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

Carolina Caramelo
carolina.caramelo@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisboa, Portugal

March 2023

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 programming 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.

1. Introduction

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 has facilitated the systematic generation and study of psychoactive drugs. Third, many common psychiatric diseases and depressive disorders, increasingly present in the general population, 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 clinical phase 2 trials that have shown potential long-term outcomes in positively impacting the symptomatology of these patients [2]. 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 focus on understanding the influence of psychedelics within the context of internal models and neural simulation of the psychedelic experience. In psychedelic-assisted psychotherapy (PAP), reported experiences have led to long-term conceptual re-organization of the individuals' perspectives [3]. 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 [3]. Trying to simulate the internal psychedelic experience through an adequate computational framework [4], while simultaneously exploring it in the context of machine learning (ML) model performance enhancement, is the focus of this work.

2. Background

2.1. Psychedelics and the serotonergic system

Psychedelic drugs are classified into classic psychedelics and atypical/non-traditional/non-classic psychedelics based on their neuro-receptor affinities and chemical structure, which together determine the primary mode of action of the substance [5]. Following the discovery of LSD and the identification of serotonin (5-HT),interest grew in the potential role of the interaction between psychedelic drugs and 5-HT systems [6]. Evidence suggests that hallucinogens mainly act as agonists at the 5-HT2A receptor (5-HT2AR), which in humans is highly expressed in the apical dendrites of excitatory glutamatergic layer 5 pyramidal (L5p) neurons in the cortex [7], being a predominantly cortical receptor and the most abundant 5-HT receptor in the cortex, especially in the prefrontal cortex (PFC) [7, 8]. 5-HT2AR antagonists have shown to substantially reduce or abolish the subjective effects of psilocybin, LSD, and DMT in humans [9, 10, 11]. While the activation of this receptor is suggested to serve as a necessary intermediary of the distinctive subjective effects of classic psychedelic substances, it is not the sole cause of these effects, as it is also involved in changes in glutamate transmission [12], thalamocortical network modulation [13], and neuroplasticity [14].

2.2. Psychological and clinical implications

The effects of classic psychedelics are strongly reliant on the user's expectations (*set*) and the context (*setting*) in which the usage occurs, making the experience subjective. These are determining factors for the therapy's success [15].

Psychometrically validated questionnaires have been used to compare the effects of psychedelics, reporting alterations in perceptual, emotional and cognitive domains. Usual reports of *peak* experiences include both pleasant and negative feelings of ego-disintegration [16]. Overall, PAP clinical studies have been showing that psychedelics impact on fundamental aspects of the experienced sense of self [17] and have been proven to function quickly and have long-lasting effects after only a few sessions/doses in people with different psychological disorders. Nevertheless, it requires cautious generalization of findings and more controlled research [2]. It is unclear whether awareness of psychedelic-induced experiences is necessary for therapeutic success, and modern neuroimaging may provide insights into the compounds' action at higher brain levels and their potential in therapy [13].

2.3. Alterations in whole-brain functional organization

Neuroimaging studies have revealed alterations in whole-brain functional organization, connectivity and dynamics under the influence of psychedelics [18, 19]. Main results include reduced connectivity within the Default Mode Network[1] and increased between-network connectivity [18, 19, 20], shifting the brain towards an increased global functional integration. It is suggested that these changes lead to a greater repertoire of brain states and increased entropy, aligned with subjective reports of altered perception and self-processing [18, 19, 20, 10]. The decreased connectivity and activity of central brain networks may facilitate a state of unconstrained cognition and desynchronized cortical activity, approximating the brain to criticality[2] and potentially dismantling reinforced negative thought patterns [18, 19, 20, 10, 21, 3].

2.4. Computational theories

Psychedelic neural correlates and psychological findings have been unified in a computational theoretical framework substantiated on hierarchical predictive coding [3]. Relaxed beliefs under psychedelics (REBUS) and the anarchic brain is a unifying model, based on the principle that psychedelics affect cortical-activity, leading to the relaxation of the precision weighting of one's high-level priors, i.e beliefs, by liberating bottom-up information and constricting top-down information flow [3], resulting in a decreased ability of high-level expectations exert hierarchical control over and be resistant to the impact of lower-level brain regions. This model explains the range of subjective phenomena associated with psychedelic experiences including ego dissolution [22], peak experiences [23], near-death-like experiences [24], the sense of anxiety and uncertainty [21], between others. Moreover, REBUS proposes that under psychedelics there the brain's energy landscape flattens, making attract-

[1] High-level brain network active in resting state mode and associated with the sense of 'ego" [18].

