The oneirogen hypothesis: modeling the hallucinatory effects of classical psychedelics in terms of replay-dependent plasticity mechanisms

Colin Bredenberg[1,2], Fabrice Normandin[1], Blake Richards[1,3,*], and Guillaume Lajoie[1,2,*]

[1]Mila - Quebec AI Institute
[2]University of Montreal
[3]McGill University
[*]indicates equal contribution

Abstract

Classical psychedelics induce complex visual hallucinations in humans, generating percepts that are coherent at a low level, but which have surreal, dream-like qualities at a high level. While there are many hypotheses as to how classical psychedelics could induce these effects, there are no concrete mechanistic models that capture the variety of observed effects in humans, while remaining consistent with the known pharmacological effects of classical psychedelics on neural circuits. In this work, we propose the "oneirogen hypothesis", which posits that the perceptual effects of classical psychedelics are a result of their pharmacological actions inducing neural activity states that truly are more similar to dream-like states. We simulate classical psychedelics' effects via manipulating neural network models trained on perceptual tasks with the Wake-Sleep algorithm. This established machine learning algorithm leverages two activity phases: a perceptual phase (wake) where sensory inputs are encoded, and a generative phase (dream) where the network internally generates activity consistent with stimulus-evoked responses. We simulate the action of psychedelics by partially shifting the model to the 'Sleep' state, which entails a greater influence of top-down connections, in line with the impact of psychedelics on apical dendrites. The effects resulting from this manipulation capture a number of experimentally observed phenomena including the emergence of hallucinations, increases in stimulus-conditioned variability, and large increases in synaptic plasticity. We further provide a number of testable predictions which could be used to validate or invalidate our oneirogen hypothesis.

1 Introduction

Classical psychedelics—including psilocybin, mescaline, DMT, and LSD—are a family of hallucinogenic compounds with a common mechanism of action: they are agonists for the 5-HT2a serotonin receptor commonly expressed on the apical dendrites of cortical pyramidal neurons [1] and on parvalbumin (PV) interneurons [2]. These drugs induce numerous effects in human subjects, including: complex visual, auditory, and tactile hallucinations; intense spiritual experiences; long-lasting alterations in mood; changes in personality; and increases in synaptic plasticity [3, 4, 5]. Recently, they have been explored clinically as potential treatments for depression and anxiety [6], as well as PTSD [7].

The 5-HT2a receptor plays a critical role in psychedelic-induced hallucinations. Indeed, behavioral measures of hallucinatory drug effects are induced selectively by cellular membrane-permeable 5-HT2a agonists [8], and perceptual effects of classical psychedelics are largely eliminated by blocking 5-HT2a receptors in the cortex [9, 8] (though 5-HT2a agonists with mixed receptor selectivity are in some cases characterized by primarily non-hallucinatory effects [10, 11]). However, very little is understood about *why* highly structured hallucinations and changes in synaptic plasticity emerge from activating cortical 5-HT2a receptors: to explain this, it is necessary to develop mechanistic theories that are capable of linking changes in neuron-level properties (receptor agonism) to changes in perception and behavior. Psychedelic drug users and therapists have long noted the 'dream-like' qualities of psychedelic drug

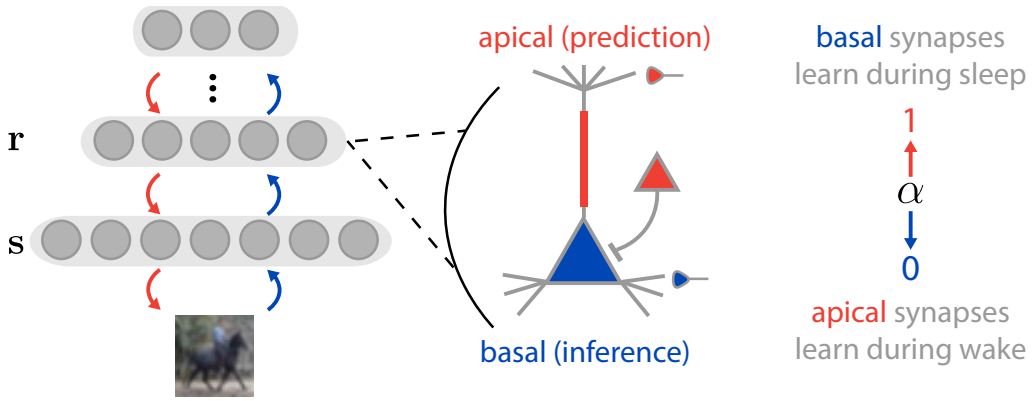

Figure 1: **Mapping the Wake-Sleep algorithm onto cortical architecture**. Left: Network architecture. We model early sensory processing in the cortex with a multilayer network, $\mathbf{r}$, receiving stimuli $\mathbf{s}$. Center: individual pyramidal neurons receive top-down inputs (red) at the apical dendritic compartment, and bottom-up inputs at the basal dendritic compartment (blue). 5-HT2a receptors are expressed on the apical dendritic shaft (red bar), and on PV interneurons (red triangle); both sites may play a role in gating basal input. Right: Over the course of Wake-Sleep training, basal inputs dominate activity during the Wake phase ($\alpha = 0$) and are used to train apical synapses, whereas apical inputs dominate activity during the Sleep phase ($\alpha = 1$) and are used to train basal synapses.

hallucinations, which are realistic but untethered from the external world; this observation leads naturally to speculation that these drugs are 'oneirogens,' or dream-manifesting compounds [12]. However, beyond perceptual phenomenology (and some evidence pointing to the effects of psychedelics on sleep cycles [13, 14, 15]), we lack a mechanistic proposal that could explain the similarity between dreams and psychedelic drug experiences. Here, we articulate the 'oneirogen hypothesis', which describes one such potential mechanistic explanation. We propose that classical psychedelics induce a dream-like state by shifting the balance between bottom-up pathways transmitting sensory information and top-down pathways ordinarily used to create replay sequences in the brain. Replay sequences have been shown to be important for learning during sleep [16, 17, 18, 19, 20]: we propose that mechanisms supporting replay-dependent learning during sleep are key to explaining the increases in plasticity caused by psychedelic drug administration. In total, our model of the functional effect of psychedelics on pyramidal neurons could provide a explanation for the perceptual psychedelic experience in terms of learning mechanisms for consolidation during sleep [21], and cortical 'replay' phenomena [22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32].

To explore the oneirogen hypothesis concretely, we use the aptly named Wake-Sleep algorithm [33], which has historically been used to train artificial neural networks (ANNs) that possess both a bottom-up "recognition" pathway and a top-down "generative" pathway to learn a representation of incoming sensory data. It enables unsupervised learning in ANNs by alternating between periods of "waking perception" (wherein bottom-up recognition pathways drive activity) and "dreaming sequences" (wherein top-down generative pathways drive activity). With these alternate periods of distinct activity, connectivity parameters in each pathway are adjusted to match the activity of the opposite pathway. This way, the top-down pathway learns to generate activity consistent with that induced by sensory inputs, and the bottom-up pathway learns better representations thanks to generated activity.

In this work, we show that within a neural network trained via Wake-Sleep, it is possible to model the action of classical psychedelics (i.e. 5-HT2a receptor agonism) by shifting the balance during the wake state from the bottom-up pathways to the top-down pathways, thereby making the 'wake' network states more 'dream-like'. Specifically, we model the effects of classical psychedelics by manipulating the relative influence of top-down and bottom-up connections in neural networks trained with the Wake-Sleep algorithm on images. Doing so, we capture a number of effects observed in experiments on individuals under the influence of psychedelics, including: the emergence of closed-eye hallucinations, increases in stimulus-conditioned variability, and large increases in synaptic plasticity. This data suggests that the oneirogen hypothesis may indeed help to explain why 5-HT2a agonists have the functional effects that they do. We subsequently identify several testable predictions that could be used to further validate the oneirogen hypothesis.

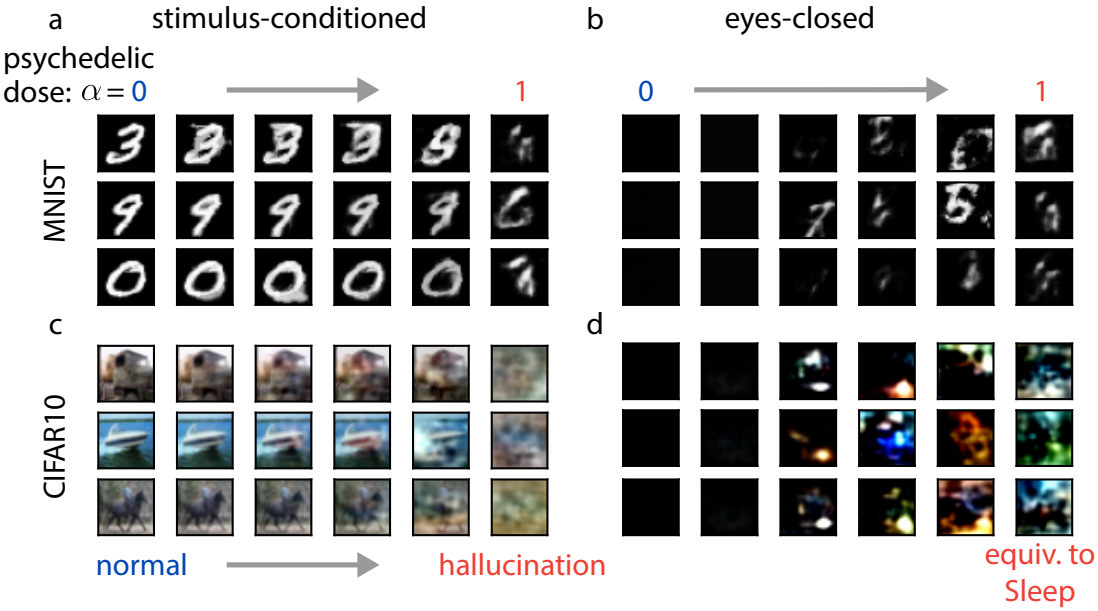

Figure 2: **Visualizing the effects of psychedelics in the model**. We model the effects of classical psychedelics by progressively increasing α from 0 to 1 in our model, where $\alpha = 1$ is equivalent to the Sleep phase. We visualize the effects of psychedelics on the network representation by inspecting the stimulus layer $\mathbf{s}$. **a)** Example stimulus-layer activity (rows) in response to an MNIST digit presentation as psychedelic dose increases (columns, left to right). **b)** Same as (a) but for 'eyes-closed' conditions where an entirely black image is presented. **c-d)** Same as (a-b), but for the CIFAR10 dataset.

2 Results

Mapping the Wake-Sleep algorithm onto cortical architecture. The Wake-Sleep algorithm allows ANNs to optimize a global, unsupervised objective function for sensory representation learning—the Evidence Lower Bound (ELBO)—through local synaptic modifications to a bottom-up recognition pathway and a top-down generative pathway. As a precursor to the variational autoencoder [34, 35], the Wake-Sleep algorithm provides a mechanism for learning a probabilistic latent representation $\mathbf{r}$ responding to incoming sensory stimuli $\mathbf{s}$, which obeys representational characteristics that are ideal for a neural system (e.g. sparsity and metabolic efficiency [36], compression and coding efficiency [37, 38], or disentanglement [39, 40]). To do this, Wake-Sleep optimizes the ELBO through an approximation of the Expectation Maximization (EM) algorithm [41] to train the two pathways (Figure 1a).[1]

Notably, the Wake-Sleep algorithm requires two phases of activity (i.e. "Wake" and "Sleep"), where the network phase is controlled by a global state variable $\alpha \in [0, 1]$ that regulates the balance between the bottom-up and top-down pathways. In the Wake phase ($\alpha = 0$), the network processes real sensory stimuli drawn from the environment, and network activity is sampled based on the bottom-up inputs (corresponding to the approximate inference distribution). In the Sleep phase ($\alpha = 1$), the network internally samples neural activity from its generative model, which then produces generated activity in the stimulus layer $\mathbf{s}$. We use this structure of the Wake-Sleep algorithm as a concrete model to express oneirogen hypothesis. Specifically, we use changes to the value of α as a means of modeling a 5-HT2a agonist-induced shift to a more dream-like state, as we detail below.

Within the Wake-Sleep algorithm, neurons alternate between 'Wake' and 'Sleep' modes, where activity during each mode is dominated by the bottom-up and top-down pathways, respectively. We can determine the neural activity for a given intermediate layer l with the following equation:

$$\mathbf{r}^{(l)} = f\left(h(\mathbf{r}), \mu(\mathbf{r}), \alpha\right) + f(\sigma_b, \sigma_p, \alpha)\boldsymbol{\eta}, \tag{1}$$

where $h(\mathbf{r})$ defines bottom-up input, $\mu(\mathbf{r})$ defines top-down input, $f(h, \mu, \alpha)$ is any interpolation function such that $f(h, \mu, 0) = h$ and $f(h, \mu, 1) = \mu$, σ_b and σ_p define the bottom-up and top-down activity standard deviations, and $\boldsymbol{\eta} \sim \mathcal{N}(0, 1)$ adds random noise to the neural activity (see Methods for more detail). Here, for notational conciseness we treat $\mathbf{r}$ as a concatenated vector of all $\mathbf{r}^{(l)}$ vectors from each

[1] For readers who are unfamiliar with the Wake-Sleep algorithm, a tutorial can be found here [42].

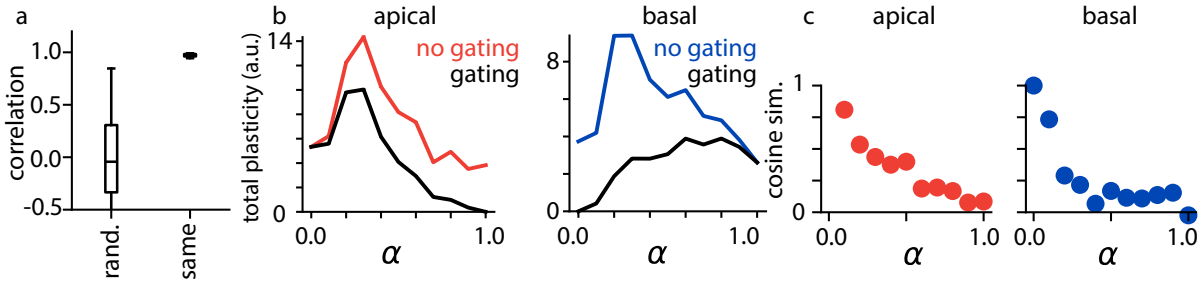

Figure 3: **Effects of psychedelics on single model neurons**. **a)** Correlations between the apical and basal dendritic compartments of either the same network neuron or between randomly selected neurons. **b)** Total plasticity for apical (left) and basal (right) synapses as α increases in the model when plasticity is either gated or not gated by α. Error bars indicate +/- 1 s.e.m. **c)** Cosine similarity between plasticity induced under psychedelic conditions compared to baseline for apical (left) and basal (right) synapses.

layer. This equation means that α controls whether bottom-up inputs or top-down inputs control the dynamics of individual neural units.

Thus, as α moves from a value of 0 to a value of 1, the activity of the neurons shifts from being driven by the bottom-up recognition pathway to being driven by the top-down generative pathway. How could this occur in the brain? Realistically, each neuron in cortex would have its own α variable defining the relative influence of top-down and bottom-up inputs on its spiking activity; here, for simplicity, we will assign the entire network a single α value reflecting the 'mean' relative top-down/bottom-up influence averaged across neurons, as determined by the network state (Wake, Sleep, dose-dependent psychedelic administration). In the cortex, excitatory pyramidal neurons receive inputs from distinct sources: inputs that are from 'higher order' cortical areas target the apical dendrites, whereas inputs that are from 'lower order' cortical or sensory subcortical areas target the basal dendrites [43]. Thus, we can capture the core idea behind the oneirogen hypothesis using the Wake-Sleep algorithm, by postulating that the bottom-up basal synapses are predominantly driving neural activity during the Wake phase (when α is low), while top-down apical synapses are predominantly driving neural activity during the Sleep phase (when α is high; Figure 1) [44]; this is in agreement with several recent theoretical studies that have proposed that apical dendrites could serve as a site for integrating top-down learning signals [45, 46, 47, 48, 49, 50], particularly those which propose that the top-down signal corresponds to a predictive or generative model of neural activity [51, 52]. This proposed change in α does indeed appear to occur during both slow-wave (SW) [53, 54] and rapid eye movement (REM) [55, 56, 57] sleep, where apical dendritic inputs have been observed to exert increased influence on neural activity that is critical for plasticity induction and consolidation of learned behaviors; during REM sleep, this increased influence has been shown to be mediated by potentiation of basal dendrite-targeting PV inhibitory interneurons [57].

