TÉCNICO LISBOA

Diffusion Augmentation in Latent Program Spaces as a Cognitive Model of Psychedelic Action

Carolina Santos Carvalho Caramelo

Thesis to obtain the Master of Science Degree in

Biomedical Engineering

Supervisor(s): Dr. Daniel Ciarán McNamee
Prof. Dr. Cláudia Alexandra Martins Lobato da Silva

Examination Committee

Chairperson: Prof. Patrícia Margarida Piedade Figueiredo
Supervisor: Dr. Daniel Ciarán McNamee
Member of the Committee: Prof. Hugo Humberto Plácido da Silva

March 2023

Declaração

Declaro que o presente documento é um trabalho original da minha autoria e que cumpre todos os requisitos do Código de Conduta e Boas Práticas da Universidade de Lisboa.

Declaration

I declare that this document is an original work of my own authorship and it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Prefácio

O trabalho apresentado nesta tese foi realizado no Champalimaud Centre for the Unknown, Fundação Champalimaud (Lisboa, Portugal), durante o período Março 2022 - Março 2023, com a supervisão do Dr. Daniel Ciarán McNamee. Esta tese foi co-supervisada no Instituto Superior Técnico pela Prof. Cláudia Lobato da Silva.

Preface

The work presented in this thesis was performed at Champalimaud Centre for the Unknown, Fundação Champalimaud (Lisbon, Portugal), during the period March 2022 - March 2023, under the supervision of Dr. Daniel Ciarán McNamee. The thesis was co-supervised at Instituto Superior Técnico by Prof. Cláudia Lobato da Silva.

Acknowledgments

"From here on love was the only consideration . . . It was and is the only purpose. Love seemed to emanate from a single point of light. . . and it vibrated."

This document is the result of a journey of one year? Five years? Twenty-three years? A journey that was solely accomplished by being nourished from the energy and love of those who walk side by side with me. My gratitude goes to the people who allowed me to fail, who taught me to be better person, and who reminded me in the toughest moments of how wonderful it is to dance through this life.

To the people who brought this work to life. To Daniel, a special thank you for welcoming me at the beginning of his adventure in Lisbon. Thank you for the tireless support, encouragement in the toughest moments, and for being a mentor during the past year, not only supervising this thesis work, but also introducing me to countless new topics. I will leave the Natural Intelligence lab knowing that science and research are sometimes unpredictable but wonderful things. I also want to thank Prof. Cláudia Lobato da Silva for the indispensable guidance, help, and, most importantly, constant enthusiasm. To all my CCU colleagues who accompanied me on my journey this year, thank you for the enthusiastic support and patience in listening to my "now and then" existential crises. A special thanks to Gonçalo for being a key piece in this puzzle, for his guidance and endless discussions, for the kind words on the toughest days.

To all those who made my time at Instituto Superior Técnico an experience I wish that could last forever. "E vai um brinde? Vai!" To you MOF, Dulce, João, Francisco, Leonor, Inês, Tiago, Francisca, Nuns, Jorge, Diogo, Marta, Sara, and Pimenta. The world is not ready for us.

The last two years have been the most challenging of my life. I am grateful to those who did not let go of my hand when riding this two-year roller coaster. To Carolina, Inês, and Diogo, for teaching me the importance of friendship in leadership. To Nono and Giullia, for showing me that home can be on the other side of the world and that love has no nationality, ethnicity, or religion.

Thank you to my best friends, who are my lifeboat in this turbulent sea that is uncovering life and its mysteries. Thank you Mariana and Teresa for believing in me more than I believe in myself. Also, thank you to all my oldest friends who grew up with me and accompanied me for more years than I can count, Bea, Mada, Tomás, David, Raquel, Bruna, and Rita, the person I am today is a little bit of you.

My purest and most profound gratitude goes to the greatest stream of affection, support, and love for which words fail me, to my family. Thank you to my grandparents and aunt for always believing that the sky is not the limit for my dreams. To Tomás, thank you for being the calm and the help in times of despair, the best friend I could have ever asked for, without you, life would not have the same color. The work here presented is also yours. To my parents, thank you for being the reason behind all the opportunities and success I have had throughout my life. Thank you for buying me the books, for teaching me the values, and most importantly for the unconditional love. I hope one day I can give back everything you have given me. Finally, to my sister, I hope we continue to inspire each other, I cannot wait to see you fly and conquer the world (meanwhile I will still be wearing your clothes). (Caju, I love you.)

Resumo

A investigação das drogas psicadélicas encontra-se numa fase de renascença devido ao potencial destas substâncias no tratamento de pacientes com doenças do foro psicológico, como a depressão e o transtorno obsessivo-compulsivo. Os resultados mostram que os psicadélicos conduzem os pacientes a terem experiências profundas, que podem catalisar mudanças psicológicas duradouras. No entanto, ainda existe uma enorme lacuna quando se trata de relacionar as interações neurofarmacológicas destas substâncias com alterações na atividade de populações neuronais, com os efeitos subjetivos que estas provocam e ainda com os resultados positivos observados na psicoterapia assistida por psicadélicos (PAP). Investigar modelos computacionais, no âmbito da neurociência cognitiva, poderá ser um passo essencial para a melhor compreensão do funcionamento destas drogas.

Nesta tese, propomos um método computacional baseado num modelo de Bayesian Program Learning (BPL) que pretende simular o efeito dos psicadélicos no cérebro. Tendo encontrado inspiração na hipótese de que, durante a experiência psicadélica, os modelos internos que as pessoas têm do mundo passam por algum tipo de modulação, ainda não completamente compreendida, permitindo-lhes formular ”novas perspectivas” após a experiência, este trabalho aborda a experiência psicadélica como uma experiência internamente conduzida. A metadologia desenvolvida estabelece uma analogia entre os efeitos das drogas psicadélicas e uma *pipeline* de programação probabilística, através 1) do aumento de dados de treino por meio de um procedimento de perturbações difusivas num espaço latente generativo e 2) da avaliação do seu impacto na *performance* do modelo na realização de uma tarefa de classificação. Para ilustrar o impacto da perturbação difusiva na tarefa de classificação, foram utilizados diferentes hiperparâmetros.

Os resultados mostram que a *framework* desenvolvida resulta num ligeiro melhoramento da *performance* do modelo em comparação com a experiência de controlo realizada, sugerindo que a abordagem conceptualizada deve ser futuramente explorada e refinada não apenas no contexto de modelos de machine learning (ML), mas também nos domínio da investigação dos psicadélicos e da ciência cognitiva.

Palavras-chave: Drogas psicadélicas, Psicoterapia assistida por psicadélicos, Modelos internos, Bayesian Program Learning, Perturbações difusivas, Aumento de dados.

Abstract

Psychedelic drugs are now undergoing a renaissance in research for their potential therapeutic applications. For the past years, numerous studies have demonstrated their effectiveness in treating mental health disorders such as depression and obsessive-compulsive disorders, leading to profound experiences that catalyze lasting psychological change. However, there is still an enormous gap when it comes to linking psychedelic neuropharmacological interactions to large-scale changes in neural populations activity, network connectivity, reported subjective effects, and the positive observed outcomes in psychedelic-assisted psychotherapy (PAP). Investigating computational models in cognitive neuroscience could be a promising research avenue to pursue in this domain.

In this thesis we propose a computational framework based on a Bayesian Program Learning (BPL) model that attempts to simulate the psychedelic action on the brain. Inspired by the hypothesis that people's internal models go through some, not yet understood, modulation allowing them to formulate "new perspectives" about the world post experience, this work approaches the psychedelic experience as internally driven. Our method establishes an analogy between psychedelic drug effects and a probabilistic program induction pipeline by 1) performing data augmentation through a generative latent space diffusion-based perturbation procedure and 2) evaluating its impact on the model's performance in a one-shot classification task. To illustrate the impact of the diffusive perturbation in the classification task, different hyperparameters were used.

Results show that the developed framework results in slightly improved model performance comparing to a control computational experiment, nevertheless, suggesting that our approach is worthwhile for exploration not only within the field of machine learning (ML), but also in the domains of psychedelic and cognitive research.

Keywords: Psychedelic drugs, Psychedelic-assisted psychotherapy, Internal models, Bayesian Program Learning, Diffusion-based perturbations, Data augmentation.

Contents

List of Tables

List of Figures

List of Acronyms

5-HT 5-hydroxytryptamine

5-HT$_{2A}$R 5-hydroxytryptamine 2A receptor

5-MeO-DMT 5-methoxy-dimethytryptamin

ACC Anterior cingulate cortex

AED Anxious Ego Dissolution

ALBUS Altered Beliefs Unders Psychedelics

AMPA a-amino-3-hydroxy-5-methylisoxazole-4-propionate

APZ-OAV Standardized psychometric assessment scale

ASC Altered states of consciousness

ASL Arterial spin labeling

BDNF Brain derived neurotrophic factor

BOLD Blood oxygen level dependent

BPL Bayesian Program Learning

CBF Cerebral blood flow

CDF Cumulative distribution function

ConvNet convolutional Neural Network

CPP Critical period plasticity

CRP Chinese Restaurant Process

CSTC Cortico-striato-thalamo-cortical

DA Dopamine

DL-BPL Diffusive Latents for Bayesian Program Learning

DMN Default mode network

DOI 2,5-dimethoxy-4-iodoam- phetamine

DOM 1-(2,5-dimethoxy-4-methylphenyl)-2-aminopropane

DP Dirichlet Process

EEG Electroencephalogram

EPSPs Excitatory postsynaptic potentials

FC Functional connectivity

fMRI Functional magnetic resonance imaging

GNS Generative neuro-symbolic

GPCR G protein-coupled receptor

GSK-3 Glycogen synthase kinase 3

HBM Hierarchical Bayesian Model

HPC Hierarchical Predictive Coding

JSD Jensen Shannon Distance

KLD Kullback-Leibler Divergence

L5p layer 5 pyramidal neurons

LSD Lysergic acid diethylamide

MCMC Markov chain Monte Carlo

MDMA 3,4-methylendioxymethampheamine

MEG Magnetoencephalography

ML Machine Learning

mPFC Medial prefrontal cortex

mRNA Messenger Ribonucleic Acid

MTL Medial temporal lobes

mTOR Mammalian target of rapamycin

N,N-DMT N,N-dimethyltryptamine

NMDA N-methyl-D-aspartate

OB Oceanic Boundlessness

OCD Obsessive compulsive disorder

P50 Event related potential occurring approximately 50 ms after the presentation of a stimulus

PAP Psychedelic-assisted psychotherapy

PCC Posterior cingulate cortex

PCP Phencyclidine

PFC Prefrontal cortex

PTSD Post-traumatic stress disorder

REBUS Relaxed beliefs under psychedelics

RSFC Resting state functional connectivity

RSN Resting-state network

SPECT Single-photon emission computerized tomography

TPNs Task positive networks

TrkB Tropomyosin receptor kinase B

VR Visionary Restructuralization

List of Algorithms

Chapter 1

Introduction

1.1 Motivation

Psychoactive drugs, including psychedelics, have been used by humans for thousands of years, dating back to its indigenous use for traditional medical practices [1]. Though they remain a controlled substance in nearly all legal jurisdictions, psychedelics have recently attracted much clinical research interest due to at least three factors. First, political campaigns have successfully led to a more relaxed regulatory framework, allowing for the use of psychedelics in public research [1]. Second, developments in synthetic pharmacology have facilitated the systematic generation and study of psychoactive drugs. Third, many common psychiatric diseases and depressive disorders, which are increasingly present in the general population and constitute one of the three leading causes of years lived with disability worldwide [2], remain resistant to current pharmacological intervention despite decades of clinical research and drug prescriptions. In particular, psychedelic compounds have attracted much interest in their potential therapeutic benefits for depression, anxiety and post-traumatic stress disorder (PTSD), resulting from a series of phase 2 clinical trials that have shown potential long-term outcomes in positively impacting the symptomatology of patients that carry these psychological disorders [2–4]. However, the neuro-computational effects of psychedelics remains poorly understood despite a wealth of knowledge regarding their molecular action in the brain. Researchers around the world are engaged in an effort to understand how these substances impact the computations, algorithms, and biological mechanisms of the human brain. This work focuses on understanding the influence of psychedelics within the context of internal models and neural simulation of the psychedelic experience. In psychedelic-assisted psychotherapy (PAP), in quiet and dark settings with minimal sensory input, a wide variety of strikingly rich and seemingly nonsensical internal visualizations have been reported, sometimes leading to long-term conceptual re-organization of the individuals' perspectives [5]. Computationally, these experiences can be interpreted as a dynamical simulation process associated with the sampling-based generative modeling of prior experiences and knowledge (i.e. episodic, semantic, and procedural memories), in an effort to produce novel explanatory interpretations of reality for consolidation and thus future reuse. This is pertinent to the proposed role of psychedelics in therapy, since aberrant beliefs usually associated

with disorders like depression or PTSD can be revised or even eradicated [5]. The focus of this work is trying to simulate the internal psychedelic experience through an adequate computational framework [6, 7] and separately model the sleep and wake phases [8, 9], while simultaneously exploring it in the context of machine learning (ML) model performance enhancement.

1.2 Overview and problem formulation

1.2.1 The history of psychedelics

"Psychedelic" is a neologism that combines the words psych ($\Psi\upsilon\chi\acute{\eta}$,"soul") and deloun ($\delta\eta\lambda o\tilde{\upsilon}\nu$, "to make visible, to reveal"), to denote "mind-revealing" [10]. The history of use of and interest in these drugs can be traced back to ancient times, when indigenous cultures, especially in America, began exploring the effects of the hallucinogenic brew *ayahuasca*. This ceremonial use of psychedelics has historically placed a strong focus on environmental context and psychological factors such as having a clear aim and an open, inquisitive mindset, as well as the importance of ceremony and rituals when using these drugs [1].

After lysergic acid diethylamide (LSD) was first synthesized in 1943 [10], reports on the subjective effects of the drug started to emerge, resulting in the first wave of interest from psychologists and psychiatrists into the therapeutic potential of these drugs [11]. Subsequently, this led to an active discussion around psychedelic research in the 1950s, despite the socio-political issues that have surrounded the subject. Since then, psychedelic drugs keep to rise interest in a wide range of fields such as molecular biology [12], neurophysiology and neuropharmacology [13], cognitive neurosciences [14], chemistry [15], anthropology [16], philosophy [17], psychology [18], sociology and arts [19].

1.2.2 A renewed interest in psychedelics in the modern era

A huge impact of these drugs was felt during the 1950s and 1960s in Western culture [1]. This was the first phase of sustained psychedelic scientific research, however, as more mainstream and counter-cultural forces embraced drugs, their societal impact grew exponentially, leading to the popularization of these drugs and resulting in LSD and related drugs to be classified as Schedule I in the United States (Controlled Substances Act, 1970) and in a similar category in most other countries [1], making them illegal.

Following a 25-year interregnum [10], research into psychedelics has been revived, with some referring to the present renaissance as the "third wave" [20]. A modern psychiatric view has emerged focusing on the potential mechanisms through which psychedelics might exert therapeutic effects [21], when used in an assisted-therapy environment [10] within the framework of psychological disorders such as depression, addiction and anxiety [22], as well as alcohol and other drug abuse disorders that present a large burden on individuals, families and country's healthcare systems [11]. After 15 years of small clinical trials providing evidence for the efficacy of these drugs in treating the above mentioned disorders [23, 24], a significant effort continues to be made aimed at understanding the neurobiological

and neuropharmacological mechanisms underlying psychedelic action [21, 25], and how it can lead to structural and functional changes in cortical neurons, e.g. via its neural plasticity-promoting properties [26]. More fundamentally, psychedelics are now viewed as a research tool for molecularly perturbing the normal functioning brain in order to understand its functional properties [27].

1.2.3 Modern experimental neuroscience tools in psychedelic research

Investigations regarding how psychedelics affect large-scale brain activity and connectivity profiles amongst different brain regions have been performed in order to characterize the atypical states of consciousness in which these substances result [28, 29]. Neuroimaging techniques such as arterial spin labeling (ASL) functional magnetic resonance imaging (fMRI) [27, 30], blood-oxygen-level-dependent (BOLD) fMRI, resting-state fMRI [14, 27, 29], as well as magnetoencephalography (MEG) [14, 27], have revealed alterations in whole brain organization that may be responsible for the acute psychedelic experience, namely variability of spontaneous brain activity fluctuations and connectivity, decreased functional connectivity and decreased oscillatory power in brain regions that are normally highly metabolically active, functionally connected and synchronous in their activity [28, 29]. These tools play a relevant role in analysing and characterizing brain networks in the temporal domain (dynamic functional connectivity), including resting-state networks (RSNs) (see Section 2.2.3 for details) [31, 32], during the psychedelic experience.

1.2.4 Computational hypothesis

To date, the above mentioned studies have suggested decreases in the activity and connectivity in brain's key connector hubs [30], proposing a "disintegration" [27] of central brain networks and enabling a state of unconstrained cognition [14, 27], leading to higher-level computational theories regarding how psychedelics might be affecting the brain, namely the RElaxed Beliefs Under pSychedelics (REBUS) framework [5].

The REBUS model aims to explain a wide range of phenomena associated with the psychedelic experience, based on the fundamental idea that, under psychedelics, there is a sensitization of high-level belief priors to bottom-up signaling[1], i.e there is a flattening of top-down belief priors with respect to perceptual expectations, and this occurrence enables the potential revision of this priors [5]. Depression, obsessive-compulsive disorder, end-of-life existential distress, addictions and eating disorders, are examples of psychiatric disorders that manifest implicit beliefs or biases that have become overly dominant and resistant to revision [33], which can be referred to pathological priors [5]. For instance, patients with a diagnosis of depression often show a negative cognitive bias, characterized by pessimism, low cognitive flexibility, inflexible thought patterns, and negative fixations regarding the "self" and the future, triggering depressive episodes, which can be interpreted as "attractor states" (stereotyped cognitive states with "gravitational pull") [28]. Considering this framework of thought, psychedelic therapy has in view to take advantage of this belief-relaxation opportunity to achieve a healthy revision of problematic

[1] Bottom-up signaling corresponds to processing of external sensory inputs [5].

beliefs [33]. Overall, not only classic psychedelics, but also dissociative psychedelics are known to have rapid onset antidepressant and anti-addictive effects [26]. Accordingly, it has been shown that serotonergic psychedelics increase neuritogenesis, spinogenesis and synaptogenesis [25], promoting dendritic branching and dendritic spine formation, representing the addition of new synapses neuronal circuitry, that might even compete with old, ”aberrant” synapses [22] existing in psychiatric disorders. It is hypothesized that psychedelics initiate a cascade of neurobiological changes that manifest at multiple scales and ultimately culminate in the relaxation of high-level beliefs [33], however, a big question remains, concerning whether the subjective effects of psychedelics are actually necessary for their therapeutic effects [25] or if the latter relies only on the biological mechanisms of these drugs.

Inspired by this high-level psychedelic computational theory and motivated by the idea that psychedelic drugs when administered in a assisted-therapy context might induce alterations in one's internal models of the world, we aim to conceptualize and then explore a probabilistic induction computational framework establishing an analogy with psychedelic action at a high cognitive level.

1.3 Objectives and Contributions

The main contributions of this thesis work are:

- Development of a comprehensive state-of-the-art review on psychedelic drugs, covering several levels of abstraction regarding the action and therapy utilization of these substances.
- Conception of a theoretical computational model framework analogous to the psychedelic action on the brain by leveraging a program induction model. The framework consists of a data augmentation procedure done by implementing diffusive perturbations in a generative latent space and of a one-shot classification task.

1.4 Thesis Outline

The document is organized as follows. Chapter 1 provides the Introduction, depicting the motivation of the work and overview of the problem; Chapter 2 contains an extensive review of psychedelic drugs exploring from their molecular mechanisms of action to high-level computational theories about these substances. Chapter 3 details the proposed computational approach for the simulation of psychedelic action. We introduce key machine learning concepts, formalizing the computational framework and establishing an analogy between the work pipeline and the action of psychedelic drugs. In Chapter 4 we describe, analyse and discuss the performed computational experiments and resort to an alternative pipeline after facing some study limitations. Chapter 5 concludes the thesis work by summarizing the main take-aways and pointing out directions for future research.

Chapter 2

The neuroscience of psychedelics: a state-of-the-art review

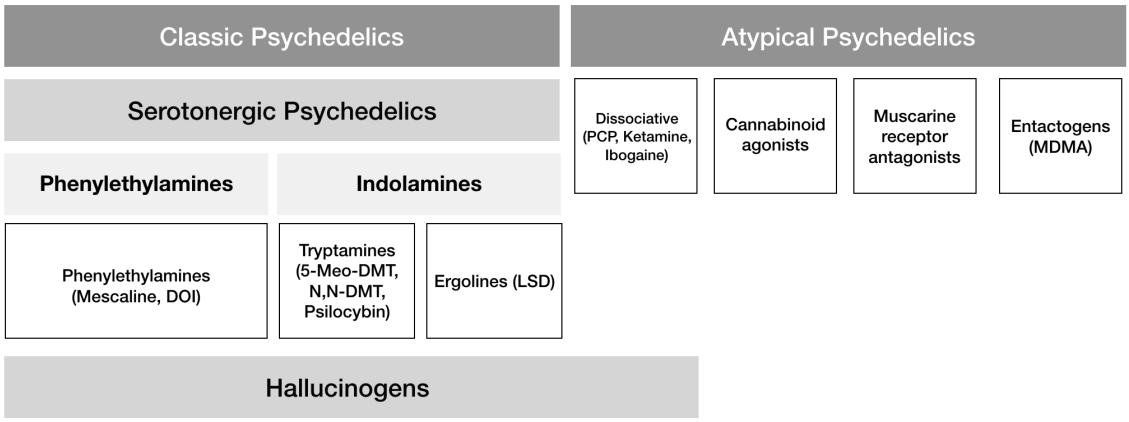
2.1 Circuit-level mechanisms of psychedelic action

What is happening in our brains when psychedelic substances are administered? The pharmacological and physiological impacts of psychedelic drug molecules on neurons and networks of neurons have been studied for some years now [34]. Although some of these effects at a molecular level are well understood, there is still a deep lack of knowledge when it comes to understanding how these induce the phenomena that people experience. In particular, these drugs have the capacity to produce intense acute experiences, and long-term alterations in neurobiology, by activating many neuromodulatory systems simultaneously. In doing so, a wide range of neural circuits which underpin fundamental perceptual and cognitive brain functions such as memory and executive decision-making are impacted [35]. Furthermore, it has been suggested that these mechanisms may be leveraged for unique therapeutic approaches to psychiatric disorders [26]. Therefore, the development of an understanding of the neuronal targets and causal effects of psychedelics, as well as establishing connections to higher cognitive functionality, is of great interest. This Section will explore the molecular properties of psychedelic drugs, as well as the circuit-level mechanisms of their action.

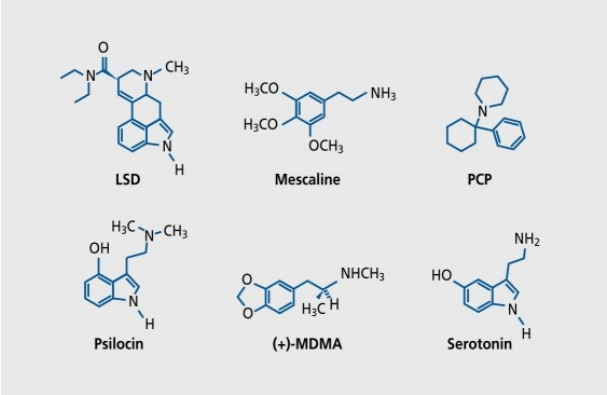
2.1.1 Classification of psychedelic compounds according to their action and subjective effect

Psychedelic drugs are classified (Figure 2.1a) into classic psychedelics and atypical/non-traditional/non-classic psychedelics based on their neuro-receptor affinities and chemical structure (Figure 2.1b), which together determine the primary mode of action of psychedelics [26]. Classic psychedelics can be further decomposed into three broad classes [37, 38]:

- *phenylethylamines* such as 3,4,5-trimethoxy-phenethylamine (derived from the peyote cactus and more commonly known as mescaline) and 2,5-dimethoxy-4-iodoam-phetamine (DOI),



(a) Psychedelic drugs' classification.



(b) Psychedelic drugs' chemical structure.

Figure 2.1: **Psychedelic drugs' classification and chemical structure.** **(a)** Classification of psychedelic drugs according to their neuro-receptor affinity, chemical structure and subjective effects. **(b)** Chemical structure of some hallucinogenic drugs compared to the serotonin chemical structure. Adapted from Vollenweider (2022) [36].

- *tryptamines* such as 5-methoxy-dimethytryptamine (5-MeO-DMT), N,N-dimethyltryptamine (N,N-DMT, that can be found in the plant ayahuasca), N,N-dimethyl-4-phosphoryloxy-tryptamine (from the psilocybe genus of mushroom, where psilocybin, a naturally occurring plant alkaloid, can be found [39]), and
- *ergolines* such as lysergic acid diethylamide (LSD), which is derived from lysergic acid extracted from ergot fungus [26].

The latter two classes are collectively referred to as *indolamines*.

Atypical psychedelics can be further categorized into *dissociative psychedelics*, which are N-methyl-D-aspartate receptor NMDA antagonists[1], including *arylcyclohexylamines* [36], such as phencyclidine (PCP), ketamine and ibogaine, as well as cannabinoid agonists (e.g., Δ9-tetrahydrocannabinol), muscarinic receptor antagonists (e.g., scopolamine), and entactogens (e.g., 3,4-methylendioxymethampheamine [MDMA], also known as "ecstasy") [26].

Indolamines and *phenylethylamines* are also called serotonergic hallucinogens, since they have been

[1] A receptor antagonist is a class of receptor ligand or drug that binds to a receptor and inhibits it rather than activating it like an agonist would. This stops or dulls a biological response.

proven to act upon the serotonergic system, as well as to induce hallucinations when taken [36]. *Aryl-cyclohexylamines* are the only compounds in the atypical psychedelics class that have also shown hallucinogenic properties. Entactogens, structurally resembling serotonergic hallucinogens, also induce psychedelic-like symptoms, however, do not cause hallucinations [36].

2.1.2 Psychedelics and the serotonergic system

Following the discovery of LSD and the identification of serotonin (5-HT), the observation that this strong psychoactive chemical had the capacity to interact with 5-HT systems sparked a great deal of interest in its role in psychedelic action [35]. Indeed, it has been shown that LSD modulates the serotonergic system via several different pathways [40]. More generally, serotonergic hallucinogens have been demonstrated to act upon serotonin receptors 5-HT_1, 5-HT_2 (namely 5-HT_{2A}, 5-HT_{2B} and 5-HT_{2C} receptors), 5-HT_6, and 5-HT_7 receptors, and partly upon adrenergic α_2 receptors and dopamine (DA) receptors D_1 and D_2 [36].

Glennon et al. (1983) [41] proposed the hypothesis that hallucinogenic drugs acted specifically at 5-HT_2 receptor sub-types based on drug discrimination studies in rats, which showed that the 5-HT_2 antagonists ketanserin and pirenperone blocked the discriminative stimulus effects of phenethylamine and tryptamine hallucinogens, including LSD [42–44]. Combined with early investigation showing that 5-HT [45, 46] antagonists inhibit 5-HT_2 receptors suppressing the discriminative stimulation of mescaline, a consistent picture emerged that the most critical mechanism in mediating the psychedelic effects is agonist or partial agonist activity at 5-HT_{2A} receptors (5-HT_{2A}R) [47].

Some of the most compelling evidence that hallucinogens have agonist activity at 5-HT_{2A}R, was obtained from two clinical studies. The first investigation, showed that 5-HT_{2A} and 5-HT_{2C} antagonist cyproheptadine antagonized the subjective effects of N,N-DMT in certain patients [48]. In the second study Vollenweider et al. (1998) [49] reported that the relatively 5-HT_{2A}-selective antagonists ketanserin and ritanserin prevented the hallucinatory effects of psilocybin as measured by the standardized psychometric assessment scale (APZ-OAV)[2] of Dittrich's altered states of consciousness (ASC) questionnaire [50], where ketanserin significantly reduced the psilocybin-induced increase in the the APZ-OAV score, leading to the conclusion that 5-HT_{2A}R blockade reduces most of the effects of psilocybin in human subjects.

The 5-HT_{2A} receptor

The 5-HT_{2A}R is a G protein-coupled receptor (GPCR) and one of the fourteen different 5-HT receptor sub-types that are expressed in the mammalian brain [51]. It is distinguished by being the main excitatory GPCR of the serotonin receptor family [52]. Remarkably, 5-HT_{2A}R antagonists have shown to substantially reduce or abolish the subjective effects of psilocybin, LSD, and N,N-DMT in humans [53–59]. Furthermore, in rodents, head twitch responses[3] induced by the administration of DOI were

[2]The use of questionnaire-based reports to access human subjective experiences of altered states of consciousness will further be explored in future sections.

[3]Fast head side-to-side movements considered to be a behavioural marker of an experience homologous to a human hallucination [60, 61].

also reduced by 5-HT_{2A}R blockage [62].

In humans 5-HT_{2A}R are highly expressed in the apical dendrites of excitatory glutamatergic layer 5 pyramidal (L5p) neurons in the cortex [63], being a predominantly cortical receptor and the most abundant 5-HT receptor in the cortex [64]. Moreover, it is particularly enriched in the prefrontal cortex (PFC), considered to be a high-level associative cortex region, with most of its cells in expressing 5-HT_{2A}R mRNA [65–67], and adjacent cortical regions. Other brain regions belonging to the default mode network (DMN) [68], whose importance for the problem's context is explained later, also exhibited a high expression of these receptors. Expression in these regions overcomes 5-HT_{2A}R expression in sub-cortical structures such as the basal ganglia, the thalamus and hippocampus [69].

In vitro electrophysiological recordings, after the administration of DOI or LSD in rat, have demonstrated an increase in the frequency and amplitude of spontaneous excitatory postsynaptic potentials (EPSPs) and excitatory postsynaptic currents in L5p neurons in the medial PFC, and also in other cortical regions by activating 5-HT_{2A}R [70, 71]. 5-HT_{2A}R activation has been shown to have depolarizing effects on neurons, turning it into a more excitable state [47, 52], however, it is not defining that this will have an overall excitatory effect on the brain, especially if the excited neurons are inhibitory.

In vivo studies in rodent's PFC, DOI had a significant net-excitatory effect on most pyramidal neurons studied, however a smaller proportion of L5p neurons were also suppressed via activation of GABAergic interneurons [72]. Moreover, another notable study, has shown that in the rat orbitofrontal cortex and anterior cingulate cortex, a smaller dosage of DOI elicited a significant activation of neuronal populations, but greater doses tended to suppress these regions [73, 74]. It seems reasonable to suggest that depending on the dose, the specific drug administered, and, possibly, on the density of 5-HT_{2A}R in distinct neuronal populations, psychedelics appear to have diverse modulatory effects across the brain's cortical areas [21].

It is evident that 5-HT_{2A}R activation serves as a necessary (if not sufficient) intermediary of the distinctive subjective effects of classic psychedelic substances, but this does not imply that its activation is the only neurochemical cause of all subjective effects. The activation of this serotonergic receptor has also revealed to be implicated in changes in glutamate transmission, as well as in influencing thalamocortical networks and neuroplasticity.

2.1.3 Alterations on cortical glutamate transmission

Glutamate is the most abundant neurotransmitter in our brain. The belief that hallucinogens enhance glutamatergic transmission in the cortex, is an overarching narrative that has been evolving for the past years [71], however, the intricacies of the process by which hallucinogens enhance cortical glutamate following 5-HT_{2A}R activation remain a source of debate. Research based on the suggestion that the presynaptic action of 5-HT involved glutamate release, have concluded that treatment with LSD or DOI increased L5p neuron activity in the PFC and was mediated by an increase in glutamate release and subsequent activation of postsynaptic a-amino-3-hydroxy-5-methylisoxazole-4-propionate (AMPA)

receptors[4] [70, 74–76].

The PFC L5p neurons receive excitatory glutamatergic input from different cortical regions, as well as from thalamic projections, and send output to both the cortex and the thalamus. According to recent research, activation of presynaptic 5-HT_{2A}R on these thalamocortical afferents also contributes to psychedelic-induced glutamatergic transmission regulation in the PFC [70], which has been supported by showing that stimulation of presynaptic 5-HT_{2A}R in thalamocortical synapses by DOI promotes N-methyl-D-aspartate (NMDA)[5] receptor-mediated transmission [78].

Furthermore, L5p neurons are known to couple bottom-up cortico-thalamic and top-down cortico-cortical loops of informational streams [79, 80], suggesting that alterations in these might lead to changes in these loops. Recent electrophysiological and neuroimaging studies of the human brain in its resting state suggest psychedelic-induced alterations in thalamic gating.

2.1.4 Alterations in thalamic gating

Psychedelic drugs are known to affect the neurons comprised in central brain networks responsible for bottom-up sensory input via the thalamus to the cortex and top-down cortico-striato-thalamic, cortico-thalamic and/or cortico-cortical control of information [21]. It has been proposed that hallucinogens disrupt information processing in cortico-striato-thalamo-cortical (CSTC) feedback loops, which are circuits linking information between the basal ganglia, thalamus, and cortex [81]. These feedback loops are known to be involved in memory, learning, and self–nonself discrimination by linking cortically processed exteroceptive perception with internal stimuli such as proprioceptive information [36]. Within this circuitry, the thalamus, highly modulated by serotonergic afferents, is essential in the gating of internal and external sensory and cognitive information flow to the cortex [81, 82], thus, the psychedelic-induced disruption would lead to an inability to screen out, i.e. to “gate”, extraneous stimuli and to selectively focus on significant elements of the environment [81].

Thalamic gating is influenced by glutamatergic cortico-striatal and cortico-thalamic pathways that project to particular and non-specific nuclei of the thalamus, as well as serotonergic and dopaminergic neurons in the raphe and ventral tegmentum, which project to multiple CSTC loop components [83]. Several lines of evidence suggest that disruptions of thalamic gating, such has the one evoked by psychedelics, happens by stimulating 5-HT_{2A}R in various locations of the CSTC loop [81], resulting in a neurotransmitter imbalance [36] and leading to an overload of the feedforward information of the cortex, consequently disrupting cortico-cortical integration of distributed neuronal activity [81, 82].

The sensory information processed by the thalamus would typically cause mediodorsal thalamic projections to fire. Hallucinogens acting directly on these terminals cause glutamate release in the absence of sufficient sensory input. Furthermore, the effects of extracellular glutamate would be amplified since pyramidal cells would now be hyperexcitable. As a result, hallucinogens might dramatically increase the sensitivity/excitability of cortical processing while also stimulating glutamate release from thalamic

[4] AMPA receptors are responsible for the bulk of fast excitatory synaptic transmission throughout the central nervous system and underlie much of the plasticity mechanisms of excitatory transmission that is expressed in the brain [70].

[5] Critical receptors for establishing, maintaining, and modifying glutamatergic synapses [77].

afferents, which generally indicate the processing of incoming sensory information. That is, for incoming sensory inputs from the thalamus, the signal-to-noise ratio in the cortex would be extremely low. Such logic is often compatible with empirical findings that hallucinogens cause highly magnified or distorted incoming sensory inputs [35]. These changes in sensory processing, might be underlying cognitive disturbances and even "ego-dissolution"[6] that are usually experienced during psychedelic states [81, 82]. Additionally, negative symptoms such as emotional and social disengagement might also be the outcome, being seen as efforts to protect the brain against input overload [36].

Several studies on LSD, MDMA [85] and DOI in animals [86], in addition to some other studies regarding the actions of LSD in humans [87–90], have allowed to construct a framework on how psychedelics affect cortico-cortical and cortico-thalamic circuits. Enhancing extracellular glutamate levels in the PFC by stimulating postsynaptic 5-HT_{2A}R on L5p and L6p (projecting to L5p) neurons, as well as presynaptic 5-HT_{2A}R on thalamocortical afferents, has a net excitatory impact on L5p neurons and promotes synaptic plasticity via AMPA and NMDA receptor-dependent pathways, and therefore affecting the thalamo-cortical broadcasting system, and thus consciousness as a whole, by simultaneously producing sensory "flooding" due to reduced thalamic gating of interoceptive and enteroceptive inputs, and by altering the meaning of percepts due to disrupted cortical–cortical interactions [21]. Neuroimaging studies, as well as studies studying startle response and prepulse inhibition in humans[7], have been supporting this framework by investigating the functional and effective connectivity of major connector hubs of the CSTC model. A phenomenon know as hyperfrontality, which describes an increased cerebral glucose metabolism in the PFC region, specifically in prefrontal and tempomedial areas, that has been identified in the previous mentioned studies, has also been shown to correlate with altered thalamic-cortical circuitry, under the influence of psychedelics, and recognized to be underlying symptoms of the psychedelic state such as hallucinations and distorted perception, such as "visionary restructuralization" and "auditory alterations" [21, 34].

Alterations in thalamic gating, however, may not be a specific signature of the psychedelic state, since functional alterations in the organization of these loops have also been seen in psychotic disorders like schizophrenia [81]. Nevertheless, these alterations may be a significant part of the observed subjective effects of these substances, which, in combination with alterations in the functional architecture and connectivity of the cortex, which is explored in Section 2.2, leads to psychedelic experiences. Further research is needed to understand if psychedelic experience characteristics are attributable to a disruption of more particular thalamo-cortical projections and cortico-cortical interactions [21].

2.1.5 Neuroplasticity

A confluence of activity in a diversity of neuronal circuits scattered across the brain ultimately controls behavior. Circuits that drive disruptive behaviour are potentiated in disease states, whereas circuits that drive more constructive behaviors are downregulated [91]. It is not by chance that disorders such as de-

[6] Experience of a compromised sense of "self" [84].

[7] Prepulse inhibition, happens when a weak prepulse stimulus comes before the startle stimulus, and is the suppression of a startle reflex response, a reflectory reaction, to a startle stimulus.It is used to evaluate sensorimotor gating in a variety of species, including rodents and humans.

pression, PTSD and addiction are all associated to imbalances in similar brain circuits [92–94] and have a high comorbidity rate [95]. Adding to this, the neurotrophic hypothesis of depression proposes that the loss of trophic support (provided by trophic factors, i.e. protein molecules that support cell survival) in brain regions such as the PFC and the hippocampus, causes atrophy, such as the retraction of dendrites and loss of dendritic spines and synapses, fundamental components of neurons, which crucially affects mood-regulating circuits, leading to the behavioral characteristics of the disease [92, 94]. Moreover, PFC atrophy has been shown to culminate in an inability to weaken and/or strengthen pathologic and beneficial circuits, respectively [92, 94, 96, 97]. Compounds capable of promoting structural and functional neural plasticity, promoting the reorganization of neural circuits to produce positive behavior in the PFC can potentially counteract these structural changes, such as neurite retraction, loss of dendritic spines, and synapses elimination, and therefore be a promising solution for these type of disorders [98–100].

New pathways for psychobiological therapy

Psychoplastogens are a novel family of fast-acting medicines that have been shown to promote structural and functional neural plasticity in the brain. Psychedelics, ketamine, and numerous other recently identified fast-acting antidepressants are examples of psychoplastogenic substances. Their application in psychiatry signifies a paradigm shift in existent approaches to treating brain illnesses, as a greater emphasis is placed on attaining targeted regulation of neuronal circuits rather than correcting "chemical imbalances" [91].

Psychedelics have proven to be a great promise in this field. Changes in mood [101] and brain function [10] have been known to occur after the initial effects of classic serotonergic psychedelics. Research has shown that serotonergic psychedelics have rapid and long-lasting antidepressant and anxiolytic effects in clinical trials after a single dose [10, 102, 103] including trials in treatment-resistant patients [10, 39, 104]. N,N-DMT, LSD, DOI and psilocybin were shown to be capable of boosting neuritogenesis and all have induced similar effects, such as increase dendritic arbor complexity and dendritic spine density in the PFC and hippocampus [25].

Neurotrophism

Studies have revealed that psychedelics raising glutamate levels in the brain [35] boosts brain derived neurotrophic factor (BDNF) expression, a key molecule involved in brain plasticity changes related to learning and memory, which encourages growth and differentiation of new neurons and synapses [105], as well as immediate-early genes linked with plasticity gene expression in vivo [106, 107].

The significance of BDNF in neuritogenesis and spinogenesis is well understood [108], therefore, understanding how BDNF signaling pathways play a role in the plasticity-promoting effects of classic psychedelics is of great interest. BDNF's high-affinity receptor tropomyosin receptor kinase B (TrkB) activation is known to enhance signaling via mTOR (the mammalian target of rapamycin) [109], which is important for structural plasticity [110], the production of synaptogenesis-related proteins [111], and already shown to be implicated in psychedelic effects [112]. The mTOR inhibitor rapamycin was demon-

strated to prevent psychedelic-induced neuritogenesis, indicating that mTOR activation is also involved in the plasticity-promoting effects of classic serotonergic psychedelics. Besides this, the discovery that LSD and DOI enhance glutamate [113] and BDNF [114] levels in the rat cortex has led to the idea that psychedelics can also improve neuroplasticity by increasing AMPA receptor activation [115], which itself has been demonstrated to stimulate the release of BDNF, in both animals and humans [116].

Given that N,N-DMT, LSD, and DOI, as well as ketamine, all seem to stimulate dendritic branching and dendritic spine formation. It may be possible to conclude that psychedelics' therapeutic effects are at least partially mediated by neural network reconfiguration, since the addition of new synapses to the neural circuitry is represented by the creation of new dendritic spines [117]. The new synaptic connections will survive or vanish in response to activity, and their existence or absence will influence the activity patterns of the neurons on which they sit [118]. Some new synapses may eventually outcompete old aberrant synapses, and the neural circuit may stop the abnormal firing that underpins mental diseases as a result of such reciprocal structural and functional alterations. It is hypothesized that the durability of clinical improvements might be explained by this physical shift in neural circuit connectivity [22].