[2] A transition zone between ordered and disordered states of consciousness [21]

ing brain states that encode beliefs less stable and influential, and the brain enters a mode that allows for potentially lasting re-vision of priors, resulting in broader processing of the inner and outer world, potentially leading to long-term benefits for mental health [3].

This work proposes a computational framework analogous to the psychedelic action on the brain, focusing on the high-level internal narrative of the psychedelic experience, modeled in terms of program induction.

3. Implementation

3.1. Internal models within the probabilistic framework

Internal models are like "small-scale models" of the external world that facilitate humans to imagine various behavioral options and their consequences in a given environment without actually committing any actions [25]. Internal models consist of prior distributions $P(y)$ over sensory signals, and recognition models $P(z|y)$) that compute latent world states z when given sensory input. Generative models explain how sensory data is produced and can be represented as either the product of a state prior $P(z)$ and a conditional distribution of sensory inputs given latent world states $P(y|z)$ or as the joint distribution between sensory input and latent variables $P(y, z)$ [26]. To determine the probability of latent states that may have produced the observed input, Bayes' rule can be applied to invert the generative model given sensory input: $P(y|z) = \frac{P(z|y)P(y)}{P(z)} = \frac{P(z|y)P(y)}{\sum_{y \in Y} P(z,y)}$ [27].

This study investigates how the psychedelic experience affects one's internal models at a cognitive level, specifically their generative and inference phases. Internal models will be conceptualized as abstract semantic and knowledge representations, modeled using Bayesian Program Learning (BPL).

3.2. Bayesian Program Learning

Probabilistic generative models can shed light on cognitive processes such as the way people can form rich and hierarchical models of the world with just a few examples [4]. For instance, Hierarchical Bayesian Models (HBMs) are useful for modeling human learning, learning by discovering the underlying structure in data [28], with higher levels representing general beliefs and lower levels corresponding to observable data [28]. The Bayesian Program Learning (BPL) framework is a HBM model that learns abstract, flexible representations of concepts (handwritten characters) from few examples by representing them as simple programs that best explain observed examples under a Bayesian criterion [4].

3.2.1 BPL model

The BPL model, trained on 30 handwritten character alphabets from the omniglot data set, learns stochastic programs for creating new character concepts, that can be observed in Figure 1.

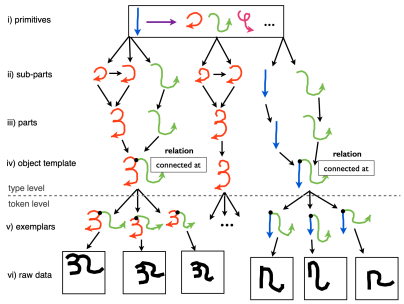


Figure 1: **Bayesian Program learning generative model.** Illustration of the generative process underlying handwritten characters. New types are generated by choosing primitive actions from a learned library (i), combining these sub-parts (*sub-strokes*) (ii) to make parts (*strokes*)(iii), and combining 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) [4].

Characters are composed by *strokes* (parts) comprised by *sub-strokes* (sub-parts), which are connected by spatial *relations* between them. It describes a generative model that is able to sample new character *types* by combining parts and subparts in new ways. The character type is itself a procedure for generating new exemplars of the correspondent concept producing new *tokens* of that same concept. BPL model is 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 [4]. Note that, constructing character "types" involves sampling primitive structures, which are shared and re-utilized across the different characters as *sub-strokes*. The model's joint distribution, on types Ψ, a set of M tokens of the corresponding type $\theta^{(1)}, ..., \theta^{(M)}$ and binary images $I^{(1)}, ..., I^{(M)}$ is

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

The process of generating a character type, more

specifically *strokes*, is the most relevant for this work's purpose. 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 $R = R_1, ..., R_\kappa$. 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 $P(\Psi) = P(\kappa)\prod_{i=1}^{\kappa} P(S_i)P(R_i|S_1, ..., S_{i-1})$, where a *stroke* is a motor routine comprising *sub-strokes* - $S_i = s_{i1}, ..., s_{in_i}$. Generating a *stroke* involves sampling and building a sequence of z_{ij} character primitive indexes, corresponding to the *sub-strokes*, as described by the first-order Markov Process