Next we ask: can we model the effects of classical psychedelics in terms of changes in α? Notably, 5-HT2a receptors are expressed in the apical dendrites of pyramidal neurons [1] and PV interneurons [2], and have an excitatory effect that positively modulates glutamatergic transmission due to apical dendritic inputs [58, 59]; further, classical psychedelic administration has been shown to have an inhibitory effect on glutamatergic transmission due to basal dendritic inputs [60]. These data suggest that 5-HT2a agonists could have a push-pull effect on cortical pyramidal neurons, increasing the relative influence of apical dendrites and decreasing the relative influence of basal dendrites [61] in much the same way as has been observed during SW and REM sleep. Hence, we can model these effects by increasing the α value in a Wake-Sleep trained network, and then ask whether the networks exhibit other phenomena that match the known impact of classical psychedelics on neural activity. We note that with this mapping of the Wake-Sleep algorithm to models of basal and apical processing, synaptic modifications at both apical and basal synapses correspond to minimizing a local prediction-error between top-down and bottom-up inputs (see Methods).

Modeling Hallucinations. To see whether a transition from waking to a more dream-like state would induce hallucinatory effects in our model, we trained multilayer neural networks with branched dendritic arbors (see Methods) on the MNIST digits dataset [62] using the Wake-Sleep algorithm and subsequently simulated hallucinatory activity by varying α (see Methods; Eq. (8)). We could visualize the effects of our simulated psychedelic with snapshots of the stimulus layer $\mathbf{s}$ at a fixed point in time for various values

of α (Figure 2; see Supplemental Materials for videos). As α increased, we observed that network activity gradually deformed away from the ground-truth stimulus in a highly structured way, adding strokes to the original digit that were not originally present. At the highest values of α tested, we found that network states were wholly divorced from the ground-truth stimulus, but retained many characteristics of the MNIST digits on which the network was trained (e.g. smooth strokes and the rough form of digits). These results emphasize that hallucinations induced by a shift to a more dream-like state in these models are heavily influenced by the training dataset, which for an animal would correspond to the statistics of the sensory environment in which it learns its sensory representation. To emphasize this point, we further trained our networks on the CIFAR10 natural images dataset [63] (Figure 2c), to provide an example of a more naturalistic training dataset. In this case, our model was not powerful enough to reproduce realistic natural images—instead, we found that our modeled hallucinatory activity corresponded to 'ripple' effects, which are similar to the 'breathing' and 'rippling' phenomena reported by psychedelic drug users at low doses [3].

These simulations were produced with a complex, multicompartmental neuron model; however, we found similar results with two alternative network architectures, one with within-layer recurrence (Supplemental Figure S1a) and one which used a simpler single compartment neuron model (Supplemental Figure S1b). We found that our single compartment model produced qualitatively less realistic generated images than the multicompartment and recurrent models, justifying our use of the more complex models (Supplemental Figure S2). To demonstrate the importance of a learned top-down pathway to produce complex, structured hallucinations in the earliest layers of our network, we generated model hallucinations from two control networks: an untrained model and a trained network where psychedelic activity was alternatively modeled by a simple increase in the variance of individual neurons (we will refer to this latter control as the noise-based hallucination protocol). We found that hallucinations under these control conditions resembled additive white noise, rather than structured digit-like shapes (Supplemental Figure S1c-d).

Psychedelic drug users also report observing the emergence of hallucinations while their eyes are closed [3]. Interestingly, we found that our model recapitulated these phenomena: as α increased, networks trained on MNIST gradually began revealing increasingly complex and digit-like patterns (Figure 2b), whereas CIFAR10-trained networks again predominantly produced 'ripple' hallucinations (Figure 2d).

Effects of psychedelics on single neurons. Having recapitulated hallucinatory phenomena in stimulus space, we next explored how our proposed mechanism affected neural activity in our network model, in order to establish markers that could be used to experimentally validate or invalidate the oneirogen hypothesis. To start, we investigated the effects of learning and psychedelic drug administration on the activity of single neurons in the model. As noted previously, the learning algorithm used here trains synapses so that top-down inputs to apical dendritic compartments match bottom-up inputs to basal dendritic compartments. As a consequence, we observed that after training, inputs to apical and basal dendritic compartments were much more correlated on the same neuron than they were for random neurons (Figure 3a), which was not observed in untrained models (Supplemental Figure S3a). This form of strongly correlated tuning has been observed in both cortex and the hippocampus [64, 65].

There are many indicators that psychedelic drug administration in humans and animals can induce marked, long-lasting changes in behavior, as well as large increases in synaptic plasticity [4, 66, 67, 8, 5]. In Wake-Sleep learning, apical synapses learn during the Wake phase, whereas basal synapses learn during the Sleep phase—thus, plasticity at apical synapses is gated by $(1 - \alpha)$, whereas plasticity at basal synapses is gated by α (see Methods). However, learning is still theoretically possible without this explicit gating, though it may be noisier and less efficient; furthermore, it is conceivable that classical psychedelics could increase the relative influence of apical inputs on the activity of a neuron without affecting this gating mechanism. As a consequence, we modeled the dose-dependent effects of psychedelics on plasticity both with and without gating (Figure 3b). Consistent with recent experimental results [4], for intermediate doses we found large increases in plasticity at both apical and basal synapses under *both* conditions, where plasticity was measured as a mean change in normalized synaptic strength across weight parameters in our network (see Methods). In our model, we found that the total evoked plasticity peaked at roughly $\alpha = 0.5$; we further found that if gating was affected by psychedelics, apical plasticity would eventually be quenched at very high drug doses. We also found that plasticity induced by psychedelic drug administration gradually became unaligned from the weight updates that would have occurred in the absence of the drug (Figure 3c), indicating that these results were not simply due to modulation of the effective learning rate of the underlying plasticity. Rather, as has been suggested by other theoretical studies [68], plasticity in the model likely increased because aberrant hallucinatory

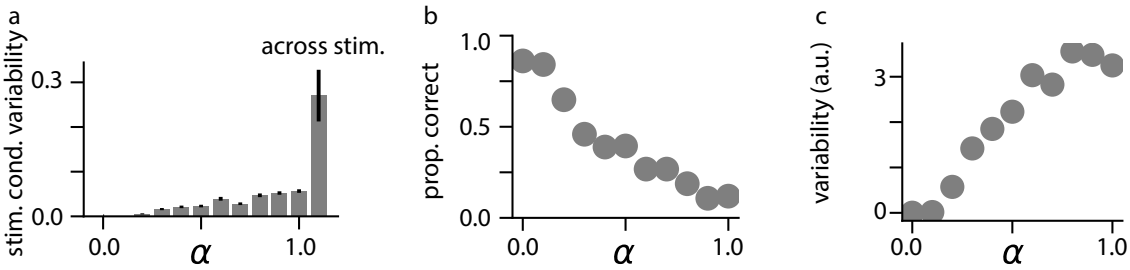

Figure 4: **Effects of psychedelics on neural variability**. **a)** Stimulus-conditioned variability for neurons in the network as α increases, as compared to variability in neural activity across stimuli (rightmost bar). Error bars indicate +/- 1 s.e.m. **b)** Proportion correct for a classifier trained to detect the label of presented MNIST digits as α increases. **c)** Variability in the logit outputs of the trained classifier as α increases.

activity pulled the learning mechanism out of a local optimum in which plasticity was minimal, producing much more plasticity across the network. Importantly, we observed these increases in plasticity in all network architectures and training datasets we explored, including for our noise-based hallucination protocol (Supplemental S4), demonstrating that changes in apical dendritic influence within a Wake-Sleep learning framework are sufficient, but *not necessary* to induce increases in synaptic plasticity: for trained networks, it would seem that even simple increases in neural variability can have similar effects.

Effects of psychedelics on neural variability. Having observed that increasing our modeled drug dosage caused heightened fluctuations and deviations from the ground-truth stimulus in the sensory layer of our network (Figure 2), we next investigated whether variability was affected at the level of individual neurons in higher layers of the model. Indeed, we found that for a fixed stimulus, neural variability increased markedly as the simulated psychedelic drug dose increased (Figure 4a). This result is consistent with the data supporting the Entropic Brain Theory [69, 70, 71, 72], in which neural activity in resting state fMRI recordings becomes increasingly 'entropic' (i.e. variable) under the influence of psychedelics; however, it is important to note that our noise-based hallucination protocol also produced these effects (Supplemental Figure S5a). Though most experimental data supporting the Entropic Brain Theory is taken from recordings with relatively poor spatial resolution, averaging activity over large cortical areas, our model predicts that this increase in variability should be reflected at the level of individual neurons; this increase in variability after psychedelic administration has been recently observed in auditory cortical neurons for active mice [73], but whether this phenomenon is general across tasks and cortical areas remains to be seen. We further found that this increase in variability corresponded to a decrease in ability to identify the stimulus being presented to the network: we trained a classifier to identify which MNIST digit was presented to our networks on Wake neural activity (see Methods), and found that the accuracy of our classifier decreased (Figure 4b) while the output variability of the classifier increased (Figure 4c) in response to drug administration.

Within our model, this increase in variability is quite sensible: in the ordinary Wake state, neural activity is constrained to correspond to the singular sensory stimulus being presented, whereas during Sleep states, neural activity is completely unconstrained by any particular sensory stimulus, reflecting instead the full distribution of *possible* sensory stimuli. As increasing α in our model interpolates between Wake and Sleep states, we can expect intermediate values of α to produce network states which are less constrained by the particular sensory stimulus being presented, reflected in increased neural variability.

Network-level effects of psychedelics. We next investigated the effects of psychedelics on network-level and inter-areal dynamics within our model. We first identified an important negative result: the pairwise correlation structure between neurons was largely preserved across psychedelic doses (Figure 5a-b), as was the effective dimensionality of population activity (Figure 5c). This was sensible, because a network that has been well-trained with the Wake-Sleep algorithm will have the same marginal distribution of network states in the Wake mode as in the Sleep mode—thus, pairwise correlations between neurons should also not differ (as measures of the second order moments of the marginal distribution). We found empirically that even for intermediate values of α in which activity is a mixture of Wake and Sleep modes, these correlations are largely unchanged; in contrast, we observed large changes in correlation structure for untrained networks, and increases in effective dimensionality for both untrained

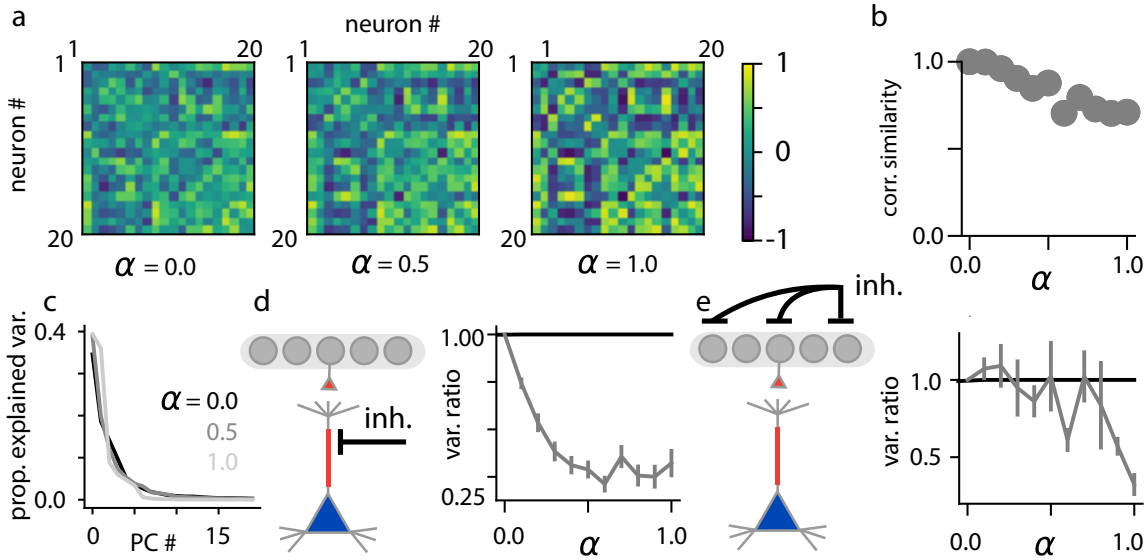

Figure 5: **Network-level effects of psychedelics**. **a)** Pairwise correlation matrices computed for neurons in layer 2 across stimuli for $\alpha = 0$ (left), $\alpha = 0.5$ (center), and $\alpha = 1.0$ (right). **b)** Correlation similarity metric between the pairwise correlation matrices of the network in the absence of hallucination ($\alpha = 0$) as compared to hallucinating network states ($\alpha > 0$). **c)** Proportion of explained variability as a function of principal component (PC) number for $\alpha \in \{0, 0.5, 1\}$. **d)** Ratio of across-stimulus variance in individual stimulus layer neurons when the apical dendrites have been inactivated, versus baseline conditions across different α values. **e)** Ratio of across-stimulus variance in individual neurons in the stimulus layer when neurons at the deepest network layer have been inactivated, versus baseline conditions across different α values. Error bars indicate +/- 1 s.e.m.

networks and for our simple noise-based hallucination protocol, suggesting that these results are more specific to our trained models in which hallucinations are caused by an increase in apical dendritic influence (Supplemental Figure S6a-b). Interestingly, these results are consistent with a recent study that has shown only minimal functional connectivity and effective dimensionality changes in task-engaged humans being presented audiovisual stimuli under the influence of psilocybin [72].

However, though the pairwise correlations between single neurons are largely preserved, the causal influence between lower and higher layers of our model network changes considerably both during hallucination and Sleep modes. Because psychedelic drug administration increases influence of apical dendritic inputs on neural activity in our model, we found that silencing apical dendritic activity reduced across-stimulus neural variability more as the psychedelic drug dose increases (Figure 5d). Further, we found that as α increased, inactivating the deepest network layer induced a large reduction in variability in the stimulus layer relative to baseline (Figure 5e), revealing that within our model, increases in top-down influence are responsible for much of the observed stimulus-conditioned variability at larger drug doses. These inactivations had no impact on neural variability in our noise-based hallucination protocol, but were observed for all network architectures and datasets that we tested in which hallucinations were caused by an increase in apical dendritic influence (Supplemental Figure S6), suggesting that these results are quite specific to our model. Further, these inactivations have not yet been performed in animals, and consequently constitute a critical testable prediction of our model.

Modeling hallucinations in large-scale pretrained networks. While our trained model is capable of capturing several effects of classical psychedelics, it also has a clear limitation: our top-down generative model does not have sufficient expressive power to induce complex hallucinations of naturalistic stimuli, producing instead 'ripples,' or 'breathing' effects that preserve lower-order statistical features of the input data (Figure 5b). While psychedelic drug users do report these phenomena, they also report observing much more complex hallucinations, including of people, animals, and scenes [74, 75].

Generative models trained through backpropagation have been much more successful in producing more complex generated sensory stimuli [35, 34, 76], and furthermore, hierarchical variational autoencoder models have a nearly identical top-down/bottom-up model architecture as our Wake-Sleep-trained

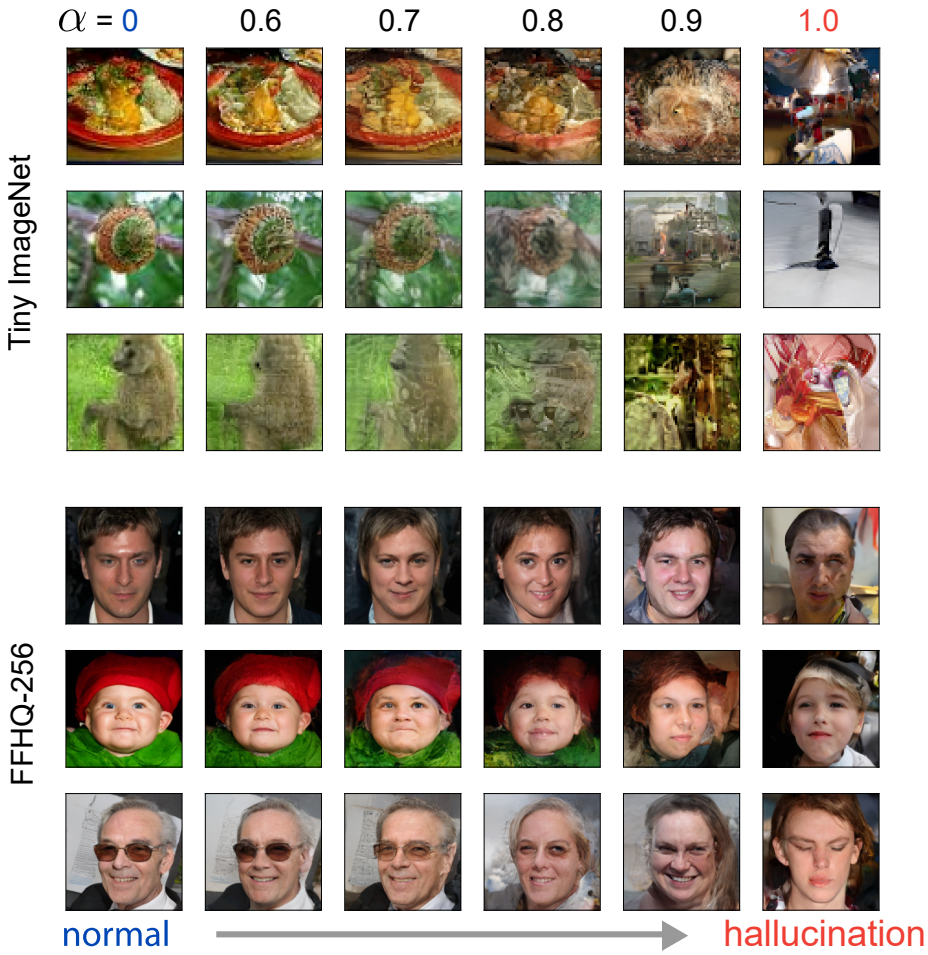

Figure 6: **Visualizing the effects of psychedelics in pretrained VDVAE models**. Decoded outputs of a pretrained VDVAE model trained on Tiny ImageNet (Top) and FFHQ-256 (Bottom) based on hallucinations generated in the top 35 layers of the model. Image samples vary along rows, and hallucination intensity, parameterized by α, increases along columns.

networks [77, 78]. Therefore, to see whether our proposed mechanism would induce complex, structured hallucinations in more powerful models, we induced hallucinations in Very Deep Variational Autoencoder (VDVAE) models [79] that were pretrained through backpropagation on a large natural images dataset, Tiny ImageNet [80], and a large corpus of human faces, FFHQ-256 [81]. These models have a few key differences compared to our Wake-Sleep-trained models: 1) they are trained through backpropagation, which is well-known to be biologically implausible [82]; 2) they exploit parameter sharing across spatial positions in convolutional layers for increased data efficiency during training, at the expense of further biological realism [83]; 3) the 'Wake' stage inference process of these models incorporates inputs from both bottom-up *and* top-down sources, which both improves performance [77] and is more biologically realistic [84, 43]; 4) the models are trained on more complex, higher-resolution datasets. Finally, to induce more 'abstract' hallucinations, we increased the α parameter in these models selectively for higher layers of the network, whereas for the Wake-Sleep-trained models we increased α evenly across layers (see Methods). Combined, these differences make for an effective model of high-level hallucination effects, at the expense of some biological realism.