Despite this, more clinical study is needed to investigate if the neuroplastic effects of psychedelics found in animal studies can be repeated in humans and are responsible for the long-term symptom reductions [21]. The debate about whether or not the ingestion of the psychedelics alone is responsible for not only the clinical outcomes, but also for the acute subjective experience in the psychedelic treatment remains [119–121].

Critical periods

An interesting recent perspective on neuroplasticity in the context of the psychedelic-assisted psychotherapy (PAP), proposes that advances in psychiatric treatment may result from understanding the implicated interventions that "release the brakes that retard" adult neuroplasticity [122], which produce the heightened sensitivity to the environment found during particular times of earlier development.

It is clear that "neuroplasticity" is a broad word that encompasses a wide range of phenomena. One specific type of neuroplasticity that should be taken into consideration in the context of PAP is critical period plasticity (CPP) [119]. A critical period is a period of time during which environmental input is required for the proper development of a brain circuit. During a critical period, the brain's plasticity is increased, and experiences have a strong impact on developing stable neurocircuitry. The brain's malleability generates both a sensitivity to environmental shocks or deprivations as well as a remarkable ability to swiftly and robustly learn abilities throughout this developmental phase. Neuronal alterations are still conceivable when crucial periods close, although they are more limited [119]. CPP represents periods in which the brain is the most receptive to external stimuli, more prone to be impacted in a lifetime manner [123, 124] and to develop sensitive stages of higher-order functioning, such as attachment, emotion regulation, and social cognition [125].

The fact that during well-defined temporal windows of opportunity, targeted enrichment in developmental domains is most successful, contributes to the belief that in a CPP framework, psychedelics' therapeutic mechanism could be understood as the pharmacological properties of psychedelics putting

the brain in a critical period “open state”, while the psychotherapeutic aspect might retrieve appropriate engrams, such as traumatic memories [25], and whose clinical and effectiveness outcomes have been linked to the type and level of psychological support provided during the event [126].

CPP may be the starting point in future work to a different level of observation missing from psychedelic research. Studies in rodents have corroborated that psychedelics might reopen a psychosocial CPP [127], but more research is needed. Adding to this, 5-HT_{2A}R might be a bridge to connect psychedelics and CPP, since, as previously mentioned, the synaptic plasticity have been found to be dependent of 5-HT_{2A}R signaling [21, 25, 91], and was recently found to be related in key development periods [40]. However, there is now competing evidence that psychedelic induced plasticity might be independent of 5-HT_{2A}R [128]. More studies are required to fully assess the contribution of neuroplasticity and CPP re-opening to the mechanism of PAP, and understand how the the adult’s brain can approximate to a child’s one during the psychedelic experience, leading to one’s opening to the surrounding environment and learning and exploration attitude towards it [119].

2.1.6 Psychological and clinical implications

Unlike every other central nervous system drug class, where the action is usually predictable regardless of circumstance, the effects of classic psychedelics are strongly reliant on the user’s expectations (referred as the *set*) and the context (referred as the *setting*) in which the usage occurs, the psychedelic experience can therefore be interpreted as a subjective experience. While the *set* comprises factors such as personality, previous experiences and the pre-dose mood, the *setting* is defined by the session environment and the external stimuli presented during the session, for instance, the light and music. PAP sessions are oriented by professionals and always built taking into consideration all these determining factors for the therapy’s success [20].

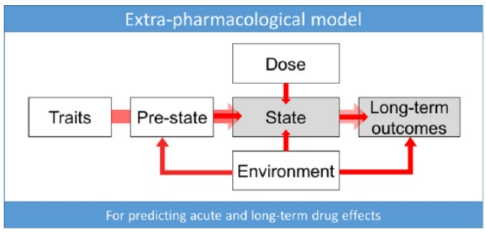


Figure 2.2: **Extra-pharmacological model.** Extra-pharmacological factors which can influence the course of the psychedelic experience. Traits are characteristic of the subject and might not only be biological, but also psychological, such as personality. Pre-state, also referred to as *set*, refers to the previous moments regarding the experience, including management of the anticipatory anxiety, of expectations and the mind set pre-experience. State refers to the quality of the psychedelic experience, which might be measured using neuroimaging tools and via subjective rating scales. Dose means the drug dosage, which might strongly influence not only the moment of the experience, but also the long-term outcomes. The environment or *setting* is dictated by the environment variables, such as the light, music and room decoration. Finally, the long-term outcomes translate the psychiatric condition symptoms of the subject, assessed through rating-scales, and also factors such as personality and perspective [20]. Adapted from Carhart-Harris (2017) [40].

Evaluating the impact of psychedelics on the human experience

Despite their chemical differences, classic psychedelics produce strikingly similar subjective effects [60]. Some examples of the cognitive, perceptual, emotional, and social relatedness effects of the psychedelics, as well as their primary pharmacological mechanisms of action, are provided in Figure 2.3.

Class and Compound	Primary Mechanism of Action	Effects					Other Compounds
		Cognition	Perception	Negative Emotions	Positive Emotions	Social Relatedness	
Classic psychedelics							
LSD, psilocybin, and ayahuasca (DMT)	Serotonin 5-HT_{2A} and 5-HT_{2C} receptor agonist	Increased cognitive flexibility (53), creative thinking (51), and insightfulness (52); distractibility and disorganized behavior (49, 51, 53, 62)	Changes in visual perception (51, 53); mystical experiences (6, 12, 34, 52); paranoia (53); hallucinations, depersonalization, derealization (51, 62, 69)	Anxiety (29, 51, 69); labile mood with anxiety (34)	Increase in well-being and life satisfaction (70); positive mood (60, 71) or blissful state (52, 53, 69)	Enhanced empathy (50); prosocial attitudes and behaviors (34); openness and trust (69)	Mescaline
Entactogens							
MDMA	Serotonin 5-HT_{2A} agonist; mixed serotonin, norepinephrine, and dopamine reuptake inhibition and release	Deficits in spatial memory (111); mild impairment on psychomotor tasks (92)	Changes in body perception, slight visual and auditory alterations, no hallucinations (92)	Distrust and hostility (103); anxiety (93, 101, 103, 105)	Increased trust and sense of a greater meaning in life (100); euphoria (92, 103) and well-being (92)	Increased connectedness toward others (91, 99, 102); increased empathy (96, 100, 103)	MDA, MDEA
Dissociative anesthetics							
Ketamine	NMDA antagonist	Deficits in vigilance, verbal fluency, delayed recall, and tests of frontal lobe function (121)	Derealization, depersonalization (8, 120, 121, 124); illusions in all sensory domains and perceptual alterations (121)	Amotivation, emotional dulling, hostility (121); anxiety (121, 123)	Improved mood (7, 8, 120, 123)	Emotional withdrawal (121)	Dextromethorphan, phencyclidine (PCP), and nitrous oxide

Figure 2.3: Primary pharmacological mechanisms of action of the psychedelic compounds and their cognitive, perceptual, emotional, and social relatedness effects. Adapted from Reiff (2020) [4] (and references therein).

Several research have used psychometrically validated questionnaires to compare the effects of hallucinogens and other drug types, such as the Altered States of Consciousness questionnaire (APZ), an instrument that has been widely used to assess the subjective response to hallucinogens, and other variations of the same such as APZ-OAV and the 5-Dimension Altered States of Consciousness (5D-ASC).

The common core of drug-induced ASC, according to Dittrich [50], may be defined by three dimensions of the APZ and APZ-OAV questionnaire [60]: Oceanic Boundlessness (OB), which reflects a pleasant state of positive depersonalization and derealization, a positive mood, a mania-like experience and an altered sense of time; Anxious Ego Dissolution (AED) that measures dysphoric effects, like "ego-dissolution", delusions, loss of self control and anxiety; Visionary Restructuralization (VR), which represents the existence of elementary hallucinations and pseudohallucinations, synesthesia, changed meaning of percepts, facilitated recollection and imagination.

Most classic psychedelics have shown to increase OB, AED and VR scores significantly [60]. People going through PAP clinical studies usually report alteration in perceptual (visual and auditory hallucinations), emotional (intensified feelings, euphoria and an increase in consciousness of their emotions) and cognitive (thought disorder, increased creativity) domains. Usual reports of more strong experiences, also called *peak* experiences, include pleasant feelings of ego-disintegration. However, ego disintegration may also be associated with negative feelings associated with loss of autonomy, self-control, and thought disorder [36]. This evidences that psychedelics impact on fundamental aspects of the experienced sense of self [129], often conceptualized as a loosening of self-boundaries, oneness, unity or

ego-dissolution [130]. Increased both positive and negative mood, emotional excitation and sensitivity, leading to emotional breakthroughs (overcoming challenging emotions or memories and thus the experience of emotional release) [53, 89, 130, 131] are also typically reported.

Overall, PAP clinical studies have been showing that psychedelics have been proven to function quickly and have long-lasting effects after only a few sessions/doses in people with psychological disorders, such as treatment-resistant depression, anxiety, addiction, PTSD or obsessive compulsive disorder (OCD), which exhibit a negative cognitive bias, characterized by pessimism, poor cognitive flexibility, rigid thought patterns and negative fixations regarding "self" and the future [28]. The administration of psychedelics in a controlled *setting*, and after careful attention to the *set*, has shown to reduce people's symptoms, making them potentially useful therapeutic agents and a revolutionary therapy model in psychiatry [10, 23, 28, 39, 132]. These substances show significant potential for reducing depressive symptoms at various time points (1 day, 1 week, 3 weeks, or even 6 months) after the therapy sessions [3, 133], even though at time points further away from the session the results are less conclusive and the percentage of patients in symptom remission decreases [133]. Psychedelic therapy also has limitations and there is a need to be cautious when generalizing findings, highlighting the need for more rigorous and controlled research [3, 133, 134]. Multiple confounders and biases have been identified in psychedelic trials, including difficulty in blinding, patient biases and expectancy, highly selected patient populations, and exclusion of patients with known risk factors. Besides this, the actual processes underpinning any therapeutic benefits of psychedelics remain mysterious, and promising clinical outcomes must be repeated in studies with larger cohorts [2–4, 133]. In particular, it is unclear whether these observed therapeutic benefits stem from the direct impact of psychedelics on brain activity or from the cognitive and psychological consequences of being in a different state of awareness. To put it another way, it is still uncertain if conscious awareness of psychedelic-induced subjective experiences is required for therapeutic success [21].

Despite these limitations, the potential benefits of psychedelic drugs in treating mental health disorders cannot be ignored. Unlike traditional antidepressant drugs, which often require weeks or months to take effect, psychedelic drugs can produce rapid and long-lasting improvements in mood and behavior after just one or a few doses [2, 133, 134]. Moreover, potential side-effects of psychedelics are well-tolerated and generally occur immediately following treatment [4]. In addition, PAP provides a unique therapeutic experience that can help patients gain new insights and perspectives on their lives and their mental health [3]. Future research should focus on addressing the limitations of previous studies, including the need for more rigorous methodology and controlled settings, more diverse and representative patient populations, more studies focusing on the abuse potential of psychedelics, and a better understanding of the mechanisms underlying the therapeutic effects of these substances [2–4, 133, 134]. With continued research and development, psychedelic drugs might have the potential to revolutionize the treatment of mental health disorders and provide new hope for patients who have not found relief with traditional therapies .

2.1.7 Summary

Classic psychedelics have been shown to stimulate 5-HT_{2A}R thus modulating brain's neuroplasticity and the way in which the cortex processes information. However, rodent studies, despite unraveling the molecular mechanisms by which psychedelic drugs seem to act, have little impact in understanding what are the brain circuit phenomena underlying psychedelic subjective effects in humans reported in PAP. The use of modern neuroimaging technology may be a potential starting point to help unravel how these compounds work at a higher brain level and bring to light their emerging potential in therapeutic efficacy.

2.2 Alterations in whole-brain functional organization

The neuromodulatory effects of psychedelics explored in the previous sections manifest changes in brain connectivity, which can be measured with neuroimaging techniques such as functional magnetic resonance imaging (fMRI). In the past years, numerous different neuroimaging approaches have been used for the measurement of brain area's connectivity [14, 28–30], showing relevant alterations in whole-brain functional organization and dynamics under the influence of psychedelics. Different analytic approaches reveal distinct aspects of the whole-brain cross-regional communication patterns. For example, functional connectivity approaches provide correlations between signals of different brain regions, while effective connectivity techniques, such as dynamical causal modelling, facilitate the inference of the influence that a neural region exerts over another [135].

Several neuroimaging and electrophysiological studies of the human brain in its resting state have revealed alterations in its normal functioning. Psychedelic-induced system-level alterations have spawned several of the hypotheses about the neural foundations of psychedelic states [21], such as the thalamic gating hypothesis explored in Section 2.1.4. Other studies indicate a reduced functional segmentation of large-scale brain networks and increasing global functional connectivity, shifting the brain towards a more global functional integration [14, 30, 54, 136] and to a more entropic state [27, 137].

2.2.1 The wake state and the psychedelic state: primary and secondary consciousness

According to the ideas of Sigmund Freud, there exist two distinct modes of cognition, namely the primary and secondary processes [138]. He described a form of cognition defined by a primitive, animistic style of thinking in some non-ordinary situations such as dreaming and psychosis, in which the flow of "neural energy" is generally "free", and dubbed it the *primary process*. Similarly, Freud observed the lack of specific functions in non-ordinary states that are typically present in waking cognition. He attributed these duties to the ego that gives rise to the *secondary process* of the mind [138].

It is proposed that the distinguishing feature of *primary states* is high entropy (uncertainty) in specific elements of brain function, entropy that is suppressed in the normal waking consciousness, conferring it characteristics associated with meta cognitive functions, such as the capacity for the formation of a

mature ego, self-reflection, theory-of-mind and mental time-travel [139]. In this line of thought, according to Karl Friston's free-energy principle [140] the mind is believed to have developed (by secondary consciousness sustained by the ego) to analyze the environment as accurately as possible by finessing its representations of the world such that surprise and uncertainty (i.e., entropy) are reduced, process that depends on the brain's capacity to organize itself into coherent, hierarchically structured systems [140, 141].

One example of a considered primal or fundamental state of consciousness that preceded the emergence of contemporary, adult, human, normal waking consciousness, is the psychedelic state [138]. Primary states, such as the psychedelic state, may exhibit "criticality", which is the property of being poised at a "critical" point in a transition zone between order and disorder state. It is believed that the brain can explore the widest range of its potential dynamical states in this critical zone [27]. Therefore, it has been suggested that the psychedelic state is fundamentally distinct from healthy adult humans' typical waking consciousness [27].

A number of large-scale intrinsic brain networks have been discovered in fMRI research during unconstrained "resting" states (usually lying quietly with eyes closed or fixating on a cross) [31, 142]. The so-called default-mode network (DMN) is a brain network that is particularly significant in the context of our work. It is thought that to reach primary states, the organization of this network must collapse and, additionally, there must be a decoupling between it and the medial temporal lobes, which are usually significantly coupled [27].

2.2.2 The Default Mode Network

The DMN was first proposed in a work by Marcus Raichle (2001) [143], which looked at a pattern of blood flow, glucose metabolism, and oxygen consumption in the resting state that consistently decreased during goal-directed cognition, representing a default mode of brain operation. In other words, the DMN is a high-level distributed system whose activity is inversely connected to activity in cortical regions that support task or stimulus-bound processing, referred as task positive networks (TPNs), whose activity increases during consistent task performance, implying great focus and relative decrease in off-task attentional lapses [144]. It has been shown that the DMN's implicated regions receive higher blood flow [145] and expend more energy [143] than other brain regions, supporting the known fact that these are densely connected [146] hubs for high-level information integration and routing [147], hosting the highest number of cortico-cortical connections in the brain.

The medial prefrontal cortex (mPFC), posterior cingulate cortex (PCC) and medial temporal lobes (MTL) are some of the implicated brain regions in the DMN [144, 148]. The implication of MTL is of special importance since this region comprises key structures , such as the hippocampus, the amygdala, the parahippocampal gyrus and the entorhinal cortex, which play an important role in memory and emotional processing [138].

A coherent sense of "self" or "ego" [138] has been suggested to originate from the development of self-organized activity in the DMN [27]. Furthermore, it has been hypothesized that coupling within

the DMN, especially between the MTL and DMN, is necessary for the sense of self and for secondary consciousness[8], and, consequently, that the decoupling between these networks will result in a predominance of primary consciousness[9] [27]. The fact that DMN resting-state functional connectivity is implicated in introspective thought [149, 150] led to the suggestion that hyper-activity and connectivity in the DMN can be related to a style of concerted introspection, typically seen in depression [138].

L5p neuron cells where 5-HT_{2A}R are located are known to fire with an inherent alpha frequency [151]. These alpha oscillations are thought to be related to temporal framing in perceptual processing [152], but, more intriguingly, a positive relationship has been found between self-reflection and alpha power [153], as well as alpha synchronization during rest [154]. Decreases in alpha power in the PCC following psilocybin administration have been discovered, showing a positive association with assessments of the subjective experience reporting "I experienced a disintegration of my ego". Although alpha frequencies accounted the most variance, scores on this item also linked favorably with decreases in delta, theta, beta, and low gamma power. Therefore, DMN has been related to self-reflective and introspective functions [155], leading to the hypothesis that psychedelics induce the primary state of consciousness by letting go of the ego's customary grip on reality [156], inducing an increase in brain's entropy.

In summary, the DMN is composed of a collection of high-level cortical nodes, which exchange neuronal impulses with sub-cortical systems and other association and poly-modal cortex, particularly those involved in emotional learning and memory. The effects of psychedelics in this network and in the inverse coupling of DMN-TPNs will be discussed in the next section. The main result being the decrease in the activity and connectivity of the DMN consistent with unconstrained and explorative thinking as observed in the psychedelic state.

2.2.3 Resting state brain studies: the effects of psychedelics in whole-brain functional organization

The study of fast changes in brain dynamics and functional connectivity[10] (FC) is of great interest in neuroimaging. It has been hypothesized that the neural correlates of the psychedelic experience can be derived from the dynamics and variability of spontaneous brain connectivity and activity fluctuations, which can be measured using tools such as fMRI, magnetoencephalography (MEG) and electroencephalogram (EEG). Several studies reporting an increase in the integration of sensory and somatomotor brain networks, as well as the disintegration of networks implied in associative brain regions have been linked to LSD and psilocybin administration.

A study using psilocybin and a task-free ASL and BOLD fMRI protocol was performed in humans in order to capture the transition from normal waking consciousness to the psychedelic state [30]. As expected, psilocybin caused substantial alterations in consciousness, showing especially relevant declines in cerebral blood flow (CBF) and BOLD signal in high-level association regions, such as PCC, the

[8] Secondary consciousness is the proeminent mode of cognition in the normal waking cousciousness. It respects reality by attentively observing and gaining knowledge from its interactions [140].

[9] Primary consciousness can be defined as a mode of cognition characterized by the so called "primary states", such as REM sleep, the on-set phase of psychosis, and the psychedelic state [27].

[10] Functional global brain connectivity is a data-driven approach, defined as the statistical relationships between different brain regions' signals over time, measuring connectivity between each voxel and the rest of the brain [21].

anterior cingulate cortex (ACC) and the mPFC, being strongly correlated with the strength of the subjective effects people reported afterwards [30]. Furthermore, a FC analysis showed that psilocybin induced a relevant decrease in the positive coupling between the mPFC and PCC, consequently implying that the felt subjective effects of the drug resulted from the decrease in the activity and connectivity of these brain's key connector hubs, promoting a state of unconstrained cognition [30]. Decreases in gamma power[11], involved in resting-state brain activity, already reported in rat studies after psilocybin infusion, have also been supportive of these results [158, 159].

Notably, the regions which showed the most consistent decrease in activity after psilocybin administration, such as the PCC and mPFC, are the ones that exhibit disproportionately high activity under normal conditions, as it has been previously mentioned [143], having an important role in consciousness and in high-level constructs, such as the "ego" [138]. Findings suggest that reciprocal connectivity between these two association regions is disrupted after psilocybin intake, implying a rebalancing of hierarchical activity in high-level modes [30]. Besides this, activity in [160] and connectivity [161] with mPFC are known to be enhanced in depression and have been shown to return to normal following successful therapy [162]. Psilocybin deactivating the mPFC consistently, as a result of 5-HT_{2A}R activation, has shown improvements in subjective well-being and trait openness, reducing depression ratings, months after an intense experience in psilocybin therapy [163].

Furthermore, another study developed under resting state conditions used three complementary neuroimaging techniques - ASL, BOLD and MEG - to conclude about alterations in brain activity under the influence of LSD [14]. The results revealed alterations in visual brain regions, all strongly correlated with visual hallucination ratings. Additionally, decreased RSFC between the parahippocampus and the retrosplenial cortex were associated with perceptions of "ego-dissolution" and "altered meaning", showing that this circuit is also critical for the preservation of sense of self and its processing of meaning [14]. Besides this, the MEG results showed decreased oscillatory power under LSD in four frequency bands in the PCC, in lower-frequency bands (i.e., 1–30 Hz), making it possible to establish not only a relation between "ego-dissolution" effects and decreased delta and alpha power, but also between simple hallucinations and decreased alpha power [14], also supported by previous studies [30, 164, 165].

Given this, it seems plausible to assume that the action of psilocybin, LSD, and other psychedelics can be underlying the desynchrony and the loss of oscillatory power in higher-level cortical regions, most likely through 5-HT_{2A}R excitation of deep L5p neurons in those same brain regions [14, 30, 164], thus contributing to the reduced stability and integrity of well-established brain networks [30], and, at the same time, causing network breakdown and desegregation by concurrently reducing the degree of separateness or segregation between them [166]. For instance, coupling between the MTL and the other cortical regions of the DMN, showed to be necessary for the maintenance of adult normal waking consciousness, is disrupted leading to desynchronization in DMN's activity [27]. Importantly, these results are consistent with the more general premise that psychedelics cause cortical brain activity to become more "entropic" [27].

Two psilocybin trials in treatment-resistant depression patients [28, 167] demonstrated functional dy-

[11] Gamma power is involved in sensitive drive and in a large range of cognitive phenomena such as attention, learning and working memory [157].

namics changes in DMN, and in other two brain networks impaired in depression, the executive network (EN) and salience network (SN), suggesting that decreased modularity or increased flexibility of these networks following psilocybin therapy might be a key component of its therapeutic mechanism of action [28].

An even more recent FC study administrating psilocybin and LSD, concluded that these might induce alterations in cortical information processing [168], being these results coherent with what we have seen in Section 2.1.4 regarding psychedelic-induced alterations in thalamic gating. Two other neuroimaging investigations demonstrated enhanced thalamic FC following LSD treatment, especially between the thalamus and the sensory and sensory somato-motor cortical areas [54, 169]. Another research studying the CSTC model's primary hubs [170], showed that LSD improves 'bottom-up' thalamo-cortical information flow to some cortical areas, it decreases information flow to others, supporting the concept that psychedelics somehow disturb thalamic gating [170].

Summing up, research has demonstrated alterations in FC between nodes of different intrinsic brain networks after the administration of psychedelics, as well as changes in regional and interregional entropy/complexity [24, 29, 54, 136, 164], even though results have often shown some inconsistencies [29, 171], concluding that more investigation is needed. Reduced functional connectivity in or between regions of the DMN has been the most consistent finding in all of the different studies that have explored FC within the nodes of intrinsic brain networks [14, 30, 54, 136, 164, 172]. Overall, findings suggest that psychedelic administration shifts the brain towards an increased global functional integration, as reflected by an increase in between-network functional connectivity, brain networks that usually show anti-correlation become active simultaneously (act as "one"). Furthermore, an expansion of the overall repertoire of explored functional connectivity motifs (increased the number of possible states to find the brain in) is also observed in the brain under psychedelic drugs, which can be interpreted as an increase in the brain's entropy [14, 30, 54, 136, 164, 171, 172]. This seem to be in consensus with the reported subjective effects during the psychedelic experience, such as altered perceptual, emotional and self processing.

2.2.4 Summary

To date, previous studies' results suggesting decreases in the activity and connectivity in the brain's key connector hubs, proposing a "disintegration" of central brain networks and enabling a state of unconstrained cognition and desynchronized cortical activity, could be the kernel for brain's approximation to criticality, potentially dismantling reinforced patterns of negative thought and behavior by breaking down the stable spatio-temporal patterns of brain activity upon which they rest, when in the psychedelic state [5, 27]. The next section will address how such psychedelic neural correlates and psychological findings can be unified in a computational theoretical framework substantiated on hierarchical predictive coding and the free energy principle [5, 27, 140].

2.3 Computational-level psychedelic action theories

Recent hypotheses in cognitive neuroscience postulate that psychedelics interfere with the integrity of neurobiological information-processing systems in order to produce their effects. An unifying framework has been developed in order to formulate the action of psychedelics by integrating theoretical frameworks such as the entropic brain, the free-energy principle and predictive processing [5]. Relaxed beliefs under psychedelics (REBUS) and the anarchic brain is a unifying model, based on the principle that psychedelics affect spontaneous cortical-activity, leading to the relaxation of the precision weighting of one's high-level priors, i.e beliefs, liberating bottom-up information flow and constricting top-down information flow [5].The next sections will cover the important concepts to understand this model, culminating in its explanation and relevance in the therapeutic context.

2.3.1 Free energy principle

The free-energy principle, introduced by Friston in 2007, states that self-organizing systems in balance with their surroundings must decrease their free energy, which is a measure of uncertainty. Biological systems, like animals and brains, resist disorder and minimize entropy to maintain stability. Free energy is a function of sensory states and recognition density, which represents the causes of the sensory input. Suppressing free energy can be achieved by changing sensory input or altering the recognition density by changing internal states. This explanation of systems is based on their innate need to enhance their internal probabilistic models and the sampling of their environments [140]. Hierarchical predictive coding is an important brain theory to understand the free energy principle and REBUS.

2.3.2 Hierarchical Predictive Coding and the Bayesian Brain

Helmholtz's proposal that the brain functions as an inference machine has inspired many theories in neuroscience, including the free energy principle [140], predictive coding [173], and the Bayesian brain [174]. Hierarchical Predictive Coding (HPC) suggest that the brain reduces prediction errors [175] or free energy by using internal hierarchical models to forecast sensory input. Top-down signals from higher-order cortical structures provide a "best guess" about hidden causes. High-level areas try to "explain" lower-level states by blocking lower-level activity until they receive top-down feedback signals that fit the bottom-up evidence [176], however, until they do so, lower levels won't "shut up". Prediction errors signal when expectation and evidence are out of sync and are sent upward to be "explained away" by higher levels of cortical processing [177]. This method of processing is similar to Bayesian inference.

The concept of hierarchy is crucial for the brain to create top-down expectations regarding sensory inputs. As the hierarchy progresses, the complexity of representations also increases (e.g. from sensations, through perceptions, and then to concepts) [138]. The brain associates multiple sets of potential causes with physiological effects and chooses the set that explains them the most effectively [178]. This process is based on internal or generative models that use Bayesian probability theory (further explored

in Section 3.1.3 and 3.1.4), where the brain has a model of the world and improves it through sensory inputs [140]. The prior constraints in this process that allow narrowing down the hypothesis space are called inductive biases or priors, representing one's beliefs about the world [178, 179]. The basic assumption is that the brain has a model of the world (internal models) [180, 181] and that it tries to improve through sensory inputs [140, 176].

The main idea underlying HPC is that the brain is considered an inference engine and a neural generative model, trying to optimize probabilistic representations of sensory input by updating beliefs. This idea has been used to explain various subjective and behavioral occurrences, including visual illusions and psychopathological disorders [182]. The process involves maximizing the internal model of the sensorium and the world over different spatial and temporal scales. The free-energy principle, which involves a bound on surprise, underlies the Bayesian brain hypothesis and can be implemented using various schemes in this field [140].

Psychedelics affect the high-level functionality of the brain's functional structure, including top-down predictive function (as seen in Section 2.2). This suggests that psychedelics dysregulate the brain's highest levels, which can affect its ability to constrain emotion and perception [5]. The REBUS framework combines this idea with the entropic brain hypothesis. [5]

2.3.3 Entropic brain

The entropic brain theory and the free-energy principle both use information theory metrics related to Shannon's entropy. The entropic brain hypothesis, initially theoretical [27], has now been supported by empirical neuroimaging investigations and behavioral complexity/entropy measurements [14, 24, 30], which augment its neurobiological evidence [5]. This theory suggests that the entropy of spontaneous brain activity reflects the richness of subjective experience in any given state of consciousness, characterizing the difference between psychedelic and normal waking states [27, 183].

Psychedelics disrupt brain functions that maintain sub-critical brain dynamics, bringing the brain closer to criticality (see Section 2.2.1), which is related to the hypothesis that a lower-entropy brain state is sustained via the ego. The entropic brain theory proposes that this ego function is characterized by the intrinsic functional connectivity of the DMN and its coupling with the MTLs [27, 138], which maps onto the subjective experience of ego-dissolution induced by psychedelics, as supported by previous neuroimaging studies (see Section 2.2.3).

2.3.4 Relaxed Beliefs Under Psychedelics and the Anarchic brain model

It has been reviewed how psychedelic substances activate 5-HT_{2A}Rs on deep-layer pyramidal neurons in high-level brain regions, leading to dysregulation of high-level components. This is supported by evidence of psychedelics' effect on high-level cellular, oscillatory, and network features, which are important for top-down predictive functions. The REBUS model aims to provide a unified framework for these mechanisms of action in the brain [5].

REBUS posits that one of classic psychedelics' main effects is to decrease the precision weighting

of one's beliefs, i.e high-level priors, and, consequently, their ability to exert hierarchical control over and be resistant to the impact of lower-level brain regions. It is suggested that the precision weighting encoded by the highest levels of the brain's functional hierarchy and their related dynamics become less precise as a result of the dysregulatory effect that psychedelics have in the brain. Precision is equivalent to inverse variance and represents felt confidence. The closer data aligns with a model, the smaller prediction errors, leading to greater confidence. The theory suggests that relaxation of precision is strongest at the highest levels of the brain's architecture, particularly the ones connected to self-hood, identity, or ego [5], and the effect of relaxing high-level priors will have an impact on making these priors less confident, subsequently on the correct functioning of the rest of the hierarchy compromising its structure and integrity [5].

Hierarchical predictive coding suggests that high-level beliefs restrict the rest of the brain's hierarchy, suppressing lower components. Under psychedelics, relaxing these high-level priors enables lower-level prediction errors to impact higher levels (that are normally unable to update beliefs due to the top-down suppressive influence of heavily-weighted priors), resulting in a decrease in top-down control and potential increase in bottom-up information flow from intrinsic systems like the limbic system. The "anarchic brain" concept arises from this hypothesis [5].

It is argued that this simple model can account for the entire range of subjective phenomena associated with psychedelic experiences, including: ego dissolution [84], the unitive and largely synonymous peak experience [184], near-death-like experiences [185], a sense of anxiety and uncertainty [27], heightened suggestibility [186], sensitivity to context [20], emotional lability [39], insight [187], paranoid and delusional thinking [39], psychological age regression and vivid autobio-graphical recollection, recourse to magical thinking [27], altered time perception, a sense of the ineffable [1], entity encounters and sensed presence [185], eyes-closed dreamlike visions, geometric hallucinations [188], and more.

Another way to put things is that the model proposes that psychedelics work by flattening or opening up the brain's energy landscape, making attracting brain states that encode beliefs less stable and influential. This leads to greater freedom for the brain to spontaneously transition between states in an uncertain way, narrowing down the number of usual dominating "attractor" states. This increased randomness and unpredictability results in an increase in brain entropy and synaptic effectiveness and plasticity [25], and also seems to reflect on the subjective experience. Empirical evidence supports this idea by demonstrating a flattened energy landscape under psychedelics [189] and an enhanced repertoire of connection motifs [29]. This altered brain activity can be experienced subjectively as a widened global state of awareness or a sensation of the mind expanding [189]. However, it can also be aversive and disconcerting, as small perturbations to the system can have large repercussions in a flattened energy landscape, as evidenced by related themes of context sensitivity [20, 39]. Furthermore, it is proposed that psychedelics when used in the right therapeutic context, can help to relax pathologically overweighted and aberrant priors associated with mental illness, such as depression or PTSD, which are strongly attached with ruminant thoughts. By revising these beliefs during the psychedelic experience, the revised priors can resonate more harmoniously with suppressed knowledge, resulting in a broader perspective of the inner and outer world [5]. This recalibration of beliefs may have long-term benefits

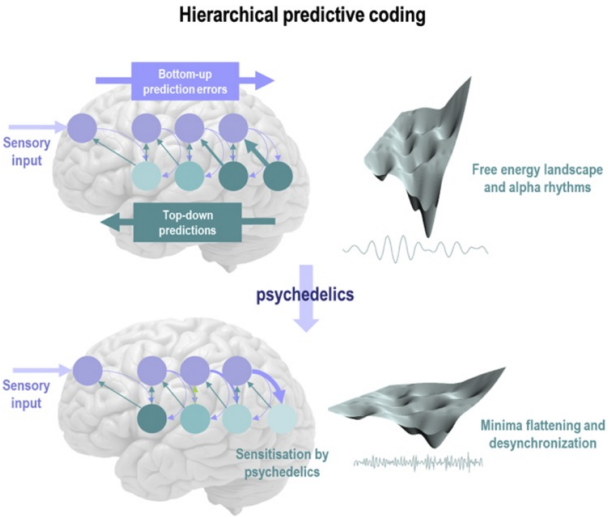


Figure 2.4: **Effect of psychedelics on hierarchical predictive coding.** "Sensory input arrives at the sensory epithelia and is compared with descending predictions. The ensuing prediction error (blue circles; e.g., neuronal populations of superficial pyramidal cells) is then passed forward into hierarchies, to update expectations at higher levels (blue arrows). These posterior expectations (teal circles; e.g., deep pyramidal cells) then generate predictions of the representations in lower levels, via descending predictions (teal arrows). The recurrent neuronal message passing (i.e., neuronal dynamics) tries to minimize the amplitude of prediction errors at each and every level of the hierarchy, thereby providing the best explanation for sensory input at multiple levels of hierarchical abstraction. Crucially, this process depends upon the precision (ascribed importance or salience) afforded to the ascending prediction errors (surprise) and the precision (felt confidence) of posterior beliefs. The basic idea pursued is that psychedelics act preferentially via stimulating 5-HT2ARs on deep pyramidal cells within the visual cortex as well as at higher levels of the cortical hierarchy. Deep-layer pyramidal neurons are thought to encode posterior expectations, priors, or beliefs. The resulting disinhibition or sensitization of these units lightens to precision of higher-level expectations so that (by implication of the model) they are more sensitive to ascending prediction errors (surprise/ascending information), as indicated by the thick blue arrow in the lower panel. Computationally, this process corresponds to reducing the precision of higher-level prior beliefs and an implicit reduction in the curvature of the free-energy landscape that contains neuronal dynamics. Effectively, this can be thought of as a flattening of local minima, enabling neuronal dynamics to escape their basins of attraction and—when in flat minima—express long-range correlations and desynchronized activity". Adapted from Carhart-Harris et al. (2019) [5].

for mental health if handled correctly and may be the fundamental foundation of successful psychedelic therapy, resulting in changes in personality towards increased openness [5, 190].

In summary, REBUS hypothesizes that global brain function can be viewed as entering a mode or state under psychedelics that: (1) features a lightening or relaxation of precision weighting on priors, and (2) allows for a potentially lasting revision of such priors, via the release of prediction error that impacts on the sensitized priors [5]. This framework defends that psychedelics alter system functioning at a level that encodes the precision of priors, beliefs, or assumptions. Subjective effects may be most tangibly felt at the perceptual level, particularly within the visual domain, but as the functioning of higher levels of the global hierarchy becomes significantly disrupted, effects will become more profound, potentially accounting for phenomena such as the dissolution of ego boundaries and potential long-term revision of high-level priors, which might lead to long-term reduced symptoms in the case of people with psychological and psychiatric disorders [5]. Even though at the end of the psychedelic experience the brain reverts to its default mode of efficient free-energy minimization, it might not happen as previously did

[25].

2.3.5 Summary

Within the existing computational theories that have been developed over the years, the REBUS model tries to capture the concepts of some of the best known cognitive neuroscience theories and incorporate them with the evidence resulting from studies at the molecular, circuit and whole-brain functional organization level, as well as clinical studies' evidence regarding the subjective effects reported by people. However, some consider the model to be too bold, and several criticisms have been made, evidencing missing pieces of the puzzle that is the functioning of psychedelic drugs in the brain [191, 192]. REBUS in the context of this work will further be discussed in Section 3.2.2.

This review has tried to capture the big picture of what is the action of psychedelics in the brain. We have looked into the molecular mechanisms, into the whole brain functional organization, focusing on high-level association networks that underlie metacognition and consciousness. We then have established the bridge between what is believed to be happening in brain circuitry and a unifying model that tries to coherently link these alterations to the psychedelic subjective experience. Nevertheless, the understanding of psychedelic action still has a lot of gaps. How can psychedelic drugs produce such a broad diversity of subjective effects? What bridges the pharmacological interactions at neuronal receptors with large-scale changes in the activity of neural populations, changes in brain network connectivity and systems-level of global brain dynamics? Is the underlying cause of the observed clinical improvements after PAP a result of the evoked pharmaco-neurophysiological cascade or the lived subjective experience? All these questions remain to be answered and are, somehow, limitations of the REBUS model.

The next chapter focuses on describing our methodological approach that takes an analogy of the psychedelic action on the brain, focusing on the high-level internal narrative rather than perceptual modulation during the psychedelic experience, and models it as a program induction model.

Chapter 3

Methods

This Chapter describes the concepts, tools, and methods used in our approach to the presented problem. First, the cognitive concept of internal models is described as well as its computational formalization in terms of probabilistic models, then Bayesian inference, and generative models. Second, the Bayesian Program Learning (BPL) framework, which forms the basis for our computational approach, is described. Third, this Chapter culminates in a proposed analogy between our innovative adaptation of BPL and the high-level action of psychedelics in the brain. Finally, a detailed elaboration of our proposed model is presented.

3.1 Internal models within the probabilistic framework

How does a human guide behavior and their decisions? The hypothesis that the nervous system builds predictive models of the physical world is a core focus of cognitive neuroscience. One can interpret the organism's internal models like "small-scale models" of external reality which facilitate the imagination of numerous behavioral options and their consequences in a given environment, and thus conclude which one is the best course of action without having to actually commit any such actions [193]. However, it is not yet completely clear what comprises an internal model. Internal models have been described in different branches of neuroscience research motivated by diverging computational approaches and, moreover, it is believed that the brain maintains a panoply of internal models and that there are a variety of complex interactions between them [193]. In the present study, we will investigate the structure of internal models in the nervous system from a computational cognitive perspective [194]. Therefore, this section will provide a broad view on what an internal model comprises within a probabilistic framework and how this is relevant for the study of our problem.

3.1.1 What is an internal model?

Internal models of one's body or environments are envisioned in a variety of theories that explain how the brain understands, predicts, and influences the outside world. We summarize the representational

components that can be considered part of an internal models and embed them within the probabilistic framework as follows:

- **Prior models**: the statistical structure of the world is far from uniformly distributed and therefore animals and humans learn internal models which comprise prior distributions $p(y)$ over sensory signals y; priors $p(u)$ motor signals u; and priors $p(x)$ over states x of the world.
- **Perceptual inference models**: When sensory input, $p(z|y)$, is provided, a class of internal models known as recognition models compute latent world states z. Generative models, in turn, are models that explain how sensory data is produced. A generative model can be generated from the product of a state prior $p(z)$ and the conditional distribution of sensory inputs given latent world states, $p(y|z)$, or it can be represented by the joint distribution between sensory input and latent variables, $p(y, z)$. In order to determine the probability over the latent states that may have produced the observed input, the generative model can be inverted using Bayes' rule given sensory input (see Section 3.1.3 for a complete contextualized description of Bayesian inference.).
- **Forward dynamical models**: A forward dynamical model is typically thought of as a neural network that can take the current estimated state, x_0, and forecast future states. This may simulate the system's passive dynamics, $p(x|x_0)$, or it could use the motor's current output to forecast the state development, $p(x|x_0, u)$.
- **Cognitive maps, latent structure representations and mental models:** The conditional probability distributions $p(z^n|z^1, ..., z^{n-1})$ of a graphical model can be used to compactly represent abstract relational structures between state variables (potentially related to different objects in the real world) [193].

In Chapter 2, we reviewed how psychedelics act on the brain at multiple levels of resolution. Based on this, we built our computational approach revolving around the question of how our internal projections of the world at a cognitive level can be altered in the psychedelic experience.

In the presented cognitive perspective, internal models will be conceptualized as abstract semantic and knowledge representations. By developing a Bayesian Program Learning (BPL) model applied to such internal representations, this work will focus on what happens to one's internal models, specifically the structure of it's generative and inference phases, in an analogy to the psychedelic experience, attempting to establish some computational intuitions about it. In the next few sections, some important concepts required for contextualizing and understanding of the BPL model used are introduced.

3.1.2 Probabilistic models

The probabilistic framework in machine learning (ML) infers (learns) models to explain observable data, enabling machines to predict future data and make decisions based on those predictions [195]. One can distinguish a probabilistic model from a deterministic one.

Considering features x (independent variables) and a response y (dependent variable). A model with parameters θ denoted by $g_\theta(x)$ is a function predicting the unobserved data based on input features and

can be learned by finding the parameters that minimize an error between the model predictions and the observed data. Considering a data generating distribution F, the model $g_\theta(x)$ can be trained by finding θ to minimize the squared error (for example) taking the expectation over the random variables drawn from F

$$\theta^* = \arg\min_\theta \mathbb{E}_{x,y\sim F}[(g_\theta(x) - y)^2] \tag{3.1}$$

The model prediction $g_\theta(x)$ will ideally be close to the target value of y given the observed features x. This model can be classified as deterministic and can be used in order to make predictions [196]. However, this model does not describe the distribution of the target variable. On the contrary, a probabilistic model expresses a probability distribution. Considering the above described variables, instead of being a deterministic function, a probabilistic model would be a probability distribution $p_\theta(y|x)$ (i.e. a predictive probabilistic model). This model can be trained from data by maximizing the likelihood of the observations,

$$\theta^* = \arg\max_\theta \mathbb{E}_{x,y\sim F}[\log p_\theta(y|x)] \tag{3.2}$$

Following training, the model p_θ would be closer to the conditional distributions of y given the features x in the data generating distributions [196]. Note that a key distinction that plays a fundamental role in probabilistic models is the uncertainty of the target variable estimates. To infer the unobserved quantities knowing the observed data, basic rules of probability theory are applied transforming the prior probability distributions (defined before observing the data) into posterior distributions (after observing the data). This process is known as Bayesian learning [195].