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

Psychedelic research suggests that modifications to the underlying structure of thought, defined by the connections between entities such as physical objects or people, have a positive impact on people with mental health issues, who were seen to have significant symptom improvements, as well as reporting "new perspectives on life" and "new ways of seeing things", [1], seemingly suggesting that something is happening in the domain of people's beliefs. Interpreting the BPL primitives as concepts that themselves form new concepts, the present study develops a ML pipeline integrating an analogy with psychedelic action on the brain, with the objective of harnessing the potential benefits of data augmentation via diffusive probabilistic programming performance perturbations. The proposed approach seeks to enhance BPL model performance on a one-shot classification task by incorporating perturbations in the generative model latent space, thereby not only simulating real-world variability, but also the psychedelic experience, attempting to increase the diversity of the training data. The four main phases defined within this pipeline are described along with its psychedelic analogy.

3.2.2 Perturbation phase: Diffusion-based perturbations

The psychedelic experience can be seen as a "perturbation" of one's priors, leading to new perspectives and mental constructs, sometimes, resulting in the rearrangement of these [3]. Perturbing the BPL generative model priors, altering its prior knowledge, can introduce novelty when generating new character images by introducing new or removing frequent character primitive transitions. The process in equation 2, part of the generative process of sampling a new character *stroke*, will be the one to be perturbed. This joint distribution represents a first-order Markov Process and depends on two probability distributions:

- $P(z_{i1})$: distribution from where the first primitive index z_{i1} of each *stroke* is sampled (referred as s). It is a 1×1212 vector comprising the probabilities of first sampling the 1212 existent primitives.
- $P(z_{ij}|z_{i(j-1)})$: distribution from where the primitive indexes z_{ij} of the following *sub-stroke* primitives in a *stroke*, are sampled (referred as pT). The distribution is the normalization of a 1212×1212 Markov matrix (pT_M), describing the probabilities of transitioning from one primitive to the other only depending on the previous state.

Inspired by computational theories on psychedelic effects [3], diffusion-based approaches [29], and statistical thermodynamics, a mathematical framework was created for the perturbation. Applying a diffusion heat kernel and softmax function to the original priors s and pT, the perturbed priors ρ_{start} and ρ_{pT}, respectively, are obtained. The diffusive perturbation process is given by

$$\rho_{start}(\beta, x) \propto \frac{e^{-\beta L_s(x)}}{\sum_{x'} e^{-\beta L_s(x')}} \tag{3}$$

$$\rho_{pT}(\beta, x, y) \propto \frac{e^{-\beta L(x,y)}}{\sum_{x',y'} e^{-\beta L(x',y')}} \tag{4}$$

where $L_s(x)$ can be interpreted as a prior on the generative process and $L(x, y)$ as a distance function in the primitive space, closely resembling s and pT_M matrices, respectively.

$$L_s(x) := -\log(s) \tag{5}$$

$$L(x, y) := -\log(pT_M) \tag{6}$$

The β parameter in equations 3 and 4 is what shapes the perturbation, acting as a scaling factor for L, defining the extent of the diffusion, and determining the distance/attention landscape between different primitives. When viewed as a temperature constant, higher values of β (lower temperature values) result in a higher attention to the original distances between primitives, contributing to a stiffening of the original prior. Conversely, lower values ofβ (higher temperature values) reduce attention to the original distances, resulting in a flattened prior landscape and making rare primitive transitions more likely to observe in *stroke* samples. 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 priors. The perturbed model will be designated as "Diffusive latents for Bayesian Program Learning" (DL-BPL). 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 $\beta = 1e-3, 0.2, 0.5, 0.8$

3.2.3 Generative phase: Data augmentation via a generative alphabet procedure

The generative phase can be understood as the simulation of one's experience under the influence of psychedelic drugs. After perturbing the model, each DL-BPL model produced a new set of character images arranged in alphabets, which can be defined as groups of related characters or concepts. These alphabets aim to maintain a similar organization to the original omniglot data set and reflect the organization of related mental constructs in our minds. The relatedness between characters in an alphabet may be due to shared experiences, memories, or contexts, implying an underlying structure to our thoughts and ideas during a psychedelic experience, even if subtle. Generating a new alphabet requires adding an extra hierarchy layer to BPL's generative model, which involves adding a prior to re-use the structural components within a set of related characters [4]. The additional level is created using the Dirichlet Process (DP), a tool in ML and statistics that facilitates this process. The new data set generated by each DL-BPL model consists of 30 alphabets, each containing 25 character images and 20 exemplars (tokens) of each character image. Following this, the inference phase was initiated.