We found that hallucinations generated by these pretrained models were much richer and more complex: increasing α in the Tiny ImageNet VDVAE caused the emergence of textural patterns and geometric shapes, while the FFHQ-256 VDVAE caused increasingly bizarre changes in facial features (Fig. 6). Both models were also capable of reproducing closed-eyes hallucinations (Fig. S7), where the content of these hallucinations was shaped by their respective training datasets.

To investigate the nature of hallucinations generated by the Tiny ImageNet VDVAE, we examined

the Laplacian pyramid of decoded hallucination images at varying α values (Fig. S8a). Essentially, a Laplacian pyramid decomposes an image into levels of decreasing resolution features, with each level encoding the residual produced by downsampling to the next-lowest resolution (level 0 corresponds to the base 64x64 pixel image, while level 5 corresponds to a 4x4 reduced-resolution set of features). We found that low-level pyramid features varied considerably at low α levels, while high-level pyramid features did not begin to vary until higher α doses (Fig. S8b). This suggests that hallucinations within our model obey a fine-to-coarse structure, where low-dose hallucinations are confined to high-frequency, spatially localized changes, and progressively increasing doses begin to cause variations in more global image features.

Lastly, we were able to replicate our previous network-level results on the Tiny ImageNet VDVAE. We found that increasing psychedelic dose α caused an increase in stimulus-conditioned variance within the model (Fig. S8c), and that across-stimulus correlation structure between network units was largely preserved across doses (Fig. S8d). Further, we found that the ratio of before- and after-inactivation across-stimulus variance decreased as the psychedelic dose α increased (though somewhat paradoxically, inactivation caused an *increase* in variance for $\alpha = 0$, likely due to the influence of top-down inputs during inference for this model). Combined, these results show that key testable predictions from our Wake-Sleep-trained model are preserved in the VDVAE, while this latter model is capable of producing some of the more complex hallucinations characteristic of psychedelic experience.

3 Discussion

Experimental results captured by our model. In this study, we have examined a hypothetical mechanism explaining how the 5-HT2a receptor agonism of classical psychedelics could induce the highly structured hallucinations reported by people who have consumed these drugs. Specifically, we have explored the 'oneirogen hypothesis', which postulates that 5-HT2a agonists have the effects that they do because they shift the neocortex to a more dream-like state, wherein activity is more strongly driven by top-down inputs to apical dendrites than normally occurs during waking. To provide a concrete model to explore the 'oneirogen hypothesis' we used the classic Wake-Sleep algorithm, which learns by toggling between a Wake phase, where activity is driven by bottom-up sensory inputs, and a Sleep phase, where activity is driven by top-down generative signals. We modeled the 'oneirogen hypothesis' by simulating psychedelic administration as an increase in a neuronal state variable (α) that switches neural activity between these two phases, such that the simulated psychedelic caused the network to enter a state somewhere between the Wake and Sleep phases, making activity during the Wake phase less tied to actual sensory inputs by increasing the relative influence of the top-down, apical compartment in the models (depending on the "dosage"). This formulation is consistent with anatomical wiring data [43], as well as several recent theoretical studies which propose a specialized learning role for top-down projections to the apical dendrites of pyramidal neurons [45, 46, 47, 48, 49, 50]. It is also consistent with the known cellular mechanism of action of classical psychedelics [1, 58, 59, 9] and experiments that demonstrate a reduced responsivity to bottom-up stimuli in cortex after psychedelic drug administration [85, 86, 87]. Using this model, we were able to produce both stimulus-conditioned and "closed-eye" hallucinations that are consistent with the low-level effects reported by psychedelic drug users [3], and we were also able to recapitulate the large increases in plasticity observed at both apical and basal synapses at moderate psychedelic doses [4].

Our model uses a particular functional form of synaptic plasticity at both apical and basal synapses, reminiscent of the classical delta rule [88], which seeks to minimize a prediction error between inputs in apical and basal synapses. There are many theoretical models of learning that propose similar forms of plasticity [46, 47, 51], so while this plasticity is a necessary prediction of our model, it is not *sufficient* to validate it. Experimentally, plasticity dynamics which could, theoretically, minimize such a prediction error have been observed in cortex [89, 90]; we found that plasticity rules of this kind induce strong correlations between inputs to the apical and basal dendritic compartments of pyramidal neurons, which has been observed in both the hippocampus and cortex [64, 65]. Psychedelic administration within our model induced large increases in plasticity, which has also been observed experimentally [4, 5]. Within our model, this plasticity should not be interpreted as 'learning,' since it arises from aberrant network activity and does not necessarily produce behavioral or perceptual improvements; it is likely closer to 'noise,' that may still be useful for helping neural networks escape from local minima in the loss optimization landscape for synaptic weights, with possible implications for individuals suffering from post-traumatic stress disorder, early life trauma, or the negative effects of sensory deprivation. Further work will be required to analyze the relationship within our model between psychedelic dosage, usage frequency, and

Testable Predictions	OH	EC	REBUS	DD
1. Psychedelic administration increases stimulus-conditioned variability of neurons.	✓	✓	✓	✓
2. Psychedelic administration preserves pairwise across-stimulus correlations between neurons.	✓	✗	✗	✗
3. Silencing apical dendritic compartments decreases neural variability more after psychedelic administration.	✓	n/a	n/a	n/a
4. Silencing higher-order cortical areas affects lower-order cortical activity more after psychedelic administration.	✓	✗	✗	✓
5. Psychedelic drug effects are mediated by the same circuitry responsible for inducing generative replay dynamics in cortex.	✓	✗	✗	✗

Table 1: **Summarizing testable predictions of the 'oneirogen hypothesis'.** Models: OH - oneirogen hypothesis; EC - Ermentrout & Cowan [94]; REBUS - Relaxed Beliefs Under Psychedelics [69]; DD - DeepDream [95]. Key: ✓- model is consistent with the prediction; ✗- model is inconsistent with the prediction; n/a - model is neither inconsistent nor consistent with the prediction.

the long-term stability of learned representations in neural networks.

Interestingly, we also found that increasing the influence of apical dendrites in the model increased stimulus-conditioned variability in our individual neurons. In cortex, this effect has recently been shown at the level of single auditory neurons [73]; further, there have been numerous studies reporting similar increases in asynchronous variability [71] (or, analogously, sample entropy [70] and Lempel-Ziv complexity [91]) in resting-state human brain recordings, previously modeled using Entropic Brain Theory. This theory proposes that many of the effects of classical psychedelics on perception and learning can be explained in terms of increases in variability induced by drug administration (e.g. the increase in variability could introduce novel patterns of thinking, or perturb learning to allow it to break out of 'local minima'). Our results are broadly consistent with this perspective, to which we have added explanatory layers that are both normative and mechanistic [92, 93]: namely, we speculate that this variability under ordinary conditions results from an ethologically important mechanism underlying generative replay for unsupervised learning during sleep or quiescence, and we propose that mechanistically this increase in variability is caused by the increased influence of top-down synapses that are not tied to incoming sensory stimuli. Alternatively, such entropy increases could be caused by increases in attention or self-reflective thought, as supported by recent studies showing that task engagement significantly attenuates psychedelic-induced entropy increases [72]; though our model does not include cognitive or attention components, such an interpretation is potentially consistent with and complementary to our framework.

Testable predictions. While our results are broadly consistent with existing experimental evidence, there are many unconfirmed aspects of our model which could be tested to validate or invalidate it (summarized in Table 1). As mentioned in the previous section, our model predicts that *single neurons* should increase variability in response to psychedelic drug administration in any cortical area affected by psychedelic drugs, an effect that has not yet been investigated systematically throughout cortex or across task conditions. Second, we propose that psychedelic drugs should not push network dynamics into wildly different operating regimes than normal wakefulness, beyond any differences observed between wakefulness and replay (dreams) during sleep. In particular, we found that our simulated psychedelic drug administration did not perturb pairwise correlations between neurons within local circuits when averaged across an ecologically representative set of stimuli.

Within our model, psychedelic drug administration is expected to increase the relative influence of top-down projections. This prediction appears to be supported by slice experiments [58, 59, 60], but to our knowledge this change in functional connectivity has not yet been shown via *in vivo* manipulations. This could be explored experimentally in several ways: first, we have shown that apical dendrite-targeted silencing experiments can identify the amount of influence apical dendritic inputs exert on neuronal dynamics; second, we have shown that increases in top-down influence can in principle be identified with interareal silencing experiments. We caution that interpreting results in this second vein may be difficult, as establishing a clean distinction between a 'higher order' and 'lower order' cortical area may

be much more difficult in a densely recurrent system such as the brain, compared to our simplified and fully observable network model.

Interestingly, if psychedelic drugs are genuinely co-opting circuitry ordinarily reserved for generative replay during periods of offline quiescence or sleep, we would expect that the same changes in functional connectivity observed during psychedelic drug administration would also occur during periods of replay. Replay been observed and dreams have been documented during both SW [23, 25] and REM [31, 32] sleep, with REM dreams exhibiting greater degrees of bizarreness, possibly indicating a more 'generative' form of replay [96]. During SW sleep, increased top-down influence has been observed from secondary motor cortex to primary somatosensory cortex [54], and from hippocampus to prefrontal cortex [25]; however, it should be noted that increased hippocampal-to-prefrontal functional coupling was not observed after classical psychedelic administration [97]. During REM sleep, increased top-down influence (or apical dendritic influence) has been observed in prefrontal, visual [56], and motor [55] cortices, with some top-down inputs originating from higher-order thalamic nuclei [57, 98]; similarly, multiple non-invasive imaging studies have observed increases in top-down functional coupling from higher-order thalamic nuclei after psychedelic administration [99, 100]. Therefore, increases in top-down coupling appear broadly consistent between REM sleep and classical psychedelic administration, while psychedelic states appear inconsistent with the hippocampal-cortical coupling during SW sleep; this latter result could potentially be explained in terms of a recent complimentary learning systems model [101], in which SW sleep is responsible for orchestrating hippocampus-cortex-coupled *episodic* replay while REM sleep is responsible for orchestrating hippocampus-cortex-decoupled *generative* replay, but more experiments and theoretical work will likely be necessary to fully characterize this additional complexity. Given these data, it seems as though REM sleep replay is a moderately stronger candidate for sharing a mechanism of action with classical psychedelics, though it remains possible that replay events during SW sleep occur via a similarly shared mechanism.

To summarize, though we have provided a candidate explanation for several of the hallucinatory effects of psychedelic drugs with a model that displays a strong correspondence with existing empirical evidence, our model rests on a number of testable assumptions. Our goal here has been to articulate these assumptions as clearly as possible, to facilitate experimental efforts to test them.

Comparisons to alternative models. Here, we review prominent existing hypotheses as to how psychedelic drugs could induce hallucinations in neural networks and compare to our model (summarized in Table 1). The first alternative proposed that incredibly complex, geometric patterns formed by DMT administration could be attributed to pattern-formation effects in visual cortex caused by a disruption of the balance between excitation and inhibition in locally-coupled topographic recurrent neural networks [94, 102]. Our work differs from this approach in several respects. First, rather than disrupting E-I balance, we propose that psychedelics increase the relative influence of apical dendrites and top-down projections on the dynamics of neural activity. Second, though their model is able to generate geometric patterns, it is not able to generate patterns that are statistically related to the features of the sensory environment (e.g. MNIST digits). Lastly, for simplicity we avoided including topographic (or convolutional) recurrent connectivity in our model; however, it would be a very fruitful direction for future research to extend our work to generative modeling of temporal video sequences, as in [103, 104]. With such a development, it is conceivable that our model could directly generalize these pattern formation-based approaches.

Perhaps more closely related to our model is the 'relaxed beliefs under psychedelics' (REBUS) model, which proposes to explain the effects of classical psychedelics in terms of predictive coding theory [69]. Similar to the Wake-Sleep algorithm, predictive coding theory [105] models sensory representation learning with neural dynamics and local synaptic modifications that collectively optimize an ELBO objective function. However, at a mechanistic level, there are numerous differences, the most easily distinguishable feature being that the Wake-Sleep algorithm *requires* periods of offline 'generative replay' to train bottom-up synapses in its network, whereas predictive coding learning occurs concomitantly with stimulus presentation. Furthermore, the REBUS model of psychedelic effects is described at a computational level, in terms of a decrease in the 'precision-weighting of top-down priors.' While it is more difficult to map the REBUS model directly onto cortical microcircuitry, and the hallucinatory effects of such a model have, to our knowledge, not been directly analyzed, it has been shown that the proposed mechanism causes an *increase* in bottom-up information flow between cortical areas [106], in direct contrast to the effects that we have shown in our model (Figure 5c-d); there is some evidence supporting this idea [107], but noninvasive imaging studies are inconsistent on this question, with many studies showing by contrast an increase in top-down functional connectivity caused by classical psychedelic administration [99, 100],

and with invasive recordings showing a decrease in the influence of bottom-up inputs [85, 86, 87]. Because interareal causal influence can be difficult to analyze statistically due to dense recurrent connectivity (i.e. correlation does not imply causation), we stress that it would be more effective to distinguish between the REBUS model and our 'oneirogen hypothesis' by performing direct interventions on inputs to the apical and basal dendritic compartments of pyramidal neurons in cortex, and by exploring whether psychedelic drugs affect the same circuitry that induces 'generative replay' during periods of sleep and quiescence. More consistent with our model, a recent non-mechanistic approach based on the DeepDream algorithm has been used to generate realistic hallucinations via increased influence from a top-down learning signal [95]; however, this model proposes no relationship between psychedelics and replay during sleep.

Lastly, it should be noted that the Wake-Sleep algorithm and our choice of network architecture constitute one particular model within a family of related models, all of which satisfy our key criteria for a good model of the 'oneirogen hypothesis,' namely that 1) the model has well-defined top-down and bottom-up pathways, 2) it learns a generative model of incoming sensory inputs, and 3) it uses periods of offline replay for learning through local synaptic plasticity. For example, in the Supplemental Materials, we have replicated all of our essential results for two alternative network architectures, also learned via the Wake-Sleep algorithm: one model uses within-layer recurrence to improve generative performance, while the other model uses a simpler single compartment neuron model. Furthermore, the closely related Contrastive Divergence learning algorithm for Boltzmann Machines [108] also involves alternations between Wake and generative Sleep phases, learns through local synaptic plasticity, and has been used to model hallucination disorders like Charles Bonnet Syndrome [109], though Boltzmann machines are computationally more cumbersome to train and require more non-biological network features than the Wake-Sleep algorithm. We feel as though it is important to recognize that models that satisfy these three criteria are more similar than they are different, and that it may be quite difficult to experimentally distinguish between them.

Limitations. While our model is capable of capturing several effects of classical psychedelics, it also has several clear limitations. Firstly, while we have been able to model complex hallucination phenomena with backpropagation-trained networks, hallucinations generated by Wake-Sleep-trained networks were generally simpler, likely because the Wake-Sleep algorithm is well-known to be a less effective representation learning and generative modeling algorithm than backpropagation [35], despite its superior biological realism. This suggests that while it is quite possible for generative modeling approaches to produce complex hallucinations through non-biological means, algorithmic or architectural improvements may be necessary in order to make the performance of the more plausible Wake-Sleep algorithm closer to that achieved by state-of-the-art models.