3.1.3 The Bayesian Framework

The idea behind Bayesian modeling is simple, yet powerful. There are two main rules underlying probability theory. The sum rule:

$$P(x) = \sum_{y\epsilon Y} P(x, y) \tag{3.3}$$

and the product rule:

$$P(x, y) = P(x)P(y|x) = P(y)P(x|y) \tag{3.4}$$

Where x and y correspond to observed and uncertain quantities, taking values in some sets X and Y, respectively. $P(x)$ expresses to the probability of x, which can be a statement about the frequency of observing a particular value or a subject belief about it. $P(x, y)$ represents the joint probability of observing x and y, while $P(y|x)$ is the probability of y conditioned on the observation of the value x. The marginal probability of x can be obtained summing the joint over the variable y, or integrating in the case of y being continuous. On the other side, according to the product rule, the joint may be broken down into its component parts, the marginal and the conditional. A consequence of these two laws is the Bayes' rule as a corollary:

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)} = \frac{P(x|y)P(y)}{\sum_{y \epsilon Y} P(x,y)} \tag{3.5}$$

In order to apply Bayesian theory to ML: replacing x with D to represent observed data, y with θ to represent unknown parameters of a model, and conditioning all terms on m, the class of probabilistic models that is being considered. To perform learning, we therefore get:

$$P(\theta|D,m) = \frac{P(D|\theta,m)P(\theta|m)}{P(D|m)} \tag{3.6}$$

$P(D|m)$ expresses the likelihood of the parameters θ in the model m, $P(\theta|m)$ the prior probability of θ and $P(\theta|D,m)$ the posterior of θ given the data D. But what is exactly learning? Learning is the process of converting existing prior knowledge or presumptions about the parameters $P(\theta|m)$ into posterior knowledge about the parameters $P(\theta|D,m)$ via the use of data D. This posterior will now serve as the prior for new data. By only using the sum and product rule to obtain the prediction, a learned model may be used to forecast or predict fresh test data, D_{test} [195, 197–200]:

$$P(D_{test}|D,m) = \int P(D_{test}|\theta,D,m)P(\theta|D,m)d\theta \tag{3.7}$$

3.1.4 Bayesian inference

Modeling requires data, a model, and a probabilistic inference algorithm. We want to understand how the above presented mathematical framework enables optimal decision making under uncertainty. The inference problem involves computing conditional probabilities of variables of interest, given some observed variables, whilst marginalizing out all other variables. Filling in missing data, computing parameter posteriors, and evaluating expectations can all be framed as inference problems [197].

In Bayesian inference each possible value of a latent state variable z is assigned to a probability, the latter representing how strong is one's beliefs that a given z value corresponds to the true state of the world [200]. For example, with respect to the Bayesian brain hypothesis (see Section 3.1.1) suggests that the brain encodes a prior $p(z)$ corresponding to one's believes about the state z, before receiving any sensory information. It is also hypothesized that the brain encodes a probabilistic internal model translating the dependency of the sensory signals y and the latent state z, this is known as the generative model (in the next section we will go further into explaining what a generative model consists of) [8]. This probabilistic model is used to compute the likelihood $p(y|z)$ once the sensory information is received, quantifying the probability of observing the signals (the data) y if a certain state z is true [193]. Then, as it was previously seen, using Bayes' rule it is possible to compute the posterior probability distribution $p(z|y) = \frac{p(y|z)p(z)}{p(y)}$, combining the prior and the likelihood in a statistically optimal manner [193].

Bayesian principles apply to complex cognitive models, allowing normative accounts of of how people generalize from small samples of a variable [201], world's causal structure inference [202], and conceptual regulation of connections between state and sensory variables [203]. However, Bayesian learning

poses computational challenges, since it involves marginalizing all the variables in the model except for the variables of interest, requiring inference approximations such as Markov Chain Monte Carlo (MCMC), variational approximations, expectation propagation, and sequential Monte Carlo [204–207].

3.1.5 Generative Models

Generative models are a class of models that can generate synthetic data points in the input space, learning to closely resemble observed data distribution [208]. Producing data from a probabilistic model provides insight into the model's prior assumptions and learning process [195]. A generative model is defined by specifying a joint probability distribution over all variables (observations and parameters) of a model. Considering the simple case where y is the observation and z the set of latent variables the generative model is given by $P(y, z) = P(y|z)P(z)$ [208]. Generative models have two main goals: finding the ideal causes to represent a specific piece of data D_i, and identifying the optimal model M to describe the entire collection of data $D = D_1, D_2, ..., D_n$. This type of unsupervised learning estimates probabilistic models for the input data and then generates new samples from it; it attempts to learn abstract representations of the input, being the high-level representations a self-organization of the input into "disentangled" concepts and their relationships [209].

The Bayesian brain hypothesis suggests that the cerebral cortex contains a generative model, with one issue being perception/inference and the other being adaptation/learning [210]. Learning involves lossy compression of data to prioritize generalization over retention, and humans are interested in high-level abstract concepts underlying perceived data. Internal models allow us to reason, play with ideas, and imagine hypothetical outcomes [209]. This motivates the next section introducing Bayesian Program Learning model.

3.2 Bayesian Program Learning

Children learn about the world through unsupervised learning, searching for patterns and structure in unlabeled information [209]. Interacting with the world allows them to test their beliefs, update their internal models, and improve their predictions. Embodied cognition through interactive play is likely a crucial component of human intelligence, presenting a significant challenge in approximating ML to human cognitive processes. Studying probabilistic generative models can provide insight into basic cognitive processes [194]. The way people bridge prior experiences to learn new things suggests they build rich causal, compositional, and hierarchical models of the world [6, 7]. Unlike ML algorithms, that usually require tens or hundreds of examples in order to learn and perform with similar accuracy as people do, people can reproduce and explain concepts, parse objects into parts and relations, and create new abstract categories with just a handful of examples [6, 7]. Causality, compositionality, and hierarchy can be explained using a Form-Structure-Data framework, which models human learning by discovering the underlying structure in data [211]. It has been proposed that discoveries regarding structural form can be understood computationally using Hierarchical Bayesian Models (HBMs) which perform probabilistic

inferences about the organizing principles of a data set, with higher levels representing graph structures and lower levels corresponding to observable data. This framework can be extended to multiple levels of abstraction, corresponding to applying a probabilistic model on top of this symbolic structure, i.e developing generative models of generative models [6, 211].

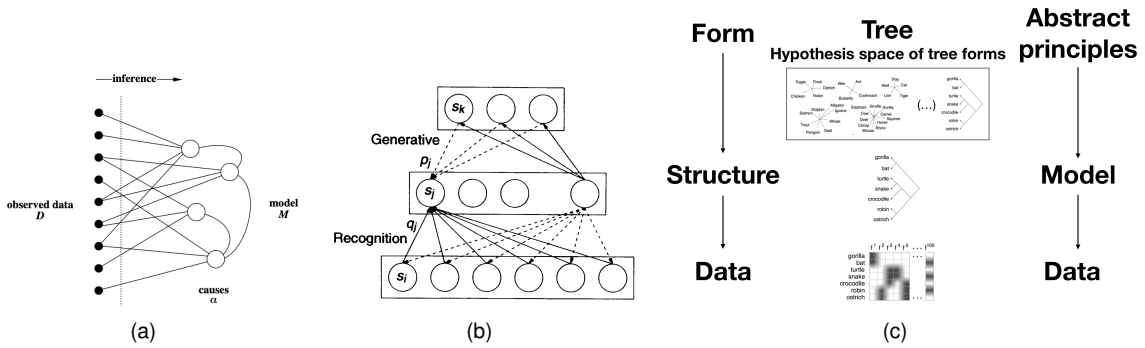


Figure 3.1: **Perception to cognition. (a) A generative model**. M correlates causes α in order to generate new data D. Inference uncovers the latent causes α in the observed data D. Adapted from Olshausen 2006 [210]. **(b) Hierarchical Bayesian model.** A Helmholtz machine is a type of hierarchical Bayesian model. It is a neural network that is designed to learn and generate sensory data by minimizing the difference between the predicted and actual sensory inputs. The Helmholtz machine includes a hierarchical structure, with each layer modeling increasingly abstract features of the sensory input, similar to other hierarchical Bayesian models. It incorporates both a bottom-up recognition network that generates predictions of sensory input, and a top-down generative network that generates sensory data from abstract features. The model is based on the idea that the brain is constantly predicting and updating its expectations about the sensory input it is receiving, and that the best explanation for the sensory input is a combination of these predictions and the actual sensory data. Adapted from Hinton 1995 [8]. **(c) Form-Structure-Data framework.** A hierarchical model where a tree is an example of a structural form, an abstract principal in an hypothesis space of structural forms. A sample from the hypothesis space of *trees* is shown, representing a model/structure that best describes the data. Adapted from Kemp and Tenenbaum 2008 [211]. This figure illustrates the concept of hierarchical models, going from a perception-based (generative/inference (recognition)) modeling, on the left, to a higher level cognitive modeling perspective, describing the process of underlying the form and structure of data, approximating these models to human learning, on the right.

There is a basic common sense understanding that differ people from machines, our view on the brain cannot simply correspond to a pattern-recognition device but as an explanation engine. To computationally understand the mind, we need to consider how knowledge is causally and compositionally constructed and how it changes over time [212]. Probabilistic programs like HBMs, which are hypotheses spaces of hypotheses spaces, priors on priors, capture general beliefs that apply across objects and situations, generating hypotheses and models for specific cases [212]. Higher-level inferences explain how priors guide future learning and can themselves be learned, allowing for abstract knowledge construction and fast inferences about new instances, the ability to perform one-shot learning of new concepts and also learn how to learn [179], representing how humans continue to learn and change their perspectives over time [212].

The Bayesian Program Learning (BPL) framework introduced by Lake et al. in 2015 [6] is a HBM computational model that tries to better capture the above mentioned human learning abilities for a large class of visual concepts - handwritten characters from different alphabets. In the context of this work, we extend these characters to a more abstract level, such as more broad concepts, structured knowledge.

The BPL model learns abstract, rich and flexible representations of concepts (handwritten characters) from just a few examples, by representing them as simple programs that best explain observed examples under a Bayesian criterion [6]. This model incorporates compositionality[1], causality[2], and learning to learn[3] to construct a good representation of new concepts from existing primitive elements. It outperformed numerous deep learning models and achieved human-level performance on a difficult one-shot classification problem [6]. The model represents concepts as simple probabilistic programs, probabilistic generative models that are expressed as structured procedures in an abstract description language [6, 195, 214]. The next section describes BPL as a starting point for this thesis work implementation.

3.2.1 Bayesian Program Learning model

The BPL model learns stochastic programs for creating new character concepts, which can be observed in Figure 3.2. Characters are composed by *strokes* (parts) that themselves are comprised by *sub-strokes* (sub-parts), which are connected by spatial *relations* between them. BPL describes a generative model that is able to sample new character "types" by combining parts and sub-parts in new ways. Creating a new character type corresponds to levels i-iv of Figure 3.2. The character type is itself a procedure for generating new exemplars of the correspondent concept producing new "tokens" of that same concept. This process is illustrated in levels v-vi of Figure 3.2. Hierarchical BPL can then be thought of as a generative model of generative models since it specifies a process for producing concepts, where each one of this concepts is a structured generative model in and of itself. The token-level variables are rendered in the raw data (images) format in the last stage, represented in level vi of Figure 3.2 [6, 7]. Constructing character types involves sampling primitive structures, which are shared and re-utilized across the different characters as *sub-strokes* and *strokes*.

The model's joint distribution, which provide a summary of the uncertainty associated with each one of the variables [208], on types Ψ, a set of M tokens of the corresponding type $\theta^{(1)}, ..., \theta^{(M)}$ and binary images $I^{(1)}, ..., I^{(M)}$ can be written as

$$P(\Psi, \theta^{(1)}, ..., \theta^{(M)}, I^{(1)}, ..., I^{(M)}) = P(\Psi) \prod_{m=1}^{M} P(I^{(m)}|\theta^{(m)}) P(\theta^{(m)}|\Psi) \tag{3.8}$$

Data set and Learning

Training BPL model was performed using the omniglot data set, which comprising 1623 different characters belonging to 50 different alphabets, and 20 different examples of each one of those characters [6]. The alphabets include writing systems from historically significant and currently spoken natural languages (such as Hebrew, Korean, and Greek), as well as imagined writing systems created for television series and video games. The alphabets were transformed into handwritten form using human participants, that were asked to draw at least on alphabet using the computer mouse on Amazon Mechanical

[1] Model captures abilities like conceptual combination and imagination, capturing key notions of parts and relations that may play important roles in perception, learning and organization of concepts [7].

[2] Model's knowledge of the underlying causal process that produces examples of a certain concept category [7].

[3] Having previously learned about parts and relations common to many similar type of concepts (inductive biases [213]), might help the model to construct a good representation of new concepts from existing primitive elements [7].

Turk.

The model was trained using a random split out of omniglot, comprising 30 alphabets (964 characters), referred to as the background set. The background set included the six most common alphabets (determined by Google hits): Latin, Greek, Japanese, Korean, Hebrew and Tagalog. The remaining 20 alphabets from omniglot were used as the evaluation set in a one-shot classification task that we will later explore.

In a method known as learning-to-learn, the model hyperparameters, including the library of primitives (first level i of Figure 3.2), as well as the empirical distributions over the other model variables, were learned when learning related concepts. Learning-to-learn can be thought of as, if given a task, a training experience and a performance measure a computer program is able to learn if with experience it enhances its performance at that task [215, 216].

Training the BPL model involved learning the models of primitives, starting positions, relations, token variability, and image transformations, however, we will only review the learning process for the primitives. For further information regarding other variables in this task, see [6, 7] as a reference.

The data collected from the human participants was composed of a set of time series with $[x, y, time]$

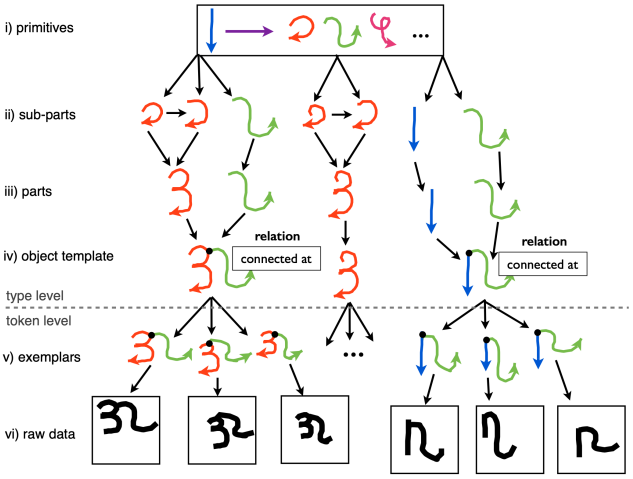


Figure 3.2: **Bayesian Program learning generative model.** Illustration of the generative process underlying handwritten characters. New types are generated by choosing primitive actions from a learned library (i), combining these sub-parts (*sub-strokes*) (ii) to make parts (*strokes*)(iii), and combining parts to define simple programs/character "types" (iv). These programs can generate different tokens, which are different examples of the same concept (v). Exemplars are finally rendered as binary images (vi). Adapted from Lake et al. (2015) [6].

Figure 3.3: Some character examples from the omniglot data set. Adapted from Lake et al. (2015) [6].

coordinates that demonstrate how the artwork was made, as well as how the *strokes* are segmented. Using the drawing data, data was standardized, all pen trajectories were normalized in time to have 50 millisecond sampling interval as approximated by linear interpolation. A pause was indicated when the pen moved less than one pixel between two places. Importantly, the segments that were taken out between pairs of pauses were designated as *sub-strokes*. The result was about $55,000$ *sub-stroke* trajectories. After this, each *sub-stroke* was fit by a spline and represented by its five control points in $\mathbb{R}^{10}$. *Sub-strokes* were divided into 1212 primitive components using a diagonal Gaussian Mixture Model fitted with expectation maximization, and minor mixture components were eliminated. The parameters that define each primitive $z, \mu_z, \sum_z, \alpha_z$ and β_z could then be fit into a maximum likelihood estimation. The transition probabilities between primitives (transition probabilities that define the transitions between primitives when defining the sequence of primitives that is generated in order to create a character type, and, consequently, a character token and a character image) $P(z_{ij}|z_{i(j-1)})$ were estimated by smoothed empirical counts. Finally, the assignment of the *sub-strokes* to the most likely primitive was done, and the regularization was determined using cross-validation by omitting 25% of the characters from the background set as validation data [6].

Generating character types

A character type Ψ is described by: the set of κ *strokes* (parts) $S = S_1, ..., S_\kappa$ and the set of *spatial relations* between them $R = R_1, ..., R_\kappa$. It is possible then to define a character type by $\Psi = \kappa, S, R$. A character type Ψ is then an abstract set of parts, sub-parts and relations that work towards to define the causal structure of the handwritten process of a person. The joint distribution of the character types can be written as

$$P(\Psi) = P(\kappa) \prod_{i=1}^{\kappa} P(S_i) P(R_i | S_1, ..., S_{i-1}) \tag{3.9}$$

The generative process that defines the probability distribution $P(\Psi)$ for a character type is shown in Algorithm 1.

Algorithm 1 Generative process of a character type

procedure GENERATETYPE ▷ Generate a new character type
  $\kappa \leftarrow P(\kappa)$ ▷ Sample the number of strokes
  for $i = 1...\kappa$ **do**
    $n_i \leftarrow P(n_i|\kappa)$ ▷ Sample the number of sub-strokes
    $S_i \leftarrow GENERATESTROKE(i, n_i)$ ▷ Sample stroke
    $\xi_i \leftarrow P(\xi_i)$ ▷ Sample relation to previous strokes
    $R_i \leftarrow P(R_i|\xi_i, S_1, ..., S_{i-1})$ ▷ Sample relation details
  end for
  $\Psi \leftarrow \kappa, R, S$
  return $@GENERATETOKEN(\Psi)$ ▷ Return handle to stochastic program
end procedure

In order to generate a new character type it is also necessary to sample the number of *strokes* κ, sampled from a normal distribution $P(\kappa)$ learned from the empirical frequencies counted during the

training of the model. Conditioned on κ, the number of *sub-strokes* n_i is sampled for each *stroke* $i = 1, .., \kappa$, also from their empirical distributions resultant from measures from the training (background) set. After sampling these hyperparameters, a *stroke* is generated.

Generating Strokes

In the handwritten characters, used in BPL, a *stroke* is considered to be initiated by pressing the pen down and finished when the pen is lifted. A *stroke* is then a motor routine comprising simple movements- the *sub-strokes* - $S_i = s_{i1}, ..., s_{in_i}$. *Sub-strokes* are separated by brief pauses of the pen, without lifting it up [6].

Each one of the *sub-strokes* is modeled by an uniform cubic *b-spline* and can be described by three variables $s_{ij} = z_{ij}, x_{ij}, y_{ij}$. The joint distribution of *strokes* is then

$$P(S_i) = P(z_i) \prod_{j=1}^{n_i} P(x_{ij}|z_{ij}) P(y_{ij}|z_{ij}) \tag{3.10}$$

where n_i is the number of *sub-strokes* and $z_{ij} \epsilon \mathbb{N}$ is a discrete class representing the indexes of each character primitive in the library of primitives, described by the distribution

$$P(z_i) = P(z_{i1}) \prod_{j=2}^{n_i} P(z_{ij}|z_{i(j-1)}) \tag{3.11}$$

representing a first-order Markov Process which has been previously learned from empirical data (see Section 3.2.1).

It is defined that each *sub-stroke* has five control $x_{ij} \epsilon \mathbb{R}^{10}$ points which are sampled from a Gaussian distribution $P(x_{ij}|z_{ij}) = \mathcal{N}(\mu_{z_{ij}}, \sum_{z_{ij}})$ and live in an abstract space which is not embedded in the image frame. Other relevant variable is the type-level scale y_{ij}, which is relative to the image frame and sampled from the distribution $P(y_{ij}|z_{ij}) = Gamma(\alpha_{z_{ij}}, \beta_{z_{ij}})$.

It is possible to conclude that a template for a certain *stroke* S_i is built first by sampling a sequence of discrete primitive actions, learned from the background data (resulting in a sequence of indexed primitives), taking into consideration that the probability of the next sampled primitive depends on the previous one, and then *strokes* are parameterized as splines by sampling the control points and the scale parameters for each one of the *sub-strokes*' primitive [6]. The process of sampling the primitive indexes that comprise a *stroke* will be the one we will further explore in this thesis work. The algorithm to generate a *stroke* is demonstrated in Algorithm 2.

Another important variable regarding character types Ψ are relations R_i which describe how the beginning parts are positioned relatively to the previous sampled parts, and can be of the type *independent* (resulting in parts that are not connected to each other), *along* (part is attached to some coordinate of the previous sampled part), *start* and *end* (part is attached to the start or end, respectively, of the previous sampled part). This model variable will not be explored deeply since it is not the focus of the implementation of this thesis work. For more details about *stroke* relations see references [6, 7].

Algorithm 2 Generate the ith stroke with n_i sub-strokes

procedure GENERATESTROKE(i, n_i) ▷ Generate a character stroke
  $z_{i,1} \leftarrow P(z_{i1})$ ▷ Sample the identity of the first sub-stroke
  for $j = 2...n_i$ **do**
    $z_{ij} \leftarrow P(z_{ij}|z_{i(j-1)})$ ▷ Sample the identities of the other sub-strokes
  end for
  for $j = 1...n_i$ **do**
    $x_{ij} \leftarrow P(x_{ij}|z_{ij})$ ▷ Sample sub-stroke's control points
    $y_{ij} \leftarrow P(y_{ij}|z_{ij})$ ▷ Sample sub-stroke's scale
    $s_{ij} \leftarrow x_{ij}, y_{ij}, z_{ij}$
  end for
  $S_i \leftarrow s_{i1}, ..., s_{in_i}$ ▷ Complete stroke definition
  return S_i
end procedure

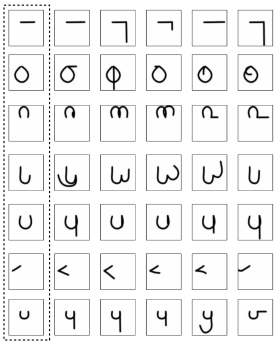


Figure 3.4: **Examples of likely primitive sequences.** The leftmost column shows the start seed, which are individual primitives in image space. The next five columns show the most likely continuations when the next primitive is sampled and added to the character. Adapted from Lake B.(2014) [7].

Generating character tokens

The second level of abstraction of BPL corresponds to the generation of tokens, which emerge by running the simple programs of character types. Character tokens $\theta^{(m)}$ are originated by executing the parts and the relations and modeling the way the ink flows from the pen to the page, resembling the process of handwriting. There are several steps to go through when generating a character token. In order to create a token-level *stroke* trajectory $S^{(m)}$, motor noise is added to the control points and to the scale of the *sub-strokes*, then the precise start location, $L^{(m)}$, is sampled, as well as global transformations including an affine warp, $A^{(m)}$, and adaptive noise parameters that will later facilitate the probabilistic inference. The last step consists of creating a binary image, $I^{(m)}$, by a stochastic rendering function, done by lining the *stroke* trajectories with grey scale ink and interpreting each pixel value as an independent Bernoulli probability distribution [6, 7].

The token-level variables $\theta^{(m)} = \{L^{(m)}, x^{(m)}, y^{(m)}, R^{(m)}, A^{(m)}, \sigma_b^{(m)}, \epsilon^{(m)}\}$, where $x^{(m)}, y^{(m)}, R^{(m)}$ are token-level control points, scale and relations, respectively and $\sigma_b^{(m)}, \epsilon^{(m)}$ are the amount of image's blur and pixel noise, respectively, are described by the distribution

$$P(\theta^{(m)}|\psi) = P(L^{(m)}|\theta_{\setminus L^{(m)}}, \psi) \prod_i P(R_i^{(m)}|R_i) P(y_i^{(m)}|y_i) P(x_i^{(m)}|x_i) P(A^{(m)}, \sigma_b^{(m)}, \epsilon^{(m)}) \tag{3.12}$$

The pseudo-code for the generation of character tokens is given by Algorithm 3.

Algorithm 3 Run the stochastic program of type Ψ to originate an image

procedure GENERATETOKEN(ψ) ▷ Generate a character token
  for $i = 1...\kappa$ **do**
    $R_i^{(m)} \leftarrow R_i$ ▷ Directly copy the type-level relation
    if $\xi_i^{(m)} =$ "$along''$ **then**
      $\tau_i \leftarrow P(\tau_i^{(m)}|\tau_i)$ ▷ Add variability to the attachment along the spline
    end if
    $L_i^{(m)} \leftarrow P(L_i^{(m)}|R_i^{(m)}, T_1^{(m)}, ..., T_{i-1}^{(m)})$ ▷ Sample stroke´s starting location
    for $j = 1...n_i$ **do**
      $x_{ij}^{(m)} \leftarrow P(x_{ij}^{(m)}|x_{ij})$ ▷ Add variability to control points
      $y_{ij}^{(m)} \leftarrow P(y_{ij}^{(m)}|y_{ij})$ ▷ Add variability to the sub-stroke scale
    end for
    $T_i^{(m)} \leftarrow f(L_i^{(m)}, x_i^{(m)}, y_i^{(m)})$ ▷ Compose a stroke's pen trajectory
  end for
  $A^{(m)} \leftarrow P(A^{(m)})$ ▷ Sample global image transformation
  $\epsilon^{(m)} \leftarrow P(\epsilon^{(m)})$ ▷ Sample the amount of pixel noise
  $\sigma_b^{(m)} \leftarrow P(\sigma_b^{(m)})$ ▷ Sample the amount of blur
  $I^{(m)} \leftarrow P(I^{(m)}|T^{(m)}, A^{(m)}, \sigma_b^{(m)}, \epsilon^{(m)})$ ▷ Render and sample the binary image
  return $I^{(m)}$
end procedure

The exploration of *stroke* relations R_i, *stroke* trajectories $T_i^{(m)}$ and image transformations $A^{(m)}$ goes beyond the scope of this work. See references [6, 7] for more information.

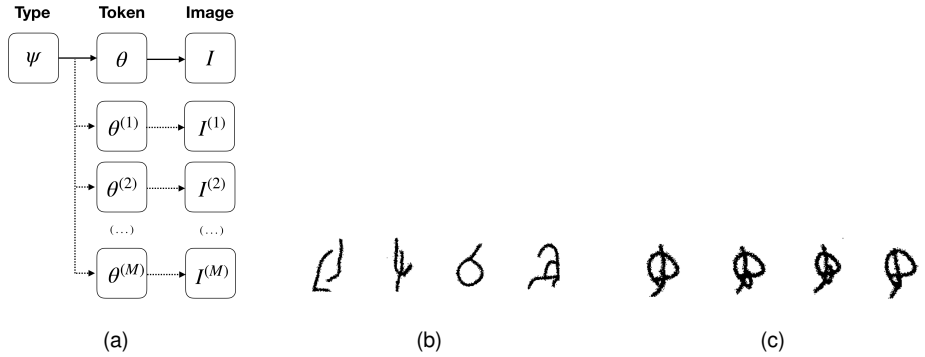


Figure 3.5: **BPL hierarchical generative model. (a)** Generative process. **(b) Character images.** Four examples of character images of different types deriving from the BPL generative process. **(c) Character tokens.** Four examples of character images of the same type but different tokens deriving from the BPL generative process.

3.2.2 Why and How? A Psychedelic Analogy

A state-of-the-art cognitive model based on probabilistic program induction principles known as the BPL model was described. This leads to a critical juncture questioning the following: "Why is this model a suitable computational analogy for characterizing the cognitive implications of psychedelic action on the brain?". This section is devoted to addressing this question.

The structure of thought is determined by the inferred relationships that underlie the connections between a set of entities. Entities and their connections may be, for example, physical objects and their

interactive properties (see Figure 3.1c), or people and their idiosyncratic relationships. Recognizing these underlying connections supports an essential cognitive computation, namely it potentiates rapid inferences when facing new, and unseen situations [201, 203]. That is, new entities are interpreted through the lens of previously learned internal models of the external world.

The unavoidable reliance on subjective reports is one of many challenges faced in understanding what happens to a person's internal models during the psychedelic experience. Despite this, such reports suggest modifications to the underlying structure of thought which apparently have a positive impact on people with mental health issues [4, 133] (see Section 2.1.6). People with disorders such as depression, anxiety, PTSD and OCD, were seen to have significant symptom improvements, as well as reporting "new perspectives on life" and "new ways of seeing things", after having a psychedelic experience in the therapy context [1], seemingly suggesting that something is happening in the domain of people's beliefs. Moreover, in Section 2.3.4, REBUS computational theory was described [5], suggesting that during the psychedelic experience people are less restricted from their a priory beliefs, being more susceptible to access mental states that in their normal waking state they would not access. This phenomenon leads us to hypothesize that previously held rigid beliefs about the structure of the world become more flexible i.e., the relationships that held together a specific set of entities in a given model of the world might become less strict, more interconnected with other entities and, from these interactions, completely new objects and entities might be inferred. Together, this motivates the question "Can psychedelic experience be a door to formulating new concepts and perspectives, resulting in new priors and have a permanent influence on how people do inference on the world?". We try to gain some intuition about this by leveraging a computational model.

The BPL framework was chosen to study this, since it makes an effort to approximate the learning and subsequent generation and inference processes of a human, being capable of generalize in ways that are mainly indistinguishable from people [6]. Besides this, the model primitives can be interpreted as concepts that themselves form new concepts (characters). The present study aims to integrate the aforementioned analogy into a machine learning pipeline, with the objective of harnessing the potential benefits of data augmentation via diffusive perturbations in BPL performance. The proposed approach seeks to enhance the generalizability and robustness of the BPL model by incorporating perturbations in its probabilistic program generative model, thereby not only simulating real-world variability, but also the psychedelic experience, attempting to increase the diversity of the training data. Through this method, we aim to improve the model's generalization performance on a one-shot classification task. Four main phases were therefore defined within this thesis' pipeline:

- **Perturbation phase:** One can think of the psychedelic action in the brain as the "perturbation" of people's priors. The perturbation phase will correspond to perturbing the high hierarchy level of the BPL's generative model, altering its prior knowledge. The hypothesis surrounds the idea that this perturbation will lead to a change on how new character concepts will be generated, i.e, on which primitive connections will be sampled when originating new concepts, comparing to the original data set. The perturbed model will be designated as "Diffusive latents for Bayesian Program Learning" (DL-BPL).

- **Generative phase:** The generative phase will correspond to the generation of a new data set leveraging the perturbed model. The new perturbed generative DL-BPL model will be used to generate a new ensemble of perturbed characters/images organized in the same way as omniglot, in groups of different alphabets. The generative phase can be understood as the simulation of one's experience under the influence of psychedelic drugs, representing the concepts/ideas/thoughts that one's brain generates.

- **Inference phase:** After generating a new data set, it will be used to augment the omniglot data set. The inference phase corresponds to invert the generative model and perform inference on the character images of the augmented data set. The inferred latent variables will then be used to update the DL-BPL model priors. Applying this to the psychedelic realm, the inference phase will correspond to the constructed perceptions following the psychedelic experience, and the consequent experience consolidation arising from it (prior's revision/update). This step can be thought of as perceptual inference.

- **Classification phase:** As a last step, a one-shot classification task will be performed on a validation data set, in order to evaluate the DL-BPL's performance. We hope this phase will bring elucidation to a few questions such as how does inference performance of the original model differ from the perturbed model inference? How does perturbing the highest hierarchy level of the model influence its classification performance when faced with something it has never seen? These questions can be projected into the plan of our study: How can a person's high-level priors be influenced after the psychedelic experience? How different is the person's generalization about the world and about life?

Psychotherapy and its possible benefits for mental health patients is one of the major motivations behind the investment and research that has existed around psychedelic drugs. Our hypothesis resides in the thought: a better generalization of the model may correspond to less rigid high-level priors, to a decreasing in the weighting of people's high-level priors, representing a wider perspective on the world, resulting from formulation of new perspectives and revision of older ones, all of this allowing a recovery of the symptomatology that is characteristic of this people. Figure 3.6 describes the pipeline.

Comparing REBUS and DL-BPL

Here we discuss the similarities and differences between REBUS and DL-BPL. REBUS is grounded in the hierarchical predictive coding theory, which is based on Bayesian learning and perceptual theory, suggesting that psychedelic substances induce a state of heightened flexibility, relaxing higher-level priors and increase sensitivity to bottom-up information. This state could provide a window of opportunity for patients to modify rigid behaviors and thoughts during PAP. In contrast, DL-BPL aims to navigate towards a higher-level cognitive theory of the psychedelic experience. Our work focuses on perturbing/altering cognitive structures, leading to the generation of new programs or entities that can generate new data under the same ”umbrella” structure, modeling this cognitive mechanism in terms of program

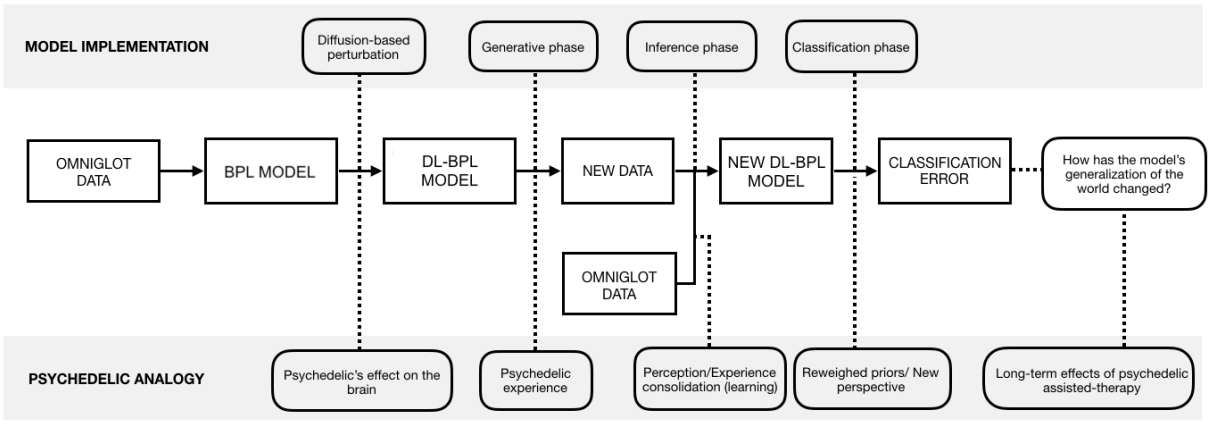


Figure 3.6: **Thesis pipeline**. Thesis work pipeline describing the psychedelic analogy with the method's implementation leveraging the BPL model.

induction. Our interpretation is that both REBUS and our approach share a common foundation in Bayesian theory and the fundamental difference resides on the idea of perspective modulation. While DL-BPL allows the generation of new concept programs themselves giving rise to different concepts, which can be interpreted as the formulation of new perspectives to be applied in several contexts post PAP, REBUS does not formalize how new cognitive structures might appear and consolidate after the psychedelic experience. The next section describes the pipeline phases in detail.

3.3 Implementation

3.3.1 Perturbation phase: Model perturbation

The first step of the above described pipeline consisted in perturbing the BPL generative model. As it has been previously described, the model consists of a set of probability distributions that in form of a joint distribution give a specification of the BPL's generative model.

Accordingly to the psychedelic analogy we are trying to make, this first step represents the perturbation of our priors of the world when under the influence of psychedelics. The idea we are here reinforcing is that, when under the influence of these substances, one's mind is able to build new perspectives, navigating through a journey of uncertainty, turning away from usual patterns of thought and exploring new mental constructs, sometimes, resulting in the rearrangement of one's priors.

Transferring this idea to the universe of the BPL model, this would correspond to modulate the generative model's priors, in order to introduce some novelty into the stochastic process of generating new character images, new data. This novelty can be represented by introducing new transitions/connections between character primitives that were not frequently seen in the generated character images, or on the other side by removing certain transitions that were more frequent to be observed during this generative process. Therefore, the change we are pursuing is restricted to the way the primitives relate to each other in the original model. We are not altering the existing model library of primitives and other existing hyperparameters, but inducing an update in the model's priors regarding primitive's transitions, which

might correspond to an update in one's beliefs about certain known concepts.

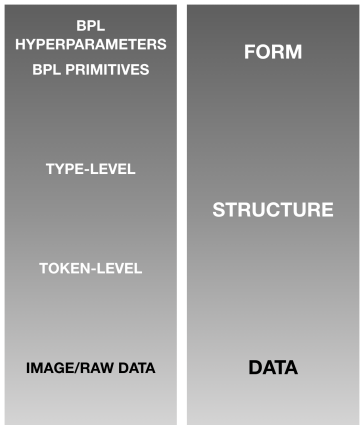


Figure 3.7: **Form vs. structure.** Referring to the Form-Structure-Data framework previously mentioned [211], the induced perturbation in BPL model will not influence the concept's form, the hyper parameters of BPL will remain unchanged, including the already existent 1212 model primitives. However, the goal is that, by perturbing the model, the potential structures under this form can be altered, giving rise to concepts/cognitive objects can emerge, or, on the other side, remove other concepts, as the connections within these concepts are modified.

The process describing the transition probabilities between the available library of primitives is part of the generative process of sampling a new character *stroke* as it was seen in Section 3.2.1. This process, described in Equation 3.13, was perturbed causing a change in the way primitives (which themselves can be interpreted as concepts) are sampled during the generation of new data, originating new correlations between these, and giving rise to new character concepts.

$$P(z_i) = P(z_{i1}) \prod_{j=2}^{n_i} P(z_{ij}|z_{i(j-1)}) \tag{3.13}$$

The joint distribution in Equation 3.13 represents a first-order Markov Process and depends on two probability distributions:

- $P(z_{i1})$: a probability distribution from where the first primitive index z_{i1} of each *stroke* is sampled, that will be referred as s matrix. This is in fact a 1×1212 vector comprising the probabilities of first sampling the 1212 existent primitives.

- $P(z_{ij}|z_{i(j-1)})$: a probability distribution from where the primitive indexes z_{ij} of the following *stroke* primitives, corresponding to the *sub-strokes*, are sampled, which will be referred as the pT matrix. The sampling of the indexes exclusively depends on the previous sampled index. The pT matrix is the normalization of a 1212×1212 Markov matrix, describing the probabilities of transitioning from one primitive to the other only depending on the previous state. The Markov matrix version of pT will be referred as pT_M.

Both of these matrices describe the model prior on which sampled character primitives and transitions between these are the most and less probable to happen when generating new character types and, therefore, both probability distributions were taken in consideration when implementing the per-

turbation. The process by which the perturbation of the process represented in Equation 3.13 was implemented is described in the next section.

Diffusion-based perturbations

The implemented perturbation took inspiration from the existing computational theories about the psychedelic effect in the brain [5], from state-of-the-art diffusion-based approaches to generative modeling [217, 218], and also referencing to statistical thermodynamics, drawing ideas from the diffusion heat kernel as well as from Boltzmann statistics. Specifically, a mathematical framework was developed in order to incorporate the perturbation of the process in Equation 3.13, as well as the context of the computational theories behind psychedelic drugs.

Understanding the structure of the primitive space is a huge part of understanding how we can perturb this space. A kernel provides a global similarity metric which specifies the local geometry of the considered data [219], expressing the prior beliefs about the existent correlations in a data space [220]. On the other side, a Markov chain describes the directions of propagation based on the kernel values [219].

Setting

$$v(x) = \int_X k(x, y) d\mu(y) \tag{3.14}$$

as the local measure of volume (or degree in a graph) and

$$p(x, y) = \frac{k(x, y)}{v(x)} \tag{3.15}$$

one can define $p(x, y)$ as the transition kernel of a Markov chain on a space X [219].

Taking this into consideration, the different model primitives can be interpreted as data points in a space. s can be understood as a prior on the characters generative process and pT matrix as the adjacency matrix that defines a Markov process between primitives [221]. That being said the s and pT matrices can be thought of weight functions, kernels, the latter translating the transitions between the primitives x and y, i.e the probability of the stochastic generative process to make a step from x to y.

$$s(x) \propto \frac{k(x)}{v(x)} \tag{3.16}$$

$$pT(x, y) \propto \frac{k(x, y)}{v(x)} \quad \text{with} \quad pT_M(x, y) \propto k(x, y) \tag{3.17}$$

A kernel over the discrete structure of primitives following the matrix exponentiation idea was constructed, resorting to a class of exponential kernels, based on the heat equation, referred as diffusion kernels [220].

Considering the primitive space a discrete space of finite dimension 1212, the kernel can be represented by an 1212×1212 matrix with rows and columns indexed by the elements of the space [220].