3.2.4 Inference phase: Learning a new model prior

During this phase, the goal was to simulate the formulated perspectives and their consolidation resulting from a psychedelic experience, producing a new model prior described by equation 2. Posterior inference was performed on the 30 newly generated alphabets for each DL-BPL model, using the latent variables representing the indexes of sampled character primitives to compute "intermediate" ρ^*_{start} and ρ^*_{pT} distributions. The images were processed and fed to the BPL model, and posterior inference involved parsing these into *strokes* and *sub-strokes*. A frequency matrix describing the frequency of primitive transitions in the character images was computed for each generated data set, then normalized. ρ^*_{start} was computed based on the first *sub-strokes*in each *stroke*, while every other transition between subsequent *sub-strokes* in each *stroke* was used for ρ^*_{pT} calculation. Secondly, an integrative estimation combining the computed priors and the original priors was calculated, resulting in the final priors for the classification task e_{start} and e_{pT}:

- The entries of ρ^*_{start} and ρ^*_{pT} with a non-zero value were selected, corresponding to the first sampled primitives in a character *stroke* and the transitions between the primitives of the remaining *sub-strokes* in a *stroke*, with a probability of occurring different from zero, respectively.
- Identified novel entries by comparing the above selected entries with the original priors. Novelty in the generated characters was defined as the observation of character primitives that showed a the five lowest sampling probabilities in the original priors and a higher probability than that in the intermediate priors
- Estimation was computed by **(1)** replacing the corresponding entries of the final priors with the identified novel transitions; **(2)** replacing the final priors entries correspondent to zero entries in the intermediate distributions with the original priors correspondent values; **(3)** obtaining the remaining entries in the final priors by computing weighted sum[3] between the correspondent values of the original and computed priors.
- Estimated final priors were normalized and updated in the DL-BPL library.

3.2.5 Classification phase: Evaluating model's performance

In the last step of the pipeline, a one-shot classification task from Lake et al.(2015) [4] was performed to evaluate the perturbed DL-BPL model's performance, analogous to ask "How different is the person's generalization about the world after the psychedelic experience?". 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 two-way Bayesian classification rule [4]

$$\begin{aligned} &\arg\max_c \log P(I^{(T)}|I^{(c)}) = \\ &\arg\max_c \log P(I^{(T)}|I^{(c)})^2 = \\ &\arg\max_c \log \left[\frac{P(I^{(c)}|I^{(T)})}{P(I^{(c)})} P(I^{(T)}|I^{(c)})\right] \end{aligned} \tag{7}$$

with $P(I^{(c)}) \approx \sum_i \tilde{w_i}$[4]. Classification was performed in the omniglot evaluation data set images, consisting of the 20 alphabets. The task involved 20 runs of 20 within-alphabet images classification series.

[3] Weights were calculated proportionally to the number of characters in the omniglot and the number of characters in the perturbed alphabets, respectively.

[4] Motor program inference score.

4. Results and discussion

4.1. Diffusion-based perturbations

This work proposes the use of diffusion-based perturbations on model priors to generate a "flattening" effect [3], weakening expectations regarding the sampled primitives when generating a new character, leading to more flexible inference and higher classification performance. This is achieved through diffusive perturbations in the latent space of character primitives, with the goal to preserve the essential structure of the probability distribution in the latent space of *characters* while reducing its concentration.