Our model also oversimplifies several aspects of biology. In particular, we do not use neurons that respect Dale's law [110, 111], and the majority of our efforts to map the Wake-Sleep algorithm onto biology focus on excitatory pyramidal neurons. Furthermore, though we do observe that neural dynamics can tolerate a significant amount of top-down input before disrupting perception, experiments and theoretical studies have shown that inputs to apical dendrites of pyramidal neurons do play an important role in waking perception [43, 98, 112], and are not *just* learning signals. We focused on clear distinctions between basally-driven Wake modes and apically-driven Sleep modes during training for computational efficiency reasons, and also due to the fact that parameter sharing across inference and generative networks in the Wake-Sleep algorithm is theoretically under-explored (though it is supported in closely related predictive coding approaches [105] and Boltzmann machines [108]). Future elaborations on our model could incorporate feedback control [113], attention [114], or multimodal sensory inputs [115] into top-down projections; such inputs could help explore how psychedelic hallucinations interact with attentional or feedback control systems in the brain, and have been shown to interact constructively with top-down learning signals in prior models [116, 117, 118]. Our use of VDVAEs is a positive step in this direction, but ideally such network architectures would be made compatible with the Wake-Sleep algorithm.

Lastly, our modeling focus has been exclusively on cortical plasticity and hallucination effects: it should be noted that our model has little bearing on other important features of the psychedelic experience of potential therapeutic relevance, because we have not included the effects of psychedelics on subcortical structures including the serotonergic system [119], which plays an important role in regulating mood and may be where psychedelics exert some of their antidepressant effects. Many studies of the effects of psychedelics on fear extinction focus on the hippocampus or the amygdala [120, 121, 122, 123]. These areas receive extensive innervation directly from serotonergic synapses originating from the dorsal raphe nucleus, which have been shown to play an important role in emotional learning [124]; because clas-

sical psychedelics may play a more direct role in modulating this serotonergic innervation, it is possible that fear conditioning results (in addition to the anxiolytic effects of psychedelics) cannot be attributed to a shift in balance between apical and basal synapses induced by psychedelic administration.

Conclusions. Here we have proposed a hypothesis for the mechanism of action of psychedelic drugs in terms of its excitatory effects on the apical dendrites of pyramidal neurons, which we propose pushes network dynamics into a state normally reserved for offline replay and learning; we have also proposed a number of testable predictions which could be used to validate or invalidate our hypothesis. If validated, our model would describe a mechanism by which psychedelic drug administration causes ordinary sensory perception to become literally more dream-like; it further suggests that the plasticity increases observed during both sleep and psychedelic experience could occur via a common mechanism dedicated to sensory representation learning in the brain. Beyond classical psychedelics, further studying the balance between apical and basal dendritic inputs to pyramidal neurons in connection to replay during sleep may be relevant for explaining the hallucinatory effects of other drugs (such as ketamine) or mental disorders like schizophrenia [125].

4 Methods

Model architecture and training. To model the effects of psychedelics on neural network dynamics and plasticity, we first constructed a simple model of the early visual system by training neural networks on two different image datasets (MNIST [62] and CIFAR10 [63]). Networks were trained with the Wake-Sleep algorithm [33], which requires, for each layer, two modes of stochastic network activity: a 'generative mode', and an 'inference mode'. For the 'inference' mode we must specify a probability distribution $b(\mathbf{r}^{(l)}|\mathbf{r}^{(l-1)})$, while for the 'generative' mode we must specify a separate distribution $p(\mathbf{r}^{(l)}|\mathbf{r}^{(l+1)})$[2]. Notice here that activity in 'inference' mode is conditioned on 'bottom-up' network states ($\mathbf{r}^{(l-1)}$), while activity in generative mode is conditioned on 'top-down' network states ($\mathbf{r}^{(l+1)}$) (Figure 1a).

The 'inference mode' specifies a probability distribution over neural activity, conditioned on the next-lower layer (where the lowest layer is the stimulus layer, i.e. $\mathbf{r}^{(0)} = \mathbf{s}$)—mechanistically it corresponds to activity generated by feedforward projections. To increase the expressive power of our neural units, we use multicompartmental neuron models similar to [126] with N_d dendritic compartments, whose voltages are summed nonlinearly to form the full input to the basal dendrites. For $l > 0$, layer activity is sampled from the distribution $\mathbf{r}^{(l)} \sim \mathcal{N}(h(\mathbf{r}^{(l-1)}), \sigma_b^2\mathbf{I})$, where for neuron i in layer l, $h_i(\mathbf{r}^{(l-1)})$ is given by:

$$h_i(\mathbf{r}^{(l-1)}) = \phi\left(\sum_{n=0}^{N_d} w_{in}^{(l)}\phi_d\left(\mathbf{W}_{in}^{(l)}\mathbf{r}^{(l-1)} + c_{in}^{(l)}\right) + b_i^{(l)}\right), \tag{2}$$

where $\mathbf{W}_{in}^{(l)}$ is a $1 \times N^{(l-1)}$ matrix of synaptic weights onto dendrite n, c_{in} is the corresponding bias for the nth dendritic compartment, $w_{in}^{(l)}$ is the strictly positive weight given to the nth dendritic branch (roughly corresponding to a conductance), and $b_i^{(l)}$ is the bias for the entire basal compartment. $\phi_d(\cdot)$ and $\phi(\cdot)$ are nonlinearities for the dendritic branches and the total basal compartment, respectively: both are the sequential composition of the tanh nonlinearity, followed by batch normalization [127]. For the dendritic branch nonlinearities, we allow for learnable affine parameters (scale and bias), but for the entire basal dendritic compartment we constrain activity to be zero-mean and unit variance across batches in order to prevent indeterminacy between apical and basal scale parameters. For the final inference layer $\mathbf{r}^{(L)}$, as in the variational autoencoder [34], we parameterize both the mean *and* a diagonal covariance matrix of the inference distribution: $\mathbf{r}^{(L)} \sim \mathcal{N}\left(h(\mathbf{r}^{(L-1)}), \mathrm{diag}(h_2(\mathbf{r}^{(L-1)}))\right)$, where $h_2(\cdot)$ is also a multicompartmental model, in this case replacing the final batch normalization with an exponential nonlinearity to ensure positivity.

The 'generative' mode specifies a probability distribution over neural activity, conditioned on the next-higher layer—it corresponds mechanistically to activity generated by feedback projections. The highest layer, $\mathbf{r}^{(L)}$ is sampled from an $N^{(L)}$-dimensional independent standard normal distribution, $\mathbf{r}^{(L)} \sim \mathcal{N}(0, \mathbf{I})$, and all subsequent layers are sampled from the distribution $\mathbf{r}^{(l)} \sim \mathcal{N}(\mu(\mathbf{r}^{(l+1)}), \sigma_p^2\mathbf{I})$,

[2] As a notational convention, we will use letters when referring to mathematical objects from the generative, top-down distribution, and their vertical reflection when referring to the inference, bottom-up distribution (e.g. p and b)

where for the ith neuron, $\mu_i(\mathbf{r}^{(l+1)})$ is given by:

$$\mu_i(\mathbf{r}^{(l+1)}) = \phi\left(\sum_{n=0}^{N_d} m_{in}^{(l)} \phi_d\left(\mathbf{M}_{in}^{(l)}\mathbf{r}^{(l+1)} + d_{in}^{(l)}\right) + a_i^{(l)}\right), \tag{3}$$

where $\mathbf{M}_{in}^{(l)}$ is a $1 \times N^{(l+1)}$ matrix of synaptic weights onto apical dendritic branch n, $d_{in}^{(l)}$ is the corresponding bias for the nth dendritic compartment, $m_{in}^{(l)}$ is the strictly positive weight given to the nth dendritic branch, and $a_i^{(l)}$ is the bias for the entire apical compartment. Again, $\phi_d(\cdot)$ and $\phi(\cdot)$ are nonlinearities, identical to the inference (basal) pathway.

While the neuron model used here is more complicated than is normally used for single-unit neuron models, functions of this kind could feasibly be implemented by nonlinear dendritic computations [126]; we further found that using this nonlinearity qualitatively improved generative performance (Supplemental Figure S2). Given these parameterized probability distributions, we then determined the neural activity for each layer l according to Eq. 1. Our network trained on MNIST was composed of 3 layers, with widths [32, 16, 6], listed in ascending order. A full list of network hyperparameters for both our MNIST and CIFAR10-trained networks can be found in the Supplemental Methods.

All synaptic weights and parameters in our networks were trained via the Wake-Sleep algorithm [33], which is known to produce 'local' parameter updates for a wide range of neuron models (and rate or spike-based output distributions), though the specific functional form of the update may vary depending on the neuron model chosen [128]. These updates, for reasonable choices of neural network architecture, can be interpreted as predictions for how synaptic plasticity should look in the brain, if learning were really occurring via the Wake-Sleep algorithm or some approximation thereof.

Consider a generic inference (basal dendrite) parameter for neuron i, $\theta_b \in \{w_{in}^{(l)}, \mathbf{W}_{in}^{(l)}, b_i^{(l)}, c_{in}^{(l)} : n = 0, ..., N_d\}$. The Wake-Sleep algorithm gives the following update, for a single stimulus presentation:

$$\Delta\theta_b = (\alpha)\eta \frac{\left(\mathbf{r}_i^{(l)} - h_i(\mathbf{r}^{(l-1)}, \theta_b)\right)}{\sigma_b^2} \frac{\partial h_i(\mathbf{r}^{(l-1)}, \theta_b)}{\partial \theta_b}, \tag{4}$$

where η is a learning rate, and the gate α ensures that learning only occurs during sleep mode. Further, for reasons of computational efficiency, we average weight updates across a batch of 512 stimulus presentations; similar results could in principal be obtained with purely online updates [129], but we opted to present stimuli in batches here in order to parallelize computations. $\frac{\partial h_i(\mathbf{r}^{(l-1)}, \theta_b)}{\partial \theta_b}$ changes depending on the parameter θ, reflecting that particular parameter's contribution to basal dendritic activity. For a dendritic branch weight $w_{in}^{(l)}$, we have:

$$\frac{\partial h_i(\mathbf{r}^{(l-1)}, w_{in}^{(l)})}{\partial w_{in}^{(l)}} = \phi'(\mathbf{v}_i^{total})\phi_d(\mathbf{v}_{in}), \tag{5}$$

where $\mathbf{v}_i^{total}$ is the total input to the basal dendritic compartment, and $\mathbf{v}_{in} = \mathbf{W}_{in}^{(l)}\mathbf{r}^{(l-1)} + c_{in}^{(l)}$ is the total input to the nth dendritic branch. This update has the functional form of a classical 'delta' learning rule [88], where a compartmental prediction error between local dendritic activity and neuronal firing rate is multiplicatively combined with branch-specific input to provide changes in the conductance for the nth branch. Similarly, for the jth synapse on the nth dendritic branch, $\mathbf{W}_{inj}^{(l)}$, we have:

$$\frac{\partial h_i\left(\mathbf{r}^{(l-1)}, \mathbf{W}_{inj}^{(l)}\right)}{\partial \mathbf{W}_{inj}^{(l)}} = \phi'(\mathbf{v}_i^{total}) w_{in}^{(l)} \phi'(\mathbf{v}_{in}) \mathbf{r}_j^{(l-1)}. \tag{6}$$

Unlike for simple one-compartment neuron models, computation of parameter updates for dendritic synapses $\mathbf{W}_{inj}^{(l)}$ requires weighting the 'delta' error by the conductance of the corresponding dendritic branch (w_{in}), which could be approximated by the passive diffusion of signaling molecules from the principal basal dendritic compartment back along dendritic branches to individual synapses.

For generative parameters ($\theta_p \in \{m_{in}^{(l)}, \mathbf{M}_{in}^{(l)}, a_i^{(l)}, d_{in}^{(l)} : n = 0, ..., N_d\}$), we have a nearly identical update for a single stimulus presentation:

$$\Delta\theta_p = (1-\alpha)\eta \frac{\left(\mathbf{r}_i^{(l)} - \mu_i(\mathbf{r}^{(l+1)}, \theta_p)\right)}{\sigma_p^2} \frac{\partial \mu_i(\mathbf{r}^{(l+1)}, \theta_p)}{\partial \theta_p}, \tag{7}$$

where now input in the apical dendritic compartment, $\mu_i(\mathbf{r}^{(l+1)})$, is being compared to the activity of the neuron as a whole to determine the magnitude and sign of plasticity. The $(1-\alpha)$ gate in this case ensures that plasticity only occurs during the Wake mode. We provide pseudocode (Algorithm 1) for our Wake-Sleep implementation, as well as a full list of algorithm and optimizer hyperparameters (Tables S1 and S2) in the Supplemental Materials[3].

Modeling Hallucinations. During training, neural network activity is either dominated entirely by bottom-up inputs (Wake, $\alpha = 0$) or by top-down inputs (Sleep, $\alpha = 1$). As a consequence, sampling neural activity is computationally low-cost, and can be performed in a single time step. During Wake, one can take a sampled stimulus variable $\mathbf{s}$, determine the activity at layer 1, then 2, and so on until layer L, while during Sleep, one can sample a latent network state in layer L and traverse the layers in reverse order, down to the stimulus layer. However, this is not possible if $\alpha \notin \{0, 1\}$, because activity in each layer l should depend simultaneously on layer $l+1$ and layer $l-1$. For this reason, we chose to model hallucinatory neural activity *dynamically*, as follows:

$$\mathbf{r}_t^{(l)} = (1 - \frac{1}{\tau})\mathbf{r}_{t-1}^{(l)} + \frac{1}{\tau} f\left(h(\mathbf{r}_{t-1}), \mu(\mathbf{r}_{t-1}), \alpha\right) + \frac{f(\sigma_b, \sigma_p, \alpha)}{\sqrt{\tau}} \boldsymbol{\eta}_{t-1}, \tag{8}$$

where τ is a time constant that determines how much of the previous network state is retained, and $\boldsymbol{\eta}_{t-1} \sim \mathcal{N}(0, \mathbf{I})$. Critically, if we take $\tau = 1$ these dynamics reduce to the sampling procedure used during training (Eq. 1). *A priori*, the choice of interpolation function $f(a, b, \alpha)$ is arbitrary. We selected the following function:

$$f(a, b, \alpha) = \kappa \log \left[(1-\alpha) \exp \frac{a}{\kappa} + \alpha \exp \frac{b}{\kappa} \right], \tag{9}$$

where $\kappa = 0.35$ is a free parameter. This function is equivalent to linear interpolation as $\kappa \to \infty$, and is equivalent to the maximum function between arguments a and b as $\kappa \to 0$ if $\alpha = 0.5$. By selecting $\kappa = 0.35$, we are biasing the system towards registering positive inputs from apical *or* basal sources (in the inclusive sense). We found that this produced 'hallucinatory' percepts in stimulus space that did not reduce the intensity of input stimuli as α increased; rather, inputs maintained their intensity and hallucinations were added on top if they were of greater intensity than the ground-truth image. All simulations were run for 800 timesteps, with $\tau = 0.1$. As a control, we compared our results to network dynamics produced purely by increases in noise, without increases in apical dendritic influence (which we refer to as our noise-based hallucination protocol). For these control simulations, we produced network activity time series with the following equation:

$$\mathbf{r}_t^{(l)} = (1 - \frac{1}{\tau})\mathbf{r}_{t-1}^{(l)} + \frac{1}{\tau} h(\mathbf{r}_{t-1}) + \frac{\sigma_b + \alpha}{\sqrt{\tau}} \boldsymbol{\eta}_{t-1}, \tag{10}$$

so that the standard deviation of the injected noise increased linearly with α.

Apical and Basal Alignment. To measure the alignment between inputs in the apical and basal dendritic compartments of our model neurons, we computed the 'Wake' neural responses to the full test dataset and measured the activity in both the basal and apical compartments of our neurons ($h(\mathbf{r}^{(l-1)})$ and $\mu(\mathbf{r}^{(l+1)})$, respectively). We then calculated the correlation coefficient between apical and basal compartments for the same neuron, compared to the correlation between compartments for two randomly selected neurons.

Quantifying plasticity. To quantify the total amount of plasticity induced in our model system by the administration of psychedelic drugs, we measured the change in *relative* parameter strength (averaging across all synapses in the network and an ensemble of 512 test images). For each test image, we simulated network dynamics according to Eq. 8. Subsequently, for each parameter θ, we calculated the net amount of plasticity induced by viewing all test images, $\Delta\theta$. We subsequently reported the relative change:

$$\Delta\theta_{rel} = \frac{|\Delta\theta|}{|\theta| + \epsilon}, \tag{11}$$

under conditions in which α values gate plasticity (as in ordinary Wake-Sleep) and under conditions in which psychedelic drug administration does not also affect plasticity gating. Here, we took $\epsilon = 10^{-2}$ to avoid numerical instabilities.

[3]Code for reproducing all results from Wake-Sleep-trained models this study is available here: https://github.com/colinbredenberg/oneirogen-hypothesis.

Classifier training. As we trained our neural network using the Wake-Sleep algorithm, we simultaneously trained a *separate* classifier network based on Wake-phase neural activity in the second network layer on a cross-entropy loss, to identify the stimulus class of the input to the system. For our classifier, we used a multilayer perceptron neural network with a single 256-unit hidden layer and $\tanh(\cdot)$ nonlinearities.