The exponentiation of a square matrix L is

$$e^{\beta L} = \lim_{n \to \infty} \left(1 + \frac{\beta L}{n}\right)^n \tag{3.18}$$

where the limit always exists [220] and its given by

$$e^{\beta L} = I + \beta L + \frac{\beta^2}{2!} L^2 + \frac{\beta^3}{3!} L^3 + ... \tag{3.19}$$

It has been shown that any infinitely divisible kernel can be expressed in this exponential form [220]

$$k(x, y) = e^{\beta L} \tag{3.20}$$

Differentiating Equation 3.20 with respect to β results in the differential equation

$$\frac{d}{d\beta} k = Lk \tag{3.21}$$

Equation 3.21 is also known as the heat or diffusion equation, describing how heat or gases diffuse in time through a continuous medium, and the resulting kernels are called diffusion or heat kernels [220]. L can be defined as a close relative of the adjacency matrix and β a parameter defining the extent of the diffusion, to specify the length scale [222], which can also be thought of as a temperature constant, making the analogy with Boltzmann statistics. Consequently, one can now make a relation between the BPL's model distributions defining the sampling of the primitive's indexes sequences of *strokes* and the diffusion kernel.

$$s(x) \propto e^{-L_s(x)} \quad \text{with} \quad \beta = 1 \tag{3.22}$$

$$pT_M(x, y) \propto e^{-L(x,y)} \quad \text{with} \quad \beta = 1 \tag{3.23}$$

$$pT(x, y) \propto \frac{e^{-L(x,y)}}{\sum_{x',y'} e^{-L(x',y')}} \tag{3.24}$$

Where $L_s(x)$ and $L(x, y)$ can be interpreted as distance matrices in the primitive space, closely resembling s and pT_M matrices, respectively.

$$L_s(x) := -\log(s) \tag{3.25}$$

$$L(x, y) := -\log(pT_M) \tag{3.26}$$

The perturbation of the original transition matrices resides in altering these by the β factor. The

perturbed matrices, which will be defined from now on as ρ_{start} and ρ_{pT}, are therefore defined as

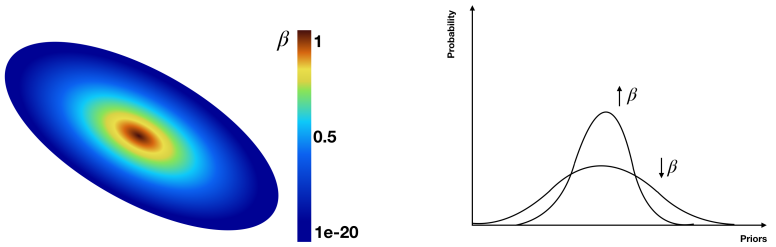
$$\rho_{start}(\beta, x) \propto \frac{e^{-\beta L_s(x)}}{\sum_{x'} e^{-\beta L_s(x')}} \tag{3.27}$$

$$\rho_{pT}(\beta, x, y) \propto \frac{e^{-\beta L(x,y)}}{\sum_{x',y'} e^{-\beta L(x',y')}} \tag{3.28}$$

The β parameter is what shapes the perturbation. As it has been mentioned above, the β parameter can be interpreted as the temperature constant of the system, being a based scale for the calculated distance L and, consequently,it is what will determine the distance/attention landscape between the different existent primitives.

The effect of this parameter in the sampled primitive transitions between primitives when generating a new character can be thought as of the effect of temperature in a room's air particles. A higher β value (lower temperature value) will result in a higher attention to the original distances existing between primitives, contributing to a stiffening of the original prior (corresponding to air particles transitioning to the nearest energy level). On the other hand, a lower β value (higher temperature value) will translate in a decrease in the attention to those original distances, resulting in a flattening of the prior landscape and making more likely to observe rare primitive transitions in *stroke* samples (corresponding to the increase in entropy of the air particles in the room, being more energy states available) when comparing to the original model samples.

Taking REBUS theory into consideration, the regime that will be explored when perturbing s and pT matrices is the one where $\beta < 1$ resulting in a flattening of the model's priors.



(a) (b)

Figure 3.8: **Effect of perturbation hyperparameter β. (a)** Heat kernel illustration. The concentration of distributional data decreases with decreasing β. **(b)** β parameter effect in the priors' landscape.

Note that, the original pT distribution does not comprise any probability value equal to zero, however the s distribution does, therefore, in order to be able to perturb these distribution values, all of the zero values in s were changed to have the small value of $1e-20$.

In order to try to establish a trend line between the above β value spectrum and the classification phase results, four perturbations were made for $\beta = 1e-3, 0.2, 0.5, 0.8$.

3.3.2 Generative phase: data augmentation via a generative alphabet procedure

The second phase of the illustrated pipeline in Figure 3.6 corresponds to generating a new group of characters for every different perturbation, i.e, for every β value mentioned at the end of Section 3.3.1 a DL-BPL model perturbation was induced, generating a new perturbed set of new character images organized in the form of alphabets.

In this context, an alphabet can be defined by a group of related concepts/characters. One way of thinking about these alphabets of character images, besides being an attempt to maintain a similar organization to the original data set, is that sets of related concepts may also be an effort to mimic what happens in our minds, representing mental constructs that are related to each other. This connection might arise because they were part of the same experience, part of the same memory, or context, suggesting some kind of underlying structure to our thoughts and ideas, even if subtle, when having a psychedelic experience.

The generation of a new alphabet requires an additional change in BPL's generative model, corresponding to adding an additional hierarchy layer to it [6]. The main change resides in the fact that an alphabet is firstly generated and only afterwards the new character types are generated from that alphabet, as it is illustrated in Figure 3.9. After the generation of new character types, generating tokens of the same type corresponds to the same exact process described in Algorithm 3. The additional alphabet level corresponds to adding a prior on top of the already existing ones. By introducing this level the model will be biased to re-use the structural components within a set of characters that are related. In other words, the generated character *strokes* will be memorized and stored for successive *stroke* samplings, reflecting a bias to re-use the parts that have already been sampled instead of of generating completely new parts. It is then possible to conclude that the new prior will favor re-using existing *strokes* [6].

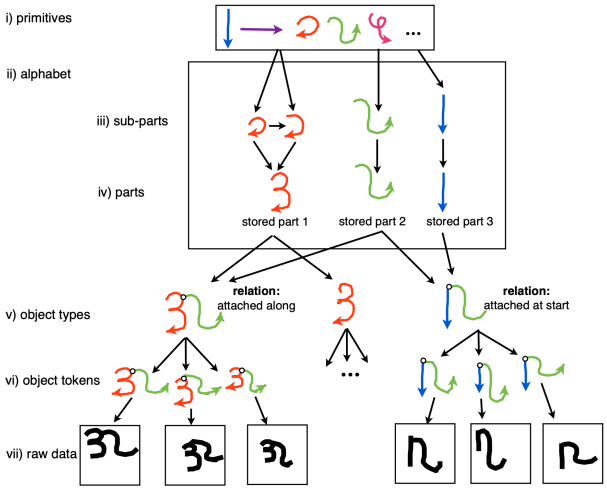


Figure 3.9: **Generative procedure for originating characters from a particular alphabet.** New *strokes* (parts) are generated by choosing primitives from the library (level i) and by combining these *sub-strokes* (sub-parts) (level iii) create parts (level iv). Characters from the same alphabet are produced by re-combining parts to define simple character types (level v). The model to generate tokens and images are the same as in original BPL (level vi-vii). Adapted from Lake et al. 2015 [6].

Generating a set of related concepts does not require the sampling of concrete variables but rather defines the transformations that links the various programs within an alphabet together using a useful tool in ML and statistics called the Dirichlet Process (DP) [223, 224].

Dirichlet Process

The DP is a stochastic process that is applied to Bayesian non-parametric models of data. It is a distribution over distributions, meaning that each draw from a DP is a distribution in and of itself [225]. Its parameters include a base distribution P_0, which is a distribution across a space Θ and a concentration parameter $\alpha > 0$. A random distribution P taken from a DP is represented by $P \sim DP(\alpha, P_0)$ [226].

In Ferguson (1973) [223] the DP was first established, using its limited dimensional distributions to demonstrate its existence. Taking a quantifiable partition of Θ, i.e, a collection of subsets whose union is Θ, $\{T_1, T_2, ..., T_K\}$, then every measurable partition of Θ is Dirichlet-distributed if $P \sim DP(\alpha, P_0)$

$$(P(T_1), ..., P(T_K)) \sim Dir(\alpha P_0(T_1), ..., \alpha P_0(T_K)) \tag{3.29}$$

This implies that if we sample a random distribution from the DP and add the probabilities of mass in a location $T \epsilon \Theta$, and, consequently, there will typically be $P_0(T)$ mass in that region. The concentration parameter α serves as an inverse variance, causing the random probability mass $P(T)$ to cluster more closely around $P_0(T)$ for increased values of α [226].

One of the Dirichlet process's properties identified by Ferguson [223] links the Chinese Restaurant Process (CRP) metaphor and the DP. Taking a random distribution sample from a DP proceeded by repeated draws from that same random distribution,

$$P \sim DP(\alpha, P_0) \tag{3.30}$$

$$\theta_j \sim P \;\; j \epsilon \{1, .., J\}. \tag{3.31}$$

By examining $\theta_{1:J}$ joint distribution, obtained by marginalizing out the random distribution P,

$$p(\theta_1, ..., \theta_J | \alpha, P_0) = \int \left(\prod_{j=1}^{J} p(\theta_j | P) \right) dP(P | \alpha, P_0), \tag{3.32}$$

it was demonstrated that θ_i displays a clustering property under this joint distribution, sharing repeated values with positive probability. A partition of the integers from 1 to J is defined by the shared value structure, and its distribution is determined by a CRP with the parameter α. It has also been demonstrated that each variable's unique value of θ_j is drawn independently from P_0 [223].

The CRP is a metaphor to a Chinese restaurant where a new customer $J + 1$ comes into the restaurant where J customers are already seated around a set of tables. All customers at a given table share the same value of θ_j . The new customer joins a previous table l with probability equal to $m_l/(J + \alpha)$, where m_l are the number of customers at that table. The new customer can also sit at a new table with

probability $\alpha/(J+\alpha)$ where a new value for θ_{J+1} is sampled from the base distribution P_0 [226].

Generating a new data set

In order to generate an alphabet of related concepts, a DP taking a concentration parameter α and a base distribution $P(.)$ as inputs was implemented. The result is a new distribution, here defined as $P - mem(.)$, a *"memoized"* transformation of the original distribution which induces dependencies between samples that were previously independent.

Conditional samples $d_{J+1}|d_1, ..., d_J \sim P - mem(.)$ have the form

$$d_{J+1}|d_1, ..., d_J \sim \sum_{j=1}^{J} \frac{\delta(d_{J+1} - d_j)}{J+\alpha} + \frac{\alpha P(d_{J+1})}{J+\alpha}, \tag{3.33}$$

where $\delta(.)$ is representing the delta function and the distribution encourages the re-use of previous values [6].

The new algorithm is an adaptation of Algorithm 1, where the conditional probability distributions used to sample the number of *strokes* κ and the *strokes* themselves, using Algorithm 2, were passed through the higher-procedure $DP(\alpha, .)$ in order to result in the new procedures $P - mem(.)$.

The novel *"memoized"* [6, 227] techniques define a set of probability distributions with the CRP clustering property, which are used to learn the number of *strokes* κ, the quantity of *sub-strokes* n_i, sample the *strokes* S_i, and relation types ξ_i that are distinctive of a specific alphabet [6].

Algorithm 4 Generate a new alphabet

procedure GENERATEALPHABET
  $P - mem(\kappa) \leftarrow DP(\alpha, P(\kappa))$
  for $\kappa = 1...10$ **do**
    $P - mem(n_i|\kappa) \leftarrow DP(\alpha, P(n_i|\kappa))$
  end for
  for $i = 1...10$ **do**
    for $n_i = 1...10$ **do**
      $P - mem(S_i|n_i) \leftarrow DP(\alpha, GENERATESTROKE(i, n_i))$
    end for
  end for
  $P - mem(\xi_i) \leftarrow DP(\alpha, P(\xi_i)$
  $A \leftarrow \{P - mem(\kappa); \forall \kappa : P - mem(n_i|\kappa); \forall i, n_i : P - mem(S_i, n_i)\}$
  return $GENERATETYPE(A)$
end procedure

The new generated alphabet is passed to Algorithm 5 in order to link the generation of new character types together, which means that the generated character will be dependent on each other, on the contrary of what happens in Algorithm 1.

The joint probability of the J character types $\psi^{(1)}, ..., \psi^{(J)}$ with M character tokens of each type $\theta^{(1,1)}, ..., \theta^{(1,M)}, ..., \theta^{(J,1)}, ..., \theta^{(J,M)}$, and, finally, the binary images $I^{(j,m)}$ has now a new form

Algorithm 5 Generate a new character type from alphabet A

procedure GENERATETYPE(A)
  $\kappa \leftarrow P - mem(\kappa)$ ▷ Sample the number of strokes
  for $i = 1...\kappa$ **do**
    $n_i \leftarrow P - mem(n_i|\kappa)$ ▷ Sample the number of sub-strokes
    $S_i \leftarrow P - mem(S_i|n_i)$ ▷ Sample a stroke with n_i sub-strokes
    $\xi_i \leftarrow P - mem(\xi_i)$ ▷ Sample the type of stroke's relation
    $R_i \leftarrow P(R_i|\xi_i, S1..., Si-1)$ ▷ Sample the details of the relation
  end for
  $\psi \leftarrow \{\kappa, R, S\}$
  return $GENERATETOKEN(\psi)$ ▷ Return program handle
end procedure

$$P(\psi^{(1)}, ..., \psi^{(J)}, \theta^{(1,.)}, ..., \theta^{(J,.)}, ..., I^{(1,.)}, ..., I^{(J,.)}) = \prod_{j=1}^{J} P(\psi^{(j)}|\psi^{(1)}, ..., \psi^{(j-1)}) \prod_{m=1}^{M} P(I^{(j,m)}|\theta^{(j,m)}) P(\theta^{(j,m)}|\psi^{(j)}) \tag{3.34}$$

where $\theta^{(j,.)}$ is short for all of the M examples of the specific type j: $\theta^{(j,1)}, ..., \theta^{(j,M)}$.

After the implementation of the alphabet construction using the Dirichlet Process the new data sets were generated. For every different β value perturbation, each new data set was generated with a similar organization to omniglot. Each data set is comprised by 30 different alphabets, each one consisting of 25 different character images and of 20 different exemplars (tokens) of each character image. After having the new data set generated by each DL-BPL model, the inference phase was initialized.

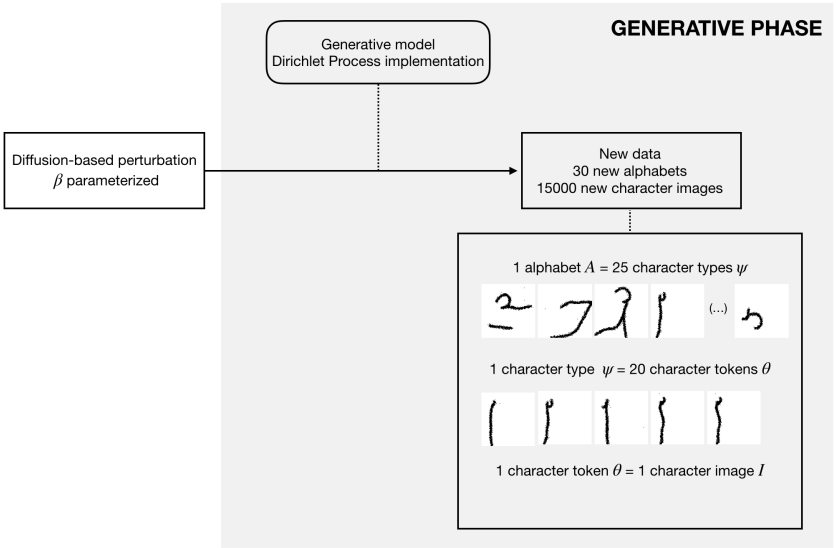


Figure 3.10: **Generative phase**. After the model perturbation with a certain β value, 30 alphabets each one with 25 character types, and each character type comprising 20 different examplars (tokens) are generated. Every new data set, for every β perturbation value, comprises 15000 new character images.

3.3.3 Inference phase: Learning a new model prior

After the generative phase, posterior inference was performed on the images of the augmented data set, including the images of the omniglot background data set, as well as the images of the 30 new generated alphabets, for each model perturbation.

The goal of this phase is to simulate one's perception of the psychedelic experience, and consequently its consolidation, resulting in new priors of DL-BPL, corresponding to the process described in Equation 3.13. For this reason, the inferred latent variables of interest to our problem correspond to the number of *strokes*, the number of *sub-strokes*, as well as the indexes of the primitives, to build a new model's prior from these. In order to achieve this, firstly, the data set images were processed to the correct format, explained in Section 3.3.3, and subsequently the inference model was applied to it.

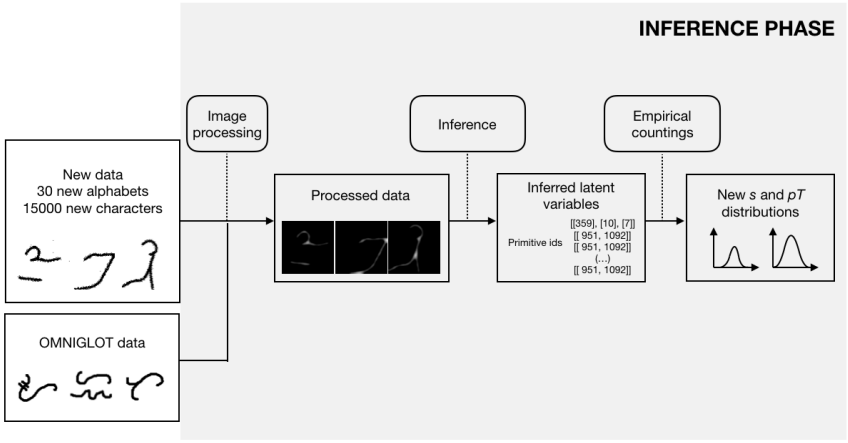


Figure 3.11: **Inference phase.** The resulting augmented data set from the generative phase was preprocessed before applying the BPL's inference algorithm to it. The inferred primitive indexes for every character image of the data set were used to compute the new model's prior regarding primitive sampling.

Image processing

Image processing consisted of transforming the 105×105 pixel data set images to an inference algorithm readable format. Image processing comprised the following steps: **(1)** Converting the generated and omniglot images to binary images, monochromatic images consisting of pixels that can only be black or white, stored as a single bit, with values 0 or 1; **(2)** Transforming each binary image to its negative; **(3)** When generated an image from BPL's generative model it usually has some pixel noise. For those images, the isolated noise pixels were removed, as well as "holes" in the character image ink were corrected; **(4)** The image edges were smoothed out by using a Gaussian kernel.

Inferring latent variables from images

Posterior inference in BPL model correspond to parse an image $I^{(m)}$ in its parts (*strokes*) and subparts (*sub-strokes*), which is a very challenge task, since it requires the exploration of a broad space of different numbers and types of *strokes*, *sub-strokes*, and *relations* [6]. In order to tackle this, taking inspi-

ration from fast human perception and from approaches for faster inference in the context of probabilistic programs [228], the BPL model uses bottom-up methods to perform fast inference.

In summary, posterior inference consists in finding a set of possible motor programs (a motor program is the set of latent variables that define a character type and token, a concept) which is an approximate fit to the inferred image, by parsing the image. The most promising motor programs are selected and improved using MCMC and continuous optimization. The final result is a set of K high probability motor programs, $\psi^{[1]}, \theta^{(m)[1]}, ..., \psi^{[K]}, \theta^{(m)[K]}$. These are the K most promising set of latent variables candidates that were found by the inference algorithm [6].

When performing inference in the data set, K was set to 1, obtaining the best parse for each image.

The motor programs approximation to the posterior is given by the equation

$$P(\psi, \theta^{(m)}|I^{(m)}) \approx \sum_{i=1}^{K} w_i \delta(\theta^{(m)} - \theta^{(m)[i]})\delta(\psi - \psi^{[i]}) \tag{3.35}$$

in which each weight w_i is inversely proportional to the motor program score, marginalizing over the type-level shape variables x and attachment points τ and restricting $\sum_i w_i = 1$,

$$w_i \propto \tilde{w_i} = P(\psi^{[i]}_{x,\tau}, \theta^{(m)[i]}, I^{(m)}). \tag{3.36}$$

The approximation can be improved if incorporating local variance surrounding type-level (the token-level variables closely track the image and allow for very little variability). This can be done without increasing computational cost, since it is not necessary to evaluate the likelihood of the image when estimating to produce conditional samples from the type-level $P(\psi|\theta^{(m)[i]}, I^{(m)}) = P(\psi|\theta^{(m)[i]})$. N samples for each motor program defined by $\theta^{(m)[i]}$, are produced by a Metropolis Hastings algorithm, designated by $\psi^{[i,1]}, ..., \psi^{[i,N]}$.

The approximation is now

$$P(\psi, \theta^{(m)}|I^{(m)}) \approx Q(\psi, \theta^{(m)}, I^{(m)}) = \sum_{i=1}^{K} w_i \delta(\theta^{(m)} - \theta^{(m)[i]}) \frac{1}{N} \sum_{j=1}^{N} \delta(\psi - \psi^{[ij]}) \tag{3.37}$$

BPL performs inference by using fast bottom-up search followed by discrete selection and continuous optimization, which will be briefly explored. For more details see references [6, 7]. The **inference algorithm** in Lake et al. (2015) [6] consists of: **(1)** The first step consists of searching for the best candidate parses by leveraging fast bottom-up methods. Inference search starts by extracting the character skeleton of the image. By applying a thinning algorithm to the raw data, and defining edges tracing the image and nodes placed at fork points, critical points in the images are identified and a character image undirected skeleton graph is built. To generate a candidate motor program, a random walk is taken in the image skeleton, visiting nodes until each edge has been transversed at least one time. The number of proposed candidate parses are usually between 10 and 150 parses; **(2)** The second step is defining *stroke* order and directions. After finding the candidates, *strokes* are sub-divided into *sub-strokes* by running an exhaustive greedy search using the prior $P(\psi)$. After classifying the *sub-strokes* accord-

ing to its primitive indexes z_i the parses decomposition are scored by the *stroke's* generative model in equation 3.38, with y_i equal to $y_i^{(m)}$; **(3)** The prior score $P(\theta^{(m)}|\psi)P(\psi)$ is finally used to select the K best parses. As mentioned, $K = 1$ is used when doing inference in the new augmented data set; **(4)** The fourth step consists in maximizing the entire joint density $P(\psi, \theta, I)$ described in equation 3.8, by dividing each parse into type and token characters $\{\psi, \theta\}$, and optimizing the continuous type-and token-level variables via gradient descent; **(5)** At the last step, having K best motor program candidates $\psi^{[1]}, \theta^{(m)[1]}, ..., \psi^{[K]}, \theta^{(m)[K]}$ a run of MCMC is run to estimate type-level local variance.

$$P(x_i^{(m)}, y_i^{(m)}, z_i) = P(z_i) \prod_{j=1}^{n_i} P(y_{ij}^{(m)}|y_{ij}) P(y_{ij}|z_{ij}) \int P(x_{ij}^{(m)}|x_{ij}) P(x_{ij}|z_{ij}) dx_{ij} \tag{3.38}$$

Constructing a new prior

In order to construct a new prior on the process to sample *stroke*'s primitives when generating a new character image, posterior inference was done in the 19280 images of the omniglot background data set and for the new 15000 images generated by the perturbed model.

The inferred *sub-stroke* primitive indexes were stored and the transition between primitive indexes empirical countings were calculated in order to compute the updated and final ρ^*_{start} and ρ^*_{pT} distributions.

The computation of the new distributions consisted of, firstly, computing a frequency matrix describing the frequency of each existent transition in the universe of the total inferred *sub-stroke* transitions and, secondly, normalizing the matrices. The first indexes of every *stroke*, correspondent to the first sampled *sub-stroke* in each *stroke*, were taken into consideration to compute ρ_{start}, while every other existent transition between the following *sub-strokes* in each *stroke* were considered in ρ_{pT}'s calculation.

$$P(z_{i1}) \leftarrow \rho^*_{start}(\beta) \tag{3.39}$$

$$P(z_{ij}|z_{i(j-1)}) \leftarrow \rho^*_{pT}(\beta) \tag{3.40}$$

This process was repeated for every β perturbation value. The new prior distributions were updated in the DL-BPL's library along with the remaining hyper-parameter model priors for the classification phase.

3.3.4 Classification phase: Evaluating the model's performance

In the last step of this work's pipeline a one-shot classification task from Lake et al. (2015) [6] was performed to evaluate the perturbed DL-BPL model's performance and inference scoring comparing it with the original BPL model.

The one-classification task entails the evaluation of the probability of a test image $I^{(T)}$ given one single training image of a new character $I^{(c)}$ correspondent to one of $c = 1, ..., C$ classes. An approximate solution for this can be computed through a Bayesian classification rule

$$\arg\max_{c} \log P(I^{(T)}|I^{(c)}) \tag{3.41}$$

BPL's inference algorithm is leveraged in order to get $K = 5$ best motor programs, each one with N samples translating type-level variability in these motor programs, of image $I^{(c)}$. The maximum score over token-level continuous parameters $\theta^{(T)}$ is obtained by running K gradient-based optimization procedures to re-fit $\theta^{(c)}$ to the test image $I^{(T)}$. When re-fitting a training image $I^{(c)}$ to a test image $I^{(T)}$, the K best $I^{(c)}$ parses are optimized with fixed type-level ψ parameters to best map the token-level $\theta^{(T)}$ variables to the test image $I^{(T)}$. The approximation of the Bayesian score is given by

$$\begin{aligned} \log P(I^{(T)}|I^{(c)}) &\approx \log \int P(I^{(T)}|\theta^{(T)}) P(\theta^{(T)}|\psi) Q(\theta^{(c)}, \psi, I^{(c)}) d\psi d\theta^{(c)} d\theta^{(T)} \\ &\approx \log \sum_{i=1}^{K} \max_{\theta^{(T)}} P(I^{(T)}|\theta^{(T)}) \frac{1}{N} \sum_{j=1}^{N} P(\theta^{(T)}|\psi^{[ij]}) \end{aligned} \tag{3.42}$$

with $Q(.,.,.)$ and w_i referencing back to 3.35.

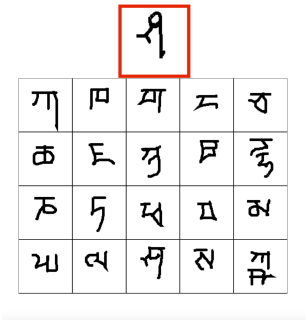
We follow Lake et al. (2015) [6] and employ a two-way Bayesian score that also takes parses of $I^{(T)}$ re-fitted to $I^{(c)}$. Consequently, the used classification rule was

$$\arg\max_{c} \log P(I^{(T)}|I^{(c)}) = \arg\max_{c} \log P(I^{(T)}|I^{(c)})^2 = \arg\max_{c} \log \left[\frac{P(I^{(c)}|I^{(T)})}{P(I^{(c)})} P(I^{(T)}|I^{(c)}) \right] \tag{3.43}$$

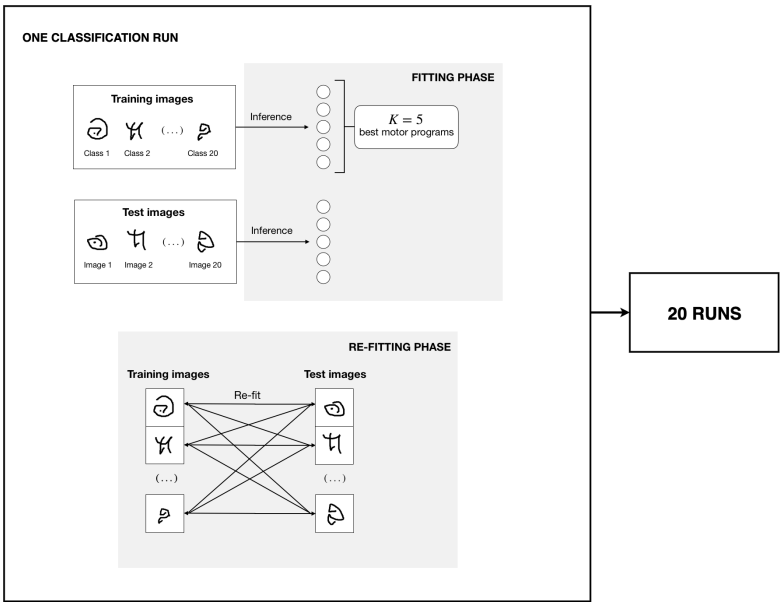
with $P(I^{(c)}) \approx \sum_i \tilde{w}_i$. From this, it is possible to conclude that the model classifies a test image by choosing the training image class that re-fits it with the highest Bayesian score.

Classification was performed in images that the model has never seen, using the evaluation omniglot data set, which is demonstrated in the supplementary materials A.1, consisting of the 20 remaining omniglot alphabets. The task involved 20 runs of 20 within-alphabet images classification series. Correctly classifying an image consists in classifying a training image as the correspondent test image of the same character type. Each run comprises 20 different training images and 20 different test images, from the same alphabet, and only one right correspondence between each one of the images. Each one of the classification run episodes consisted of feeding the inference model with one training example from the 20 training character images. Furthermore, the model has to re-fit the training image to each one of the 20 test images. Every run yields 20 episodes in total, therefore representing 400 task trials. For each of the perturbed models, the classification scores were obtained for every run episode, as well as the average classification error per run and the average classification error for the 20 runs.

Additional performed analysis are described in supplementary material A.2.



(a)



(b)

Figure 3.12: **Classification phase. (a)** Example of a classification task. The model tries to classify the image surrounded in red as one of the 20 images below. **(b)** One-shot classification can be divided into two phases. One classification run consists of the fitting phase that corresponds to infer the 5 best motorprograms of every training and test images (above). The second step corresponds to the re-fitting phase, where each training image is re-fitted to the test images, and vice-versa. For every culminating pipeline resulting from the perturbation with $\beta = 1e-3, 0.2, 0.5, 0.8$, this process was repeated 20 times (20 runs).

Chapter 4

Experimental analysis

4.1 Diffusion-based perturbations

4.1.1 Original model priors analysis

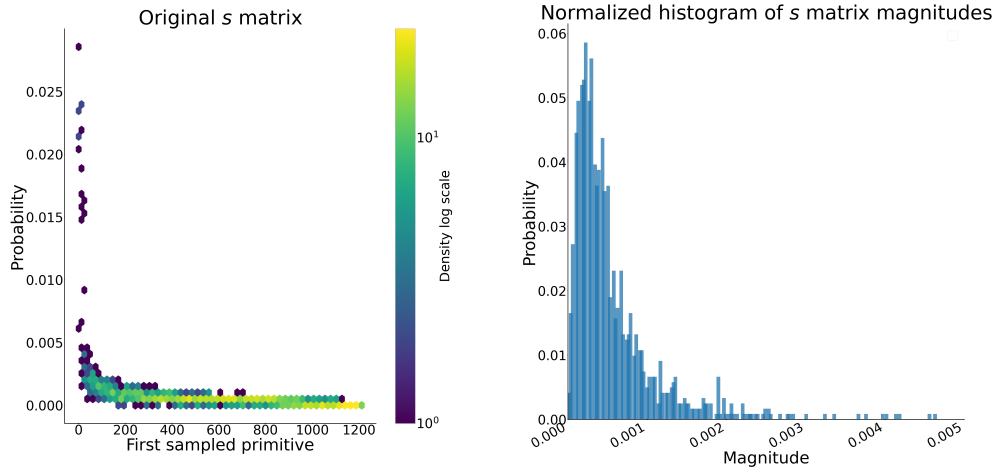
The first step of the presented computational analysis consisted in trying to understand the structure of the original model priors that determine which primitive indexes are sampled when generating a character, namely s and pT matrices. In order to achieve this, the priors were visualized and statistically analysed to give us an understanding of where the mass of each of the priors is concentrated revealing, for example, the range of probability magnitudes and frequencies in the given distributions. Recall (see Section 3.3.1) the process of sampling which primitive indexes will give rise to the new generated character is recurrently determined by

$$\begin{aligned} P(z_i) &= s_{i1} \prod_{j=2}^{n_i} pT_{i(j-1),ij} \\ &= P(z_{i1}) \prod_{j=2}^{n_i} P(z_{ij}|z_{i(j-1)}) \end{aligned}$$

The s matrix, with components $s_{i1} = P(z_{i1})$ in Equation 3.13, represents the probability distribution from where the first primitive index z_{i1} of each *stroke* is sampled. Analysing Figure 4.1, we can conclude that the mass of the matrix is highly concentrated within the smaller probability values.

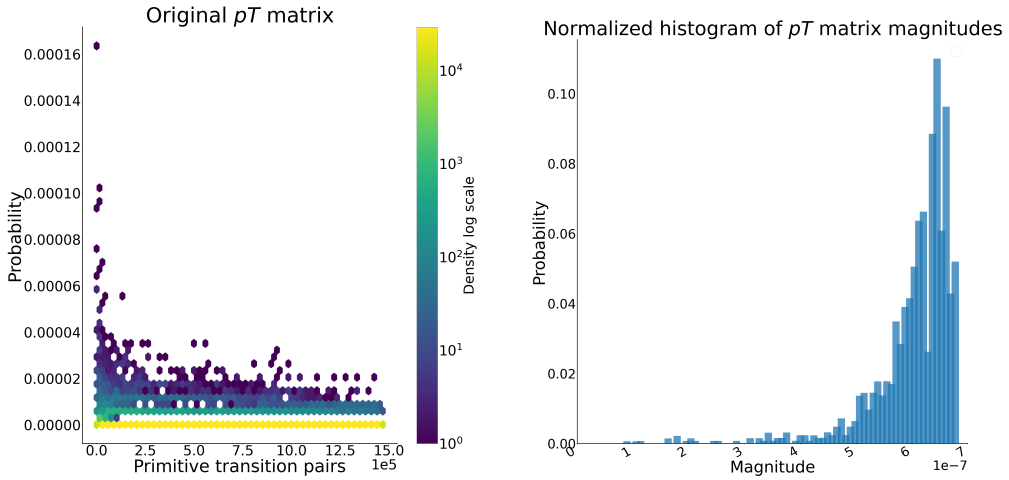
The pT matrix, $P(z_{ij}|z_{i(j-1)})$ in Equation 3.13, represents the probability distribution from where the indexes z_{ij} of the following *stroke* primitives (the ones after the first *sub-stroke* primitive), are sampled. Note that the maximum magnitude value in pT is smaller than the one in s, since it is a probability distribution of higher dimensions. Besides this, similarly to what we can observe for the s matrix, the pT matrix has its mass concentrated in the smaller probability values.

From this visualization it is possible to conclude that the highest probability values referring, in the case of the s matrix to the first primitive sampled in a *stroke*, and in the case of the pT matrix to the transitions between one primitive and another in the remaining character *sub-strokes* of a *stroke*, correspond to a very low number of primitives and transitions between primitives, respectively. This



(a) Original s matrix. (b) Histogram s matrix.

Figure 4.1: **Original s matrix (a)** structure and **(b)** histogram. On the left, it is possible to observe the original 1×1212 s matrix structure, showing the probabilities, on the y-axis, of each one the 1212 existent model primitives, on the x-axis, being the first one to be sampled in the beginning of generating a character *stroke*. The probability magnitudes in s distribution vary within a minimum value of 0 and a maximum value of 0.0286. On the right histogram, a higher occurring frequency of probabilities between 0 and 0.005 values within the 1212 entries of the s matrix is shown.



(a) Original pT matrix. (b) Histogram pT matrix.

Figure 4.2: **Original pT matrix (a)** structure and **(b)** histogram. On the left panel, it is possible to observe the structure of the 1212×1212 pT matrix. On the x-axis of the plot, the primitive transition pairs (corresponding to the matrix entries) are ordered by matrix row, totaling the $1212 \times 1212 = 1468944$ pT matrix entries, and on the y-axis the probabilities of each one of these primitive transitions to happen when generating a character are shown. The probabilities' magnitudes expressed in the pT matrix, concerning the transitions between character's primitives, vary between a minimum value of $8.5130e-8$ and a maximum value of 0.0002.It is possible to conclude from the histogram on the right side that the mass of the matrix is mainly found within probability values between $1e-7$ and $7e-7$.

leads us to assume that these primitives and transitions will then be the ones which are most likely to observe in the generative process of a character.

As discussed in Section 3.3.1, our fundamental motivation for applying a diffusion-based perturbation on these priors is to generate a "flattening" effect [5]. Our hypothesis is that this "flattening" effect, corresponding to a relaxation of the existent model priors, will weaken expectations regarding the primitives that will be sampled when generating a new character, observing a particular effect in the generative phase within our modeling framework (Section 3.3.2) such that novel primitives and primitive transitions will be generated. We expect that this will lead to more flexible inference and higher classification performance in the evaluation omniglot dataset. This computational mechanism can be seen as analogous to what is believed to happen in human perceptual processing under the influence of psychedelics [5], albeit at a higher level of cognition in contrast to previous work (see Section 3.2.2).

In the next section, we confirm our computational hypotheses by interrogating the restructured priors resulting from the diffusion-based perturbation framework.

4.1.2 Perturbed model priors analysis

Intuitively, we seek to perturb the probability distribution in the latent space of *characters* such that the essential structure is preserved but the probability mass becomes less concentrated [217, 218]. We propose to accomplish this with diffusive perturbations in the latent space of *character primitives* (specifically determined by the model priors s vector and pT matrix). Our hypothesis is that this approach will extrapolate new and distributionally-consistent combinations of character primitives with a relatively low incidence of characters inconsistent with the character data generating process.

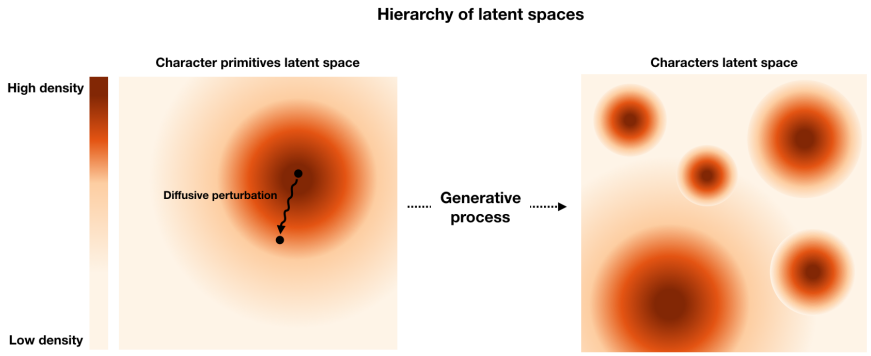
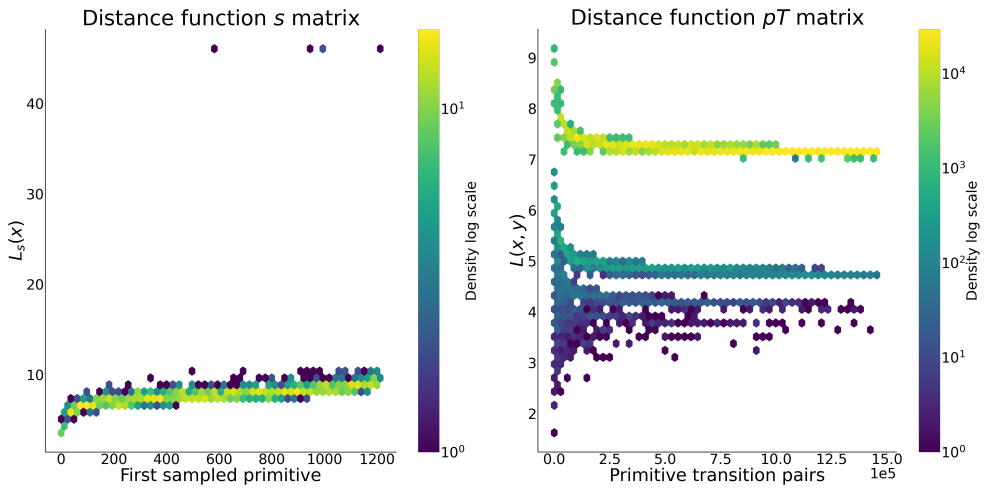


Figure 4.3: **Conceptual representation of latent spaces hierarchy.** We are hoping to induce novelty in the characters space through diffusive perturbations in the space of character primitives, by preserving its essential structure but flattening the probability density landscape. This figure represents our hypothesis that perturbations in the primitive layer will manifest as more complex and multimodal changes in the character layer.

The first step in obtaining the new perturbed priors ρ_{start} and ρ_{pT} (see equations 3.27 and 3.28) was the intermediate step of computing the distance function $L_s(x)$ and $L(x, y)$ for the s and pT priors, respectively (see equations 3.25 and 3.26).

In order to approximately implement diffusion-based perturbations, we interpret L as a "distance" between character primitives, describing the relationship between pairs of these primitives in terms of



(a) Distance function s matrix (b) Distance function pT matrix

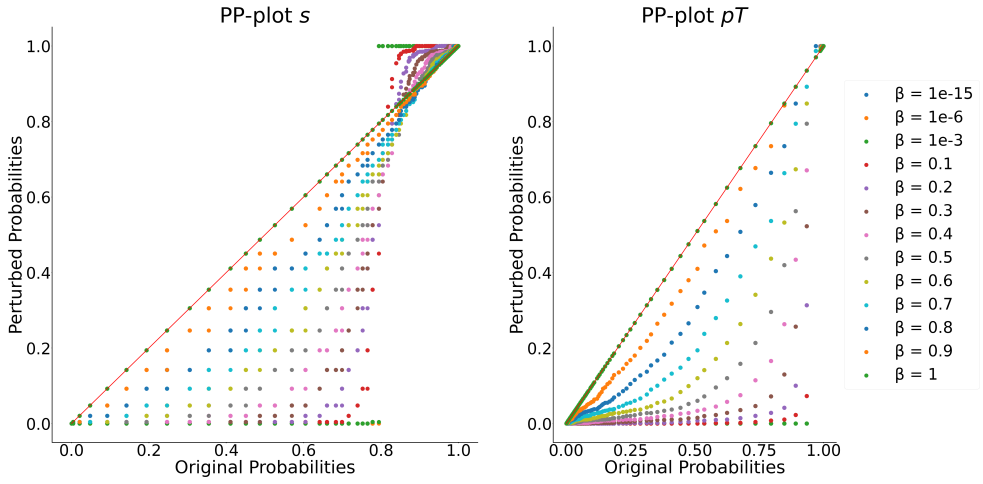
Figure 4.4: Distance functions in primitive space for original model priors.

their connectivity. Note that pT is not a symmetric matrix and therefore L does not induce a Euclidean embedding of the character primitives [219, 221]. We employ an analogous approach to perturbing the s prior.