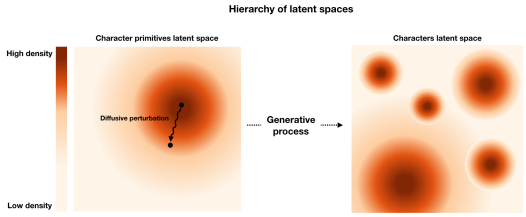
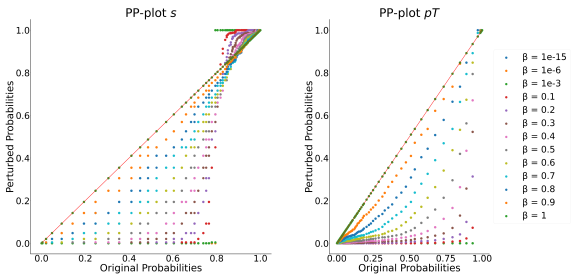


Figure 2: **Conceptual representation of latent spaces hierarchy.** We are hoping to induce novelty in the characters space through diffusive perturbations in the space of character primitives, by preserving its essential structure but flattening the probability density landscape. This figure represents our hypothesis that perturbations in the primitive layer will manifest as more complex and multimodal changes in the character layer.

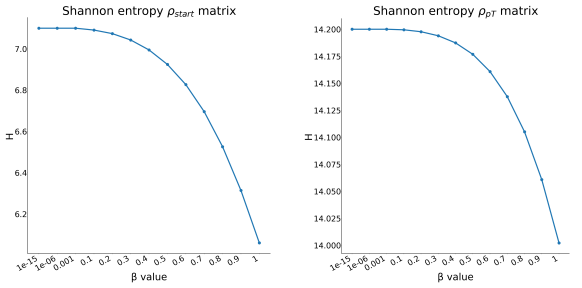
The effect of diffusion-based perturbation on s and pT priors was simulated for different β parameters. The result is shown in Figure 3, where the Probability-Probability (P-P) plots are used to compare the cumulative distribution functions (CDFs) of the original and perturbed distributions. The deviation of the points from the 45-degree line is analyzed to compare the distributions.



(a) P-P plot s and ρ_{start}. (b) P-P plot pT and ρ_{pT}.

Figure 3: **P-P plots** of **(a)** s and **(b)** pT distributions and respective distribution perturbations, for different β values.

As β decreases, the degree of deviation from the linear function increases, the distributions deviate more from the original priors. Additionally, with decreasing β, the distribution probability values globally decrease, and the matrices' mass becomes concentrated in a smaller probability value range. As β approaches $1e-15$ value, the distribution structure disintegrates, converging on a uniform distribution.



(a) (b)

Figure 4: Shannon entropy H (in bits) measure of the new perturbed priors **(a)** ρ_{start} and **(b)** ρ_{pT} for different β values.

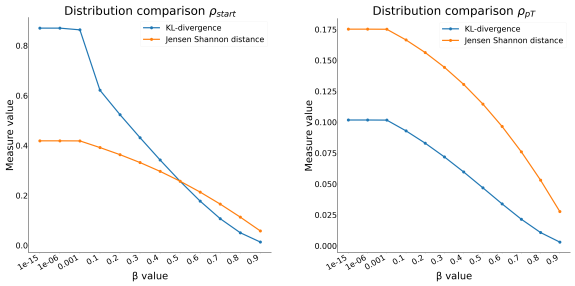
Furthermore, Shannon entropy of the distributions decreases with increasing β as seen in Figure 4. This supports the idea that a diffusion-based perturbation with a β value closer to zero corresponds to a higher entropy distribution state. Perturbing the s and pT distributions of a model approximates them to a state of increasing disorder (higher entropy), which aligns with the concept of brain approximating a higher entropy state under psychedelics [21, 3]. In the context of PAP, mental illnesses are believed to stem from reinforced attractor states, which lead to rigid thinking and behavior patterns [30]. Psychedelics have the potential to break these reinforced patterns by inducing a flattening effect [3]. Analogously, perturbing the model primitive space could change the attractor landscape and create a modified latent space with new attracting poles as illustrated in Figure 2.

Besides this, analysing Kullback–Leibler divergence (KLD), as well as Jensen-Shannon distance (JSD) values allows us to identify the differences in two data distributions, understanding how much change we are inducing when replacing the model's prior for the new perturbed priors.

Figure 5 shows a decrease in KLD and JSD values with an increasing β value, once again leading to the confirmation that there is a higher information[5] 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.