We then quantified the accuracy of the classifier on the test set, based on neural activity drawn from the final time step T of hallucination simulations with various values of α. We further measured the average variance of the 10-dimensional output logits of the neural network.

Quantifying correlation matrix similarity before and after psychedelics. To quantify how similar the pairwise correlations between neurons in our model networks were before and after the administration of psychedelics, we recorded hallucinatory network dynamics for an ensemble of 512 test images, and measured pairwise correlations between neurons in the first network layer. To compare these matrices, we then report the correlation coefficient between the flattened $N \times N$ matrices. For this metric, a value of 1 indicates that the correlation matrices are perfectly aligned, while a value of -1 indicates that pairwise correlations are fully inverted.

Quantifying interareal causality through inactivations. To quantify changes in interareal functional connectivity induced by psychedelics, we performed two different types of inactivation. In the first, we inactivated the apical dendritic compartments of all neurons in the stimulus layer, and measured how this inactivation affected across-stimulus variability of neurons relative to the fully active state. In the second method, we inactivated all neurons in the deepest layer, and measured the same effect in across-stimulus variability in the stimulus layer. For both inactivation schemes, we report the mean and standard error of the variance ratio:

$$VR = \frac{\mathrm{Var}_{inact}(\mathbf{r}^{(0)}) + \epsilon_v}{\mathrm{Var}(\mathbf{r}^{(0)}) + \epsilon_v}, \tag{12}$$

where we added $\epsilon_v = 10^{-3}$ to the denominator to prevent numerical instability and to the numerator ensure that the ratio evaluates to 1 if the two variances are equivalent.

Generating hallucinations in hierarchical variational autoencoders. To model more complex hallucination phenomena than could be observed in our simpler Wake-Sleep-trained networks, we used pretrained VDVAE [79] models trained on Tiny ImageNet [80], a 64x64 pixel variant of ImageNet, and FFHQ-256 [81], a dataset of 256x256 pixel human faces. VDVAE models are very similar to our Wake-Sleep-trained models: they are trained on the same unsupervised representation learning objective function (the ELBO), and every layer of the multilayer network models are parameterized by a bottom-up inference distribution b and a top-down generative distribution p. VDVAE models are top-down VAEs [77], which means that the inference distribution is conditioned on bottom-up stimuli *and* latent network activity at higher layers, i.e. the distribution is written $b(\mathbf{r}^{(l)}|h(\mathbf{s}, \mathbf{r}^{(l+1)}))$, where $h(\cdot)$ is a parameterized neural network. By contrast the generative distribution is conditioned only on top-down inputs, and is written $p(\mathbf{r}^{(l)}|\mu(\mathbf{r}^{(l+1)}))$, where $\mu(\mathbf{r}^{(l+1)})$ is also a parameterized neural network.

For our Wake-Sleep-trained networks, we modeled hallucinations by simulating a stochastic time series at each layer (Eq. 8), but for the VDVAE models we found this to be computationally infeasible. Instead, we modeled hallucinations with a single bottom-up and top-down pass through the network, as follows:

$$\mathbf{r}^{(l)} = (1-\alpha)\mathbf{r}_b^{(l)} + (\alpha)\mathbf{r}_p^{(l)}, \tag{13}$$

where $\mathbf{r}_b^l \sim b(\mathbf{r}^{(l)}|h(\mathbf{s}, \mathbf{r}^{(l+1)}))$ is a sample from the inference distribution, and $\mathbf{r}_p^l \sim p(\mathbf{r}^{(l)}|\mu(\mathbf{r}^{(l+1)}))$ is a sample from the generative distribution. This generation scheme is simpler and less computationally expensive than our previous method, while still producing purely Wake-stage sampling when $\alpha = 0$ and Sleep-stage sampling when $\alpha = 1$; intermediate values of α correspond to modeled hallucinatory network states[4]. Our Laplacian pyramid analysis of generated images was performed using the Pyrtools package [130].

[4]Code for reproducing results obtained with pretrained VDVAE models is available here: https://github.com/colinbredenberg/vdvae.

5 Acknowledgements and Funding

We would like to thank members of both G.L. and B.R.'s labs, as well as James M. Shine, Brandon Munn, Christopher Whyte, Veronica Chelu, Jiameng Wu, Matthew Larkum, Santiago Jaramillo, Michael Wehr, Neil Savalia, Alexandra Klein, Sarah Cook, Conor Lane, Anousheh Bakhti-Suroosh, Runchong Wang, Michael Okun, and Jordan O'Byrne for insightful discussions and feedback. This work was supported by: [GL] NSERC Discovery Grant (RGPIN-2018-04821), Canada CIFAR AI Chair Program, Canada Research Chair in Neural Computations and Interfacing (CIHR, tier 2). [BR] NSERC (Discovery Grant: RGPIN-2020-05105; Discovery Accelerator Supplement: RGPAS-2020-00031; Arthur B. McDonald Fellowship: 566355-2022) and CIFAR (Canada AI Chair; Learning in Machine and Brains Fellowship). [CB] is supported in part by the FRQNT Strategic Clusters Program (Centre UNIQUE - Quebec Neuro-AI Research Center). The authors acknowledge the material support of NVIDIA in the form of computational resources.

References

[1] Robert L Jakab and Patricia S Goldman-Rakic. 5-hydroxytryptamine2a serotonin receptors in the primate cerebral cortex: possible site of action of hallucinogenic and antipsychotic drugs in pyramidal cell apical dendrites. *Proceedings of the National Academy of Sciences*, 95(2):735–740, 1998.

[2] J De Almeida and G Mengod. Quantitative analysis of glutamatergic and GABAergic neurons expressing 5-HT2A receptors in human and monkey prefrontal cortex. *Journal of neurochemistry*, 103(2):475–486, 2007.

[3] Katrin H Preller and Franz X Vollenweider. Phenomenology, structure, and dynamic of psychedelic states. *Behavioral neurobiology of psychedelic drugs*, pages 221–256, 2018.

[4] Ling-Xiao Shao, Clara Liao, Ian Gregg, Pasha A Davoudian, Neil K Savalia, Kristina Delagarza, and Alex C Kwan. Psilocybin induces rapid and persistent growth of dendritic spines in frontal cortex in vivo. *Neuron*, 109(16):2535–2544, 2021.

[5] Steven F Grieco, Eero Castrén, Gitte M Knudsen, Alex C Kwan, David E Olson, Yi Zuo, Todd C Holmes, and Xiangmin Xu. Psychedelics and neural plasticity: therapeutic implications. *Journal of Neuroscience*, 42(45):8439–8449, 2022.

[6] Silvia Muttoni, Maddalena Ardissino, and Christopher John. Classical psychedelics for the treatment of depression and anxiety: A systematic review. *Journal of affective disorders*, 258:11–24, 2019.

[7] Erwin Krediet, Tijmen Bostoen, Joost Breeksema, Annette van Schagen, Torsten Passie, and Eric Vermetten. Reviewing the potential of psychedelics for the treatment of PTSD. *International Journal of Neuropsychopharmacology*, 23(6):385–400, 2020.

[8] Maxemiliano V Vargas, Lee E Dunlap, Chunyang Dong, Samuel J Carter, Robert J Tombari, Shekib A Jami, Lindsay P Cameron, Seona D Patel, Joseph J Hennessey, Hannah N Saeger, et al. Psychedelics promote neuroplasticity through the activation of intracellular 5-HT2A receptors. *Science*, 379(6633):700–706, 2023.

[9] Rainer Kraehenmann, Dan Pokorny, Leonie Vollenweider, Katrin H Preller, Thomas Pokorny, Erich Seifritz, and Franz X Vollenweider. Dreamlike effects of LSD on waking imagery in humans depend on serotonin 2a receptor activation. *Psychopharmacology*, 234:2031–2046, 2017.

[10] A Richard Green, Annis O Mechan, J Martin Elliott, Esther O'Shea, and M Isabel Colado. The pharmacology and clinical pharmacology of 3, 4-methylenedioxymethamphetamine (MDMA,"ecstasy"). *Pharmacological reviews*, 55(3):463–508, 2003.

[11] Danuta Marona-Lewicka, Deborah M Kurrasch-Orbaugh, Jennifer R Selken, Medhane G Cumbay, Joshua G Lisnicchia, and David E Nichols. Re-evaluation of lisuride pharmacology: 5-hydroxytryptamine1a receptor-mediated behavioral effects overlap its other properties in rats. *Psychopharmacology*, 164(1):93–107, 2002.

[12] Robin Carhart-Harris. Waves of the unconscious: the neurophysiology of dreamlike phenomena and its implications for the psychodynamic model of the mind. *Neuropsychoanalysis*, 9(2):183–211, 2007.

[13] Christopher W Thomas, Cristina Blanco-Duque, Benjamin J Bréant, Guy M Goodwin, Trevor Sharp, David M Bannerman, and Vladyslav V Vyazovskiy. Psilocin acutely alters sleep-wake architecture and cortical brain activity in laboratory mice. *Translational psychiatry*, 12(1):77, 2022.

[14] Daniela Dudysová, Karolina Janků, Michal Šmotek, Elizaveta Saifutdinova, Jana Kopřivová, Jitka Bušková, Bryce Anthony Mander, Martin Brunovskỳ, Peter Zach, Jakub Korčák, et al. The effects of daytime psilocybin administration on sleep: implications for antidepressant action. *Frontiers in pharmacology*, 11:602590, 2020.

[15] Manel J Barbanoj, Jordi Riba, S Clos, S Giménez, E Grasa, and S Romero. Daytime ayahuasca administration modulates REM and slow-wave sleep in healthy volunteers. *Psychopharmacology*, 196:315–326, 2008.

[16] Gabrielle Girardeau, Karim Benchenane, Sidney I Wiener, György Buzsáki, and Michaël B Zugaro. Selective suppression of hippocampal ripples impairs spatial memory. *Nature neuroscience*, 12(10):1222–1223, 2009.

[17] Lorena Deuker, Jan Olligs, Juergen Fell, Thorsten A. Kranz, Florian Mormann, Christian Montag, Martin Reuter, Christian E. Elger, and Nikolai Axmacher. Memory consolidation by replay of stimulus-specific neural activity. *Journal of Neuroscience*, 33(49):19373–19383, 2013.

[18] Gaetan De Lavilléon, Marie Masako Lacroix, Laure Rondi-Reig, and Karim Benchenane. Explicit memory creation during sleep demonstrates a causal role of place cells in navigation. *Nature neuroscience*, 18(4):493–495, 2015.

[19] Nicolas Maingret, Gabrielle Girardeau, Ralitsa Todorova, Marie Goutierre, and Michaël Zugaro. Hippocampo-cortical coupling mediates memory consolidation during sleep. *Nature neuroscience*, 19(7):959–964, 2016.

[20] Antonio Fernández-Ruiz, Azahara Oliva, Eliezyer Fermino de Oliveira, Florbela Rocha-Almeida, David Tingley, and György Buzsáki. Long-duration hippocampal sharp wave ripples improve memory. *Science*, 364(6445):1082–1086, 2019.

[21] Matthew P Walker and Robert Stickgold. Sleep-dependent learning and memory consolidation. *Neuron*, 44(1):121–133, 2004.

[22] Zoltán Nádasdy, Hajime Hirase, András Czurkó, Jozsef Csicsvari, and György Buzsáki. Replay and time compression of recurring spike sequences in the hippocampus. *Journal of Neuroscience*, 19(21):9497–9507, 1999.

[23] Albert K Lee and Matthew A Wilson. Memory of sequential experience in the hippocampus during slow wave sleep. *Neuron*, 36(6):1183–1194, 2002.

[24] David J Foster. Replay comes of age. *Annual review of neuroscience*, 40:581–602, 2017.

[25] Daoyun Ji and Matthew A Wilson. Coordinated memory replay in the visual cortex and hippocampus during sleep. *Nature neuroscience*, 10(1):100–107, 2007.

[26] David R Euston, Masami Tatsuno, and Bruce L McNaughton. Fast-forward playback of recent memory sequences in prefrontal cortex during sleep. *science*, 318(5853):1147–1150, 2007.

[27] Adrien Peyrache, Mehdi Khamassi, Karim Benchenane, Sidney I Wiener, and Francesco P Battaglia. Replay of rule-learning related neural patterns in the prefrontal cortex during sleep. *Nature neuroscience*, 12(7):919–926, 2009.

[28] Tal Kenet, Dmitri Bibitchkov, Misha Tsodyks, Amiram Grinvald, and Amos Arieli. Spontaneously emerging cortical representations of visual attributes. *Nature*, 425(6961):954–956, 2003.

[29] Shengjin Xu, Wanchen Jiang, Mu-ming Poo, and Yang Dan. Activity recall in a visual cortical ensemble. *Nature neuroscience*, 15(3):449–455, 2012.

[30] Kari L Hoffman and Bruce L McNaughton. Coordinated reactivation of distributed memory traces in primate neocortex. *Science*, 297(5589):2070–2073, 2002.

[31] Kenway Louie and Matthew A Wilson. Temporally structured replay of awake hippocampal ensemble activity during rapid eye movement sleep. *Neuron*, 29(1):145–156, 2001.

[32] Thomas Andrillon, Yuval Nir, Chiara Cirelli, Giulio Tononi, and Itzhak Fried. Single-neuron activity and eye movements during human rem sleep and awake vision. *Nature communications*, 6(1):7884, 2015.

[33] Geoffrey E Hinton, Peter Dayan, Brendan J Frey, and Radford M Neal. The "wake-sleep" algorithm for unsupervised neural networks. *Science*, 268(5214):1158–1161, 1995.

[34] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pages 1278–1286. PMLR, 2014.

[35] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[36] Eero P Simoncelli. Vision and the statistics of the visual environment. *Current opinion in neurobiology*, 13(2):144–149, 2003.

[37] Eero P Simoncelli and Bruno A Olshausen. Natural image statistics and neural representation. *Annual review of neuroscience*, 24(1):1193–1216, 2001.

[38] Johannes Ballé, Valero Laparra, and Eero P Simoncelli. End-to-end optimized image compression. *arXiv preprint arXiv:1611.01704*, 2016.

[39] James J DiCarlo, Davide Zoccolan, and Nicole C Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, 2012.

[40] Irina Higgins, Loic Matthey, Arka Pal, Christopher P Burgess, Xavier Glorot, Matthew M Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR (Poster)*, 3, 2017.

[41] Shiro Ikeda, Shun-ichi Amari, and Hiroyuki Nakahara. Convergence of the wake-sleep algorithm. *Advances in neural information processing systems*, 11, 1998.

[42] Kevin G Kirby. A tutorial on Helmholtz Machines. *Department of Computer Science, Northern Kentucky University*, 2006.

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

[44] Jaan Aru, Francesca Siclari, William A Phillips, and Johan F Storm. Apical drive—a cellular mechanism of dreaming? *Neuroscience & Biobehavioral Reviews*, 119:440–455, 2020.

[45] Konrad P Körding and Peter König. Supervised and unsupervised learning with two sites of synaptic integration. *Journal of computational neuroscience*, 11:207–215, 2001.

[46] Robert Urbanczik and Walter Senn. Learning by the dendritic prediction of somatic spiking. *Neuron*, 81(3):521–528, 2014.

[47] Jordan Guerguiev, Timothy P Lillicrap, and Blake A Richards. Towards deep learning with segregated dendrites. *Elife*, 6:e22901, 2017.

[48] João Sacramento, Rui Ponte Costa, Yoshua Bengio, and Walter Senn. Dendritic cortical microcircuits approximate the backpropagation algorithm. *Advances in neural information processing systems*, 31, 2018.

[49] Blake A Richards and Timothy P Lillicrap. Dendritic solutions to the credit assignment problem. *Current opinion in neurobiology*, 54:28–36, 2019.

[50] Alexandre Payeur, Jordan Guerguiev, Friedemann Zenke, Blake A Richards, and Richard Naud. Burst-dependent synaptic plasticity can coordinate learning in hierarchical circuits. *Nature neuroscience*, 24(7):1010–1019, 2021.

[51] Colin Bredenberg, Benjamin Lyo, Eero Simoncelli, and Cristina Savin. Impression learning: Online representation learning with synaptic plasticity. *Advances in Neural Information Processing Systems*, 34:11717–11729, 2021.

[52] Tom M George, Kimberly L Stachenfeld, Caswell Barry, Claudia Clopath, and Tomoki Fukai. A generative model of the hippocampal formation trained with theta driven local learning rules. *Advances in Neural Information Processing Systems*, 36, 2024.

[53] Julie Seibt, Clément J Richard, Johanna Sigl-Glöckner, Naoya Takahashi, David I Kaplan, Guy Doron, Denis De Limoges, Christina Bocklisch, and Matthew E Larkum. Cortical dendritic activity correlates with spindle-rich oscillations during sleep in rodents. *Nature communications*, 8(1):684, 2017.

[54] Daisuke Miyamoto, Daichi Hirai, CCA Fung, Ayumu Inutsuka, Maya Odagawa, Takayuki Suzuki, Roman Boehringer, Chinnakkaruppan Adaikkan, Chie Matsubara, Norio Matsuki, et al. Top-down cortical input during nrem sleep consolidates perceptual memory. *Science*, 352(6291):1315–1318, 2016.