Observing Figure 4.4a and 4.4b and making a comparison with the original priors' plots in Figures 4.1a and 4.2a, respectively, it is possible to establish a structural similarity between these plots, where higher probabilities in figures 4.1a and 4.2a correspond to smaller values in the distance function plots, and vice-versa. Accordingly, smaller distance values in Figure 4.4b plot, for instance, represent a higher connectivity between two primitive states, and vice-versa. In the case of Figure 4.4a, the distance function can be thought of as an initial generative process prior.

After having a better comprehension of the priors' structure and relations within primitives, in order to understand the diffusion-based perturbation effect on both priors, different β parameters were simulated. As mentioned in Section 3.3.1 the spectrum of interest for the β parameter is $0 < \beta < 1$. Figure A.2 in supplementary material shows the effect each diffusion-based perturbation had on the original priors while varying the β parameter. This effect is summarized in Figure 4.5 in the Probability-Probability (P-P) plots for both priors, in which the cumulative distribution functions (CDFs) of the two distributions (perturbed and original) are plotted against each other. Any evaluation point on the plot indicates what percentage of data lies at or below that point in both original and perturbed distributions (as per definition of CDF). To compare the distributions, the deviation of the points from the 45-degree line ($x = y$) is analyzed.

We can see that the points for $\beta = 1$ (no perturbation) lie within $x = y$, and, with decreasing β the degree of deviation from the linear function increases, and therefore more different the distributions are from the original priors. Furthermore, observing the structure plots (Figure A.2), one can conclude that with a decreasing β parameter value, both ρ_{start} and ρ_{pT} distributions' mass becomes concentrated in a smaller probability value range as seen in Figure 4.6 where this effect is demonstrated across the β parameters $0.8, 0.5, 0.2, 1e-3$.

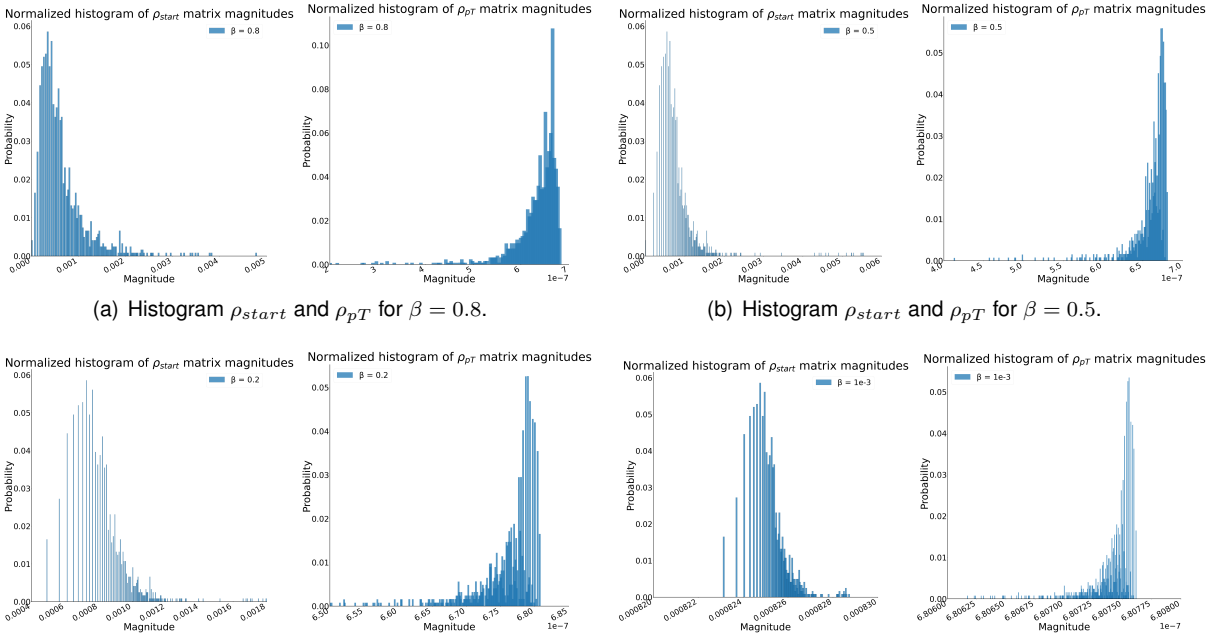


(a) P-P plot s and ρ_{start}.

(b) P-P plot pT and ρ_{pT}.

Figure 4.5: **P-P plots** of **(a)** s and **(b)** pT distributions and respective distribution perturbations, for different β values.

This analysis reflects and confirms our hypotheses regarding the effect on the original model priors. One can observe a decrease in the priors' concentration as the value of β also decreases. However, it is also possible to observe that for the lowest β values, such as $\beta = 1e-15$, the structure of the distribution begins to disintegrate, seemingly converging on a uniform distribution.



(a) Histogram ρ_{start} and ρ_{pT} for $\beta = 0.8$.

(b) Histogram ρ_{start} and ρ_{pT} for $\beta = 0.5$.

(c) Histogram ρ_{start} and ρ_{pT} for $\beta = 0.2$.

(d) Histogram ρ_{start} and ρ_{pT} for $\beta = 1e-3$.

Figure 4.6: New perturbed priors ρ_{start} (on the left) and ρ_{pT} (on the right) histograms for $\beta =$ **(a)** 0.8, **(b)** 0.5, **(c)** 0.2, **(d)** $1e-3$.

4.1.3 Additional analysis

The computation of Shannon entropy, Kullback-Leibler divergence (KLD) and Jensen Shannon distance (JSD), with respect to the original model priors and to the perturbed priors was additionally analyzed.

The Shannon entropy of both new distributions ρ_{start} and ρ_{pT} for the different β parameter values was computed. One can interpret the entropy value as "the minimum number of bits it would take to encode the distribution's information". In Figure 4.7 we can observe the decreasing value of the distributions' Shannon entropy with increasing β for both prior's analysis, confirming once again our intuition that a diffusion-based perturbation with a β value closer to zero corresponds to a higher entropy distribution state.

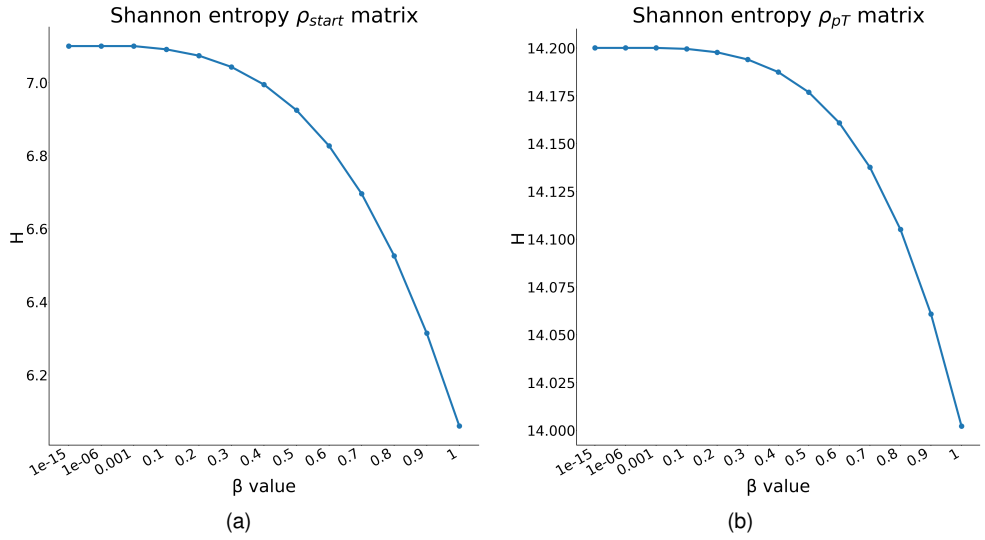


Figure 4.7: Shannon entropy H (in bits) measure of the new perturbed priors **(a)** ρ_{start} and **(b)** ρ_{pT} for $\beta = 1e-15, 1-6, 1e-3, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$.

Referencing to Section 2.2.1, while keeping in mind the psychedelic's action in the brain analogy, one can also think of entropy applied to the context of states of consciousness and their dynamics, and how the psychedelic state has shown to disrupt certain aspects of brain function, such as the repertoire of functional connectivity motifs [27]. Taking the original model priors as a baseline, analogous to the normal waking consciousness (see Section 2.2.1), one can conclude that perturbing the model's s and pT distributions is approximating them to a state of increasing disorder, of higher entropy [20, 27]. An increase in entropy in neuronal circuits may indicate that the brain is exploring a wider range of patterns of activity, potentially departing from its normal repertoire of states, i.e. attractors [229, 230]. Perturbing and flattening one's attractor landscape is believed to increase the probability of the system's state to move between attractors and hence promoting more flexible, and adaptive dynamics [231]. In fact, in the context of psychedelic therapy, it has been proposed that mental illnesses are underpinned by excessively reinforced attractors [232], which lead to rigid patterns of thinking and behaving, psychedelics may have the potential induce this flatenning effect, ultimately leading to a breaking of these reinforced patterns of thought and behavior [233, 234]. Analogously, perturbing the model primitive space might

fundamentally change the ”attractor landscape” in the space of characters, hopefully originating a modified latent space with new attracting poles, as illustrated in Figure 4.3.

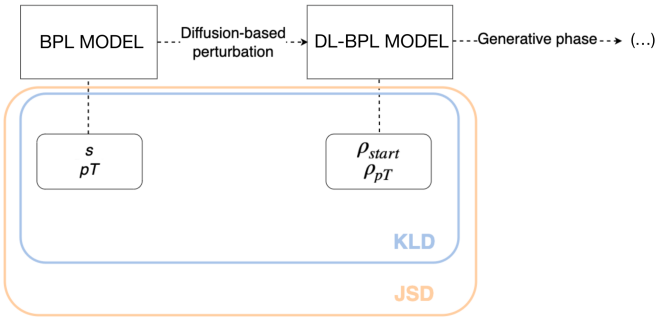


Figure 4.8: **Generative priors perturbation**. The original non-perturbed BPL model distributions s and pT are perturbed through a diffusion-based perturbation method originating the new distributions ρ_{start} and ρ_{pT}, and, consequently a new perturbed DL-BPL model, which will be the one to be used in the generative phase. KLD and JSD measures were computed in order to compare the original and new distributions.

KLD is another measure that has its origins in information theory, in which the primary goal is to quantify how much information is in data. Analysing KLD values allows us to identify the differences in two data distributions. This is a relevant measure to understand how much change we are inducing when replacing the model's prior for the new perturbed priors. Besides this, JSD, a symmetrized and smoothed version of KLD was also explored, as a better metric for comparing two data distributions. Both KLD and JSD were computed, comparing the original s distributions and the differently parameterized ρ_{start} and, similarly, comparing the original pT distribution and the differently parameterized ρ_{pT}, allowing a comparison regarding the original generative model priors with the new generative priors, which will be replaced in BPL's library,in order to generate new perturbed datasets.

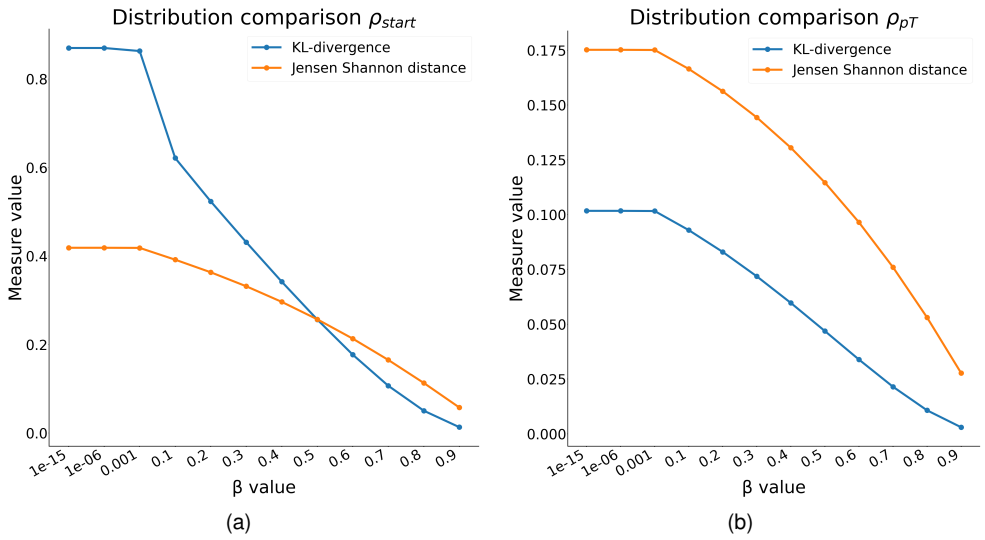


Figure 4.9: **(a)** Kullback–Leibler divergence and Jensen-Shannon distance between original s matrix and perturbed ρ_{start} matrix for different β parameter values. **(b)** Kullback–Leibler divergence and Jensen-Shannon distance between original pT matrix and perturbed ρ_{pT} matrix for different β parameter values. The threshold equaled to one represents the upper bound of JSD.

In Figure 4.9, for both plots, it is possible to see a decrease in KLD and JSD values with an increasing β value, once again leading to the confirmation that there is a higher information[1] loss between the novel distributions compared to the original ones when β gets closer to zero, and a recovering of the original distributions when β gets closer to one. The higher effect that the perturbation has on the ρ_{start} distributions than on the ρ_{pT} distributions is also evidenced, which can be due to the dimension of s distribution and its higher probability magnitude values.

In order to proceed to the next phase of the thesis pipeline, four β parameter values were chosen. The selection criterion took into consideration a trade-off between structure and entropy, working towards the goal of increasing the probability of seeing characters with not so likely primitives to emerge during the generative process sampling, without completely lesioning the priors' structure, avoiding uniformity. Even though psychedelics are believed to lead the brain closer to criticality, this does not represent a state of full disorder. The next pipeline steps were explored by perturbing the model with parameters $\beta = 1e-3, 0.2, 0.5, 0.8$, trying to cover the range $0 < \beta < 1$ but also taking into account the above described aspects. For the generative phase, the original priors in BPL's library were replaced by the computed perturbed priors ρ_{start} and ρ_{pT}, resulting in a new perturbed model for each value of β, that will be referred as DL-BPL.

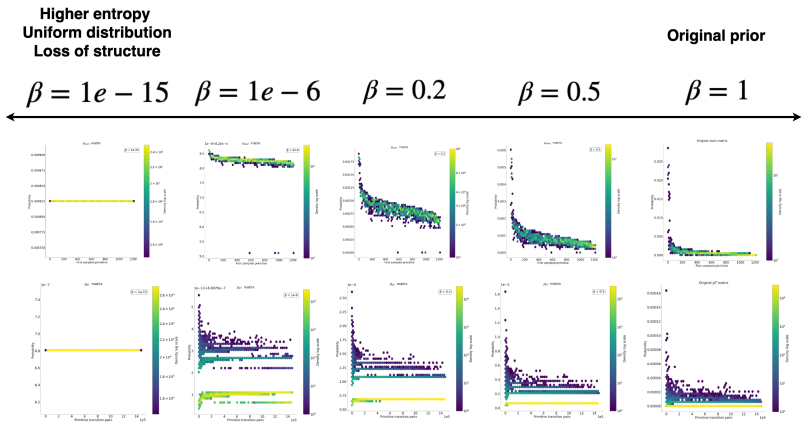


Figure 4.10: Diffusion-based perturbation spectrum for s (above) and pT (below) matrices.

Before proceeding onto the next phase, it is important to emphasize our core innovation in applying diffusive perturbations to do data augmentation in the context of probabilistic program induction. Based on the broad concept of diffusion processes, our data augmentation framework can be related to other pieces of work such as diffusion models [235]. Diffusion probabilistic models emerged as a class of generative models inspired by non-equilibrium thermodynamics [217]. The idea behind these models is to systematically and slowly destroy structure in a data distribution through an iterative forward diffusion process[2], and then to learn a reverse diffusion process that restores structure in data, yielding a generative model of the data [217, 235–237]. Besides this, state-of-the-art work [218, 236], bridging both score-based generative modeling[3] and diffusion probabilistic modeling into an unified framework, has

[1]The information contained in a probability distribution refers to the statistical properties of the distribution.

[2]The diffusion process is typically modeled using stochastic differential equations or partial differential equations. The resulting model is trained by maximizing the likelihood of the observed data under the diffusion process, which is done using stochastic gradient descent or other optimization algorithms.

[3]Score-based generative models are a class of generative models that do not explicitly model the probability density function of

implemented perturbations to data points with noise and train score-based models on the noisy data points instead, in order to increase model performance and overcome limitations in modeling complex distributions with sharp or discontinuous changes in density. Taking this into consideration, our approach introduces and explores how diffusion-based perturbations can be applied to data augmentation in a probabilistic induction model and how such perturbations manifest at distinct levels in a program hierarchy.

4.2 Generative phase

The second phase of our modeling pipeline consisted in generating an augmented dataset of perturbed characters. This generative phase was executed for each beta parameter value of the diffusive perturbation separately. Each DL-BPL model generated 30 new perturbed alphabets, as explained in Section 3.3.2.

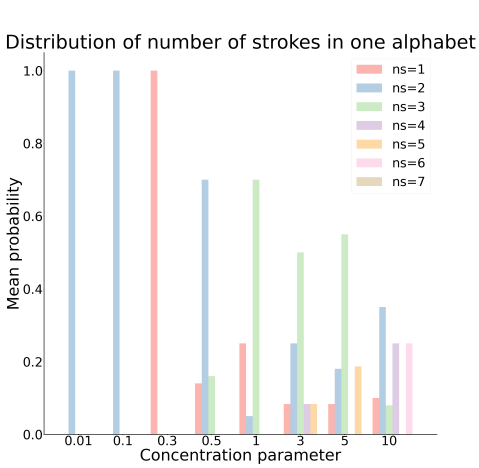
Before generating the alphabets, the DP concentration parameter, α, had to be chosen to be used in their generation. Section 3.3.2 explored the role of the α parameter in the DP, emphasizing that it can be understood as an inverse variance, such as a larger α value results in a higher concentration of mass around the mean of the DP. After implementing the DP within the DL-BPL's generative process, the concentration parameter α was optimized based on small-scale sampling tests, the objective being to select the parameter that best suited the desired features in the set of characters that will form an alphabet.

Firstly, using the original non-perturbed BPL model and different DP parameter values $\alpha = 0.01, 0.1, 0.3, 0.5, 1, 3, 5, 10$, an alphabet with 25 characters was generated. Secondly, the distribution of the number of *strokes* in one character was computed and analysed for each one of the generated alphabets. Furthermore, the percentage of primitive indexes which were repeated in the sampling process across characters in each one of the alphabets was also assessed. The resulting analysis can be seen in Figures 4.11a and 4.11b.

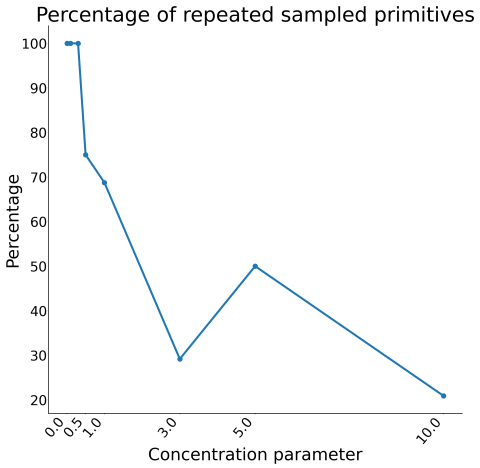
An important objective in the perturbed alphabet generation is the preservation of the essential structure across the characters it comprises, including the number of *strokes* and the primitive indexes that make up these *strokes*. However, while this structure is something to be preserved, some variability among the characters of an alphabet is desirable for flexible inference. Accordingly to this, it is observable that the α values that originate alphabets which most resemble the above requirements are $\alpha = 3$ and $\alpha = 5$, showing not only variability across characters regarding the number of *strokes*, but also preserving around 30% and 50%, respectively, of the primitives across the alphabet characters, which was considered a balanced trade-off between novelty and similarity for our purposes.

Subsequently, the interval between these α values was further explored, repeating the same sampling test, but this time for $\alpha = 3, 3.5, 4, 4.5, 5$, and generating five alphabets for each value. Figures 4.11c

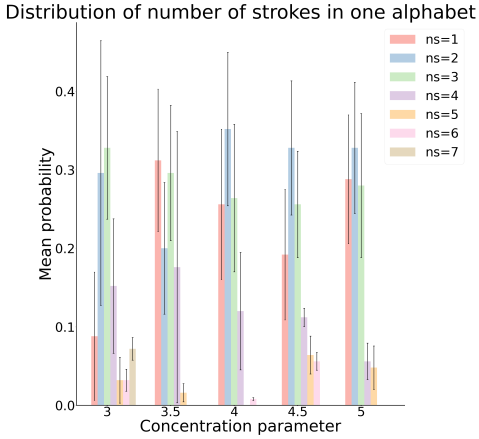
the data but instead model its gradient or score function. The score function is the gradient of the log-probability density function with respect to the data, and it is used to estimate the likelihood of a given sample via Langevin sampling. Score-based models are typically trained by maximizing the log-likelihood of the data, which can be done using gradient-based optimization algorithms such as stochastic gradient descent [218, 236].



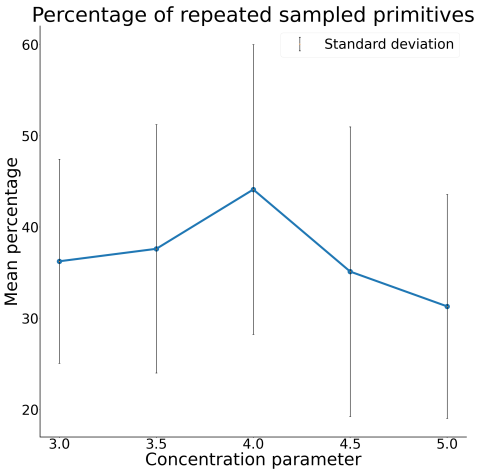
(a) Number of character strokes (ns) distribution across the alphabet characters generated with different concentration parameters $\alpha = 0.01, 0.1, 0.3, 0.5, 1, 3, 5, 10$.



(b) Percentage of repeated sampled primitive ids across the alphabet characters generated with different concentration parameters $\alpha = 0.01, 0.1, 0.3, 0.5, 1, 3, 5, 10$.



(c) Number of character strokes (ns) distribution across across the alphabet characters generated with different concentration parameters $\alpha = 3, 3.5, 4, 4.5, 5$. Error bars represent the standard deviation between analysis' runs.



(d) Percentage of repeated sampled primitive ids across the alphabet characters generated with different concentration parameters $\alpha = 3, 3.5, 4, 4.5, 5$.Error bars represent the standard deviation between analysis' runs.

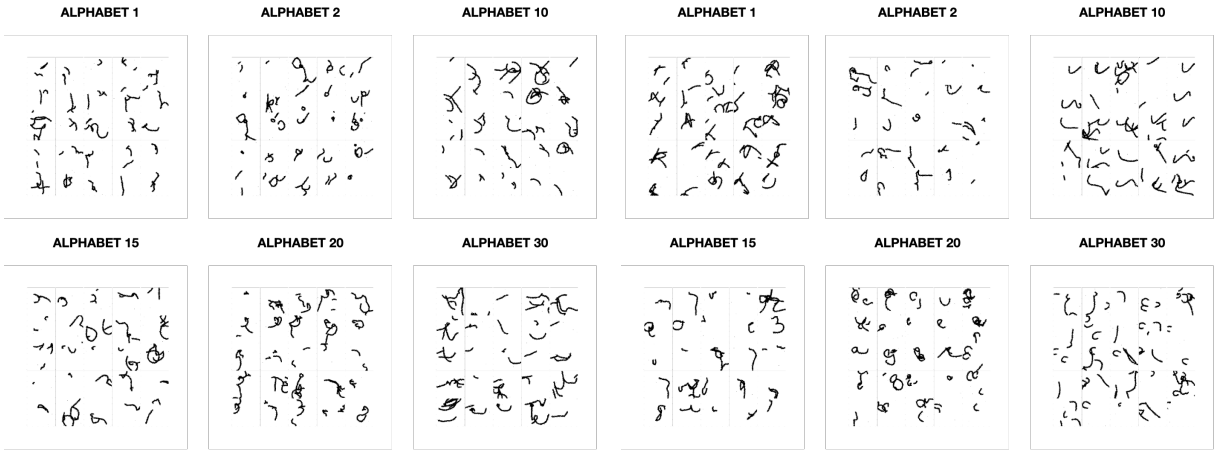
Figure 4.11: Dirichlet Process concentration parameter analysis.

and 4.11d show the obtained results. The standard deviation between the obtained results for the five alphabets generated with each α parameter were also computed, evidencing the inherent stochasticity of the generative process. It is noticeable that the latter results did not show any significant difference given the criteria expressed above. Given that, it was decided to use $\alpha = 4.5$ in the generation of every alphabet. After defining the α parameter, 30 alphabets were generated for each one of the perturbed models with $\beta = 1e-3, 0.2, 0.5, 0.8$. Figure 4.12 shows some examples of the generated alphabets.

During alphabet generation, character type parameters were optimized using gradient descent, in order to maximize the likelihood score under the prior $P(\Psi) = P(\kappa) \prod_{i=1}^{\kappa} P(S_i) P(R_i | S_1, ..., S_{i-1})$.

For each of the four different DL-BPL models, a new dataset were generated each one consisting of 30 new alphabets (15000 new characters). The omniglot dataset was then augmented with the 30 new alphabets (note that this process was also separately repeated for each one of the different perturbations

parameterized with $\beta = 1e-3, 0.2, 0.5, 0.8$).



(a) Generated alphabets by DL-BPL with $\beta = 1e-3$.

(b) Generated alphabets by DL-BPL with $\beta = 0.2$.

(c) Generated alphabets by DL-BPL with $\beta = 0.5$.

(d) Generated alphabets by DL-BPL with $\beta = 0.8$.

Figure 4.12: **Generated DL-BPL alphabets.** Some examples of the generated alphabets by the perturbed models with **(a)** $\beta = 1e-3$, **(b)** $\beta = 0.2$, **(c)** $\beta = 0.5$, **(d)** $\beta = 0.8$.

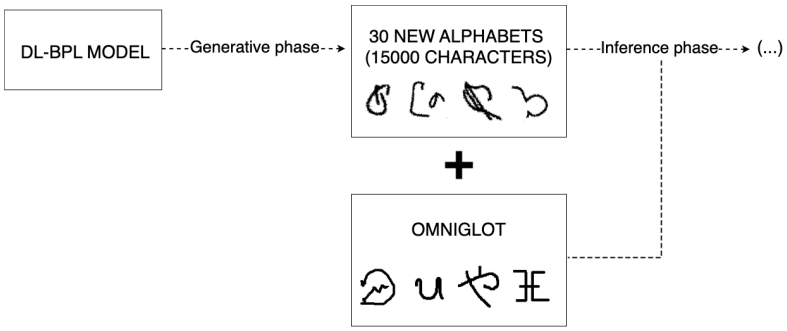


Figure 4.13: **Data augmentation**. Each one of the generative perturbed models with $\beta = 1e-3, 0.2, 0.5, 0.8$ originated a new dataset consisting of 30 alphabets, which itself comprises 15000 new characters, that were augmented to the omniglot dataset, resulting in four different datasets. The inference phase was repeated four times for each one of these four different datasets correspondent to perturbed data with the above β parameterizations.

The stochastic process of generating a new alphabet has shown some limitations in the context of this work. Despite the sampling tests that were performed and choosing a concentration parameter of $\alpha =$

4.5, it was observable that some characters within the same alphabet were still very similar, contributing to decreased variability within the entire dataset. Besides this, there are some evident differences when comparing omniglot characters with the ones generated by BPL. The generative process of BPL with no constraints is noticeably far way from the human process, seeming to represent mostly nonsensical doodles, sometimes even showing overlapping *strokes*. In contrast, omniglot characters have been drawn by humans, presenting a more structured representation. Moreover, it was also noticed that the ink of some character images was out of the frame. These are all factors which may later influence the inference process of inferring the latent primitive indexes of the alphabet characters.

The omniglot dataset was augmented with perturbed data. Then, the augmented omniglot dataset was used to infer a new generative model with the aim of establishing new priors ρ^*_{start} and ρ^*_{pT} assigning a higher probability to transitions that were previously relatively unlikely.

Note that data augmentation has been widely used in ML and Deep Learning as an approach to increase the size and diversity of training datasets without having to directly collect new data by bootstrapping from the original data. It has many applications within ML e.g. it acts as a regularizer and helps to avoid overfitting [238, 239]. Typical data augmentation algorithms in computer vision include geometric transformations (rotations, translations, flipping), color space augmentations, kernel filters applications, mixing images and random erasing. These are now being extended with the usage of deep learning approaches comprising adversarial training, generative adversarial networks, neural style transfer, and meta-learning search algorithms [238]. Other domains, such as natural language processing, can also benefit from data augmentation methodologies including swapping, deletion, random insertion, interpolation techniques among others [239]. Our methodology deviates from existing approaches as it involves the utilization of a diffusion-based perturbation framework on a generative latent space. This is achieved by embedding BPL primitives and subsequently applying a heat kernel to them, ultimately generating novel data. Within the context of this work, the execution of data augmentation necessitates an inference step that involves updating the model priors to account for new perturbed data. Our objective is to investigate the efficacy of diffusive perturbations, hyperparameter tuning, and inference-based data augmentation as a viable strategy for enhancing model performance.

4.3 Inference phase

In the inference phase of our pipeline, latent variables were inferred from the augmented datasets, with a particular focus on capturing the constitutent primitives of the dataset characters. The new probability distributions regarding the first sampled primitive in each character *stroke* (ρ^*_{start}) and the transitions between character *sub-strokes* (ρ^*_{pT}) were constructed by normalizing the empirical counts of the inferred primitives in the augmented omniglot data. Conceptually, this can be understood as a learning step, corresponding to learning of new priors ρ^*_{start} and ρ^*_{pT} for classification inference. The inference phase was run four times for the four generated datasets, each one of them corresponding to a distinct DL-BPL model perturbation. The BPL model received the processed character images and approximated them to their posterior, inferring the best parse for each one of the images, as was explained in

Section 3.3.3.

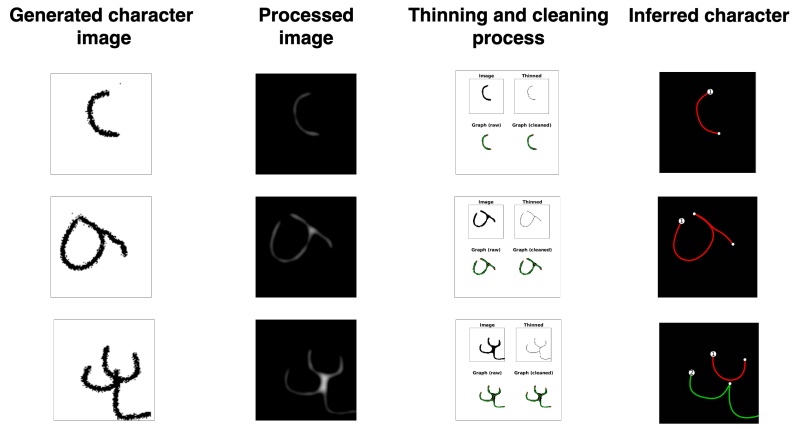


Figure 4.14: **Inference phase steps.** The augmented dataset character images (left column) were processed and fed to the original BPL model. A preprocessing algorithm is applied to the processed image (second column) as explained in 3.3.3, followed by thinning and cleaning process (third column). The final result of the inference process is a set of the model's latent variables, from which the inferred character primitive indexes are the ones of our interest. The inference result visualization (fourth column) consists of the character image parsing, where each color represents a different inferred *stroke*.)

During the inference process some of the generated character images of the perturbed datasets were excluded from the process, for various reasons, including cases such as (1) characters with only one *stroke* and *sub-stroke*, (2) produced on an extremely small scale, or (3) characters whose image ink went out of range of the image frame making them illegible. Ultimately, the number of characters, generated by each of the perturbed models and used to compute the new priors was stable and is described in Table 4.1.

Table 4.1: Number of inferred characters for the different β parameterized perturbations.

β **parameter**	**Inferred perturbed images**
$1e-3$	14715
0.2	14647
0.5	14788
0.8	14664

After inferring the four different omniglot augmented datasets and computing the primitive indexes' empirical countings, the final model priors ρ^*_{start} and ρ^*_{pT} were obtained.

The novel learned priors ρ^*_{start} and ρ^*_{pT} showed significant differences with respect to their sparsity, not only with respect to s and pT, but also with respect to ρ_{start} and ρ_{pT}, respectively. In this situation, sparsity is a measure that describes the percentage of a distribution which has a zero value. The approximate sparsity value of the original s distribution is 0.004, while the sparsity value of pT is 0. One may state that we are taking a strict perspective on sparsity, implying that a sparse matrix corresponds to have many of its entries equal to zero. However, it is significant to note that many entries in both of BPL original priors, specially in pT, are close to zero which can lead to considered it is "effectively" sparse from a finite sampling perspective.

It was possible to conclude that the sparsity values for ρ^*_{start} are one order of magnitude higher than

the original distribution. Besides this, the largest difference is evidenced in ρ^*_{pT}'s sparsity values, which are very close to a distribution sparsity of 100%, being the distribution mainly composed of zero values (Figure A.3 in supplementary material). These distribution differences are also evidenced in KLD and JSD measures, which are demonstrated in Figure 4.16.

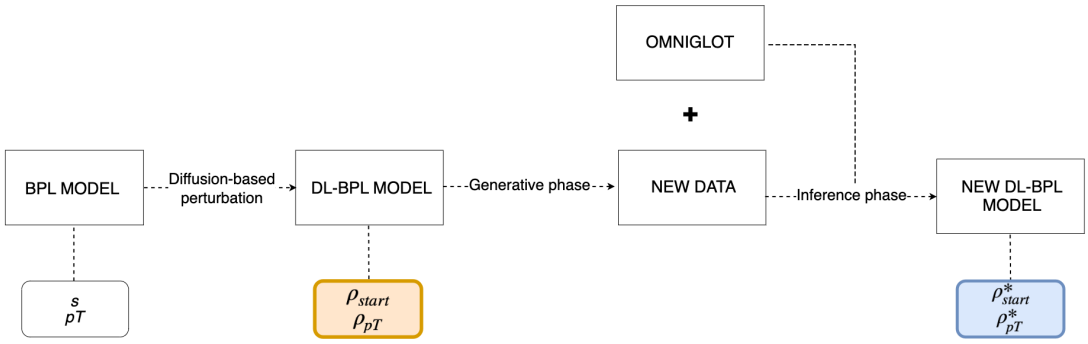


Figure 4.15: **Prior's scheme.**After perturbing the original priors (perturbed priors in orange), the generative and, subsequently, the inference phase led to the computation of the final distributions (final computed priors in blue), which were used in the classification task. In Figure 4.16 one can compare the KLD and JSD measures comparing the original prior distributions with the perturbed priors (in orange), resulting from the perturbation phase, and with the new learned priors (in blue), resulting from the inference phase.

KLD and JSD values regarding s and ρ^*_{start} distributions don't show any significant differences across β values. This leads us to the conclusion that the effect of the diffusive perturbations for the different β values was lost in the inference step. In other words, after the inference phase, when individually inferred, the different augmented datasets generated with the different β parameterized DL-BPL models, gave rise to new computed priors ρ^*_{start} that equally differ from the original s distribution. Additionally, the KLD and JSD measures seem to have a value at an approximate intermediate level of the range of the previous values measured for ρ_{start} and s as seen in Figures 4.16e and 4.16f, thus showing a bigger difference from s than some perturbed priors ρ_{start} ($\beta = 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$) and on the contrary, a lower difference than others ($\beta = 1e-15, 1e-6, 1e-3, 0.1, 0.2, 0.3$).

On the other hand, the differences in KLD and JSD measures relative to pT and ρ^*_{pT} distributions have shown to be more significant. In Figures 4.16g and 4.16h it is possible to observe that these measures have much higher values than the comparison between pT and ρ_{pT}, due to the fact that the new computed ρ^*_{pT} are highly sparse distributions, and, therefore, significantly differ from the original pT prior. Furthermore, similar to the analysis done for ρ_{start} and s, we can observe that the effect of the perturbation on the distribution for different values of β was also lost.

The observed sparsity in the originated ρ^*_{pT} distributions can be problematic regarding the model's classification task performance, jeopardizing it. For instance, the intuition is that the model might opt for inferring *strokes* with only one *sub-stroke* since there are not many "available" primitive transitions in ρ^*_{pT}. We investigated the problem of sparsity in ρ^*_{pT}. To understand the origin of this sparsity, we analyzed the statistics of the number character *strokes*, as well as the number of *sub-strokes* in each of those *strokes*, in both generated and inferred characters. Note that the statistics regarding the inferred characters do not include the excluded characters (see Table 4.1), however, working with these character numbers,

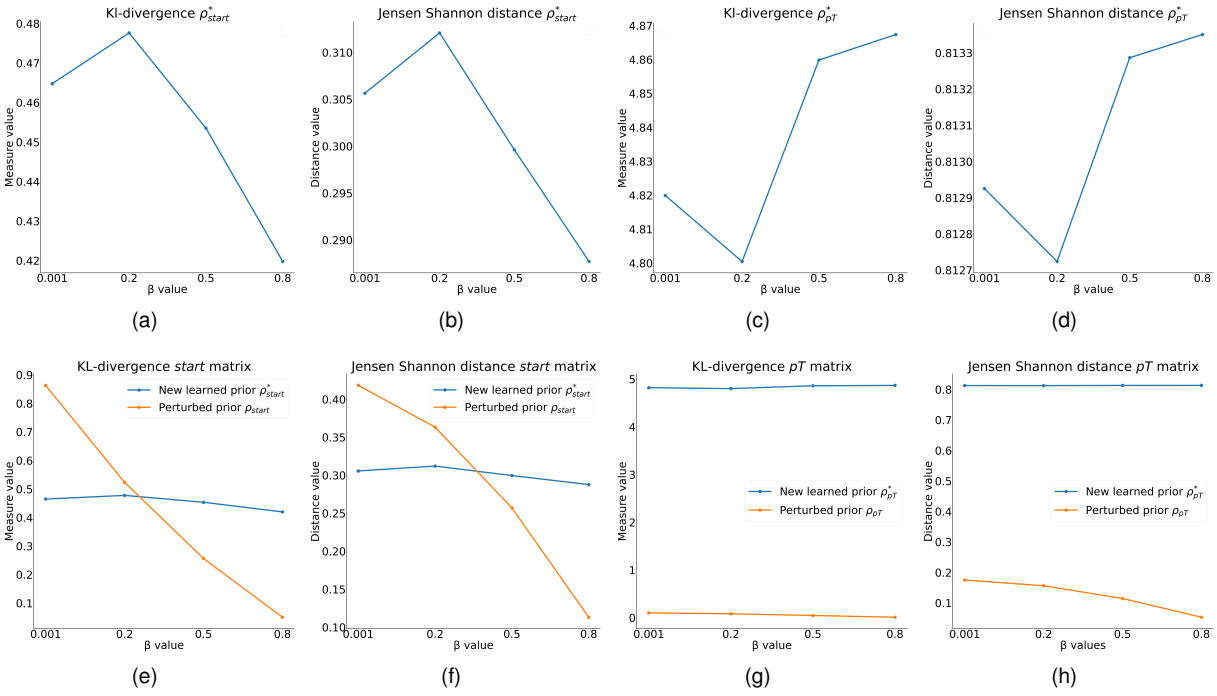


Figure 4.16: **(a)** $KL(s||\rho^*_{start})$. **(b)** $JSD(s||\rho^*_{start})$. **(c)** $KL(pT||\rho^*_{pT})$. **(d)** $JSD(pT||\rho^*_{pT})$. **(e)** Comparing $KL(s||\rho_{start})$ and $KL(s||\rho^*_{start})$. **(f)** Comparing $JSD(s||\rho_{start})$ and $JSD(s||\rho^*_{start})$. **(g)** Comparing $KL(pT||\rho_{pT})$ and $KL(pT||\rho^*_{pT})$. **(h)** Comparing $JSD(pT||\rho_{pT})$ and $JSD(pT||\rho^*_{pT})$.

still allows us to get some idea of what might be happening.