To increase the probability of seeing characters

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



(a) (b)

Figure 5: Kullback–Leibler divergence and Jensen-Shannon distance between **(a)** original s and perturbed ρ_{start}, **(b)** original pT and perturbed ρ_{pT} for different β parameter values.

with less likely primitives during the generative process sampling, the model was perturbed with $\beta = 1e-3, 0.2, 0.5, 0.8$, considering a trade-off between prior structure and entropy, avoiding completely lesioning the priors' structure and uniformity. The perturbed priors replaced the original ones in BPL's library, originating a DL-BPL model for each β. Based on the broad concept of diffusion processes [29], it is possible conclude that our core innovation in applying diffusive perturbations to do data augmentation in the context of probabilistic programming.

4.2. Generative phase

The DL-BPL models with varying β generated 30 new perturbed alphabets, each containing $15,000$ new characters. The generative process used an optimized Dirichlet Process concentration parameter $\alpha = 4.5$ to preserve the essential structure of characters within one alphabet, including the number of *strokes* and primitive indexes which comprise them. However, the desire of some variability among alphabet characters for flexible inference was also considered.

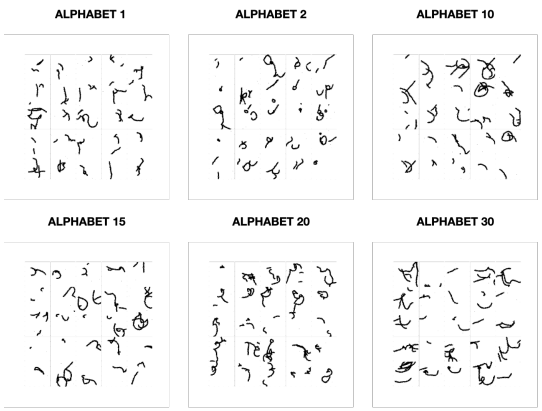


Figure 6: Some examples of the DL-BPL generated alphabets with $\beta = 1e-3$.

Despite efforts to optimize the generative process of DL-BPL models, limitations were observed, including similar characters within the same alphabet were generated, decreasing variability within the data set. Additionally, the generated characters lacked the structured representation found in human-drawn omniglot characters, and some images had ink outside of the frame, factors that can affect the inference process of latent primitive indexes. Nevertheless, this methodology is unique as it employs a diffusion-based perturbation framework on a generative latent space for data augmentation, aiming to investigate the effectiveness of diffusive perturbations, hyperparameter tuning and inference-based data augmentation as a viable strategy for enhancing model performance. This involves embedding BPL primitives, applying a heat kernel, ultimately generating new data and an inference step, necessary to update the model priors to account for this new perturbed data.

4.3. Inference phase

After inference, the estimated priors e_{start}, contrasting to ρ_{start}, demonstrate a higher similarity to the original priors with lower KLD and JSD values. On the other hand, e_{pT}, contrasting to ρ_{pT} demonstrate a higher difference to the original pT distribution, with higher KLD and JSD values. Besides this, 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, while distribution comparison between pT and e_{pT} show stable KLD and JSD values across β parameters, evidencing a "lost" perturbation effect when updating the pT prior after data augmentation. The differently parameterized DL-BPL models were updated with the estimated priors for the one-shot classification task.

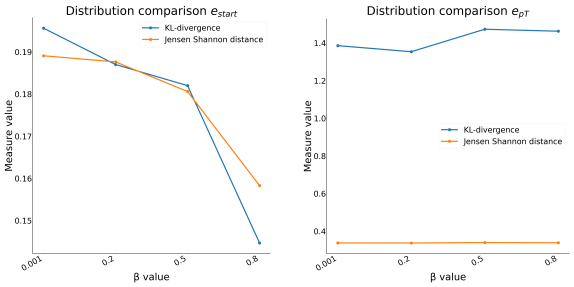


Figure 7: Kullback–Leibler divergence and Jensen-Shannon distance between **(a)** original s and e_{start}, **(b)** original pT and e_{pT} for different β parameter values.

4.4. Classification phase

The study by Lake et al. (2015) achieved a 3.3% error in a one-shot classification task using BPL [4].