[55] Wei Li, Lei Ma, Guang Yang, and Wen-Biao Gan. Rem sleep selectively prunes and maintains new synapses in development and learning. *Nature neuroscience*, 20(3):427–437, 2017.

[56] Yanmei Zhou, Cora Sau Wan Lai, Yang Bai, Wei Li, Ruohe Zhao, Guang Yang, Marcos G Frank, and Wen-Biao Gan. Rem sleep promotes experience-dependent dendritic spine elimination in the mouse cortex. *Nature communications*, 11(1):4819, 2020.

[57] Mattia Aime, Niccolò Calcini, Micaela Borsa, Tiago Campelo, Thomas Rusterholz, Andrea Sattin, Tommaso Fellin, and Antoine Adamantidis. Paradoxical somatodendritic decoupling supports cortical plasticity during rem sleep. *Science*, 376(6594):724–730, 2022.

[58] GK Aghajanian and GJ Marek. Serotonin induces excitatory postsynaptic potentials in apical dendrites of neocortical pyramidal cells. *Neuropharmacology*, 36(4-5):589–599, 1997.

[59] George K Aghajanian and Gerard J Marek. Serotonin and hallucinogens. *Neuropsychopharmacology*, 21(2):16S–23S, 1999.

[60] Viktor L Arvanov, Xiaofu Liang, Angelo Russo, and Rex Y Wang. LSD and DOB: interaction with 5-HT2A receptors to inhibit NMDA receptor-mediated transmission in the rat prefrontal cortex. *European Journal of Neuroscience*, 11(9):3064–3072, 1999.

[61] Javier Hidalgo Jiménez, Kristjan Kaup, and Jaan Aru. Electrophysiological mechanisms of psychedelic drugs: A systematic review. *bioRxiv*, pages 2025–07, 2025.

[62] Li Deng. The MNIST database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6):141–142, 2012.

[63] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[64] Lou Beaulieu-Laroche, Enrique HS Toloza, Norma J Brown, and Mark T Harnett. Widespread and highly correlated somato-dendritic activity in cortical layer 5 neurons. *Neuron*, 103(2):235–241, 2019.

[65] Justin K O'Hare, Jamie Wang, Margjele D Shala, Franck Polleux, and Attila Losonczy. Distal tuft dendrites shape and maintain new place fields. *bioRxiv*, pages 2024–02, 2024.

[66] Romain Nardou, Edward Sawyer, Young Jun Song, Makenzie Wilkinson, Yasmin Padovan-Hernandez, Júnia Lara de Deus, Noelle Wright, Carine Lama, Sehr Faltin, Loyal A Goff, et al. Psychedelics reopen the social reward learning critical period. *Nature*, 618(7966):790–798, 2023.

[67] Mario de la Fuente Revenga, Bohan Zhu, Christopher A Guevara, Lynette B Naler, Justin M Saunders, Zirui Zhou, Rudy Toneatti, Salvador Sierra, Jennifer T Wolstenholme, Patrick M Beardsley, et al. Prolonged epigenomic and synaptic plasticity alterations following single exposure to a psychedelic in mice. *Cell reports*, 37(3), 2021.

[68] Arthur Juliani, Veronica Chelu, Laura Graesser, and Adam Safron. A dual-receptor model of serotonergic psychedelics: therapeutic insights from simulated cortical dynamics. *bioRxiv*, pages 2024–04, 2024.

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

[70] Alexander V Lebedev, Mendel Kaelen, Martin Lövdén, Jonna Nilsson, Amanda Feilding, David J Nutt, and Robin L Carhart-Harris. LSD-induced entropic brain activity predicts subsequent personality change. *Human brain mapping*, 37(9):3203–3213, 2016.

[71] 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*, 8:20, 2014.

[72] Joshua S Siegel, Subha Subramanian, Demetrius Perry, Benjamin P Kay, Evan M Gordon, Timothy O Laumann, T Rick Reneau, Nicholas V Metcalf, Ravi V Chacko, Caterina Gratton, et al. Psilocybin desynchronizes the human brain. *Nature*, pages 1–8, 2024.

[73] Max Horrocks, Jennifer L Mohn, and Santiago Jaramillo. The serotonergic psychedelic DOI impairs deviance detection in the auditory cortex. *bioRxiv*, pages 2024–09, 2024.

[74] Benny Shanon. Ayahuasca visualizations a structural typology. *Journal of Consciousness Studies*, 9(2):3–30, 2002.

[75] José Luis Díaz. Sacred plants and visionary consciousness. *Phenomenology and the Cognitive Sciences*, 9:159–170, 2010.

[76] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[77] Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. Ladder variational autoencoders. *Advances in neural information processing systems*, 29, 2016.

[78] Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. *Advances in neural information processing systems*, 33:19667–19679, 2020.

[79] Rewon Child. Very deep vaes generalize autoregressive models and can outperform them on images. *arXiv preprint arXiv:2011.10650*, 2020.

[80] Jiayu Wu, Qixiang Zhang, and Guoxi Xu. Tiny imagenet challenge. *Technical report*, 2017.

[81] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[82] Timothy P Lillicrap, Adam Santoro, Luke Marris, Colin J Akerman, and Geoffrey Hinton. Backpropagation and the brain. *Nature Reviews Neuroscience*, 21(6):335–346, 2020.

[83] Roman Pogodin, Yash Mehta, Timothy Lillicrap, and Peter E Latham. Towards biologically plausible convolutional networks. *Advances in neural information processing systems*, 34:13924–13936, 2021.

[84] Ferenc Csikor, Balázs Meszéna, Bence Szabó, and Gergő Orbán. Top-down inference in an early visual cortex inspired hierarchical variational autoencoder. *arXiv preprint arXiv:2206.00436*, 2022.

[85] Edward V Evarts, William Landau, Walter Freygang jr, and Wade H Marshall. Some effects of lysergic acid diethylamide and bufotenine on electrical activity in the cat's visual system. *American Journal of Physiology-Legacy Content*, 182(3):594–598, 1955.

[86] Zohre Azimi, Ruxandra Barzan, Katharina Spoida, Tatjana Surdin, Patric Wollenweber, Melanie D Mark, Stefan Herlitze, and Dirk Jancke. Separable gain control of ongoing and evoked activity in the visual cortex by serotonergic input. *Elife*, 9:e53552, 2020.

[87] Angie M Michaiel, Philip RL Parker, and Cristopher M Niell. A hallucinogenic serotonin-2a receptor agonist reduces visual response gain and alters temporal dynamics in mouse v1. *Cell reports*, 26(13):3475–3483, 2019.

[88] Bernard Widrow and Michael A Lehr. 30 years of adaptive neural networks: perceptron, madaline, and backpropagation. *Proceedings of the IEEE*, 78(9):1415–1442, 1990.

[89] Per Jesper Sjöström and Michael Häusser. A cooperative switch determines the sign of synaptic plasticity in distal dendrites of neocortical pyramidal neurons. *Neuron*, 51(2):227–238, 2006.

[90] Robert C Froemke, Johannes J Letzkus, Björn M Kampa, Giao B Hang, and Greg J Stuart. Dendritic synapse location and neocortical spike-timing-dependent plasticity. *Frontiers in synaptic neuroscience*, 2:29, 2010.

[91] Pedro AM Mediano, Fernando E Rosas, Christopher Timmermann, Leor Roseman, David J Nutt, Amanda Feilding, Mendel Kaelen, Morten L Kringelbach, Adam B Barrett, Anil K Seth, et al. Effects of external stimulation on psychedelic state neurodynamics. *ACS Chemical Neuroscience*, 15(3):462–471, 2024.

[92] Colin Bredenberg and Cristina Savin. Desiderata for normative models of synaptic plasticity. *Neural Computation*, 36(7):1245–1285, 2024.

[93] Daniel Levenstein, Veronica A Alvarez, Asohan Amarasingham, Habiba Azab, Zhe S Chen, Richard C Gerkin, Andrea Hasenstaub, Ramakrishnan Iyer, Renaud B Jolivet, Sarah Marzen, et al. On the role of theory and modeling in neuroscience. *Journal of Neuroscience*, 43(7):1074–1088, 2023.

[94] G Bard Ermentrout and Jack D Cowan. A mathematical theory of visual hallucination patterns. *Biological cybernetics*, 34(3):137–150, 1979.

[95] Keisuke Suzuki, Warrick Roseboom, David J Schwartzman, and Anil K Seth. A deep-dream virtual reality platform for studying altered perceptual phenomenology. *Scientific reports*, 7(1):15982, 2017.

[96] Robert Stickgold, J Allen Hobson, Roar Fosse, and Magdalena Fosse. Sleep, learning, and dreams: off-line memory reprocessing. *Science*, 294(5544):1052–1057, 2001.

[97] Carli Domenico, Daniel Haggerty, Xiang Mou, and Daoyun Ji. Lsd degrades hippocampal spatial representations and suppresses hippocampal-visual cortical interactions. *Cell reports*, 36(11), 2021.

[98] Christopher J Whyte, Michelle J Redinbaugh, James M Shine, and Yuri B Saalmann. Thalamic contributions to the state and contents of consciousness. *Neuron*, 112(10):1611–1625, 2024.

[99] Andrew Gaddis, Daniel E Lidstone, Mary Beth Nebel, Roland R Griffiths, Stewart H Mostofsky, Amanda F Mejia, and Frederick S Barrett. Psilocybin induces spatially constrained alterations in thalamic functional organizaton and connectivity. *Neuroimage*, 260:119434, 2022.

[100] Stefano Delli Pizzi, Piero Chiacchiaretta, Carlo Sestieri, Antonio Ferretti, Maria Giulia Tullo, Stefania Della Penna, Giovanni Martinotti, Marco Onofrj, Leor Roseman, Christopher Timmermann, et al. Lsd-induced changes in the functional connectivity of distinct thalamic nuclei. *NeuroImage*, 283:120414, 2023.

[101] Dhairyya Singh, Kenneth A Norman, and Anna C Schapiro. A model of autonomous interactions between hippocampus and neocortex driving sleep-dependent memory consolidation. *Proceedings of the national academy of sciences*, 119(44):e2123432119, 2022.

[102] Paul C Bressloff, Jack D Cowan, Martin Golubitsky, Peter J Thomas, and Matthew C Wiener. Geometric visual hallucinations, Euclidean symmetry and the functional architecture of striate cortex. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 356(1407):299–330, 2001.

[103] T Anderson Keller and Max Welling. Neural wave machines: learning spatiotemporally structured representations with locally coupled oscillatory recurrent neural networks. In *International Conference on Machine Learning*, pages 16168–16189. PMLR, 2023.

[104] T Anderson Keller, Lyle Muller, Terrence Sejnowski, and Max Welling. Traveling waves encode the recent past and enhance sequence learning. *arXiv preprint arXiv:2309.08045*, 2023.

[105] Rajesh PN Rao and Dana H Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1):79–87, 1999.

[106] Hardik Rajpal, Pedro AM Mediano, Fernando E Rosas, Christopher B Timmermann, Stefan Brugger, Suresh Muthukumaraswamy, Anil K Seth, Daniel Bor, Robin L Carhart-Harris, and Henrik J Jensen. Psychedelics and schizophrenia: Distinct alterations to Bayesian inference. *NeuroImage*, 263:119624, 2022.

[107] Andrea Alamia, Christopher Timmermann, David J Nutt, Rufin VanRullen, and Robin L Carhart-Harris. Dmt alters cortical travelling waves. *Elife*, 9:e59784, 2020.

[108] David H Ackley, Geoffrey E Hinton, and Terrence J Sejnowski. A learning algorithm for Boltzmann machines. *Cognitive science*, 9(1):147–169, 1985.

[109] David P Reichert, Peggy Series, and Amos J Storkey. Charles bonnet syndrome: evidence for a generative model in the cortex? *PLoS computational biology*, 9(7):e1003134, 2013.

[110] Thomas L O'Donohue, William R Millington, Gail E Handelmann, Patricia C Contreras, and Bibie M Chronwall. On the 50th anniversary of Dale's law: multiple neurotransmitter neurons. *Trends in Pharmacological Sciences*, 6:305–308, 1985.

[111] Jonathan Cornford, Damjan Kalajdzievski, Marco Leite, Amélie Lamarquette, Dimitri M Kullmann, and Blake Richards. Learning to live with Dale's principle: ANNs with separate excitatory and inhibitory units. *bioRxiv*, pages 2020–11, 2020.

[112] Brandon R Munn, Eli J Müller, Vicente Medel, Sharon L Naismith, Joseph T Lizier, Robert D Sanders, and James M Shine. Neuronal connected burst cascades bridge macroscale adaptive signatures across arousal states. *Nature Communications*, 14(1):6846, 2023.

[113] Bill Podlaski and Christian K Machens. Biological credit assignment through dynamic inversion of feedforward networks. *Advances in Neural Information Processing Systems*, 33:10065–10076, 2020.

[114] Grace W Lindsay. Attention in psychology, neuroscience, and machine learning. *Frontiers in computational neuroscience*, 14:29, 2020.

[115] Nizar Islah, Guillaume Etter, Mashbayar Tugsbayar, Busra Tugce Gurbuz, Blake Richards, and Eilif B Muller. Learning to combine top-down context and feed-forward representations under ambiguity with apical and basal dendrites. *Cerebral Cortex*, 35(6):bhaf134, 2025.

[116] Aditya Gilra and Wulfram Gerstner. Predicting non-linear dynamics by stable local learning in a recurrent spiking neural network. *Elife*, 6:e28295, 2017.

[117] Alexander Meulemans, Matilde Tristany Farinha, Javier García Ordóñez, Pau Vilimelis Aceituno, João Sacramento, and Benjamin F Grewe. Credit assignment in neural networks through deep feedback control. *Advances in Neural Information Processing Systems*, 34:4674–4687, 2021.

[118] Pieter R Roelfsema and Arjen van Ooyen. Attention-gated reinforcement learning of internal representations for classification. *Neural computation*, 17(10):2176–2214, 2005.

[119] Robin L Carhart-Harris and DJ Nutt. Serotonin and brain function: a tale of two receptors. *Journal of psychopharmacology*, 31(9):1091–1120, 2017.

[120] Cristiano Bombardi and Giuseppe Di Giovanni. Functional anatomy of 5-ht2a receptors in the amygdala and hippocampal complex: relevance to memory functions. *Experimental brain research*, 230(4):427–439, 2013.

[121] Xiaolong Jiang, Guoqiang Xing, Chunhui Yang, Ajay Verma, Lei Zhang, and He Li. Stress impairs 5-ht2a receptor-mediated serotonergic facilitation of gaba release in juvenile rat basolateral amygdala. *Neuropsychopharmacology*, 34(2):410–423, 2009.

[122] Thomas J Kelly, Emma M Bonniwell, Lianwei Mu, Xiaojie Liu, Ying Hu, Vladislav Friedman, Hao Yu, Wantang Su, John D McCorvy, and Qing-song Liu. Psilocybin analog 4-oh-dipt enhances fear extinction and gabaergic inhibition of principal neurons in the basolateral amygdala. *Neuropsychopharmacology*, 49(5):854–863, 2024.

[123] Praachi Tiwari, Pasha A Davoudian, Darshana Kapri, Ratna Mahathi Vuruputuri, Lindsay A Karaba, Mukund Sharma, Giulia Zanni, Angarika Balakrishnan, Pratik R Chaudhari, Amartya Pradhan, et al. Ventral hippocampal parvalbumin interneurons gate the acute anxiolytic action of the serotonergic psychedelic doi. *Neuron*, 112(22):3697–3714, 2024.

[124] Klaus-Peter Lesch and Jonas Waider. Serotonin in the modulation of neural plasticity and networks: implications for neurodevelopmental disorders. *Neuron*, 76(1):175–191, 2012.

[125] Philip R Corlett, Chris D Frith, and Paul C Fletcher. From drugs to deprivation: a bayesian framework for understanding models of psychosis. *Psychopharmacology*, 206(4):515–530, 2009.

[126] Panayiota Poirazi, Terrence Brannon, and Bartlett W Mel. Pyramidal neuron as two-layer neural network. *Neuron*, 37(6):989–999, 2003.

[127] Sergey Ioffe. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[128] Colin Bredenberg, Ezekiel Williams, Cristina Savin, Blake Richards, and Guillaume Lajoie. Formalizing locality for normative synaptic plasticity models. *Advances in Neural Information Processing Systems*, 36, 2024.

[129] Ezekiel Williams, Colin Bredenberg, and Guillaume Lajoie. Flexible phase dynamics for bio-plausible contrastive learning. In *International Conference on Machine Learning*, pages 37042–37065. PMLR, 2023.

[130] Eero Simoncelli, Rob Young, William Broderick, Pierre-Étienne Fiquet, Zhuo Wang, Zahra Kadkhodaie, Nikhil Parthasarathy, and Brian Ward. Pyrtools: tools for multi-scale image processing, April 2025.

[131] Dana Strauss, Sara de la Salle, Jordan Sloshower, and Monnica T Williams. Research abuses against people of colour and other vulnerable groups in early psychedelic research. *Journal of Medical Ethics*, 48(10):728–737, 2022.