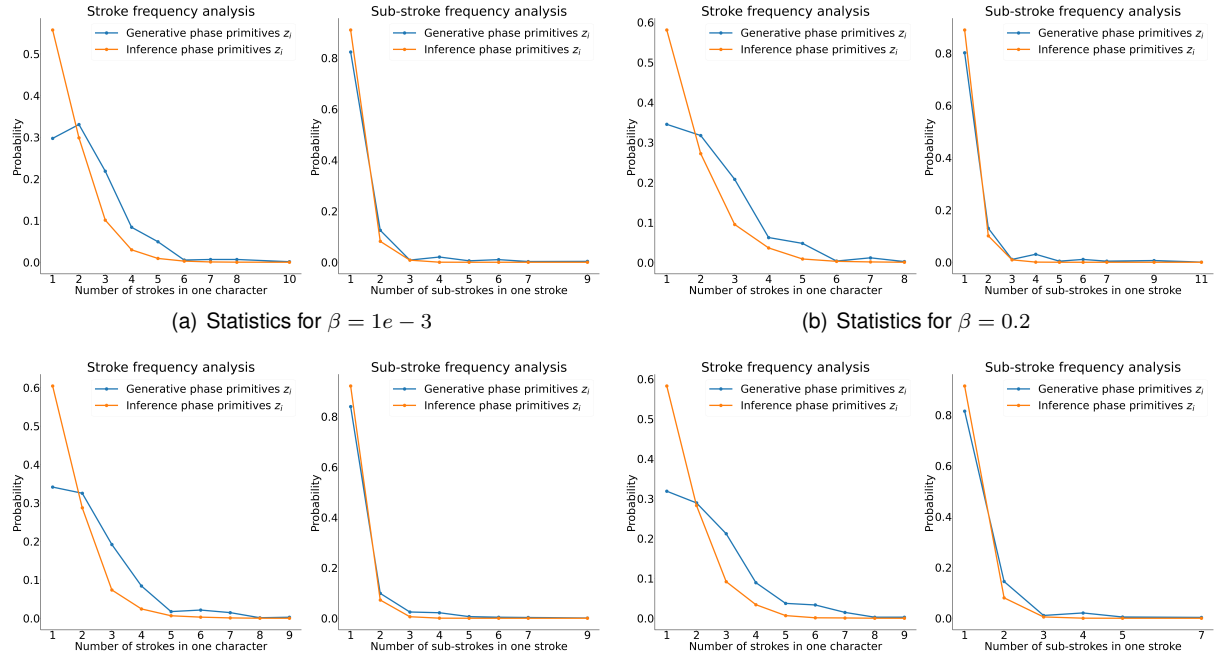


Figure 4.17: Comparing *stroke* and *sub-stroke* statistics of the generated character primitive indexes with the inferred character primitive indexes of the generated perturbed dataset for the different DL-BPL models. Despite statistical differences, the generative and inference process seem to be following the same ”behaviour”.

Our understanding of this problem led us to believe that, when inferring the character's latent vari-

ables, the percentage of character *strokes* with more than one *sub-stroke* was inferior to the percentage of character *strokes* with only one *sub-stroke*. This is because if *strokes* only have one *sub-stroke*, the counting of that one *sub-stroke* primitive will be accounted for the computation of ρ^*_{start}, and not for the computation of ρ^*_{pT}, since there is no transition to be considered.

The performed analysis came to confirm this, revealing that, for the four different β parameterized perturbations, the majority of generated characters had only one *stroke* and a prominent number of *sub-strokes* per *stroke* of one. Interestingly, we were able to conclude that the problem of sparsity originated in the generative phase, which produced an increased number of characters with only one *stroke*, as well as an increased number of *strokes* with only one *sub-stroke*. Even though the generative phase seems to be the problem seed, in Figure 4.17 these probabilities seem to be even more increased after the inference phase, evidencing some discrepancies between the statistics of generated and inferred characters. It is important to note that this issue could have been addressed by conditioning the generative phase to only generate characters with more than one *stroke*, and more than one *sub-stroke* per *stroke*. However, this conditioning was not part of our objectives in this study. Additionally, this is in agreement with our expectations of finding a higher frequency of one and two *stroke* characters, based on the statistics of the original model (see references [6, 7] for details). Nevertheless, the obtained *sub-stroke* statistics does not seem to be in such agreement, with higher number of one *stroke* characters; we would be expecting to see a higher number of two *sub-strokes* per *stroke*.

In summary, our analysis highlights the importance of examining the *stroke* and *sub-stroke* statistics of both generated and inferred characters, leading us to a clear view of how these negatively influenced the computation of ρ^*_{pT} and, on the other side, why ρ^*_{start} most closely resembles the original prior. These findings informed posterior work developing an alternative pipeline explored in Section 4.5. Despite these observations and the understanding that inference followed by empirical primitive frequency countings are strongly influenced by sample variance, we proceeded to use the newly computed distributions in the last phase of the pipeline, the classification task.

4.4 Classification phase

The ability of a model to adapt and appropriately respond to novel, previously unobserved data is referred to as generalization. Generalization evaluates how effectively a model can process new data to produce accurate predictions following trained. The final phase of our pipeline was a 20-way classification task (see Section 3.3.4 for a complete description). The classification task was used to test the generalization capabilities of each of the updated DL-BPL models we trained with different β parameter values. The classification step in our pipeline was applied to each of these DL-BPL models separately. As explored in Section 3.2.2, this one-shot classification task is interpreted as a measure of a model's generalization after a computational experiment analogous to a psychedelic experience. From this perspective, a lower one-shot classification error corresponds to better model generalization with respect to the evaluation dataset.

Considering the previous results' analysis, our first intuition was that the structure, namely the high

ρ^*_{pT} distribution sparsity, of the new model priors will hinder the classification task relative to the original model results. Nevertheless, for the purposes of completion, this analysis was performed. Each run of the 20-way classification task corresponded to classifying 20 test images into one of the classes of 20 training images (images showed in Appendix A.1). This classification consisted of a fitting step, in which both the 20 test images' and the 20 training images' latent motor programs were inferred, and a second re-fitting step, where test images were re-fitted to the training images, resulting this the decision Bayesian classification score seen in Equation 3.43 (see Section 3.3.4 for details). In the next sections, each one of these intermediate steps' results will be explored and analysed.

4.4.1 Fitting test and train images

Fitting the test and training images in each run corresponded to performing inference on these images and getting the corresponding inference score i.e. log-likelihood (Equation 3.35) of each image. Note that a control classification task was run utilizing the non-perturbed original BPL model ($\beta = 1$). Inference was done on every test and training image for the 20 runs, using the four different diffusion-perturbed models. A few examples of the obtained fitting results are shown in Figure 4.18a.

Figure A.4 in the supplementary material shows a complete example for all the perturbed models. By visual inspection, it is possible to observe a much more accurate representation of the image parses done by the original model than by the perturbed models, which seem to show greater difficulty in representing all of the $K = 5$ best image parses in an accurate way, getting further away from the image representation in the last parses. This effect might be explained by the ρ^*_{pT}'s sparsity observed in every model perturbation, which will have a big impact on characters with *strokes* with more than one *sub-stroke*.

In order to have a better understanding of how the inference results of each model will influence the final Bayesian classification score and how the inference performance of each model, the average inference scores of all of the evaluation dataset images used in the classification task were assessed, as demonstrated in Figure 4.18b.

It is possible to conclude that the average fitting score is much higher for the original BPL model, having a value of -493.35, while the obtained average fitting scores for the models resulting from the perturbations with $\beta = 1e-3, 0.2, 0.5, 0.8$, were -2508.28, -2504.91, -2506.85, -2502.58, respectively, showing very similar score results. The similarity observed in the inference scores, obtained by the perturbed models, indicates that the effect of the perturbation was lost in the inference process, which is in agreement with the conclusions drawn after the inference phase.

4.4.2 Re-fitting test and training images

The second step of the classification task was re-fitting. Individually leveraging the four different DL-BPL models, the 20 runs were executed, where the 20 training images were re-fitted to the 20 test images, and vice-versa (the Bayesian classification score is a two-way score) resulting in 20 classification matrices, as the one example in Appendix A.6, where rows are the test image examples, columns are

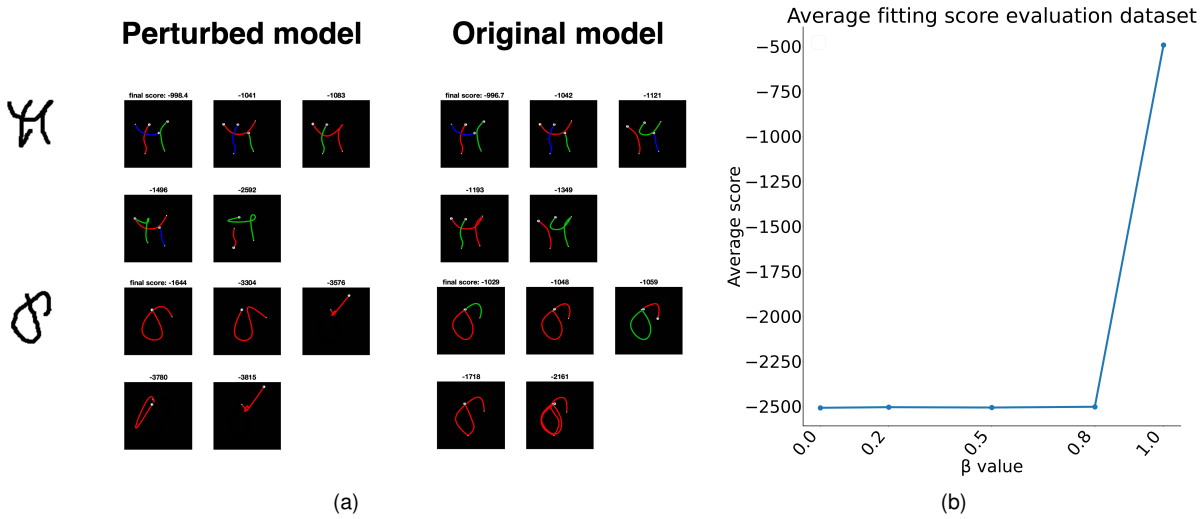


Figure 4.18: **(a)** Few examples of fitting results of the run 1 test images leveraging the original model (on the right) and the perturbed models (on the left) with $\beta = 1e-3$. **(b)** Average fitting score of evaluation dataset images for the different perturbation β parameters and for the original BPL model.

training examples and the matrix entries correspond to the final re-fitting score given by Equation 3.43. Remember that, in every run, the corresponding test images and training image classes have the same number (i.e., test image 1 belongs to training image class 1, test image 2 belongs to training image class 2, etc.). Figure 4.19 shows two examples, one with correct classification and one misclassified, using the DL-BPL perturbed with $\beta = 1e-3$.

4.4.3 One-shot classification results

The one-shot classification results for every run and every model perturbation are represented in Figure A.5 in the supplementary material. The average errors for the different model perturbations are illustrated in Figure 4.20. Lake et. al (2015) [6] performed a one-shot classification task leveraging BPL, obtaining an error of 3.3%. After studying BPL in detail and trying to replicate every step of the one-shot classification running simulation, we were not able to reproduce the author's paper results, achieving a 9% error. For the purposes of this work, we will take our control experiment as the baseline result. That being said, from these results, it is possible to conclude that the classification performance of the perturbed models was significantly worse than the original model's performance 9% classification error.

It is possible to identify several factors that may have led to these poor results across the various stages of the pipeline. After the perturbation phase and subsequent generative phase, it was possible to observe that, for the four explored perturbations, the high number of perturbed characters which mainly consisted of *strokes* with only one *sub-stroke* had a deleterious effect on the inference phase, resulting in problems when computing ρ^*_{pT}. Consequently, the sparsity of the obtained ρ^*_{pT} distributions, might have been the main problem in the fitting and re-fitting process that led to high classification error. Even though, in the fitting process it was possible to conclude that the ρ^*_{start} distribution still shows a similar structure to the original distribution, it was observed that the model experienced some

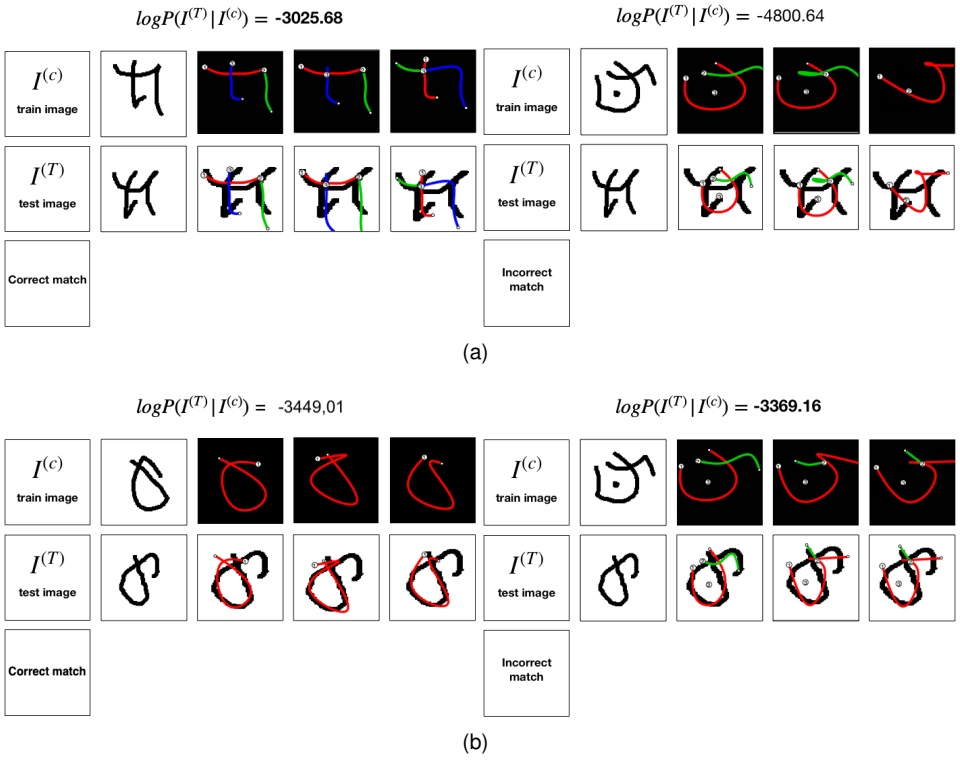
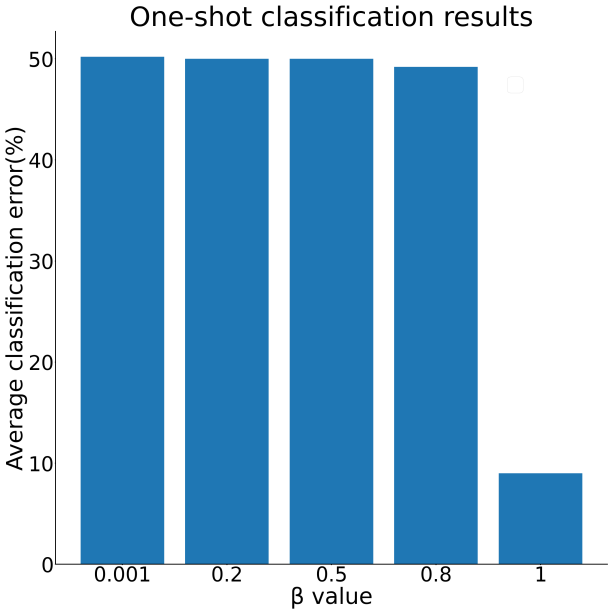


Figure 4.19: **Classification examples.** The two examples illustrate two different training images being re-fitted to a test image. The top-three first posterior parses of two different training images are shown in the top row, and their re-fits to the same test image are shown in the second row. **(a)** Correctly re-fitting classified example, the correct correspondent training image reports a higher final classification score, indicating that $I^{(T)}$ is well-explained by the motor programs of $I^{(c)}$. **(b)** Misclassified re-fitting example, this time the higher classification score does not correspond to the right training image class. Note that the higher Bayesian classification scores $logP(I^{(T)}|I^{(c)})$ are indicated in bold.



β parameter	Average one-shot classification error
$1e-3$	50.2%
0.2	50%
0.5	50%
0.8	49.2%
1	9%

Figure 4.20: Average one-classification error for each perturbed DL-BPL model and control (original non-perturbed BPL).

difficulties when inferring character motor programs after inferring the first and second best parses.

Under those circumstances, the fact that the computation of the ρ^*_{start} parameter vector obtained a reasonable structure was not enough for good classification performance in the model. Furthermore, it was noticed that, similarly and accordingly to what happened in the inference phase after the new prior's computation, the model's performance was significantly close for the four different model perturbations, evidencing the loss of the diffusion-perturbation effect. In light of these results, an alternative pipeline was developed to try to overcome what we identified as the major bottleneck in this pipeline.

4.5 Alternative pipeline

The second exploratory step in this work consisted of adapting the first developed pipeline and specifically redesigning the inference phase in order to circumvent the issue that was identified and described in the previous section. To do this, all phases of the originally developed pipeline were repeated starting from the inference phase.

An alternative method for the inference phase was developed, where the main difference consisted in not feeding the inference model with the augmented dataset, comprising the omniglot plus the 30 generated perturbed alphabets, but only with the latter. Under these circumstances, the priors ρ^*_{start} and ρ^*_{pT} were computed only with data resulting from inference of the new perturbed data, using the same previously utilized method of empirical countings and, secondly, an integrative estimation combining the computed priors and the original priors was calculated, in order to arrive at the final priors, now designated by e_{start} and e_{pT}, which were then used in the classification task. This estimate took into account the original priors, learned from the omniglot dataset [6], but also the "intermediate" distributions ρ^*_{start} and ρ^*_{pT}. Although some limitations regarding the generative phase have been previously identified, with this alteration in the inference phase, it was decided to proceed with the same data generated in the generative phase, placing no constraints on the model's generative process, other than the DP. We suggest that the development of new computational mechanisms in the generative phase of our pipeline may provide fruitful avenues for future work (see Section 5.2 for further discussion).

Given these aspects, the inference phase was performed for each of the sets of new alphabets generated with the model perturbations parameterized with $\beta = 1e-3, 0.2, 0.5, 0.8$. The estimation of e_{start} and e_{pT} consisted of:

- The entries of the matrices with a non-zero value were selected, in the case of the ρ^*_{start} corresponding to the first sampled primitives in a character *stroke* with probability different from zero, and, in the case of the ρ^*_{pT}, to the transitions between the primitives of the remaining *sub-strokes* in a *stroke*, with a probability of occurring different from zero.

- From these selected entries, those considered to be novel were identified. Novelty in the generated characters was defined as the observation of character primitives that showed a low sampling probability in the original priors. To define the "low sampling probability in the original priors", a range including the five lowest probability values for the *stroke*'s first sampled primitives, corresponding to the interval of 0 to $1.3319e-4$ in s distribution as well as the five lowest probability

values for the following *stroke*'s primitive transitions ranging from $8.5130e-8$ to $1.6140e-7$ in the pT distribution, corresponding to the following *stroke*'s primitive transitions. Besides this, to consider a first sampled *stroke* primitive or a primitive transition as being novel the thresholds $\rho^*_{start}(x) > 1.3319e-4$ and $\rho^*_{pT}(x,y) > 1.6140e-7$, respectively, were defined.

- After the identification of novel entries, the integrative estimation combining the original priors and the computed priors was performed, which consisted of: **(1)** The values of the entries considered as novel in the computed priors were replaced in the corresponding entries of the final priors' estimation; **(2)** The values of the entries of the original priors, corresponding to the entries of the computed priors with value zero, were substituted in the corresponding entries of the final priors' estimation; **(3)** Lastly, the values of the remaining entries in the final priors' estimation were obtained by a weighted sum between the values of the original and computed priors, where the weights were calculated proportionally to the number of characters in the omniglot and the number of characters in the perturbed alphabets, respectively.
- Finally, the estimated priors were normalized.

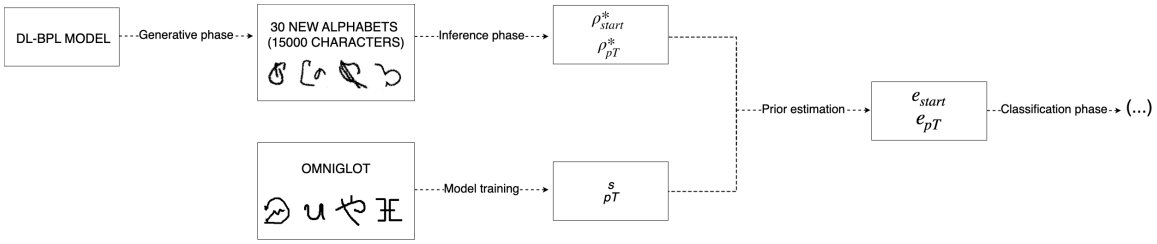


Figure 4.21: Alternative pipeline.

The following equations 4.1 and 4.2 represent the process described above.

$$e_{start}(x) \leftarrow \begin{cases} \rho^*_{start}(x), & \text{if } 0 < s(x) < 1.3319e-04 \\ & \quad \text{and } \rho^*_{start}(x) > 1.3319e-04 \\ s(x), & \text{if } \rho^*_{start}(x) = 0 \\ w_1 s(x) + w_2 \rho^*_{start}(x), & \text{otherwise} \end{cases} \tag{4.1}$$

$$e_{pT}(x,y) \leftarrow \begin{cases} \rho^*_{pT}(x,y), & \text{if } 8.5130e-08 < pT(x,y) < 1.6140e-07 \\ & \quad \text{and } \rho^*_{pT}(x,y) > 1.6140e-07 \\ pT(x,y), & \text{if } \rho^*_{pT}(x,y) = 0 \\ w_1 pT(x,y) + w_2 \rho^*_{pT}(x,y), & \text{otherwise} \end{cases} \tag{4.2}$$

The resulting weights are shown in Table A.2 in supplementary material, where w_1 is the original prior weight and w_2 the new computed prior weight. It is important to notice that the number of inferred images

are the same as the ones describes in Section 4.3. The definition of novelty and the hyperparameters it involves are a subject that should be further explored.

To explore the new alternative pipeline the data set generated with the DL-BPL models parameterized with $\beta = 0.8, 0.5, 0.2, 1e-3$. What is intended with this estimation is to get the priors that will be used in the classification task to capture the novel primitives and novel primitive transitions observed in the perturbed characters, while simultaneously retaining the structure of the original priors. When compared to the first pipeline learned priors ρ^*_{start} and ρ^*_{pT}, the new estimated priors e_{start} and e_{pT} demonstrate a higher similarity to the original priors, with lower KLD and JSD values. On one side, distribution comparison between s and e_{start} exhibited decreasing KLD and JSD values with increasing β, showing that the effect of the perturbation was preserved. On the other hand, distribution comparison between pT and e_{pT} show relatively stable KLD and JSD values across β parameters, as shown in Figure 4.22. These results seem to reflect what was desired.

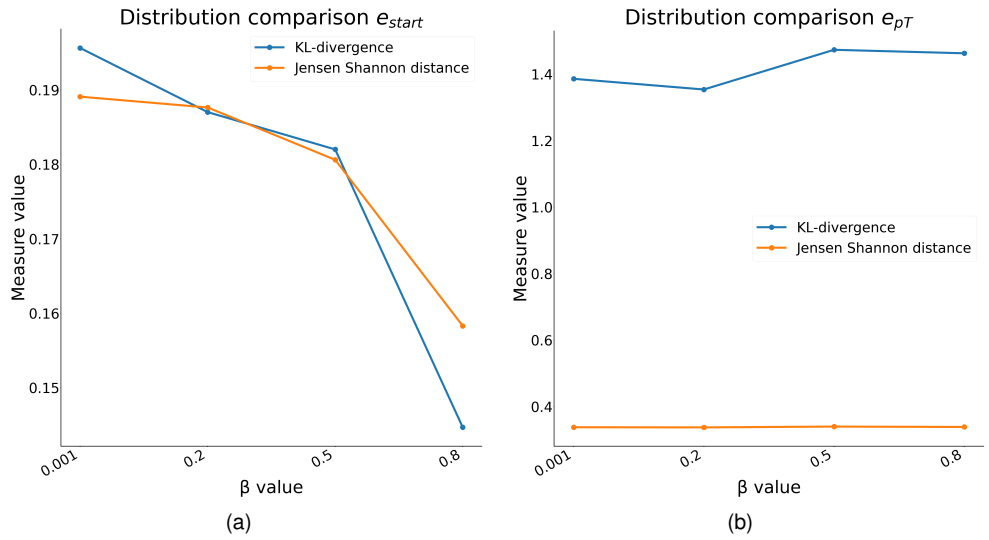


Figure 4.22: **(a)** $KL(s||e_{start})$ and $JSD(s||e_{start})$. **(b)** $KL(pT||e_{pT})$ and $JSD(pT||e_{pT})$

.

From Figure 4.23a it is possible to conclude that the average fitting scores of the evaluation images leveraging the DL-BPL are very close to the control score, without a significant difference between scores across β values. Figure 4.23b shows that the best classification average error was 7% for $\beta = 0.8$, slightly lower than the control 9% result. The average error for $\beta = 1e-3$ was 7.5%, for $\beta = 0.2$ was 8% and for $\beta = 0.5$ the obtained error was 9.5%.

The omniglot dataset was developed in the context of the work by Lake et. al (2015) [6], with the objective of studying how humans and machines perform one-shot learning. Seven years after its release, progress has been made in performing a one-shot classification task using this dataset [240]. Different study approaches, including the usage of different models and augmented datasets, have tried to tackle the challenge of obtaining a smaller one-shot classification error than BPL.

Results of this one-shot classification task include human participants which have shown to achieve an error rate of 4.5%, BPL has demonstrated comparable performance to humans, achieving an error

Figure 4.23: **Classification results alternative pipeline. (a)** Average fitting score of evaluation dataset images for the differently β parameterized DL-BPL models resulting from the alternative pipeline and for control (original non-perturbed BPL). **(b)** Average one-shot classification error for each perturbed DL-BPL and control.

rate of 3.3% [4]. In addition, Lake and colleagues (2015) [6] trained a basic convolutional neural network (ConvNet) for the identical task, which achieved a one-shot error rate of 13.5%. However, the most effective neural network model at the time was a deep Siamese ConvNet, which obtained an error rate of 8.0% after being trained with considerable data augmentation [240, 241]. Nonetheless, this error rate is still approximately twice that of human performance and Bayesian program learning (BPL). Adapted from the omniglot "3-year progress report", Figure 4.24 shows some other interesting results obtained since the omniglot release [240] by models such as matching networks (Matching Net) [242], prototypical networks (Prototypical Net) [243], model-agnostic meta-learning (MAML) [244], recursive cortical networks (RCNs) [245], variational homoencoder (VHE) [246], graph neural networks (Graph Nets) [247] and attentive recurrent comparators (ARCs) [248].

The majority of models do not integrate the compositional or causal architecture of character formation beyond implicit learning through numerous instances of character discrimination. Additionally, alternative forms of this task make it arduous to directly compare their performance with the original BPL results. A more recent result consists in the attempt of a generative neuro-symbolic (GNS) model [249] [5] to perform "within alphabet" one-shot classification, obtaining a test error rate of 5.7%, with no data augmentation, outperforming all other models that received the same background training, except for BPL [249]. This thesis work was also an attempt to improve performance on this one-shot classification task under an analogy of how psychedelics act on the brain. One important thing to mention is that we did not train the model from scratch with the augmented dataset, but instead we developed a different pipeline design. As previously mentioned, we wanted to take advantage of BPL model architecture and exclusively perturbed the priors which define primitive sampling, through an innovative diffusion-based

[4] Note that this is outcome reported in Lake's (2015) [6] scholarly work ; however, our attempts to replicate the outcome did not yield the same result.

[5] GNS is a model of handwritten character concepts, based on BPL framework.

	Original		Augmented	
	Within alphabet	Within alphabet (minimal)	Within alphabet	Between alphabet
background set				
# alphabets	30	5	30	40
# classes	964	146	3,856	4,800
2015 results				
Humans	$\leq 4.5\%$			
BPL	**3.3%**	**4.2%**		
Simple ConvNet	13.5%	23.2%		
Siamese Net			8.0%*	
2016-2018 results				
Prototypical Net	13.7%	30.1%	**6.0%**	4.0%
Matching Net				6.2%
MAML				4.2%
Graph Net				**2.6%**
ARC			**1.5%***	**2.5%***
RCN	7.3%			
VHE	18.7%			4.8%

Figure 4.24: **One-shot classification error rate across models.** "One-shot classification error rate for both within-alphabet classification [6] and between-alphabet classification [242], either with the ' "Original" background set or with an "Augmented" set that uses more alphabets (# alphabets) and character classes for learning to learn (# classes). The best results for each problem formulation are in bold , and the results for "minimal" setting are the average of two different splits." * Results used additional four-fold class augmentation by applying 90 degree rotations and additional random augmentations such as scaling, shearing, translations, etc. [249]. Adapted from Lake et al., (2019) [240].

perturbation process, ultimately augmenting the omniglot dataset in 750 new character classes, totaling 1714 classes (# characters). Besides this, all characters classes in a task episode come from the same alphabet as originally proposed [6]. The best obtained result was 7%, resulting from the perturbation with $\beta = 0.8$, showing a promising result when compared to other models that augmented the dataset in approximately twice the character classes we used and to the control test.

Concluding, our psychedelic analogy hypothesis resulted in a positive result, possibly being a first step towards high-level computational modeling of psychedelics, but also showing a promising avenue in the investigation of diffusive data augmentation in the probabilistic induction field. It is, nevertheless, imperative to emphasize that the omniglot one-shot classification task challenge involves more than just learning from a significant amount of background training and minimal inductive biases to tackle a single task. Instead, the challenge lies in learning from a limited amount of background training while considering the inductive biases that humans bring to the domain (whatever those biases may be) [240].

In addition to the conclusions drawn above, it is worth mentioning that the task of identifying novelty in the augmented data sets was found to be challenging. To gain insight into this, we adopted the definition of affinity proposed by Gontijo-Lopes et al. (2020) [250] for a post pipeline analysis. This metric measures the extent to which an augmentation alters the training data distribution learned by the model, being sensitive to properties of both the model and data distribution. Specifically, affinity is defined as the difference between the validation accuracy of a model trained on the original data set and evaluated on the original evaluation data set, and the accuracy of the same model evaluated on an augmented evaluation set (see Section A.8). We generated an evaluation data set of the same size as the omniglot evaluation data set with the DL-BPL model perturbed with a β value of 0.8, and performed the one-shot classification task on it. The resulting one-shot classification error for the DL-BPL evaluation set was 22.2%, indicating an affinity of -13.2%. These results suggest that the data generated by the diffusively perturbed model contains novelty that is out-of-distribution for the BPL model. Further analyses are

necessary to investigate the remaining beta parameterized perturbations, as well as additional analysis to explore the diversity metric (also introduced in [250]), which together could be useful in optimizing the beta parameterized diffusive augmentation.

4.5.1 Computational modeling in psychedelic research

The application of machine learning models to investigate the impact of psychedelic drugs on cognitive processes represents a relatively underexplored research area, lacking in literature. This can be attributed in part to the challenges of developing accurate computational models that reflect the complex biological mechanisms underlying the effects of these drugs.

Studies simulating the visual hallucinations phenomenology that happens during the psychedelic experience have been the most explored, including the "Deep Dream" algorithm [251] and the "Hallucination Machine" [252] that have shown to simulate biologically plausible and ecologically valid visual hallucinations, and provide a powerful tool to complement the recent resurgence of research into altered states of consciousness [252]. Inspired by these studies, more recent work presented the output of two deep convolutional neural network architectures resulting in visual features reminiscent of descriptions of psychedelic-induced visual imagery, exemplifying a psychedelic perturbation via N,N-DMT [253], moving towards the conceptualization of potential biological mechanisms of the balanced integration of exogenous and endogenous information into conscious experience/visual perception mediated by the serotonergic system [253] (see Sections 2.1.2 and 2.1.4).

The existing work on psychedelic drugs focuses on simulating visual hallucinations, but this new modeling approach aims to explore higher-order cognitive hierarchies by perturbing internal representations at abstract levels of knowledge integration and concept formulation. We develop an expansive multi-phase framework for cognitive processing (i.e. estimation, generation inference, and classification) such that the impact of psychedelic perturbations on each phase may lead to distinct and interacting effects on the subjective experience. More specifically, we propose a novel data augmentation approach, namely diffusive latent space perturbations in the context of probabilistic program models as an alternative approach to computationally formalizing how psychedelics might lead to new perspectives in PAP.

Chapter 5

Conclusions

5.1 Summary

This thesis introduced a novel framework to explore the high-level computational theories of psychedelic action in the brain based on probabilistic program induction. Our proposed pipeline has four different phases, namely diffusion-based perturbations, generative, inference and classification, that work together with the objective of improving BPL model's performance in a classification task through a data augmentation procedure originated by diffusively perturbing in the latent space of the generative component of our model.

Due to the exploratory nature of this work, comprising new perspectives in both ML and psychedelic phenomenology, our pipeline design initially encountered some limitations, namely in the inference phase, resulting in some poor classification results. Nevertheless, a detailed analysis was performed in order to identify the main bottleneck and an alternative pipeline was developed to bypass the problem.

The computational experiments corroborated our theoretical hypothesis regarding the diffusion perturbation effect on the model priors. By differentially parameterizing the diffusive perturbations we were to observe the effects of β hyperparameter value in the classification task results, despite not being able to establish a positive correlation between decreasing β and increased model performance. We were not able to reproduce the results in the original BPL paper [6], and, therefore, the control classification test using the original model served as our baseline error. With respect to this baseline, our DL-BPL pipeline obtained a lower error classification result for the highest β value, $\beta = 0.8$, evidencing how the stochasticity of the generative character process, as well as finite sampling is intrinsically related to the induced data set novelty. We believe these results, nevertheless, show this methodology is a promising avenue of investigation to further refine and explore in the context of probabilistic induction.

5.2 Limitations and future work

The present findings highlight the importance of conducting detailed analysis to improve pipeline design and hyperparameter choices. To optimize the exploration of the β perturbation parameterization,

Bayesian optimization using Gaussian processes could be employed, as testing the entire pipeline with different hyperparameters is computationally and time expensive. The process of updating the new model priors also further requires careful consideration. One idea is to condition the generative process to only produce characters with more than one *stroke* and more than one *sub-stroke* per *stroke*, which can help avoid problems encountered during the inference phase. Additionally, exploring the direct computation of priors from the generated alphabets without the need for inference could be an interesting avenue to pursue. Defining novelty in the perturbed alphabets proved to be a significant challenge. To address this, quantile analysis could be employed to define probability thresholds for Equations 4.1 and 4.2 and new methods to define novelty should be investigated. Furthermore, in addition to updating primitive sampling priors, training the model with the augmented dataset should also be considered.

Another potential area of exploration is diffusion-based perturbations in the context of recent work such as the GNS framework [249] that incorporates the hierarchy of BPL and neural networks. Besides this, exploring models with different types of data can be another area of investigation.

In the realm of computational modeling and psychedelic research, neural networks offer a promising avenue for exploring the action of psychedelic drugs at the circuit level in the brain,by approximating models to biology. Neural networks have been inspired by biological neural networks and can be more easily mathematically analyzed than their natural counterparts at the neuronal level [254]. One potential application of machine learning in this area is to mathematically model neuroplasticity, modeling the pharmacodynamic effects of psychoactive drugs on spiking neural network plasticity. By experimenting with network behavior before and after these modifications and comparing them to human neural network behavior, researchers can explore the effects of psychedelic drugs on the brain [254]. Another possible area of investigation is the hippocampus, which is thought to be involved in cognitive abilities related to past experience, spatial mapping, planning and imagination [255]. By modeling hippocampal circuits and the generative process underlying these circuits [256], researchers might be able to better understand the cognitive effects of psychedelic substances.

While psychedelics have shown tremendous potential to revolutionize therapy, the reasons behind their positive long-term outcomes are not fully understood. One possibility is that the strong subjective experiences people have during PAP sessions play a substantial role in this (see Section 2.1.6). In fact, some novel work has attempted to quantitatively map people's reports of psychedelic experiences to drug receptor binding affinities, gene transcription profiles, and brain structure, translating a person's experience into the molecular profiles responsible for that experience, allowing researchers to identify which specific psychedelic compounds could benefit specific patients [257]. In addition, we believe that putting more effort in modeling psychedelic action not only at the circuit, but also, at the high cognitive level (and not on simple visual hallucinatory phenomenology) should be taken into consideration when trying to fill the gaps in the research field.

We conclude that these types of studies may be a path to explore when it comes to improved psychiatric treatment [257] and precision medicine.

Bibliography

[1] M. Pollan. *How to change your mind: What the new science of psychedelics teaches us about consciousness, dying, addiction, depression, and transcendence*. Penguin, 2018.

[2] R. F. Leger and E. M. Unterwald. Assessing the effects of methodological differences on outcomes in the use of psychedelics in the treatment of anxiety and depressive disorders: A systematic review and meta-analysis. *Journal of Psychopharmacology*, 36(1):20–30, 2022.

[3] B. Romeo, L. Karila, C. Martelli, and A. Benyamina. Efficacy of psychedelic treatments on depressive symptoms: A meta-analysis. *Journal of Psychopharmacology*, 34(10):1079–1085, 2020.

[4] C. M. Reiff, E. E. Richman, C. B. Nemeroff, L. L. Carpenter, A. S. Widge, C. I. Rodriguez, N. H. Kalin, W. M. McDonald, W. G. on Biomarkers, and a. D. o. t. A. P. A. C. o. R. Novel Treatments. Psychedelics and psychedelic-assisted psychotherapy. *American Journal of Psychiatry*, 177(5): 391–410, 2020.

[5] R. L. Carhart-Harris and K. Friston. REBUS and the anarchic brain: toward a unified model of the brain action of psychedelics. *Pharmacological reviews*, 71(3):316–344, 2019.

[6] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

[7] B. M. Lake. *Towards more human-like concept learning in machines: Compositionality, causality, and learning-to-learn*. PhD thesis, Massachusetts Institute of Technology, 2014.

[8] G. E. Hinton, P. Dayan, B. J. Frey, and R. M. Neal. The" wake-sleep" algorithm for unsupervised neural networks. *Science*, 268(5214):1158–1161, 1995.

[9] K. Ellis, C. Wong, M. Nye, M. Sable-Meyer, L. Cary, L. Morales, L. Hewitt, A. Solar-Lezama, and J. B. Tenenbaum. Dreamcoder: Growing generalizable, interpretable knowledge with wake-sleep Bayesian program learning. *arXiv preprint arXiv:2006.08381*, 2020.

[10] R. L. Carhart-Harris and G. M. Goodwin. The therapeutic potential of psychedelic drugs: past, present, and future. *Neuropsychopharmacology*, 42(11):2105–2113, 2017.

[11] J. Gardner, A. Carter, K. O'Brien, and K. Seear. Psychedelic-assisted therapies: The past, and the need to move forward responsibly. *International Journal of Drug Policy*, 70:94–98, 2019.

[12] J. Gonzalez-Maeso and S. C. Sealfon. Agonist-trafficking and hallucinogens. *Current medicinal chemistry*, 16(8):1017–1027, 2009.

[13] P. Celada, M. V. Puig, L. Díaz-Mataix, and F. Artigas. The hallucinogen DOI reduces low-frequency oscillations in rat prefrontal cortex: reversal by antipsychotic drugs. *Biological psychiatry*, 64(5): 392–400, 2008.

[14] R. L. Carhart-Harris, S. Muthukumaraswamy, L. Roseman, M. Kaelen, W. Droog, K. Murphy, E. Tagliazucchi, E. E. Schenberg, T. Nest, C. Orban, et al. Neural correlates of the LSD experience revealed by multimodal neuroimaging. *Proceedings of the National Academy of Sciences*, 113(17):4853–4858, 2016.

[15] D. Gems. Alexander shulgin and ann shulgin, pihkal, a chemical love story. alexander shulgin and ann shulgin, tihkal, the continuation. *Theoretical Medicine and Bioethics*, 20(5):477–479, 1999.

[16] B. C. Labate and C. Cavnar. *Ayahuasca shamanism in the Amazon and beyond*. Oxford Ritual Studies, 2014.

[17] C. Letheby. The epistemic innocence of psychedelic states. *Consciousness and cognition*, 39: 28–37, 2016.

[18] R. L. Carhart-Harris, M. Kaelen, M. Bolstridge, T. Williams, L. Williams, R. Underwood, A. Feilding, and D. J. Nutt. The paradoxical psychological effects of lysergic acid diethylamide (LSD). *Psychological medicine*, 46(7):1379–1390, 2016.

[19] K. Johnson. *Are you experienced?: how psychedelic consciousness transformed modern art*. Prestel Verlag, 2011.

[20] R. L. Carhart-Harris, L. Roseman, E. Haijen, D. Erritzoe, R. Watts, I. Branchi, and M. Kaelen. Psychedelics and the essential importance of context. *Journal of Psychopharmacology*, 32(7): 725–731, 2018.

[21] F. X. Vollenweider and K. H. Preller. Psychedelic drugs: neurobiology and potential for treatment of psychiatric disorders. *Nature Reviews Neuroscience*, 21(11):611–624, 2020.

[22] K. Lukasiewicz, J. J. Baker, Y. Zuo, and J. Lu. Serotonergic psychedelics in neural plasticity. *Frontiers in Molecular Neuroscience*, page 221, 2021.

[23] R. Carhart-Harris, B. Giribaldi, R. Watts, M. Baker-Jones, A. Murphy-Beiner, R. Murphy, J. Martell, A. Blemings, D. Erritzoe, and D. J. Nutt. Trial of psilocybin versus escitalopram for depression. *New England Journal of Medicine*, 384(15):1402–1411, 2021.

[24] R. L. Carhart-Harris, M. Bolstridge, C. Day, J. Rucker, R. Watts, D. Erritzoe, M. Kaelen, B. Giribaldi, M. Bloomfield, S. Pilling, et al. Psilocybin with psychological support for treatment-resistant depression: six-month follow-up. *Psychopharmacology*, 235(2):399–408, 2018.

[25] C. Ly, A. C. Greb, L. P. Cameron, J. M. Wong, E. V. Barragan, P. C. Wilson, K. F. Burbach, S. S. Zarandi, A. Sood, M. R. Paddy, et al. Psychedelics promote structural and functional neural plasticity. *Cell reports*, 23(11):3170–3182, 2018.

[26] T. Calvey and F. M. Howells. An introduction to psychedelic neuroscience. *Progress in brain research*, 242:1–23, 2018.

[27] R. L. Carhart-Harris, R. Leech, P. J. Hellyer, M. Shanahan, A. Feilding, E. Tagliazucchi, D. R. Chialvo, and D. Nutt. The entropic brain: a theory of conscious states informed by neuroimaging research with psychedelic drugs. *Frontiers in human neuroscience*, page 20, 2014.

[28] R. E. Daws, C. Timmermann, B. Giribaldi, J. D. Sexton, M. B. Wall, D. Erritzoe, L. Roseman, D. Nutt, and R. Carhart-Harris. Increased global integration in the brain after psilocybin therapy for depression. *Nature Medicine*, pages 1–8, 2022.