However, we were unable to reproduce the results and achieved a 9% error instead, which was considered as the baseline result. From Figure 8 it is possible to observe that that the best classification average episode classification error was 7% for $\beta = 0.8$, slightly lower than the 9% control result. The average error for $\beta = 1e-3$ was 7.5%, for $\beta = 0.2$ was 8% and for $\beta = 0.5$ 9.5%, concluding that was not possible to establish a direct correlation between decreasing β value and decreasing classification error, possibly due to finite sampling variability in the data augmentation generative process.

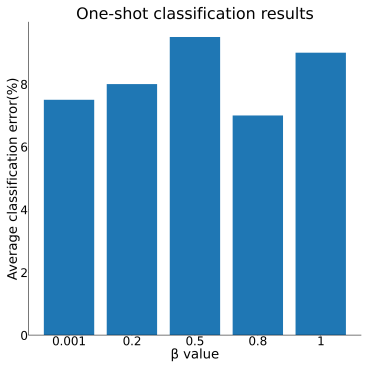


Figure 8: Average episode one-shot classification error for each perturbed DL-BPL model and control.

The omniglot data set was created for studying one-shot learning in humans and machines [4]. Several approaches have been used to improve the performance of one-shot classification tasks on this data set, including the use of different models and augmented datasets. A progress report published three years after the data set's release shows the results obtained by different models, including BPL[6]. Figure 9 shows the interesting results obtained by different models [31].

A more recent study attempts to improve the performance of a generative neuro-symbolic model (GNS)[7] obtaining a test error rate of 5.7%.The present study aimed to improve one-shot classification task performance under an analogy of how psychedelics act on the brain. The model was not trained again, but a different pipeline design was developed, perturbing the priors using a diffusion-based process, resulting in the augmentation of the omniglot dataset in 750 classes and 1714 characters. The best obtained result was 7% test error rate, resulting from the perturbation with $\beta = 0.8$, showing a promising result when compared to other models that augmented the data set and to the control experiment. This results suggests that the psychedelic analogy hypothesis could be a first step towards high-level computational modeling of psychedelics and an avenue for investigating diffusive data augmentation in the probabilistic induction field.

	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 9: **One-shot classification error rate across models.** "One-shot classification error rate for both within-alphabet classification [4] and between-alphabet classification [32], either with the "Original" background set or with an "Augmented" set that uses more alphabets and character classes for learning to learn. The best results for each problem formulation are bolded." * Results used additional four-fold class augmentation and many other augmentations such as scaling, shearing, translations, etc. [33]. Adapted from Lake et. al (2019) [31].

5. Conclusions

In contrast with existing computational work on psychedelic drugs 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 programming models as an alternative approach to computationally formalizing how psychedelics might lead to new perspectives in PAP. From a ML standpoint, this pipeline works towards improving BPL model's performance in a classification task through this data augmentation procedure. Our 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 [4], and, therefore, the

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

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

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, showing that this methodology is a promising avenue of investigation to further refine and explore in the context of probabilistic induction.

5.1. 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, namely defining novelty in the perturbed alphabets proved to be a significant challenge and should be investigated. Furthermore, in addition to updating primitive sampling priors, training the model with the augmented data set should also be considered. Concluding, the use of neural networks in computational modeling and psychedelic research offer a promising avenue for exploring the effects of psychedelic drugs on brain circuits, by approximating models to biology. It is suggested that there should be more emphasis on studying the high cognitive level effects of psychedelics rather than just their visual hallucinatory effects. We believe this studies can lead to improved psychiatric treatment and precision medicine.

Acknowledgements

The author would like to thank Dr. Daniel McNamee (Champalimaud Foundation) and Prof.Dr. Cláudia Lobato da Silva (SCERG,iBB-IST) for all the support in the development of this project.

References

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

[2] Bruno Romeo, Laurent Karila, Catherine Martelli, and Amine Benyamina. Efficacy of psychedelic treatments on depressive symptoms: A meta-analysis. *Journal of Psychopharmacology*, 34(10):1079–1085, 2020.

[3] Robin L Carhart-Harris and KJ6588209 Friston. Rebus and the anarchic brain: toward a unified model of the brain action of psychedelics. *Pharmacological reviews*, 71(3):316–344, 2019.

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

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

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

[7] Elaine T Weber and Rodrigo Andrade. Htr2a gene and 5-ht2a receptor expression in the cerebral cortex studied using genetically modified mice. *Frontiers in neuroscience*, 4:36, 2010.