[132] James A Bauml and Stacy B Schaefer. *Peyote: History, tradition, politics, and conservation.* Bloomsbury Publishing USA, 2016.

[133] Giorgio Samorini. The oldest archeological data evidencing the relationship of homo sapiens with psychoactive plants: A worldwide overview. *Journal of Psychedelic Studies*, 3(2):63–80, 2019.

[134] Marc-Antoine Dilhac, Christophe Abrassart, and Nathalie Voarino. Report of the Montréal declaration for a responsible development of artificial intelligence. 2018.

[135] Kerri Rea and Bruce Wallace. Enhancing equity-oriented care in psychedelic medicine: Utilizing the EQUIP framework. *International Journal of Drug Policy*, 98:103429, 2021.

[136] Annette J Browne, Colleen Varcoe, Marilyn Ford-Gilboe, C Nadine Wathen, and EQUIP Research Team. EQUIP healthcare: An overview of a multi-component intervention to enhance equity-oriented care in primary health care settings. *International journal for equity in health*, 14:1–11, 2015.

[137] Peter Dayan, Geoffrey E Hinton, Radford M Neal, and Richard S Zemel. The Helmholtz machine. *Neural computation*, 7(5):889–904, 1995.

[138] Tahereh Toosi and Elias Issa. Brain-like flexible visual inference by harnessing feedback feedforward alignment. *Advances in Neural Information Processing Systems*, 36, 2024.

[139] A Aldo Faisal, Luc PJ Selen, and Daniel M Wolpert. Noise in the nervous system. *Nature reviews neuroscience*, 9(4):292–303, 2008.

[140] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.

[141] Zahra Kadkhodaie and Eero Simoncelli. Stochastic solutions for linear inverse problems using the prior implicit in a denoiser. *Advances in Neural Information Processing Systems*, 34:13242–13254, 2021.

[142] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[143] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[144] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

6 Ethics Declarations

Psychedelic drug research has a long history fraught with many instances of unethical research practice [131]. Further, psychedelic drug use itself has long been stigmatized and punished through legal measures [132], often at the expense of indigenous peoples, who have long incorporated psychoactive substances into their cultural and spiritual practices [133]. In the interest of avoiding a repetition of past mistakes, we feel compelled to provide explicit guidance on how our work should be interpreted and used. To do so, we will take inspiration from two principal ethical frameworks: the Montreal Declaration on Responsible AI [134], and the EQUIP framework for equity-oriented healthcare [135, 136]. We strongly encourage anyone considering extending our research or using our work in any form of clinical setting to ensure that subsequent research adheres to these frameworks.

Below, drawing from these ethical frameworks, we will provide a set of guidelines for how our work should be interpreted and used. Though these guidelines are by no means exhaustive, our hope is that adherence to them will help promote the potential positive outcomes of our work while limiting potential negative consequences.

Guidelines for the ethical use of this study

Do:

1. Ensure that the elements of our hypothesis have been adequately tested, as outlined in our discussion, *before* using our framework in any form of clinical or therapeutic setting.
2. Use our ideas to inform further basic neuroscience research on perception, learning, sleep, and replay phenomena.
3. Explore our ideas as an opportunity to inform your own understanding of cognition, learning, and perception, with the understanding that these ideas have not yet been fully validated experimentally.
4. Feel free to ask us if you are worried that your proposed use of our work may have negative impacts.

Do not:

1. Report our results as scientific fact. We have outlined a *hypothesis*, which is designed to be tested by the experimental neuroscience community.
2. Cite or interpret our results without an adequate understanding of the evidence supporting the various claims made in this study. Feel free to ask us if you are worried that you may be misinterpreting our results.
3. Use our results to extract undue or inequitable profit. The ideas developed in this paper are the product of decades of research and public funding, built upon *centuries* of exploration of psychedelics. Any knowledge or value contained within this paper is the common heritage of all humanity, with particular recognition due to the indigenous and marginalized communities that have historically suffered and are currently suffering from oppressive government and industry policies.
4. Use our results for any application that could violate human rights or harm human beings in any way.

A Supplementary Materials

Recurrent network model. To explore the extent to which our results hold for different neuron models, and to give our generative model more expressive power than the traditional Helmholtz machine [137], we constructed a network model with a single timestep of within-layer recurrent denoising in each layer, which gives our model some similarities to denoising diffusion approaches [138]. For both our 'inference' mode and our 'generative' mode we specify both a denoised network state $\bar{\mathbf{r}}^{(l)}$ and a noise-corrupted network state $\mathbf{r}^{(l)}$ for layer l; specifying a neural network model is then equivalent to specifying, for each layer, a joint probability distribution over denoised and noise-corrupted network states for both the inference and generative modes, i.e. for the 'inference' mode we must specify a probability distribution $b(\bar{\mathbf{r}}^{(l)}, \mathbf{r}^{(l)}|\mathbf{r}^{(l-1)})$, while for the 'generative' mode we must specify a separate distribution $p(\bar{\mathbf{r}}^{(l)}, \mathbf{r}^{(l)}|\bar{\mathbf{r}}^{(l+1)})$[5]. Notice here that activity in 'inference' mode is conditioned on 'bottom-up' network states ($\mathbf{r}^{(l-1)}$), while activity in generative mode is conditioned on 'top-down' network states ($\bar{\mathbf{r}}^{(l+1)}$) (Figure 1a).

The 'inference mode' specifies a probability distribution over neural activity, conditioned on the next-lower layer (where the lowest layer is the stimulus layer, i.e. $\mathbf{r}^{(0)} = \mathbf{s}$)—mechanistically it corresponds to activity generated by feedforward projections. For $l > 0$, layer activity is sampled from the distribution $\bar{\mathbf{r}}^{(l)} \sim \mathcal{N}(h(\mathbf{r}^{(l-1)}), \sigma_b^2\mathbf{I})$, where $h(\mathbf{r}^{(l-1)})$ is given by:

$$h(\mathbf{r}^{(l-1)}) = \tanh(\mathbf{W}^{(l)}\mathbf{r}^{(l-1)} + b^{(l)}). \tag{14}$$

Subsequently, we add additional noise to get a noise-corrupted network state $\mathbf{r}^{(l)} \sim \mathcal{N}((1-\bar{\sigma}_b^2\mathbf{I})\bar{\mathbf{r}}^{(l)}, \bar{\sigma}_b^2)$; while noise corruption is a natural feature of network dynamics in the brain [139], we include it here in our model because it has been shown that *denoising* is a critical aspect of many powerful generative modeling approaches [140, 141, 142], and we have likewise found that it improves the quality of generated images in our learned networks (Supplemental Figure S2).

The 'generative' mode specifies a probability distribution over neural activity, conditioned on the next-higher layer—it corresponds mechanistically to activity generated by feedback projections. The highest layer, $\mathbf{r}^{(L)}$ is sampled from an $N^{(L)}$-dimensional independent standard normal distribution, $\mathbf{r}^{(L)} \sim \mathcal{N}(0, \mathbf{I})$, and all subsequent layers are sampled from the distribution $\mathbf{r}^{(l)} \sim \mathcal{N}(\mu(\bar{\mathbf{r}}^{(l+1)}), \sigma_p^2\mathbf{I})$, where $\mu(\bar{\mathbf{r}}^{(l+1)})$ is given by:

$$\mu(\bar{\mathbf{r}}^{(l+1)}) = \tanh(\mathbf{M}^{(l)}\bar{\mathbf{r}}^{(l+1)} + a^{(l)}), \tag{15}$$

where $m^{(l)}$ is a $N^l \times N^{(l+1)}$ weight matrix, and a is a bias term. Subsequently, the network goes through a single timestep of recurrent denoising, so that $\bar{\mathbf{r}}^{(l)} \sim \mathcal{N}(\bar{\mu}(\mathbf{r}^{(l)}), \bar{\sigma}_p^2\mathbf{I})$, where $\bar{\mu}(\mathbf{r}^{(l)})$ is given by:

$$\bar{\mu}(\mathbf{r}^{(l)}) = \mathbf{r}^{(l)} + \sigma(\mathbf{C}_1\mathbf{r}^{(l)} + \mathbf{c}_1)\tanh(\mathbf{C}_2\mathbf{r}^{(l)} + \mathbf{c}_2), \tag{16}$$

where $\sigma(\cdot)$ is a sigmoid nonlinearity that acts as a gating function similar those used in the LSTM [143] and GRU [144], $\mathbf{C}_1$ and $\mathbf{C}_2$ are $N^{(l)} \times N^{(l)}$ recurrent weight matrices, and $\mathbf{c}_1$ and $\mathbf{c}_2$ are biases. While this is a more complicated nonlinearity than is normally used for single-unit neuron models, functions of this kind could feasibly be implemented by nonlinear dendritic computations [126]; we further found that using this nonlinearity qualitatively improved generative performance. Given these parameterized probability distributions, we then determined the neural activity for each layer l according to Eq. (1). As with our multicompartmental neuron model, inference and generative parameters were updated according to Eqs. (4) and (7), respectively. Recurrent network hyperparameters are available in Table S3.

Simplified neuron model. As a control, we also tested our results using a simplified multilayer perceptron neuron model, which used neither batch normalization nor multiple dendritic branches. For the 'inference' mode within the simplified model, for $l > 0$, layer activity is sampled from the distribution $\mathbf{r}^{(l)} \sim \mathcal{N}(h(\mathbf{r}^{(l-1)}), \sigma_b^2\mathbf{I})$, where for neuron i in layer l, $h_i(\mathbf{r}^{(l-1)})$ is given by:

$$h_i(\mathbf{r}^{(l-1)}) = \tanh\left(\mathbf{W}_{i:}^{(l)}\mathbf{r}^{(l-1)} + b_i^{(l)}\right), \tag{17}$$

where $\mathbf{W}_{i:}^{(l)}$ is a $1 \times N^{(l-1)}$ matrix of basal synaptic weights onto neuron i, and b_i is the corresponding bias.

[5] As a notational convention, we will use letters when referring to mathematical objects from the generative, top-down distribution, and their vertical reflection when referring to the inference, bottom-up distribution (e.g. p and b)

The simplified 'generative' mode likewise replaces the branched neuron model used in the main text with a multilayer perceptron model. The highest layer, $\mathbf{r}^{(L)}$ is sampled from an $N^{(L)}$-dimensional independent standard normal distribution, $\mathbf{r}^{(L)} \sim \mathcal{N}(0, \mathbf{I})$, and all subsequent layers are sampled from the distribution $\mathbf{r}^{(l)} \sim \mathcal{N}(\mu(\mathbf{r}^{(l+1)}), \sigma_p^2 \mathbf{I})$, where for the ith neuron, $\mu_i(\mathbf{r}^{(l+1)})$ is given by:

$$\mu_i(\mathbf{r}^{(l+1)}) = \tanh\left(\mathbf{M}_{i:}^{(l)} \mathbf{r}^{(l+1)} + a_i^{(l)}\right), \tag{18}$$

where $\mathbf{M}_{i:}^{(l)}$ is a $1 \times N^{(l+1)}$ matrix of apical synaptic weights onto neuron i, and $a_i^{(l)}$ is the corresponding bias. As with the branched neuron model, inference and generative parameters were updated according to Equations (4) and (7), respectively. For optimization, we used the identical hyperparameters to the multicompartment neuron model (Table S1).

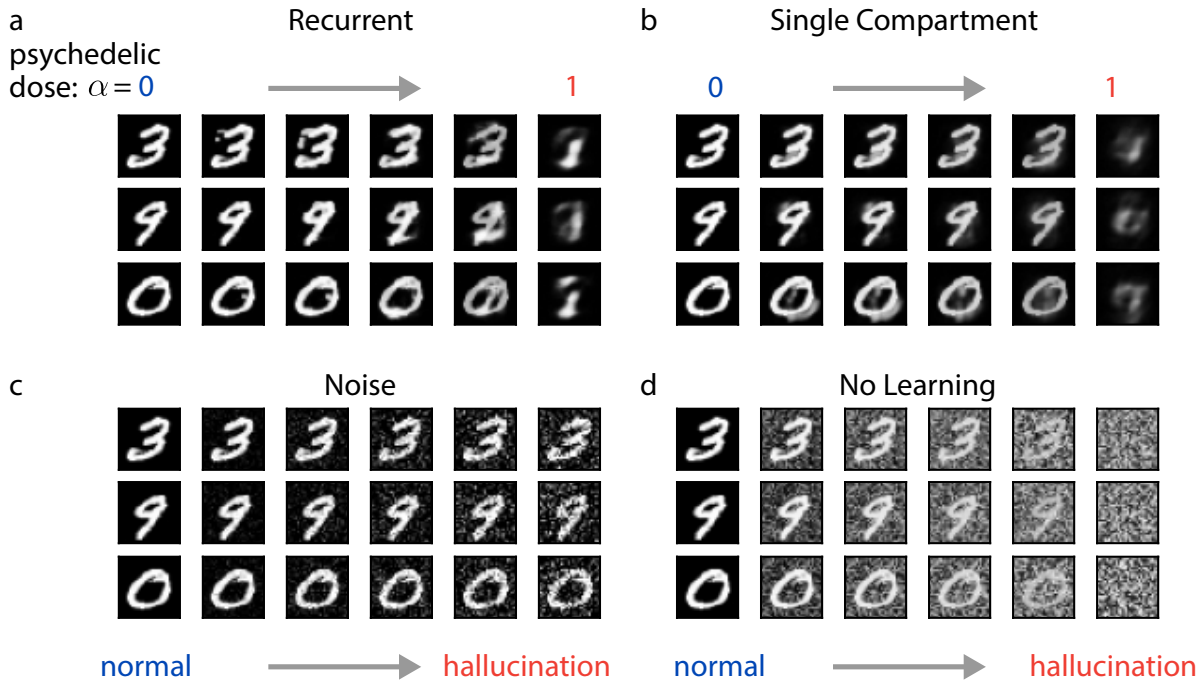

Figure S1: **Visualizing the effects of psychedelics for alternative model architectures**. We model the effects of classical psychedelics by progressively increasing α from 0 to 1 in alternative model architectures. We visualize the effects of psychedelics on the network representation by inspecting the stimulus layer **s**. **a)** Example stimulus-layer activity (rows) in response to an MNIST digit presentation as psychedelic dose increases (columns, left to right) in the recurrent network model. **b)** Same as (a) but for our single compartment neuron model. **c)** Same as (a) using the multicompartment neuron model used for our main results, but for our noise-based hallucination hallucination protocol. **d)** Same as (c), but in a network in which neither the generative nor inference pathways have been trained beyond initialization.

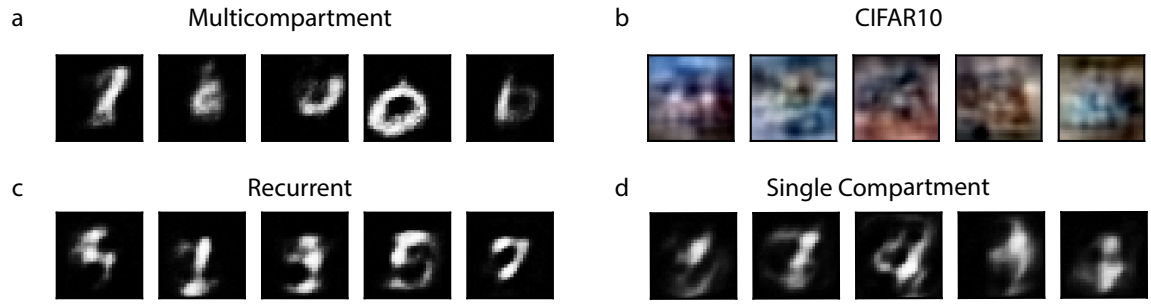

Figure S2: **Example generated images for different model architectures and datasets**. Generated images sampled from Eq. (1) with $\alpha = 1$ for: **a)** Our primary multicompartment neuron model trained on MNIST, **b)** A multicompartment neuron model trained on CIFAR10, **c)** The recurrent network model, **d)** The single compartment neuron model.

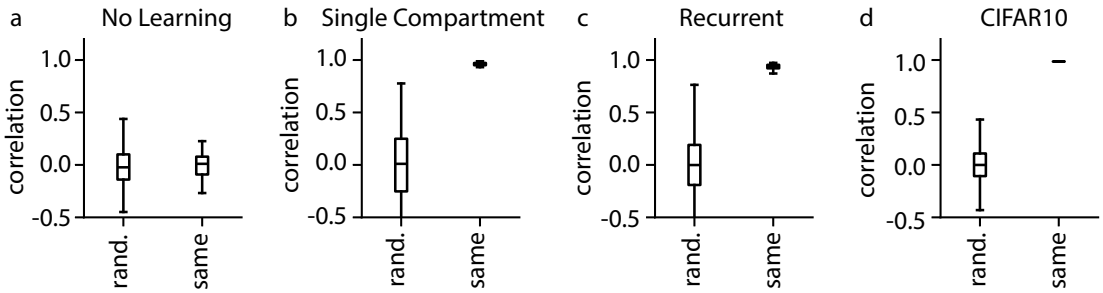

Figure S3: **Alignment between apical and basal dendritic compartments for different model architectures and datasets**. Apical-basal alignment for: **a)** An untrained multicompartment neuron model trained on MNIST, **b)** A single compartment neuron model, **c)** A recurrent network model, **d)** A multicompartment neuron model trained on CIFAR10.

