." In: *Science* 274, pp. 1724–1726.

— (1997). "Chaotic Balanced State in a Model of Cortical Circuits." In: *Neural Comput* 10, pp. 1321–1371.

Wallas, G. (1926). *The art of thought*. Solis Press.

Watt, A. and Desai, N. (2010). "Homeostatic plasticity and STDP: keeping a neuron's cool in a fluctuating world." In: *Front Synaptic Neurosci* 2.

Waxman, S. G. (2017). "Higher Cortical Functions." In: *Clinical Neuroanatomy, 28e*. New York, NY: McGraw-Hill Education.

Wei, Y. (2012). "Eigenvalue spectra of asymmetric random matrices for multicomponent neural networks." In: *Phys Rev E Stat Nonlin Soft Matter Phys* 85.

Weinberger, N. M. (2004). "Specific long-term memory traces in primary auditory cortex." In: *Nat Rev Neurosci* 5, pp. 279–290.

Wilson, H. R. and Cowan, J. D. (1972). "Excitatory andinhibitory interactions in localized populations of model neurons." In: *Biophys J* 12.

— (1973). "A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue." In: *Kybernetik* 13, pp. 55–80.

Xu, L., Zhang, J., Yin, L., Long, X., Zhang, W., and Zhang, Q. (2020). "Recent progress in efficient organic two-photon dyes for fluorescence imaging and photodynamic therapy." In: *J Mater Chem C* 8, pp. 6342–6349.

Yang, Y., Liu, D. Q., Huang, W., Deng, J., Sun, Y., Zuo, Y., and Poo, M. M. (2016). "Selective synaptic remodeling of amygdalocortical connections associated with fear memory." In: *Nat Neurosci* 19, pp. 1348–1355.

Yasumatsu, N., Matsuzaki, M., Miyazaki, T., Noguchi, J., and Kasai, H. (2008). "Principles of long-term dynamics of dendritic spines." In: *J Neurosci* 28, pp. 13592–13608.

Yuste, R. (2015). "From the neuron doctrine to neural networks." In: *Nat Rev Neurosci* 16, pp. 487–497.

Zenke, F. and Gerstner, W. (2017). "Hebbian plasticity requires compensatory processes on multiple timescales." In: *Phil Trans R Soc B* 372.

Zenke, F., Gerstner, W., and Ganguli, S. (2017). "The temporal paradox of Hebbian learning and homeostatic plasticity." In: *Curr Opin Neurobiol* 43, pp. 166–176.

Zhang, M. and Saggar, M. (2020). "Complexity of resting brain dynamics shaped by multiscale structural constraints." In: *bioRxiv*.

Zhang, W. and Linden, D. J. (2003). "The other side of the engram: experience-driven changes in neuronal intrinsic excitability." In: *Nat Rev Neurosci* 4, pp. 885–900.

Zhao, Y., Zhang, Z., Liu, X., Xiong, C., Xiao, Z., and Yan, J. (2015). "Imbalance of excitation and inhibition at threshold level in the auditory cortex." In: *Front Neural Circuits* 9.

Zheng, P., Dimitrakakis, C., and Triesch, J. (2013). "Network Self-Organization Explains the Statistics and Dynamics of Synaptic Connection Strengths in Cortex." In: *PLoS Comput Biol* 9.

Ziv, N. E. and Brenner, N. (2018). "Synaptic tenacity or lack thereof: Spontaneous remodeling of synapses." In: *Trends Neurosci* 41, pp. 89–99.

Ziv, Y., Burns, L. D., Cocker, E. D., Hamel, E. O., Ghosh, K. K., Kitch, L. J., El Gamal, A., and Schnitzer, M. J. (2013). "Long-term dynamics of CA1 hippocampal place codes." In: *Nat Neurosci* 16, pp. 264–266.

> *Ich danke der Academy, für's Erkennen von Talent*
> *Das Leben schreit nach Energie, wahrscheinlich war ich besser nie als in diesem Moment*
> *Jetzt hier im Licht, Pacino, de Niro und ich*
>
> — *Kettcar,* 2002

ACKNOWLEDGEMENTS

I want to thank everyone involved in the making of this thesis.
First and foremost I want to thank Matthias Kaschube for making all of this possible. Thank you for being a great supervisor and mentor!
Many thanks go to Jochen Triesch for agreeing to be the second supervisor of this thesis, but also for being a friendly collaborator and some pieces of good advice.
Dominik Aschauer not only provided the beautiful dataset and was a great collaborator, but also became a friend.
Thank you, Simon Rumpel, for the great collaboration, all the data, and your constant constructive criticism. I hope I haven't heard the last of it.
Bettina Hein, Amon Khavari, Thomas Lai read this thesis and provided helpful feedback.
Thanks go to all current and former members of the Kaschube and Rumpel labs for patiently answering my questions and discussing even the weirdest papers and ideas.
I want to thank Gertrud and Paul Eppler, Nicola and Mathias Weber, Philipp Eppler for their continuing support.
Thanks go to Alex and Andi and Jan and Henning and Oli and Sam and Timm and Tobi: Tanzt hart!
Betty at Studentenwerk supplied me with a seemingly limitless amount of good food.
The scientific community at FIAS was very forthcoming and helped create a very friendly and productive working environment and the FIAS administration in general – and Susanne Steiner in particular – made my life easier and more fun.
Tamo and Alex and everyone at Feinstaub who pour a beautiful beer.
Thanks to Rüdiger for being great company.
Last but far from least I want to thank Raphaela Golling for always having my back.

CURRICULUM VITAE

Name	Jens-Bastian Eppler
Date of Birth	11th December 1987
Place of Birth	Frankfurt am Main
Nationality	German

Education

since 2015	PhD student in Physics - Goethe University Frankfurt Frankfurt Institute for Advanced Studies
2013 - 2015	MSc in Physics - Goethe University Frankfurt Master thesis: *'Spontaneous Categorization in Mouse Auditory Cortex'* supervised by Prof. Dr. Matthias Kaschube
2008 - 2013	BSc in Physics - Goethe University Frankfurt Bachelor thesis: *'Construction of a TPC test chamber in order to investigate the voltage trips of the ALICE TPC'* supervised by Prof. Dr. Harald Appelshäuser 2010 - 2011 ERASMUS study period at Copenhagen University
2007 - 2008	Zivildienst - Red Cross Künzelsau
2007	Abitur - Ganerben Gymnasium Künzelsau

COLOPHON

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*". `classicthesis` is available for both LaTeX and LyX:

`https://bitbucket.org/amiede/classicthesis/`