[8] Katarina Varnäs, Christer Halldin, and Håkan Hall. Autoradiographic distribution of serotonin transporters and receptor subtypes in human brain. *Human brain mapping*, 22(3):246–260, 2004.

[9] Katrin H Preller, Marcus Herdener, Thomas Pokorny, Amanda Planzer, Rainer Kraehenmann, Philipp Stämpfli, Matthias E Liechti, Erich Seifritz, and Franz 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.

[10] Katrin H Preller, Leonhard Schilbach, Thomas Pokorny, Jan Flemming, Erich Seifritz, and Franz 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.

[11] Rainer Kraehenmann, Dan Pokorny, Helena Aicher, Katrin H Preller, Thomas Pokorny, Oliver G Bosch, Erich Seifritz, and Franz X Vollenweider. Lsd increases primary process thinking via serotonin 2a receptor activation. *Frontiers in pharmacology*, 8:814, 2017.

[12] Gerard J Marek. Interactions of hallucinogens with the glutamatergic system: permissive network effects mediated through cortical layer v pyramidal neurons. *Behavioral Neurobiology of Psychedelic Drugs*, pages 107–135, 2017.

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

[14] Lauren Lepow, Hirofumi Morishita, and Rachel Yehuda. Critical period plasticity as a

framework for psychedelic-assisted psychotherapy. *Frontiers in neuroscience*, page 1165, 2021.

[15] Robin L Carhart-Harris, Leor Roseman, Eline Haijen, David Erritzoe, Rosalind Watts, Igor Branchi, and Mendel Kaelen. Psychedelics and the essential importance of context. *Journal of Psychopharmacology*, 32(7):725–731, 2018.

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

[17] Robin L Carhart-Harris, M Bolstridge, CMJ Day, J Rucker, R Watts, DE Erritzoe, Mendel 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.

[18] Robin L Carhart-Harris, David Erritzoe, Tim Williams, James M Stone, Laurence J Reed, Alessandro Colasanti, Robin J Tyacke, Robert Leech, Andrea L Malizia, Kevin 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.

[19] Robin L Carhart-Harris, Suresh Muthukumaraswamy, Leor Roseman, Mendel Kaelen, Wouter Droog, Kevin Murphy, Enzo Tagliazucchi, Eduardo E Schenberg, Timothy Nest, Csaba 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.

[20] Felix Müller, Patrick C Dolder, André Schmidt, Matthias E Liechti, and Stefan Borgwardt. Altered network hub connectivity after acute lsd administration. *NeuroImage: Clinical*, 18:694–701, 2018.

[21] Robin Lester Carhart-Harris, Robert Leech, Peter John Hellyer, Murray Shanahan, Amanda Feilding, Enzo Tagliazucchi, Dante R Chialvo, and David Nutt. The entropic brain: a theory of conscious states informed by neuroimaging research with psychedelic drugs. *Frontiers in human neuroscience*, page 20, 2014.

[22] Matthew M Nour, Lisa Evans, David Nutt, and Robin L Carhart-Harris. Ego-dissolution and psychedelics: validation of the ego-dissolution inventory (edi). *Frontiers in human neuroscience*, 10:269, 2016.

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

[24] Christopher Timmermann, Leor Roseman, Luke Williams, David Erritzoe, Charlotte Martial, Héléna Cassol, Steven Laureys, David Nutt, and Robin Carhart-Harris. Dmt models the near-death experience. *Frontiers in psychology*, page 1424, 2018.

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

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

[27] Zoubin Ghahramani. Bayesian nonparametrics and the probabilistic approach to modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1984):20110553, 2013.

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

[29] Peiye Zhuang, Samira Abnar, Jiatao Gu, Alex Schwing, Joshua M Susskind, and Miguel Ángel Bautista. Diffusion probabilistic fields. In *International Conference on Learning Representations*, 2023.

[30] Marieke Wichers, Marieke J Schreuder, Rutger Goekoop, and Robin 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.

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

[32] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016.

[33] Reuben Feinman and Brenden M Lake. Generating new concepts with hybrid neuro-symbolic models. *arXiv preprint arXiv:2003.08978*, 2020.