[29] E. Tagliazucchi, R. Carhart-Harris, R. Leech, D. Nutt, and D. R. Chialvo. Enhanced repertoire of brain dynamical states during the psychedelic experience. *Human brain mapping*, 35(11):5442–5456, 2014.

[30] R. L. Carhart-Harris, D. Erritzoe, T. Williams, J. M. Stone, L. J. Reed, A. Colasanti, R. J. Tyacke, R. Leech, A. L. Malizia, K. Murphy, et al. Neural correlates of the psychedelic state as determined by fMRI studies with psilocybin. *Proceedings of the National Academy of Sciences*, 109(6):2138–2143, 2012.

[31] J. S. Damoiseaux, S. Rombouts, F. Barkhof, P. Scheltens, C. J. Stam, S. M. Smith, and C. F. Beckmann. Consistent resting-state networks across healthy subjects. *Proceedings of the national academy of sciences*, 103(37):13848–13853, 2006.

[32] M. P. Van Den Heuvel and H. E. H. Pol. Exploring the brain network: a review on resting-state fMRI functional connectivity. *European neuropsychopharmacology*, 20(8):519–534, 2010.

[33] R. L. Carhart-Harris. How do psychedelics work? *Current Opinion in Psychiatry*, 32(1):16–21, 2019.

[34] M. Avram, H. Rogg, A. Korda, C. Andreou, F. Müller, and S. Borgwardt. Bridging the gap? altered thalamocortical connectivity in psychotic and psychedelic states. *Frontiers in Psychiatry*, 12, 2021.

[35] D. E. Nichols. Hallucinogens. *Pharmacology & therapeutics*, 101(2):131–181, 2004.

[36] F. X. Vollenweider. Brain mechanisms of hallucinogens and entactogens. *Dialogues in clinical neuroscience*, 2022.

[37] J. P. Barsuglia, M. Polanco, R. Palmer, B. J. Malcolm, B. Kelmendi, and T. Calvey. A case report SPECT study and theoretical rationale for the sequential administration of ibogaine and 5-MeO-DMT in the treatment of alcohol use disorder. *Progress in brain research*, 242:121–158, 2018.

[38] K. S. Murnane. The renaissance in psychedelic research: what do preclinical models have to offer. *Progress in brain research*, 242:25–67, 2018.

[39] R. L. Carhart-Harris, M. Bolstridge, J. Rucker, C. M. Day, D. Erritzoe, M. Kaelen, M. Bloomfield, J. A. Rickard, B. Forbes, A. Feilding, et al. Psilocybin with psychological support for treatment-resistant depression: an open-label feasibility study. *The Lancet Psychiatry*, 3(7):619–627, 2016.

[40] R. Carhart-Harris and D. Nutt. Serotonin and brain function: a tale of two receptors. *Journal of Psychopharmacology*, 31(9):1091–1120, 2017.

[41] R. A. Glennon, R. Young, and J. A. Rosecrans. Antagonism of the effects of the hallucinogen DOM and the purported 5-HT agonist quipazine by 5-HT2 antagonists. *European journal of pharmacology*, 91(2-3):189–196, 1983.

[42] F. C. Colpaert, C. Niemegeers, and P. Janssen. A drug discrimination analysis of lysergic acid diethylamide (LSD): in vivo agonist and antagonist effects of purported 5-hydroxytryptamine antagonists and of pirenperone, a LSD-antagonist. *Journal of Pharmacology and Experimental Therapeutics*, 221(1):206–214, 1982.

[43] F. Colpaert and P. Janssen. The head-twitch response to intraperitoneal injection of 5-hydroxytryptophan in the rat: antagonist effects of purported 5-hydroxytryptamine antagonists and of pirenperone, an LSD antagonist. *Neuropharmacology*, 22(8):993–1000, 1983.

[44] P. M. Laduron, P. F. Janssen, and J. E. Leysen. In vivo binding of [3h] ketanserin on serotonin s2-receptors in rat brain. *European Journal of Pharmacology*, 81(1):43–48, 1982.

[45] R. G. Browne and B. T. Ho. Role of serotonin in the discriminative stimulus properties of mescaline. *Pharmacology Biochemistry and Behavior*, 3(3):429–435, 1975.

[46] J. Winter. Blockade of the stimulus properties of mescaline by a serotonin antagonist. *Archives internationales de pharmacodynamie et de therapie*, 214(2):250–253, 1975.

[47] D. E. Nichols. Psychedelics. *Pharmacological reviews*, 68(2):264–355, 2016.

[48] H. Meltzer, B. Wiita, B. Tricou, M. Simonovic, V. Fang, and G. Manov. Effect of serotonin precursors and serotonin agonists on plasma hormone levels. *Advances in biochemical psychopharmacology*, 34:117–139, 1982.

[49] F. Vollenweider. Advances and pathophysiological models of hallucinogenic drug actions in humans: a preamble to schizophrenia research. *Pharmacopsychiatry*, 31(S 2):92–103, 1998.

[50] A. Dittrich. The standardized psychometric assessment of altered states of consciousness (ASC) in humans. *Pharmacopsychiatry*, 31(S 2):80–84, 1998.

[51] R. A. Glennon, M. Dukat, B. Grella, S.-S. Hong, L. Costantino, M. Teitler, C. Smith, C. Egan, K. Davis, and M. V. Mattson. Binding of β-carbolines and related agents at serotonin (5-HT2

and 5-HT1a), dopamine 2 and benzodiazepine receptors. *Drug and alcohol dependence*, 60(2): 121–132, 2000.

[52] R. Andrade. Serotonergic regulation of neuronal excitability in the prefrontal cortex. *Neuropharmacology*, 61(3):382–386, 2011.

[53] K. H. Preller, M. Herdener, T. Pokorny, A. Planzer, R. Kraehenmann, P. Stämpfli, M. E. Liechti, E. Seifritz, and F. X. Vollenweider. The fabric of meaning and subjective effects in LSD-induced states depend on serotonin 2a receptor activation. *Current Biology*, 27(3):451–457, 2017.

[54] K. H. Preller, L. Schilbach, T. Pokorny, J. Flemming, E. Seifritz, and F. X. Vollenweider. Role of the 5-HT2a receptor in self-and other-initiated social interaction in lysergic acid diethylamide-induced states: A pharmacological fMRI study. *Journal of Neuroscience*, 38(14):3603–3611, 2018.

[55] R. Kraehenmann, D. Pokorny, H. Aicher, K. H. Preller, T. Pokorny, O. G. Bosch, E. Seifritz, and F. X. Vollenweider. LSD increases primary process thinking via serotonin 2a receptor activation. *Frontiers in pharmacology*, 8:814, 2017.

[56] R. Kraehenmann, D. Pokorny, L. Vollenweider, K. H. Preller, T. Pokorny, E. Seifritz, and F. X. Vollenweider. Dreamlike effects of LSD on waking imagery in humans depend on serotonin 2a receptor activation. *Psychopharmacology*, 234(13):2031–2046, 2017.

[57] F. X. Vollenweider, M. F. Vollenweider-Scherpenhuyzen, A. Bäbler, H. Vogel, and D. Hell. Psilocybin induces schizophrenia-like psychosis in humans via a serotonin-2 agonist action. *Neuroreport*, 9(17):3897–3902, 1998.

[58] K. H. Preller, T. Pokorny, A. Hock, R. Kraehenmann, P. Stämpfli, E. Seifritz, M. Scheidegger, and F. X. Vollenweider. Effects of serotonin 2a/1a receptor stimulation on social exclusion processing. *Proceedings of the National Academy of Sciences*, 113(18):5119–5124, 2016.

[59] M. Valle, A. E. Maqueda, M. Rabella, A. Rodríguez-Pujadas, R. M. Antonijoan, S. Romero, J. F. Alonso, M. À. Mañanas, S. Barker, P. Friedlander, et al. Inhibition of alpha oscillations through serotonin-2a receptor activation underlies the visual effects of ayahuasca in humans. *European Neuropsychopharmacology*, 26(7):1161–1175, 2016.

[60] A. L. Halberstadt. Recent advances in the neuropsychopharmacology of serotonergic hallucinogens. *Behavioural brain research*, 277:99–120, 2015.

[61] A. L. Halberstadt, M. Chatha, A. K. Klein, J. Wallach, and S. D. Brandt. Correlation between the potency of hallucinogens in the mouse head-twitch response assay and their behavioral and subjective effects in other species. *Neuropharmacology*, 167:107933, 2020.

[62] R. Schreiber, M. Brocco, V. Audinot, A. Gobert, S. Veiga, and M. J. Millan. (1-(2, 5-dimethoxy-4 iodophenyl)-2-aminopropane)-induced head-twitches in the rat are mediated by 5-hydroxytryptamine 5-HT_{2A}R: modulation by novel 5-HT2a/2c antagonists, DA1 antagonists and

5-HT1a agonists. *Journal of Pharmacology and Experimental Therapeutics*, 273(1):101–112, 1995.

[63] E. T. Weber and R. Andrade. Htr2a gene and 5-HT2a receptor expression in the cerebral cortex studied using genetically modified mice. *Frontiers in neuroscience*, 4:36, 2010.

[64] K. Varnäs, C. Halldin, and H. Hall. Autoradiographic distribution of serotonin transporters and receptor subtypes in human brain. *Human brain mapping*, 22(3):246–260, 2004.

[65] J. De Almeida and G. Mengod. Quantitative analysis of glutamatergic and gabaergic neurons expressing 5-HT_{2A}R in human and monkey prefrontal cortex. *Journal of neurochemistry*, 103(2): 475–486, 2007.

[66] C. P. Muller and K. A. Cunningham. *Handbook of the behavioral neurobiology of serotonin*. Academic Press, 2020.

[67] M. K. Madsen, P. M. Fisher, D. Burmester, A. Dyssegaard, D. S. Stenbæk, S. Kristiansen, S. S. Johansen, S. Lehel, K. Linnet, C. Svarer, et al. Psychedelic effects of psilocybin correlate with serotonin 2a receptor occupancy and plasma psilocin levels. *Neuropsychopharmacology*, 44(7): 1328–1334, 2019.

[68] V. Beliveau, M. Ganz, L. Feng, B. Ozenne, L. Højgaard, P. M. Fisher, C. Svarer, D. N. Greve, and G. M. Knudsen. A high-resolution in vivo atlas of the human brain's serotonin system. *Journal of Neuroscience*, 37(1):120–128, 2017.

[69] R. Gross-Isseroff, D. Salama, M. Israeli, and A. Biegon. Autoradiographic analysis of age-dependent changes in serotonin 5-HT2 receptors of the human brain postmortem. *Brain Research*, 519(1-2):223–227, 1990.

[70] G. J. Marek. Interactions of hallucinogens with the glutamatergic system: permissive network effects mediated through cortical L5p neurons. *Behavioral Neurobiology of Psychedelic Drugs*, pages 107–135, 2017.

[71] G. K. Aghajanian and G. J. Marek. Serotonin, via 5-HT2a receptors, increases epscs in layer v pyramidal cells of prefrontal cortex by an asynchronous mode of glutamate release. *Brain research*, 825(1-2):161–171, 1999.

[72] M. V. Puig, P. Celada, L. Díaz-Mataix, and F. Artigas. In vivo modulation of the activity of pyramidal neurons in the rat medial prefrontal cortex by 5-HT2a receptors: relationship to thalamocortical afferents. *Cerebral Cortex*, 13(8):870–882, 2003.

[73] J. Wood, Y. Kim, and B. Moghaddam. Disruption of prefrontal cortex large scale neuronal activity by different classes of psychotomimetic drugs. *Journal of Neuroscience*, 32(9):3022–3031, 2012.

[74] L. Lladó-Pelfort, P. Celada, M. Riga, E. Troyano-Rodriguez, N. Santana, and F. Artigas. Effects of hallucinogens on neuronal activity. *Behavioral Neurobiology of Psychedelic Drugs*, pages 75–105, 2017.

[75] J.-C. Béïque, M. Imad, L. Mladenovic, J. A. Gingrich, and R. Andrade. Mechanism of the 5-hydroxytryptamine 2a receptor-mediated facilitation of synaptic activity in prefrontal cortex. *Proceedings of the National Academy of Sciences*, 104(23):9870–9875, 2007.

[76] C. Zhang and G. J. Marek. Ampa receptor involvement in 5-hydroxytryptamine2a receptor-mediated pre-frontal cortical excitatory synaptic currents and doi-induced head shakes. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 32(1):62–71, 2008.

[77] A. C. Gambrill, G. P. Storey, and A. Barria. Dynamic regulation of NMDA receptor transmission. *Journal of neurophysiology*, 105(1):162–171, 2011.

[78] A. Barre, C. Berthoux, D. De Bundel, E. Valjent, J. Bockaert, P. Marin, and C. Bécamel. Presynaptic serotonin 2a receptors modulate thalamocortical plasticity and associative learning. *Proceedings of the National Academy of Sciences*, 113(10):E1382–E1391, 2016.

[79] J. Aru, M. Suzuki, R. Rutiku, M. E. Larkum, and T. Bachmann. Coupling the state and contents of consciousness. *Frontiers in Systems Neuroscience*, 13:43, 2019.

[80] M. Larkum. A cellular mechanism for cortical associations: an organizing principle for the cerebral cortex. *Trends in neurosciences*, 36(3):141–151, 2013.

[81] F. X. Vollenweider and M. A. Geyer. A systems model of altered consciousness: integrating natural and drug-induced psychoses. *Brain research bulletin*, 56(5):495–507, 2001.

[82] M. A. Geyer and F. X. Vollenweider. Serotonin research: contributions to understanding psychoses. *Trends in pharmacological sciences*, 29(9):445–453, 2008.

[83] N. Swerdlow, M. Geyer, and D. Braff. Neural circuit regulation of prepulse inhibition of startle in the rat: current knowledge and future challenges. *Psychopharmacology*, 156(2):194–215, 2001.

[84] M. M. Nour, L. Evans, D. Nutt, and R. L. Carhart-Harris. Ego-dissolution and psychedelics: validation of the ego-dissolution inventory. *Frontiers in human neuroscience*, 10:269, 2016.

[85] M. Carlsson and A. Carlsson. Schizophrenia: a subcortical neurotransmitter imbalance syndrome? *Schizophrenia bulletin*, 16(3):425–432, 1990.

[86] T. E. Sipes and M. A. Geyer. Doi disrupts prepulse inhibition of startle in rats via 5-HT2a receptors in the ventral pallidum. *Brain research*, 761(1):97–104, 1997.

[87] F. X. Vollenweider, P. A. Csomor, B. Knappe, M. A. Geyer, and B. B. Quednow. The effects of the preferential 5-HT2a agonist psilocybin on prepulse inhibition of startle in healthy human volunteers depend on interstimulus interval. *Neuropsychopharmacology*, 32(9):1876–1887, 2007.

[88] B. B. Quednow, M. Kometer, M. A. Geyer, and F. X. Vollenweider. Psilocybin-induced deficits in automatic and controlled inhibition are attenuated by ketanserin in healthy human volunteers. *Neuropsychopharmacology*, 37(3):630–640, 2012.

[89] Y. Schmid, F. Enzler, P. Gasser, E. Grouzmann, K. H. Preller, F. X. Vollenweider, R. Brenneisen, F. Müller, S. Borgwardt, and M. E. Liechti. Acute effects of lysergic acid diethylamide in healthy subjects. *Biological psychiatry*, 78(8):544–553, 2015.

[90] J. Riba, A. Rodríguez-Fornells, and M. J. Barbanoj. Effects of ayahuasca on sensory and sensorimotor gating in humans as measured by P50 suppression and prepulse inhibition of the startle reflex, respectively. *Psychopharmacology*, 165(1):18–28, 2002.

[91] D. E. Olson. Psychoplastogens: a promising class of plasticity-promoting neurotherapeutics. *Journal of experimental neuroscience*, 12:1179069518800508, 2018.

[92] A. F. Arnsten. Stress signalling pathways that impair prefrontal cortex structure and function. *Nature reviews neuroscience*, 10(6):410–422, 2009.

[93] J. Peters and C. Büchel. Neural representations of subjective reward value. *Behavioural brain research*, 213(2):135–141, 2010.

[94] S. J. Russo and E. J. Nestler. The brain reward circuitry in mood disorders. *Nature Reviews Neuroscience*, 14(9):609–625, 2013.

[95] T. M. Kelly and D. C. Daley. Integrated treatment of substance use and psychiatric disorders. *Social work in public health*, 28(3-4):388–406, 2013.

[96] A. E. Autry and L. M. Monteggia. Brain-derived neurotrophic factor and neuropsychiatric disorders. *Pharmacological reviews*, 64(2):238–258, 2012.

[97] D. J. Christoffel, S. A. Golden, and S. J. Russo. Structural and synaptic plasticity in stress-related disorders. 2011.

[98] E. Castrén and H. Antila. Neuronal plasticity and neurotrophic factors in drug responses. *Molecular psychiatry*, 22(8):1085–1095, 2017.

[99] S. Hayley and D. Litteljohn. Neuroplasticity and the next wave of antidepressant strategies. *Frontiers in cellular neuroscience*, 7:218, 2013.

[100] B. Kolb and A. Muhammad. Harnessing the power of neuroplasticity for intervention. *Frontiers in human neuroscience*, 8:377, 2014.

[101] A. K. Davis, F. S. Barrett, and R. R. Griffiths. Psychological flexibility mediates the relations between acute psychedelic effects and subjective decreases in depression and anxiety. *Journal of contextual behavioral science*, 15:39–45, 2020.

[102] D. E. Nichols, M. W. Johnson, and C. D. Nichols. Psychedelics as medicines: an emerging new paradigm. *Clinical Pharmacology & Therapeutics*, 101(2):209–219, 2017.

[103] R. G. Dos Santos, F. L. Osório, J. A. S. Crippa, J. Riba, A. W. Zuardi, and J. E. Hallak. Antidepressive, anxiolytic, and antiaddictive effects of ayahuasca, psilocybin and lysergic acid diethylamide

(LSD): a systematic review of clinical trials published in the last 25 years. *Therapeutic advances in psychopharmacology*, 6(3):193–213, 2016.

[104] J. J. Rucker, L. A. Jelen, S. Flynn, K. D. Frowde, and A. H. Young. Psychedelics in the treatment of unipolar mood disorders: a systematic review. *Journal of Psychopharmacology*, 30(12):1220–1229, 2016.

[105] M. Miranda, J. F. Morici, M. B. Zanoni, and P. Bekinschtein. Brain-derived neurotrophic factor: a key molecule for memory in the healthy and the pathological brain. *Frontiers in cellular neuroscience*, page 363, 2019.

[106] D. A. Martin and C. D. Nichols. Psychedelics recruit multiple cellular types and produce complex transcriptional responses within the brain. *EBioMedicine*, 11:262–277, 2016.

[107] C. D. Nichols and E. Sanders-Bush. A single dose of lysergic acid diethylamide influences gene expression patterns within the mammalian brain. *Neuropsychopharmacology*, 26(5):634–642, 2002.

[108] S. Cohen-Cory, A. H. Kidane, N. J. Shirkey, and S. Marshak. Brain-derived neurotrophic factor and the development of structural neuronal connectivity. *Developmental neurobiology*, 70(5):271–288, 2010.

[109] N. Takei and H. Nawa. mTOR signaling and its roles in normal and abnormal brain development. *Frontiers in molecular neuroscience*, 7:28, 2014.

[110] J. Jaworski and M. Sheng. The growing role of mTOR in neuronal development and plasticity. *Molecular neurobiology*, 34(3):205–219, 2006.

[111] C. A. Hoeffer and E. Klann. mTOR signaling: at the crossroads of plasticity, memory and disease. *Trends in neurosciences*, 33(2):67–75, 2010.

[112] R.-J. Liu, M. Fuchikami, J. M. Dwyer, A. E. Lepack, R. S. Duman, and G. K. Aghajanian. GSK-3 inhibition potentiates the synaptogenic and antidepressant-like effects of subthreshold doses of ketamine. *Neuropsychopharmacology*, 38(11):2268–2277, 2013.

[113] J. W. Muschamp, M. J. Regina, E. M. Hull, J. C. Winter, and R. A. Rabin. Lysergic acid diethylamide and [-]-2, 5-dimethoxy-4-methylamphetamine increase extracellular glutamate in rat prefrontal cortex. *Brain research*, 1023(1):134–140, 2004.

[114] V. A. Vaidya, G. J. Marek, G. K. Aghajanian, and R. S. Duman. 5-HT2a receptor-mediated regulation of brain-derived neurotrophic factor mRNA in the hippocampus and the neocortex. *Journal of Neuroscience*, 17(8):2785–2795, 1997.

[115] F. X. Vollenweider and M. Kometer. The neurobiology of psychedelic drugs: implications for the treatment of mood disorders. *Nature Reviews Neuroscience*, 11(9):642–651, 2010.

[116] H. Jourdi, Y.-T. Hsu, M. Zhou, Q. Qin, X. Bi, and M. Baudry. Positive AMPA receptor modulation rapidly stimulates BDNF release and increases dendritic mRNA translation. *Journal of Neuroscience*, 29(27):8688–8697, 2009.

[117] G. W. Knott, A. Holtmaat, L. Wilbrecht, E. Welker, and K. Svoboda. Spine growth precedes synapse formation in the adult neocortex in vivo. *Nature neuroscience*, 9(9):1117–1124, 2006.

[118] M. Fu and Y. Zuo. Experience-dependent structural plasticity in the cortex. *Trends in neurosciences*, 34(4):177–187, 2011.

[119] L. Lepow, H. Morishita, and R. Yehuda. Critical period plasticity as a framework for psychedelic-assisted psychotherapy. *Frontiers in neuroscience*, page 1165, 2021.

[120] B. D. Heifets and R. C. Malenka. Disruptive psychopharmacology. *JAMA psychiatry*, 76(8):775–776, 2019.

[121] D. B. Yaden and R. R. Griffiths. The subjective effects of psychedelics are necessary for their enduring therapeutic effects. *ACS Pharmacology & Translational Science*, 4(2):568–572, 2020.

[122] R. J. Davidson and B. S. McEwen. Social influences on neuroplasticity: stress and interventions to promote well-being. *Nature neuroscience*, 15(5):689–695, 2012.

[123] S. J. Lupien, B. S. McEwen, M. R. Gunnar, and C. Heim. Effects of stress throughout the lifespan on the brain, behaviour and cognition. *Nature reviews neuroscience*, 10(6):434–445, 2009.

[124] L. C. Pratchett and R. Yehuda. Foundations of posttraumatic stress disorder: does early life trauma lead to adult posttraumatic stress disorder? *Development and psychopathology*, 23(2):477–491, 2011.

[125] D. J. Piekarski, C. M. Johnson, J. R. Boivin, A. W. Thomas, W. C. Lin, K. Delevich, E. M. Galarce, and L. Wilbrecht. Does puberty mark a transition in sensitive periods for plasticity in the associative neocortex? *Brain research*, 1654:123–144, 2017.

[126] T. Noorani, A. Garcia-Romeu, T. C. Swift, R. R. Griffiths, and M. W. Johnson. Psychedelic therapy for smoking cessation: Qualitative analysis of participant accounts. *Journal of Psychopharmacology*, 32(7):756–769, 2018.

[127] R. Nardou, E. M. Lewis, R. Rothhaas, R. Xu, A. Yang, E. Boyden, and G. Dölen. Oxytocin-dependent reopening of a social reward learning critical period with MDMA. *Nature*, 569(7754): 116–120, 2019.

[128] N. Hesselgrave, T. A. Troppoli, A. B. Wulff, A. B. Cole, and S. M. Thompson. Harnessing psilocybin: antidepressant-like behavioral and synaptic actions of psilocybin are independent of 5-HT2r activation in mice. *Proceedings of the National Academy of Sciences*, 118(17):e2022489118, 2021.

[129] R. Millière, R. L. Carhart-Harris, L. Roseman, F.-M. Trautwein, and A. Berkovich-Ohana. Psychedelics, meditation, and self-consciousness. *Frontiers in psychology*, 9:1475, 2018.

[130] E. Studerus, M. Kometer, F. Hasler, and F. X. Vollenweider. Acute, subacute and long-term subjective effects of psilocybin in healthy humans: a pooled analysis of experimental studies. *Journal of psychopharmacology*, 25(11):1434–1452, 2011.

[131] L. Roseman, E. Haijen, K. Idialu-Ikato, M. Kaelen, R. Watts, and R. Carhart-Harris. Emotional breakthrough and psychedelics: Validation of the emotional breakthrough inventory. *Journal of Psychopharmacology*, 33(9):1076–1087, 2019.

[132] L. Roseman, D. Nutt, and R. Carhart-Harris. Quality of acute psychedelic experience predicts therapeutic efficacy of psilocybin for treatment-resistant depression. *Frontiers in Pharmacology*, 8, 2018. doi: 10.3389/fphar.2017.00974. URL `http://dx.doi.org/10.3389/fphar.2017.00974`.

[133] K. Ko, E. I. Kopra, A. J. Cleare, and J. J. Rucker. Psychedelic therapy for depressive symptoms: A systematic review and meta-analysis. *Journal of Affective Disorders*, 2022.

[134] M. Cavarra, A. Falzone, J. G. Ramaekers, K. P. Kuypers, and C. Mento. Psychedelic-assisted psychotherapy—a systematic review of associated psychological interventions. *Frontiers in Psychology*, page 2996, 2022.

[135] K. J. Friston. Functional and effective connectivity: a review. *Brain connectivity*, 1(1):13–36, 2011.

[136] F. Müller, P. C. Dolder, A. Schmidt, M. E. Liechti, and S. Borgwardt. Altered network hub connectivity after acute LSD administration. *NeuroImage: Clinical*, 18:694–701, 2018.

[137] T. F. Varley, R. Carhart-Harris, L. Roseman, D. K. Menon, and E. A. Stamatakis. Serotonergic psychedelics LSD & psilocybin increase the fractal dimension of cortical brain activity in spatial and temporal domains. *Neuroimage*, 220:117049, 2020.

[138] R. L. Carhart-Harris and K. J. Friston. The default-mode, ego-functions and free-energy: a neurobiological account of freudian ideas. *Brain*, 133(4):1265–1283, 2010.

[139] S. M. Fleming and R. J. Dolan. The neural basis of metacognitive ability. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1594):1338–1349, 2012.

[140] K. Friston. The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, 11(2): 127–138, 2010.

[141] D. S. Bassett, E. Bullmore, B. A. Verchinski, V. S. Mattay, D. R. Weinberger, and A. Meyer-Lindenberg. Hierarchical organization of human cortical networks in health and schizophrenia. *Journal of Neuroscience*, 28(37):9239–9248, 2008.

[142] C. F. Beckmann, M. DeLuca, J. T. Devlin, and S. M. Smith. Investigations into resting-state connectivity using independent component analysis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1457):1001–1013, 2005.

[143] M. E. Raichle, A. M. MacLeod, A. Z. Snyder, W. J. Powers, D. A. Gusnard, and G. L. Shulman. A default mode of brain function. *Proceedings of the National Academy of Sciences*, 98(2):676–682, 2001.

[144] R. L. Buckner, J. R. Andrews-Hanna, and D. L. Schacter. The brain's default network: anatomy, function, and relevance to disease. *Annals of the new York Academy of Sciences*, 1124(1):1–38, 2008.

[145] Q. Zou, C. W. Wu, E. A. Stein, Y. Zang, and Y. Yang. Static and dynamic characteristics of cerebral blood flow during the resting state. *Neuroimage*, 48(3):515–524, 2009.

[146] P. Hagmann, L. Cammoun, X. Gigandet, R. Meuli, C. J. Honey, V. J. Wedeen, and O. Sporns. Mapping the structural core of human cerebral cortex. *PLoS biology*, 6(7):e159, 2008.

[147] M. P. Van Den Heuvel, R. S. Kahn, J. Goñi, and O. Sporns. High-cost, high-capacity backbone for global brain communication. *Proceedings of the National Academy of Sciences*, 109(28):11372–11377, 2012.

[148] P. Fransson and G. Marrelec. The precuneus/posterior cingulate cortex plays a pivotal role in the default mode network: Evidence from a partial correlation network analysis. *Neuroimage*, 42(3): 1178–1184, 2008.

[149] M. G. Berman and J. Jonides. Ruminating on rumination. *Biological psychiatry*, 70(4):310–311, 2011.

[150] A. Vanhaudenhuyse, A. Demertzi, M. Schabus, Q. Noirhomme, S. Bredart, M. Boly, C. Phillips, A. Soddu, A. Luxen, G. Moonen, et al. Two distinct neuronal networks mediate the awareness of environment and of self. *Journal of cognitive neuroscience*, 23(3):570–578, 2011.

[151] F. L. da Silva. Neural mechanisms underlying brain waves: from neural membranes to networks. *Electroencephalography and clinical neurophysiology*, 79(2):81–93, 1991.

[152] M. L. Lőrincz, K. A. Kékesi, G. Juhász, V. Crunelli, and S. W. Hughes. Temporal framing of thalamic relay-mode firing by phasic inhibition during the alpha rhythm. *Neuron*, 63(5):683–696, 2009.

[153] G. G. Knyazev, J. Y. Slobodskoj-Plusnin, A. V. Bocharov, and L. V. Pylkova. The default mode network and EEG alpha oscillations: an independent component analysis. *Brain research*, 1402: 67–79, 2011.

[154] K. Jann, T. Dierks, C. Boesch, M. Kottlow, W. Strik, and T. Koenig. BOLD correlates of EEG alpha phase-locking and the fMRI default mode network. *Neuroimage*, 45(3):903–916, 2009.

[155] P. Qin and G. Northoff. How is our self related to midline regions and the default-mode network? *Neuroimage*, 57(3):1221–1233, 2011.

[156] L. Ekroot and T. M. Cover. The entropy of Markov trajectories. *IEEE Transactions on Information Theory*, 39(4):1418–1421, 1993.

[157] X. Jia and A. Kohn. Gamma rhythms in the brain. *PLoS biology*, 9(4):e1001045, 2011.

[158] T. Pálenícek et al. Comparison of the effects of hallucinogens psilocin and mescaline on quantitative EEG and sensorimotor gating–an animal model of psychosis. *Psychiatrie*, 15:44–48, 2011.

[159] A. Shmuel and D. A. Leopold. Neuronal correlates of spontaneous fluctuations in fmri signals in monkey visual cortex: Implications for functional connectivity at rest. *Human brain mapping*, 29 (7):751–761, 2008.

[160] W. C. Drevets, J. L. Price, and M. L. Furey. Brain structural and functional abnormalities in mood disorders: implications for neurocircuitry models of depression. *Brain structure and function*, 213 (1):93–118, 2008.

[161] Y. I. Sheline, J. L. Price, Z. Yan, and M. A. Mintun. Resting-state functional magnetic ressonance imaging in depression unmasks increased connectivity between networks via the dorsal nexus. *Proceedings of the National Academy of Sciences*, 107(24):11020–11025, 2010.

[162] P. E. Holtzheimer and H. S. Mayberg. Stuck in a rut: rethinking depression and its treatment. *Trends in neurosciences*, 34(1):1–9, 2011.

[163] C. S. Grob, A. L. Danforth, G. S. Chopra, M. Hagerty, C. R. McKay, A. L. Halberstadt, and G. R. Greer. Pilot study of psilocybin treatment for anxiety in patients with advanced-stage cancer. *Archives of general psychiatry*, 68(1):71–78, 2011.

[164] S. D. Muthukumaraswamy, R. L. Carhart-Harris, R. J. Moran, M. J. Brookes, T. M. Williams, D. Errtizoe, B. Sessa, A. Papadopoulos, M. Bolstridge, K. D. Singh, et al. Broadband cortical desynchronization underlies the human psychedelic state. *Journal of Neuroscience*, 33(38):15171–15183, 2013.

[165] R. L. Carhart-Harris, R. Leech, D. Erritzoe, T. M. Williams, J. M. Stone, J. Evans, D. J. Sharp, A. Feilding, R. G. Wise, and D. J. Nutt. Functional connectivity measures after psilocybin inform a novel hypothesis of early psychosis. *Schizophrenia bulletin*, 39(6):1343–1351, 2013.

[166] L. Roseman, R. Leech, A. Feilding, D. J. Nutt, and R. L. Carhart-Harris. The effects of psilocybin and MDMA on between-network resting state functional connectivity in healthy volunteers. *Frontiers in human neuroscience*, 8:204, 2014.

[167] R. L. Carhart-Harris, L. Roseman, M. Bolstridge, L. Demetriou, J. N. Pannekoek, M. B. Wall, M. Tanner, M. Kaelen, J. McGonigle, K. Murphy, et al. Psilocybin for treatment-resistant depression: fMRI-measured brain mechanisms. *Scientific reports*, 7(1):1–11, 2017.

[168] M. Girn, L. Roseman, B. Bernhardt, J. Smallwood, R. Carhart-Harris, and R. N. Spreng. Serotonergic psychedelic drugs LSD and psilocybin reduce the hierarchical differentiation of unimodal and transmodal cortex. *NeuroImage*, 256:119220, 2022.

[169] F. Müller, C. Lenz, P. Dolder, U. Lang, A. Schmidt, M. Liechti, and S. Borgwardt. Increased thalamic resting-state connectivity as a core driver of LSD-induced hallucinations. *Acta Psychiatrica Scandinavica*, 136(6):648–657, 2017.

[170] K. H. Preller, A. Razi, P. Zeidman, P. Stämpfli, K. J. Friston, and F. X. Vollenweider. Effective connectivity changes in LSD-induced altered states of consciousness in humans. *Proceedings of the National Academy of Sciences*, 116(7):2743–2748, 2019.

[171] E. Tagliazucchi, L. Roseman, M. Kaelen, C. Orban, S. D. Muthukumaraswamy, K. Murphy, H. Laufs, R. Leech, J. McGonigle, N. Crossley, et al. Increased global functional connectivity correlates with LSD-induced ego dissolution. *Current Biology*, 26(8):1043–1050, 2016.

[172] F. Palhano-Fontes, K. C. Andrade, L. F. Tofoli, A. C. Santos, J. A. S. Crippa, J. E. Hallak, S. Ribeiro, and D. B. de Araujo. The psychedelic state induced by ayahuasca modulates the activity and connectivity of the default mode network. *PloS one*, 10(2):e0118143, 2015.

[173] R. P. Rao and D. H. Ballard. Development of localized oriented receptive fields by learning a translation-invariant code for natural images. *Network: Computation in Neural Systems*, 9(2):219, 1998.

[174] D. C. Knill and A. Pouget. The Bayesian brain: the role of uncertainty in neural coding and computation. *TRENDS in Neurosciences*, 27(12):712–719, 2004.

[175] J. Hohwy. *The predictive mind*. OUP Oxford, 2013.

[176] D. Kersten, P. Mamassian, and A. Yuille. Object perception as Bayesian inference. *Annu. Rev. Psychol.*, 55:271–304, 2004.

[177] A. Clark. Radical predictive processing. *The Southern Journal of Philosophy*, 53:3–27, 2015.

[178] J. B. Tenenbaum, C. Kemp, T. L. Griffiths, and N. D. Goodman. How to grow a mind: Statistics, structure, and abstraction. *science*, 331(6022):1279–1285, 2011.

[179] C. Kemp, A. Perfors, and J. B. Tenenbaum. Learning overhypotheses with hierarchical Bayesian models. *Developmental science*, 10(3):307–321, 2007.

[180] R. Shamey and R. G. Kuehni. Helmholtz, hermann ludwig von 1821–1894. In *Pioneers of Color Science*, pages 193–196. Springer, 2020.

[181] R. L. Gregory. Perceptual illusions and brain models. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 171(1024):279–296, 1968.

[182] P. C. Fletcher and C. D. Frith. Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews Neuroscience*, 10(1):48–58, 2009.

[183] C. C. Hilgetag and M.-T. Hütt. Hierarchical modular brain connectivity is a stretch for criticality. *Trends in cognitive sciences*, 18(3):114–115, 2014.

[184] L. Roseman, D. J. Nutt, and R. L. Carhart-Harris. Quality of acute psychedelic experience predicts therapeutic efficacy of psilocybin for treatment-resistant depression. *Frontiers in pharmacology*, 8:974, 2018.

[185] C. Timmermann, L. Roseman, L. Williams, D. Erritzoe, C. Martial, H. Cassol, S. Laureys, D. Nutt, and R. Carhart-Harris. N,N-DMT models the near-death experience. *Frontiers in psychology*, page 1424, 2018.

[186] R. L. Carhart-Harris, M. Kaelen, M. Whalley, M. Bolstridge, A. Feilding, and D. J. Nutt. LSD enhances suggestibility in healthy volunteers. *Psychopharmacology*, 232(4):785–794, 2015.

[187] T. M. Carbonaro, M. W. Johnson, E. Hurwitz, and R. R. Griffiths. Double-blind comparison of the two hallucinogens psilocybin and dextromethorphan: similarities and differences in subjective experiences. *Psychopharmacology*, 235(2):521–534, 2018.

[188] P. C. Bressloff, J. D. Cowan, M. Golubitsky, P. J. Thomas, and M. C. Wiener. What geometric visual hallucinations tell us about the visual cortex. *Neural computation*, 14(3):473–491, 2002.

[189] S. Atasoy, L. Roseman, M. Kaelen, M. L. Kringelbach, G. Deco, and R. L. Carhart-Harris. Connectome-harmonic decomposition of human brain activity reveals dynamical repertoire reorganization under LSD. *Scientific reports*, 7(1):1–18, 2017.

[190] K. A. MacLean, M. W. Johnson, and R. R. Griffiths. Mystical experiences occasioned by the hallucinogen psilocybin lead to increases in the personality domain of openness. *Journal of psychopharmacology*, 25(11):1453–1461, 2011.

[191] A. Safron. On the varieties of conscious experiences: Altered beliefs under psychedelics (ALBUS). 2020.

[192] A. Safron. On the varieties of psychedelic experiences: belief dynamics and altered states of consciousness.

[193] D. McNamee and D. M. Wolpert. Internal models in biological control. *Annual review of control, robotics, and autonomous systems*, 2:339, 2019.

[194] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40:e253, 2017.

[195] Z. Ghahramani. Probabilistic machine learning and artificial intelligence. *Nature*, 521(7553):452–459, 2015.

[196] I. Y. Chen, S. Joshi, M. Ghassemi, and R. Ranganath. Probabilistic machine learning for healthcare. *arXiv preprint arXiv:2009.11087*, 2020.

[197] Z. Ghahramani. Bayesian non-parametrics and the probabilistic approach to modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1984):20110553, 2013.

[198] W. H. Jefferys and J. O. Berger. Ockham's razor and Bayesian analysis. *American scientist*, 80 (1):64–72, 1992.

[199] C. Rasmussen and Z. Ghahramani. Occam's razor. *Advances in neural information processing systems*, 13, 2000.

[200] D. J. MacKay, D. J. Mac Kay, et al. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[201] J. B. Tenenbaum and T. L. Griffiths. Generalization, similarity, and Bayesian inference. *Behavioral and brain sciences*, 24(4):629–640, 2001.

[202] T. L. Griffiths and J. B. Tenenbaum. Theory-based causal induction. *Psychological review*, 116 (4):661, 2009.

[203] J. B. Tenenbaum, T. L. Griffiths, and C. Kemp. Theory-based Bayesian models of inductive learning and reasoning. *Trends in cognitive sciences*, 10(7):309–318, 2006.

[204] R. M. Neal. *Probabilistic inference using Markov chain Monte Carlo methods*. Department of Computer Science, University of Toronto Toronto, ON, Canada, 1993.

[205] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

[206] A. Doucet, S. Godsill, and C. Andrieu. On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and computing*, 10(3):197–208, 2000.

[207] T. P. Minka. Expectation propagation for approximate Bayesian inference. *arXiv preprint arXiv:1301.2294*, 2013.

[208] C. M. Bishop and N. M. Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[209] J. M. Tomczak. Why deep generative modeling? In *Deep Generative Modeling*, pages 1–12. Springer, 2022.

[210] B. A. Olshausen. Bayesian probability theory and generative models. Technical report, Technical report, 2006.

[211] C. Kemp and J. B. Tenenbaum. The discovery of structural form. *Proceedings of the National Academy of Sciences*, 105(31):10687–10692, 2008.

[212] T. D. Ullman and J. B. Tenenbaum. Bayesian models of conceptual development: Learning as building models of the world. 2020.

[213] J. Baxter. A model of inductive bias learning. *Journal of artificial intelligence research*, 12:149–198, 2000.

[214] S. Laurence and E. Margolis. Concept nativism and neural plasticity. *The Conceptual Mind. New Directions in the Study of Concepts*, pages 117–147, 2015.

[215] S. Thrun. Learning to learn: Introduction, 1996.

[216] T. M. Mitchell and T. M. Mitchell. *Machine learning*, volume 1. McGraw-hill New York, 1997.

[217] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.

[218] Y. Song, C. Durkan, I. Murray, and S. Ermon. Maximum likelihood training of score-based diffusion models. *Advances in Neural Information Processing Systems*, 34:1415–1428, 2021.

[219] R. R. Coifman and S. Lafon. Diffusion maps. *Applied and computational harmonic analysis*, 21 (1):5–30, 2006.

[220] R. I. Kondor and J. Lafferty. Diffusion kernels on graphs and other discrete structures. In *Proceedings of the 19th international conference on machine learning*, volume 2002, pages 315–322, 2002.

[221] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the national academy of sciences*, 102(21):7426–7431, 2005.

[222] R. Kondor and J.-P. Vert. Diffusion kernels. *kernel methods in computational biology*, pages 171–192, 2004.

[223] T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230, 1973.

[224] J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica sinica*, pages 639–650, 1994.

[225] Y. W. Teh et al. Dirichlet process. *Encyclopedia of machine learning*, 1063:280–287, 2010.

[226] S. J. Gershman and D. M. Blei. A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology*, 56(1):1–12, 2012.

[227] N. Goodman, V. Mansinghka, D. M. Roy, K. Bonawitz, and J. B. Tenenbaum. Church: a language for generative models. *arXiv preprint arXiv:1206.3255*, 2012.

[228] T. D. Kulkarni, P. Kohli, J. B. Tenenbaum, and V. Mansinghka. Picture: A probabilistic programming language for scene perception. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 4390–4399, 2015.

[229] P. Bak. *How nature works: the science of self-organized criticality*. Springer Science & Business Media, 2013.

[230] G. Pruessner. *Self-organised criticality: theory, models and characterisation*. Cambridge University Press, 2012.

[231] I. Hipólito, J. Mago, F. Rosas, and R. Carhart-Harris. Pattern breaking: a complex systems approach to psychedelic medicine. 2022.

[232] M. Wichers, M. J. Schreuder, R. Goekoop, and R. N. Groen. Can we predict the direction of sudden shifts in symptoms? transdiagnostic implications from a complex systems perspective on psychopathology. *Psychological medicine*, 49(3):380–387, 2019.

[233] J. Vohryzek, J. Cabral, P. Vuust, G. Deco, and M. L. Kringelbach. Understanding brain states across spacetime informed by whole-brain modelling. *Philosophical Transactions of the Royal Society A*, 380(2227):20210247, 2022.

[234] J. Cruzat, Y. S. Perl, A. Escrichs, J. Vohryzek, C. Timmermann, L. Roseman, A. I. Luppi, A. Ibañez, D. Nutt, R. Carhart-Harris, et al. Effects of classic psychedelic drugs on turbulent signatures in brain dynamics. *Network Neuroscience*, 6(4):1104–1124, 2022.

[235] P. Zhuang, S. Abnar, J. Gu, A. Schwing, J. M. Susskind, and M. Á. Bautista. Diffusion probabilistic fields. In *International Conference on Learning Representations*, 2023.

[236] Y. Song and S. Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

[237] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[238] C. Shorten and T. M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

[239] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, and E. Hovy. A survey of data augmentation approaches for natural language processing. *arXiv preprint arXiv:2105.03075*, 2021.

[240] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum. The omniglot challenge: a 3-year progress report. *Current Opinion in Behavioral Sciences*, 29:97–104, 2019.

[241] G. Koch, R. Zemel, R. Salakhutdinov, et al. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2. Lille, 2015.

[242] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016.

[243] J. Snell, K. Swersky, and R. Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

[244] C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

[245] D. George, W. Lehrach, K. Kansky, M. Lázaro-Gredilla, C. Laan, B. Marthi, X. Lou, Z. Meng, Y. Liu, H. Wang, et al. A generative vision model that trains with high data efficiency and breaks text-based captchas. *Science*, 358(6368):eaag2612, 2017.

[246] L. B. Hewitt, M. I. Nye, A. Gane, T. Jaakkola, and J. B. Tenenbaum. The variational homoencoder: Learning to learn high capacity generative models from few examples. *arXiv preprint arXiv:1807.08919*, 2018.

[247] V. G. Satorras and J. B. Estrach. Few-shot learning with graph neural networks. In *International conference on learning representations*, 2018.

[248] P. Shyam, S. Gupta, and A. Dukkipati. Attentive recurrent comparators. In *International conference on machine learning*, pages 3173–3181. PMLR, 2017.

[249] R. Feinman and B. M. Lake. Generating new concepts with hybrid neuro-symbolic models. *arXiv preprint arXiv:2003.08978*, 2020.

[250] R. Gontijo-Lopes, S. J. Smullin, E. D. Cubuk, and E. Dyer. Affinity and diversity: Quantifying mechanisms of data augmentation. *arXiv preprint arXiv:2002.08973*, 2020.

[251] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[252] K. Suzuki, W. Roseboom, D. J. Schwartzman, and A. K. Seth. A deep-dream virtual reality platform for studying altered perceptual phenomenology. *Scientific reports*, 7(1):1–11, 2017.

[253] M. M. Schartner and C. Timmermann. Neural network models for N,N-DMT-induced visual hallucinations. *Neuroscience of Consciousness*, 2020(1):niaa024, 2020.

[254] D. Furaha and P. Michael. Research proposal: Monitoring optimization of artificial neural network stability undergoing modelled neuroplastic effects of serotonergic psychoactives.

[255] R. C. O'Reilly, C. Ranganath, and J. L. Russin. The structure of systematicity in the brain. *Current directions in psychological science*, 31(2):124–130, 2022.

[256] D. C. McNamee, K. L. Stachenfeld, M. M. Botvinick, and S. J. Gershman. Flexible modulation of sequence generation in the entorhinal–hippocampal system. *Nature neuroscience*, 24(6):851–862, 2021.

[257] G. Ballentine, S. F. Friedman, and D. Bzdok. Trips and neurotransmitters: Discovering principled patterns across 6850 hallucinogenic experiences. *Science Advances*, 8(11):eabl6989, 2022.

[258] C. E. Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27 (3):379–423, 1948.

[259] S. Kullback and R. A. Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[260] J. M. Joyce. Kullback-leibler divergence. In *International encyclopedia of statistical science*, pages 720–722. Springer, 2011.

[261] B. Fuglede and F. Topsoe. Jensen-Shannon divergence and hilbert space embedding. In *International Symposium onInformation Theory, 2004. ISIT 2004. Proceedings.*, page 31. IEEE, 2004.

[262] J. Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

[263] M. Menéndez, J. Pardo, L. Pardo, and M. Pardo. The Jensen-Shannon divergence. *Journal of the Franklin Institute*, 334(2):307–318, 1997.

[264] F. Osterreicher and I. Vajda. A new class of metric divergences on probability spaces and its applicability in statistics. *Annals of the Institute of Statistical Mathematics*, 55(3):639–653, 2003.

[265] R. Feinman. pyBPL, 2020. URL `https://github.com/rfeinman/pyBPL`.

Appendix A

Supplements

A.1 Omniglot evaluation data set

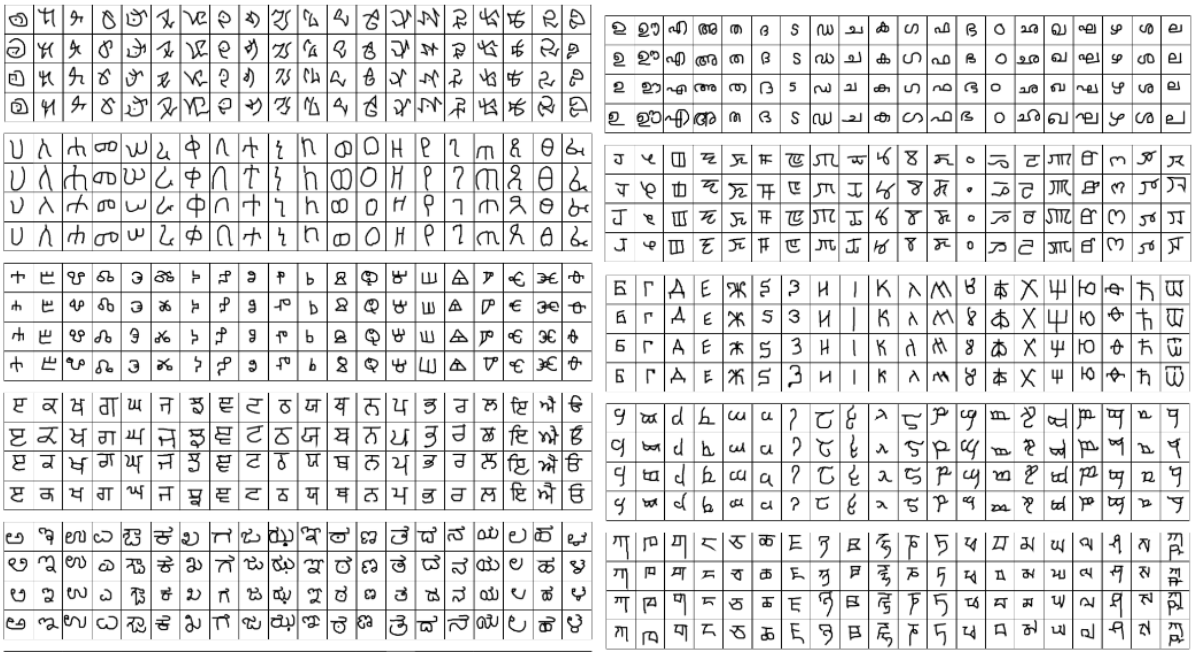


Figure A.1: **Evaluation data set.** Each two rows correspond to the training (above) and test (below) of one run. From top to bottom, and left to right, it is possible to observe the images utilized in the one-shot classification task, from run 1 to run 20.

A.2 Additional analyses

A.2.1 Shannon entropy

Shannon entropy is a measure of the amount of uncertainty or surprise in a set of data. It was introduced by Claude Shannon in 1948 [258] as a way to quantify the amount of information in a message. The entropy of a message or data set is defined as the average number of bits needed to represent each possible outcome.

The formula for Shannon entropy is given by:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i) = \mathbb{E}[-logp(X)] \tag{A.1}$$

where $H(X)$ is the entropy of the data set X, n is the number of possible outcomes, and $p(x_i)$ is the probability of outcome x_i. The entropy is measured in bits and ranges from 0 (when there is no uncertainty) to $\log n$ (when all outcomes are equally likely). Entropy in information theory is directly analogous to the entropy in statistical thermodynamics, both capturing increased ”randomness”.

Shannon entropy can also be thought as a measure of the amount of uncertainty in a probability distribution. In this case, the entropy is a measure of the average amount of information contained in each event in the distribution. The formula for Shannon entropy of a probability distribution is similar to the formula for entropy of a data set, but it is given by:

$$H(P) = -\sum_{i=1}^{n} p_i \log p_i \tag{A.2}$$

where this time P is a probability distribution.

While this is a useful measure of uncertainty in a data set or probability distribution, it does not take into account the structure or relationships between the outcomes. Other information measures, such as mutual information and Kullback-Leibler divergence (KLD), can be used to quantify the amount of information shared between two data sets or to measure the difference between two probability distributions.

A.2.2 Kullback–Leibler divergence

Kullback-Leiber divergence (KL divergence) was introduced by Kullback and Leibler in 1951 [259] and it is an information-based measure of divergence between two probability distributions.

Given two distributions P and Q defined over X, with Q continuous with respect to P. The Kullback-Leibler divergence (KLD) can be interpreted as the existent *cross-entropy* difference for Q on P defined by $H(P,Q) = -\sum_{x\epsilon X} P(x) \log Q(x)$ for discrete distributions and the *self-entropy* [258] of P $H(P) = H(P,P) = -\sum_{x\epsilon X} \log P(x)$. Therefore, $D_{KL}(P,Q) = H(P) - H(P,Q)$ is the expected difference, from the standpoint of P, between the information encoded in P and the information encoded in Q. As a result of $H(P,Q)$ being the P-expectation of the number of bits of information, beyond those encoded in Q.

Equally, KLD can be understood as the anticipated excess surprise when P is the actual distribution and Q is used as the model. A value of $D_{KL}(P,Q) = 0$ represents that the two distributions have identical quantities of information. It is important to notice that is KLD is unbounded, but always positive.

For discrete probability distribution KLD is given by the equation

$$D_{KL}(P,Q) = -\sum_{x\epsilon X} P(x) \log \frac{Q(x)}{P(x)} \tag{A.3}$$

This measure was computed between the original distributions $P(z_{i1})$ and $P(z_{ij}|z_{i(j-1)})$, and ρ_{start} and ρ_{pT}, obtained after the perturbation phase, respectively, as well as between the original distribution and ρ_{start} and ρ_{pT}, obtained after the inference phase.

It should be clearly stated that KL-divergence is not a true metric since it contradicts the triangle inequality and is not symmetric. As a result, the Kullback-Leibler ”distance” concept is erroneous [260].

For that reason the Jensen-Shannon distance (JSD) has also been computed.

A.2.3 Jensen-Shannon distance

The general Jansen-Shannon divergence is given by

$$H(\sum_{v} \alpha_v P_v) - \sum_{v} H(P_v) = \sum_{v} \alpha_v D_{KL}(P_v, \bar{P}) \tag{A.4}$$

where $\sum_v \alpha_v P_v$ is a mixture of probability distributions with $\bar{P} = \sum_v \alpha_v P_v)$ and $\sum_v H(P_v) < \infty$. Giving to the involved distributions various weights α_v based on their relative importance is one of the key characteristics of the Jensen-Shannon divergence.

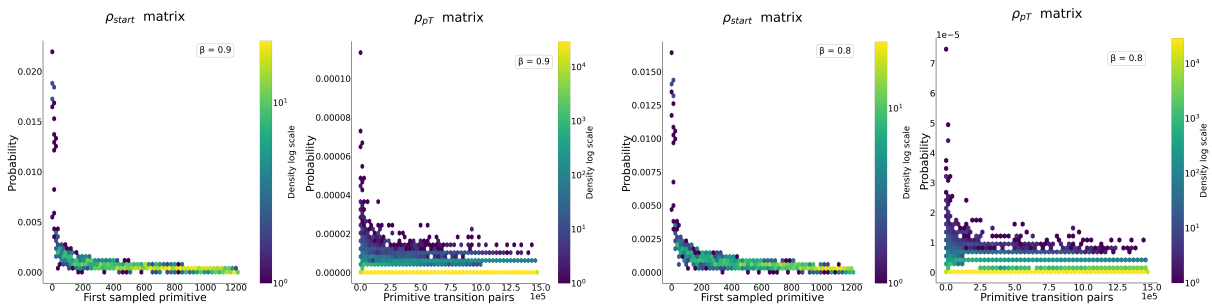
The symmetrized and smooth version of KLD is called the specific Jansen-Shannon divergence and it is given by

$$JSD(P,Q) = \frac{1}{2} D_{KL}(P||M) + \frac{1}{2} D_{KL}(Q||M) \text{ with } M = \frac{1}{2}(P+Q) \tag{A.5}$$

where M corresponds to the uniform mixture of the two probability distributions [261–264]. The square root of the JSD is a metric often referred to as Jensen–Shannon distance and it was computed for the same combination of probability distributions mentioned in section A.2.2. JSD is bounded between 0 and 1.

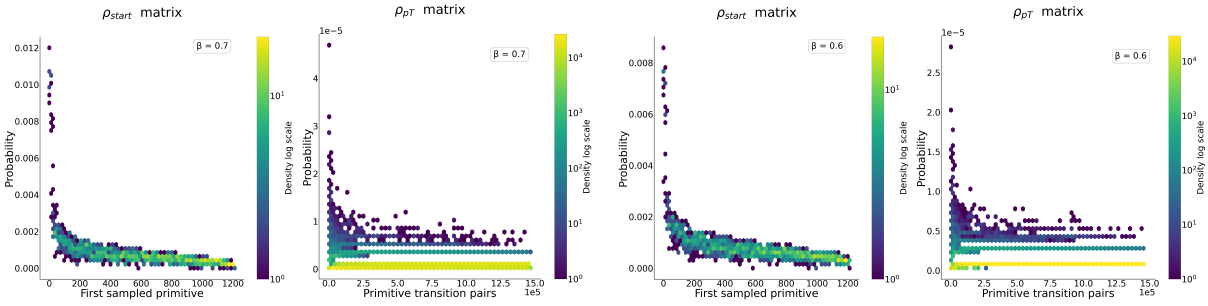
A.3 Diffusion-based perturbations effect

Diffusion-based perturbation effect on the model priors across different β simulation values.



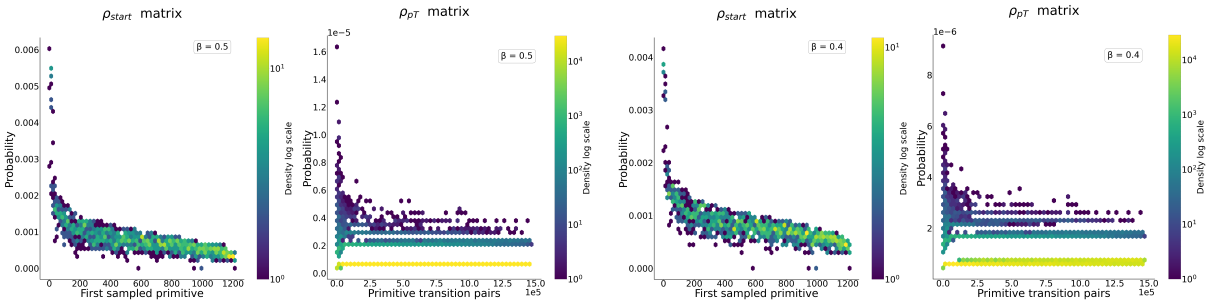
(a) ρ_{start} and ρ_{pT} for $\beta = 0.9$

(b) ρ_{start} and ρ_{pT} for $\beta = 0.8$



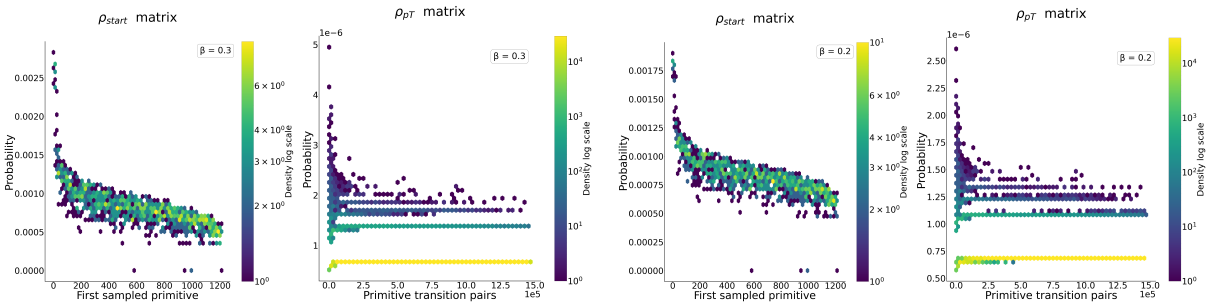
(c) ρ_{start} and ρ_{pT} for $\beta = 0.7$

(d) ρ_{start} and ρ_{pT} for $\beta = 0.6$



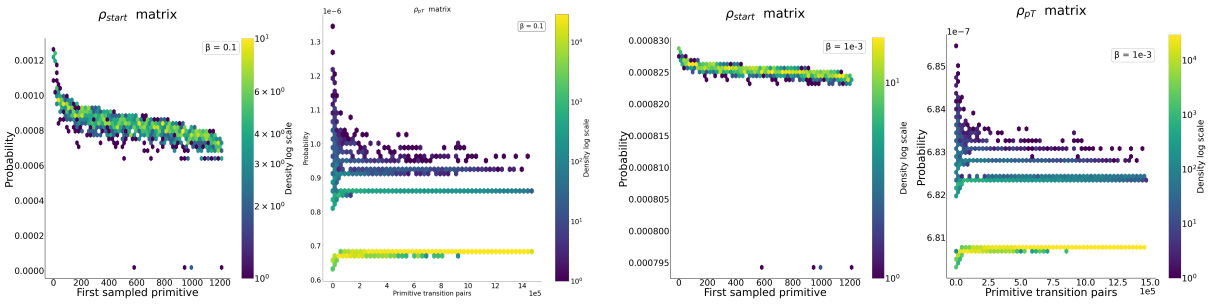
(e) ρ_{start} and ρ_{pT} for $\beta = 0.5$

(f) ρ_{start} and ρ_{pT} for $\beta = 0.4$



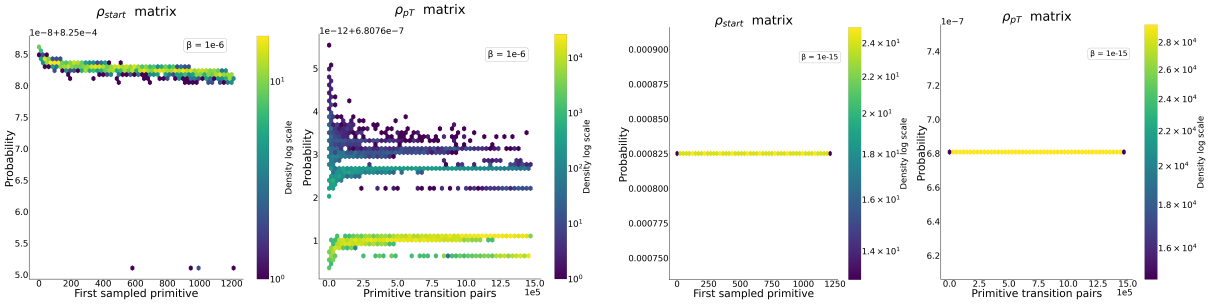
(g) ρ_{start} and ρ_{pT} for $\beta = 0.3$

(h) ρ_{start} and ρ_{pT} for $\beta = 0.2$



(i) ρ_{start} and ρ_{pT} for $\beta = 0.1$

(j) ρ_{start} and ρ_{pT} for $\beta = 1e-3$

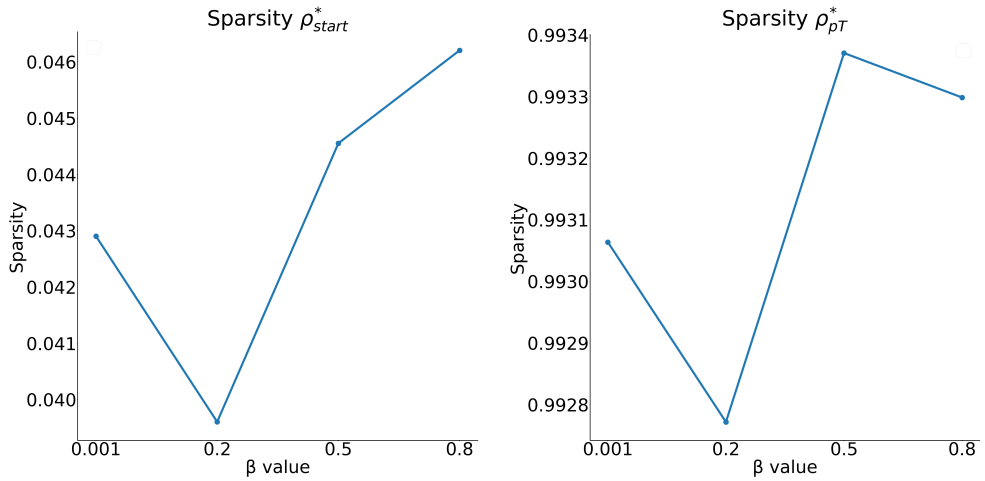


(k) ρ_{start} and ρ_{pT} for $\beta = 1e-6$

(l) ρ_{start} and ρ_{pT} for $\beta = 1e-15$

Figure A.2: Diffusion-based perturbations of the original model priors for different β parameters resulting in the new perturbed priors ρ_{start} and ρ_{pT}.

A.4 Novel priors sparsity

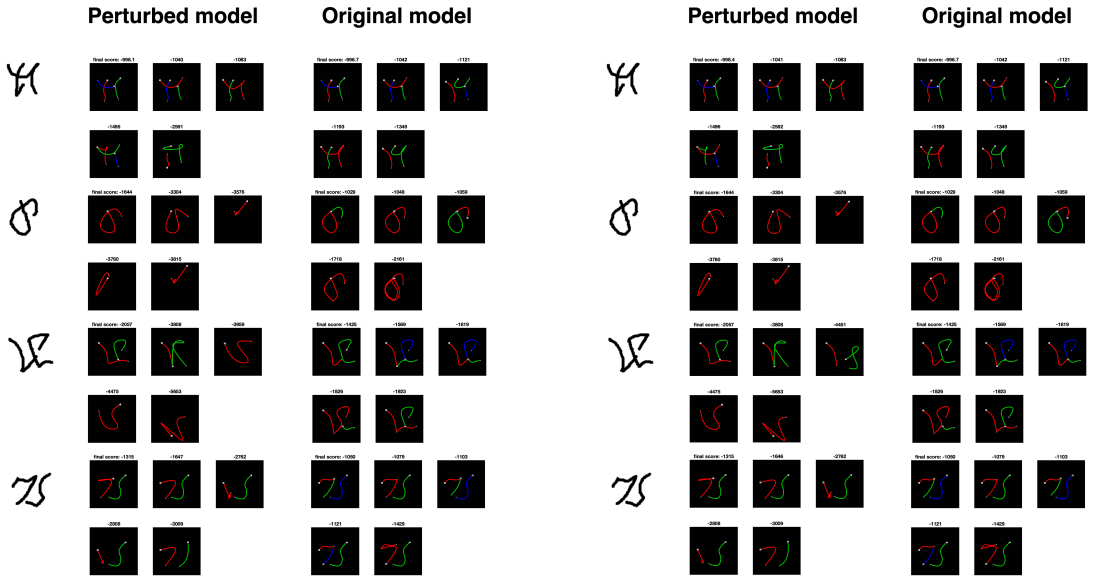


(a) ρ^*_{start} matrix sparsity for the different β values.

(b) ρ^*_{pT} matrix sparsity for the different β values.

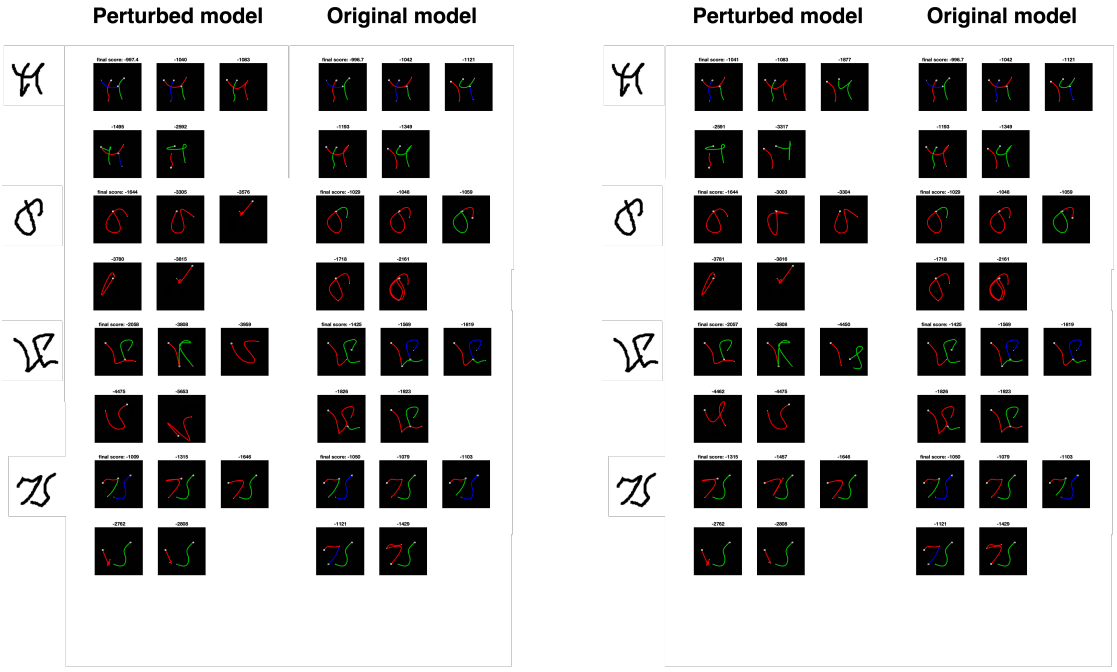
Figure A.3: New prior's sparsity values.

A.5 Fitting examples



(a) Fitting results of model resulted from $\beta = 1e-3$ perturbation.

(b) Fitting results of model resulted from $\beta = 0.2$ perturbation.

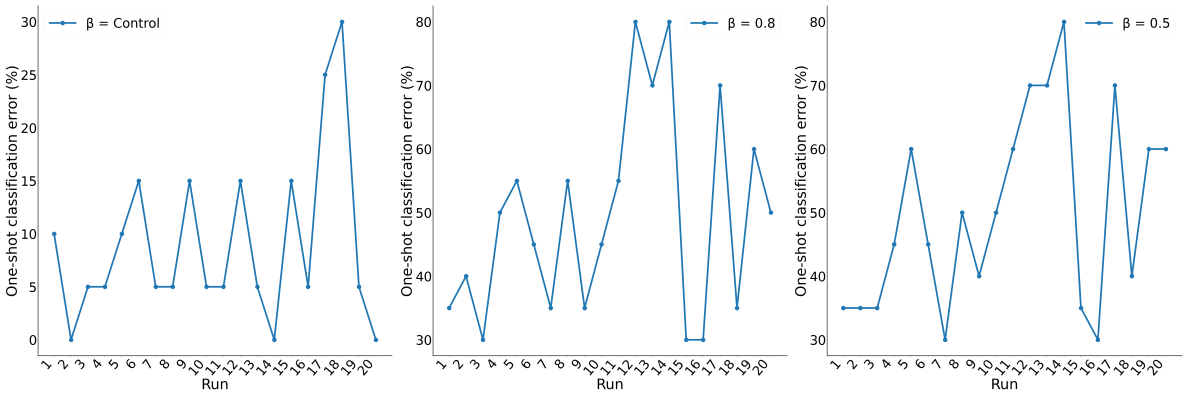


(c) Fitting results of model resulted from $\beta = 0.5$ perturbation. (d) Fitting results of model resulted from $\beta = 0.8$ perturbation.

Figure A.4: Comparing a few examples of fitting results of the run 1 test images leveraging the original model (on the right) and the perturbed models (on the left) with **(a)** $\beta = 1e-3$, **(b)** $\beta = 0.2$, **(c)** $\beta = 0.5$, **(d)** $\beta = 0.8$.

A.6 One-shot classification results

The one-shot classification results for every classification task run.



(a) (b) (c)

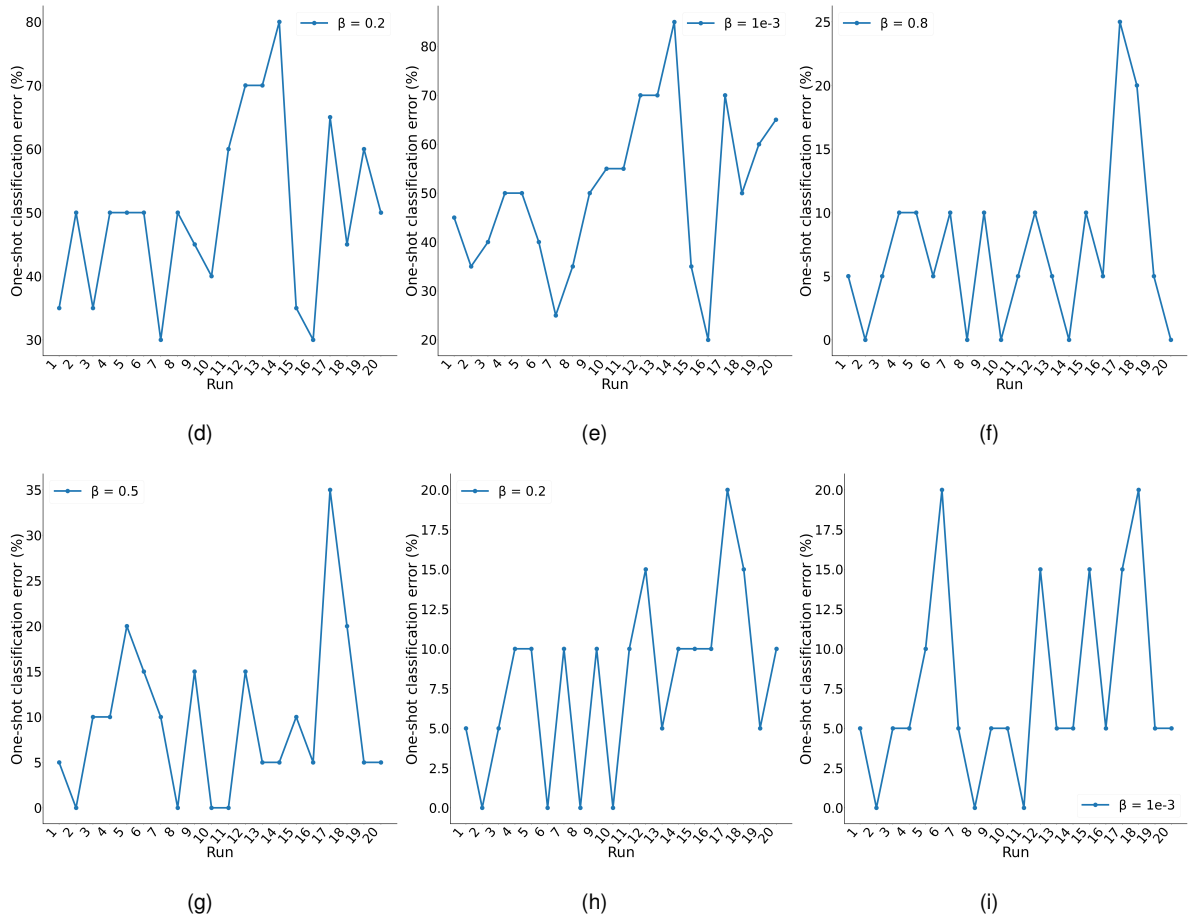


Figure A.5: One-shot classification errors for each run leveraging **(a)** the original BPL model (control) and the DL-BPL models which resulted from the the first pipeline with **(b)** $\beta = 0.8$, **(c)** $\beta = 0.5$, **(d)** $\beta = 0.2$, **(e)** $\beta = 1e-3$ and from the alternative pipeline **(f)** $\beta = 0.8$, **(g)** $\beta = 0.5$, **(h)** $\beta = 0.2$, **(i)** $\beta = 1e-3$.

Table A.1 shows an example of a scoring matrix of the Bayesian score classification rule for the classification run 1 from the pipeline using the perturbed model parameterized with $\beta = 1e-3$. In bold the highest scores are indicated corresponding to the selected classification between images.

A.7 Weight computation for prior estimation.

Weight computation for the integrative estimation between the original model priors and the inferred priors.

Table A.2: Weight computation for prior estimation.

Diffusion-based perturbation	Omniglot images	Generated perturbed images	Total images	w_1	w_2
$1e-3$	19280	14715	33995	0.567143	0.432857
0.2	19280	14647	33927	0.568279	0.431721
0.5	19280	14788	34068	0.565927	0.434073
0.8	19280	14664	33944	0.567995	0.432005

Table A.1: Bayesian scores table of Run 1 of perturbed model with $\beta = 1e-3$.

Images	Train 1	Train 2	Train 3	Train 4	Train 5	Train 6	Train 7	Train 8	Train 9	Train 10
Test 1	-1126,42	-1489,67	-977,87	-1223,25	-1849,87	-1620,51	-1542,58	**-268,19**	-698,98	-1586,17
Test 2	-4676,68	**-3025,68**	-3172,52	-3694,57	-4800,64	-3884,89	-4164,34	-3530,07	-3431,27	-4289,11
Test 3	-4131,01	-3693,66	**-1894,29**	-3519,90	-4522,39	-3461,97	-4408,65	-3172,11	-3445,96	-4317,21
Test 4	-3974,11	-4202,36	-3728,05	-3449,01	**-3369,16**	-4266,92	-4188,85	-3692,19	-3529,56	-4337,47
Test 5	-4188,71	-4426,61	**-3013,72**	-4606,27	-3600,96	-4600,87	-4614,06	-3738,29	-4194,97	-4937,44
Test 6	-3712,21	-3515,48	-2678,81	-3332,03	-4311,32	**-1896,47**	-3893,88	-3116,25	-2955,43	-4043,00
Test 7	-3998,61	-3869,27	-3129,06	-3855,39	-4126,50	-3342,03	**-2822,74**	-3442,76	-3556,73	-4417,61
Test 8	-3578,03	-3629,15	-3392,48	-3542,44	-4135,11	-4032,38	-3717,32	**-2295,60**	-2823,72	-4073,84
Test 9	-3740,08	-3933,54	-3288,67	-3352,27	-4282,95	-3824,87	-4155,47	-3659,97	**-2595,22**	-4270,09
Test 10	-4072,40	-3363,09	-3286,26	-3269,76	-3935,15	-3650,91	-3806,92	-3166,99	-3346,93	**-2604,12**
Test 11	-4134,65	-4215,72	-3390,42	-3554,45	-5042,86	-3587,02	-4225,83	-3639,29	-3158,38	-4703,88
Test 12	-1472,98	-1499,73	-1420,89	-900,75	-2230,47	-1598,73	-1862,37	-1400,92	**-761,62**	-2182,96
Test 13	-1536,12	-1824,92	-1590,25	**-983,75**	-2272,38	-2189,29	-2467,57	-1475,29	-1175,68	-2324,52
Test 14	-4103,54	-4480,14	-3556,24	-4071,75	-4159,55	-4524,73	-4568,29	**-3529,54**	-3948,66	-4270,81
Test 15	-3821,21	-3277,18	**-2701,51**	-3126,36	-4000,59	-3928,81	-3954,90	-2794,65	-3199,07	-3790,72
Test 16	-2524,95	-1878,28	-1532,61	-2427,64	-2624,19	-2441,47	-2175,67	**-1176,41**	-1713,66	-2441,82
Test 17	-3989,27	-3433,68	-3584,02	-3764,48	-5126,45	-4050,96	-4097,19	-3771,23	**-3342,92**	-4546,89
Test 18	-2340,70	-2313,40	-1803,00	-2150,18	-2471,58	-1708,98	-1923,37	-1787,15	-1611,09	-2392,44
Test 19	-4789,59	-4435,20	-3602,61	-4117,18	-5214,68	-4259,30	-4909,20	-3377,52	-3801,09	-5075,79
Test 20	-1772,34	-1255,46	**-421,98**	-1498,88	-1756,07	-1519,69	-1678,38	-473,26	-674,96	-2070,61

Images	Train 10	Train 12	Train 13	Train 14	Train 15	Train 16	Train 17	Train 18	Train 19	Train 20
Test 1	-1299,48	-952,67	-1790,04	-1749,64	-2001,17	-758,41	-1774,98	-1482,92	-598,29	-1280,12
Test 2	-3901,05	-4134,85	-4105,40	-4800,37	-4555,22	-3890,82	-4336,16	-4555,31	-3737,69	-4402,91
Test 3	-2860,51	-2938,24	-4507,07	-4194,97	-4513,57	-3914,64	-4202,47	-4394,92	-3822,13	-3781,84
Test 4	-4573,95	-4440,63	-3766,49	-4214,69	-4821,50	-4204,59	-4348,32	-4683,85	-3832,42	-4307,90
Test 5	-4205,35	-5062,43	-4178,98	-4457,46	-5076,14	-4291,66	-4866,91	-5101,13	-4187,69	-4536,48
Test 6	-3061,04	-3069,84	-3936,28	-4347,60	-4142,93	-3227,98	-4141,90	-3658,53	-3246,68	-4113,64
Test 7	-3806,79	-3950,11	-3400,57	-4416,78	-4679,62	-3581,16	-4161,47	-3784,07	-3095,61	-4069,85
Test 8	-4006,18	-3540,14	-4188,34	-4081,01	-4565,39	-3251,22	-3806,39	-3907,73	-2899,35	-3822,81
Test 9	-4104,41	-3175,36	-4255,07	-4810,34	-4306,57	-4357,37	-4244,23	-4045,27	-3517,17	-4380,32
Test 10	-3546,62	-3942,45	-3272,75	-4243,72	-4703,08	-3500,39	-4337,97	-3529,75	-3890,52	-4328,76
Test 11	**-2655,43**	-2697,30	-5002,38	-4588,00	-5082,75	-4226,39	-4124,68	-4365,07	-3543,98	-4408,71
Test 12	-1592,54	-1226,27	-2389,50	-2636,37	-2529,87	-1891,90	-2221,07	-1821,89	-951,89	-2227,48
Test 13	-2256,29	-2028,95	-2159,92	-2871,38	-2776,18	-2205,83	-3115,72	-2255,25	-1805,07	-2315,02
Test 14	-4504,51	-4335,85	-4417,52	**-2229,64**	-5079,31	-4192,51	-4872,72	-5160,81	-3989,57	-4859,46
Test 15	-4044,01	-3623,64	-3486,29	-3586,59	-3467,59	-2971,26	-3873,07	-4291,77	-3305,63	-3731,74
Test 16	-2228,08	-2337,92	-2332,01	-2103,64	-3132,99	-1258,45	-2737,58	-2686,35	-1695,18	-2181,40
Test 17	-3434,02	-3632,11	-4135,30	-4795,53	-4817,60	-4320,73	-3366,17	-3817,68	-3757,79	-4883,62
Test 18	-2062,01	-1646,04	-2233,22	-2599,06	-2489,90	-2114,47	-2304,55	**-1598,71**	-1698,12	-2602,78
Test 19	-3977,71	-3876,06	-5010,37	-5005,38	-4783,03	-3512,68	-5315,18	-4990,50	**-3065,23**	-4246,53
Test 20	-1578,34	-1219,71	-1668,72	-1437,68	-1709,06	-841,13	-2221,25	-1988,06	-1030,52	-1236,78

A.8 Affinity analysis

Following [250] affinity definition: Let D_{train} and D_{val} be training and validation datasets drawn IID from the same clean data distribution, and let $D_{val}^{'}$ be derived from D_{val} by applying a stochastic augmentation strategy, a, once to each image in D_{val}, $D_{val}^{'} = (a(x), y) : \forall (x, y) \in D_{val}$. Furthermore, let m be a model trained on D_{train} and $A(m, D)$ denote the model's accuracy when evaluated on data set D. The Affinity, $T[a; m; D_{val}]$, is given by $T[a; m; D_{val}] = A(m, D'_{val}) - A(m, D_{val})$. The metric of affinity offers several benefits, including ease of measurement as it only requires clean training of the model. Additionally, it is not affected by any potential interaction between data augmentation and training because augmentation is only utilized on the evaluation set. Furthermore, affinity serves as a distance measurement that is responsive to characteristics of both the model and data distribution.

A.9 Code

The code for this thesis purpose was developed in Python and Matlab, recurring to [265] and [6] repositories, mainly consisting of the implementation of diffusion-based perturbation framework, Dirichlet Process, image processing algorithm, integrative estimation pipeline and adaptation of BPL generative, inference, refitting and classification algorithms.