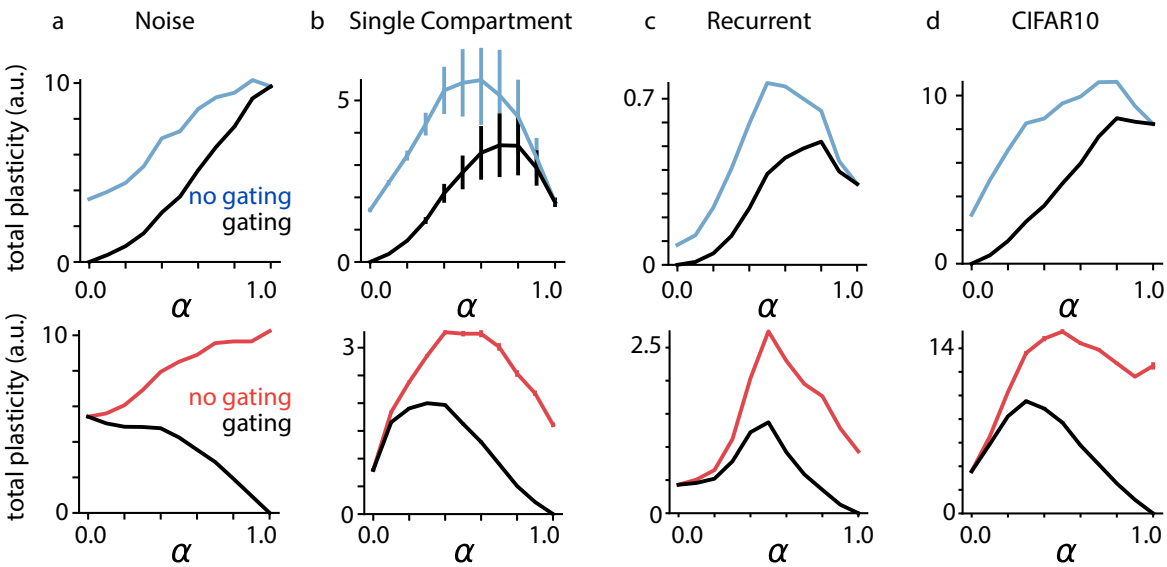

Figure S4: **Hallucination-induced synaptic plasticity for different neuron models**. **a)** Basal (top) and apical (bottom) plasticity as a function of α for a multicompartment neuron model trained on MNIST, using our noise-based hallucination protocol as a control. **b)** Same as (a) for a single compartment neuron model, using our primary hallucination protocol. **c)** Same as (b) for a recurrent network model, **d)** Same as (b) for a multicompartment neuron model trained on CIFAR10. Error bars indicate +/- 1 s.e.m.

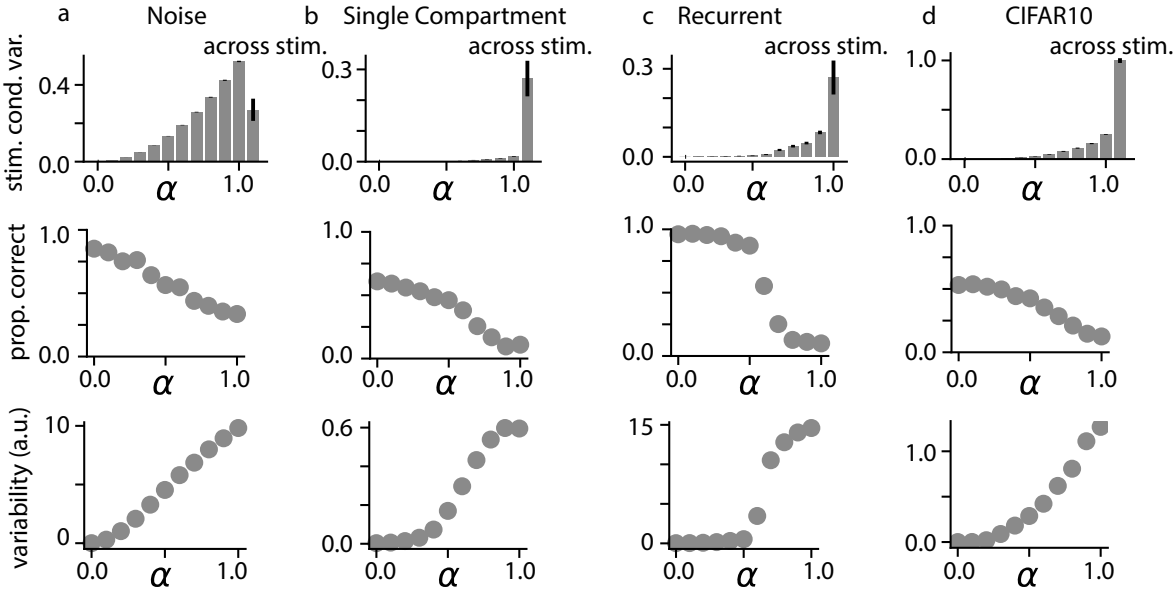

Figure S5: **Neural variability changes for different neuron models**. **a)** Stimulus-conditioned variability (top), classifier accuracy (middle), and classifier output variability (bottom) as a function of α for a multicompartment neuron model trained on MNIST, using our noise-based hallucination protocol as a control. **b)** Same as (b) for a single compartment neuron model, using our primary hallucination protocol. **c)** Same as (b) for a recurrent network model, **d)** Same as (b) for a multicompartment neuron model trained on CIFAR10. Error bars indicate +/- 1 s.e.m.

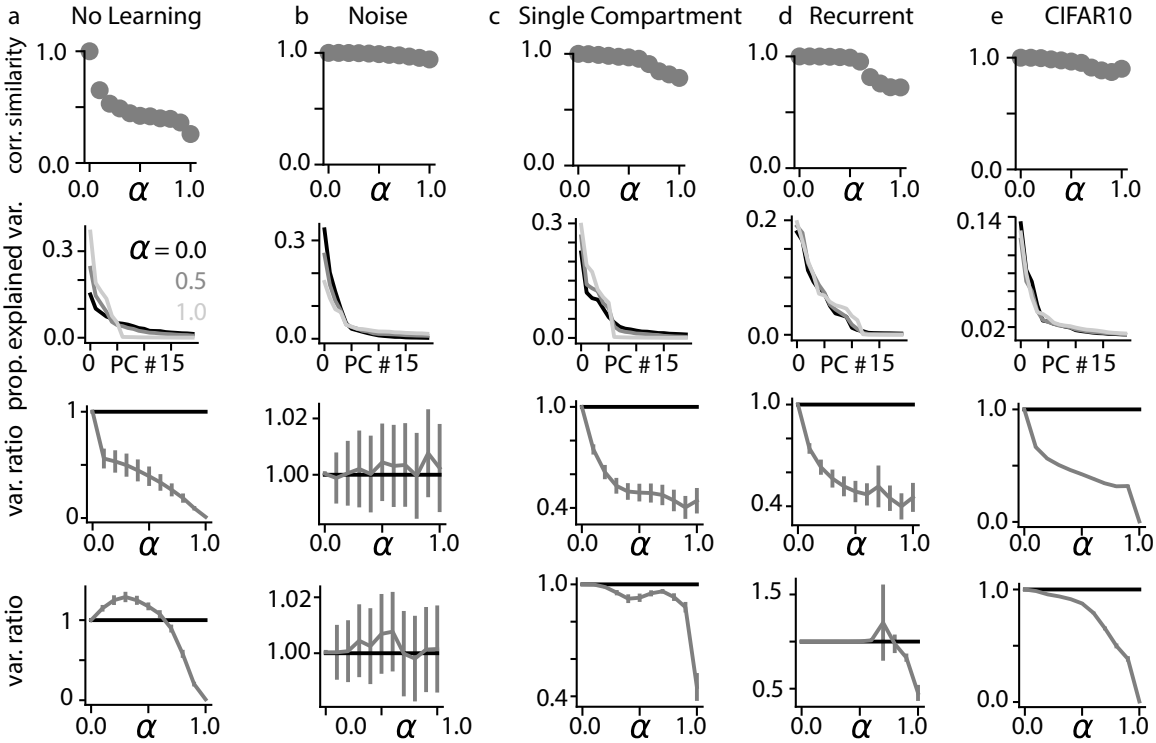

Figure S6: **Network-level effects of psychedelics for different network architectures and training datasets**. For each network architecture, we examine: correlation similarity as a function of α (top row), the proportion explained variance across stimuli as a function of principal component number (second row), the ratio of across-stimulus variance in stimulus layer neurons when apical dendrites have been inactivated compared to baseline conditions across different α values (third row), and the ratio of across-stimulus variance in stimulus layer neurons when the deepest network layer has been inactivated across different α values (fourth row). **a)** Results for an untrained multicompartment neuron. **b)** Results for a multicompartment neuron model trained on MNIST, using our noise-based hallucination protocol. **c)** Results for a single compartment neuron model. **d)** Results for a recurrent network model. **e)** Results for a multicompartment neuron model trained on CIFAR10. Error bars indicate +/- 1 s.e.m.

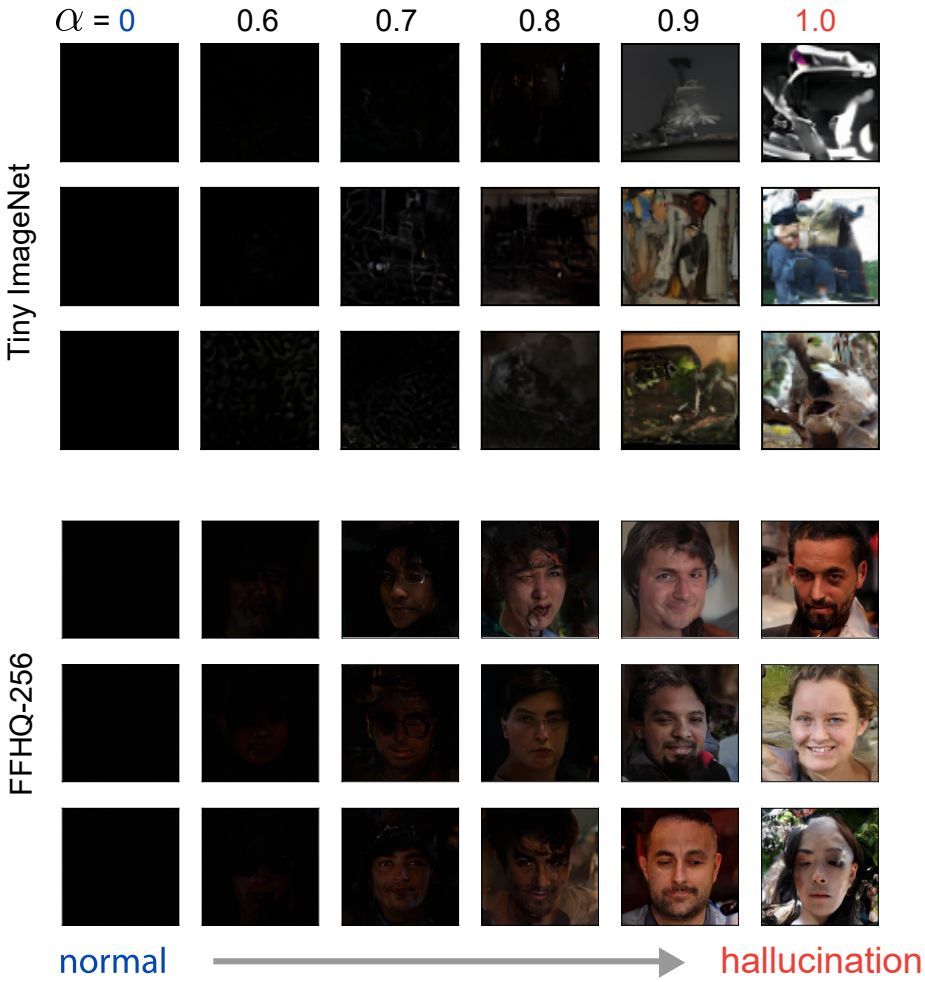

Figure S7: **Visualizing the eyes-closed effects of psychedelics in pretrained VDVAE models**. Decoded outputs of a pretrained VDVAE model trained on Tiny ImageNet (Top) and FFHQ-256 (Bottom) based on hallucinations generated in the top 35 layers of the model. Black input images were used for samples in all rows. Hallucination intensity, parameterized by α, increases along columns.

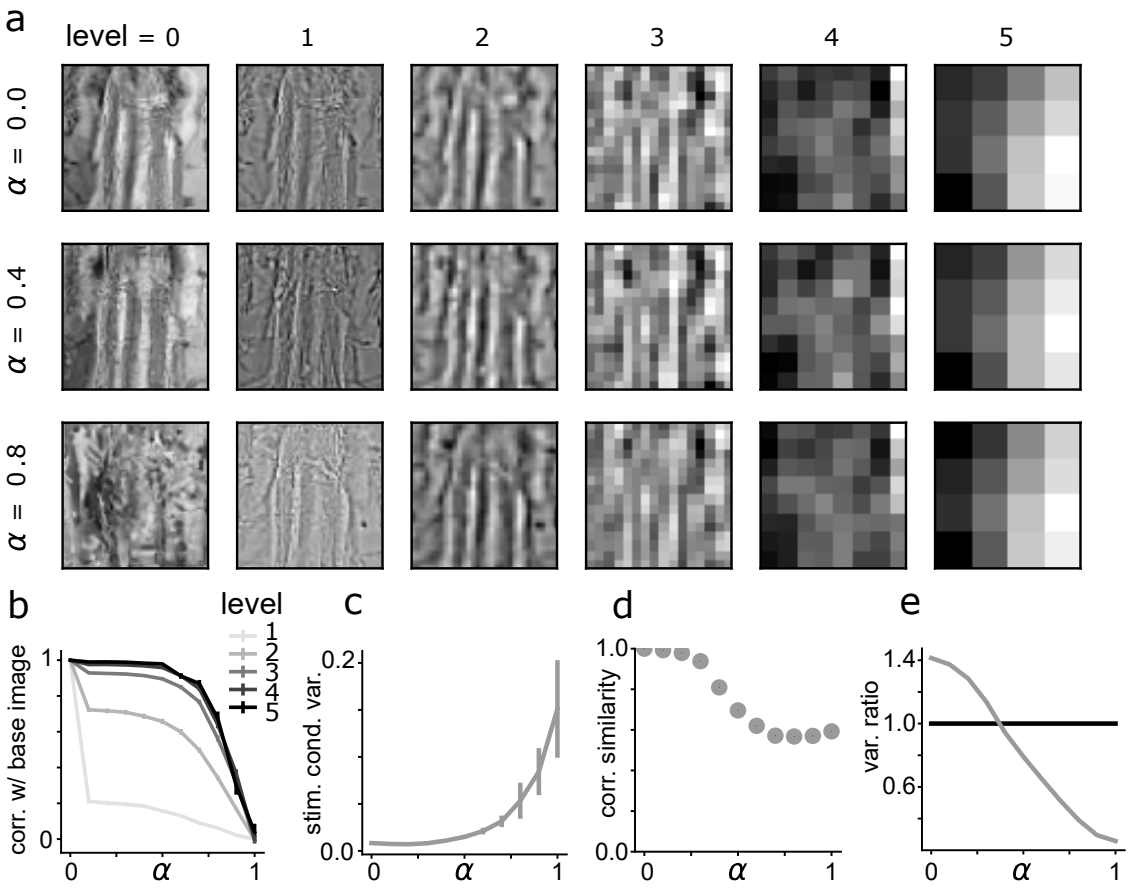

Figure S8: **Analyzing the image- and network-level effects of psychedelics in a Tiny ImageNet-pretrained VDVAE model**. **a)** Laplacian pyramid features for a grayscale example input image from the Tiny ImageNet dataset (top left). Pyramid levels increase along columns, corresponding to decreasing resolution, and hallucination intensity increases along rows. **b)** Correlation similarity across pyramid levels between the base image and a hallucinated image across different α values, averaged over 100 image samples. **c)** Stimulus-conditioned variance of units in layer 30 (descending from the top of the network) across different α values, averaged over 100 sample images and 32 distinct trials. **d)** Correlation similarity calculated between correlation matrices for units in layer 30 across different α values, averaged over spatial positions and 100 sample images. **e)** Ratio of across-stimulus variance in individual units of layer 30 when the highest 20 layers of the network have been inactivated, versus baseline conditions across different α values. Error bars indicate +/- 1 s.e.m.

Parameter	Function	Value
$\{N^{(l)}\}$	Layer Widths	$[32, 16, 6]$
σ_b	Inference standard dev.	0.01
σ_p	Generative standard dev.	0.01
N_d	Number of dendritic branches	16
L	Epoch number	50
T_w	Length of Wake phase	200
T_s	Length of Sleep phase	200
K	Batch number	512
η_p	Generative parameter learning rate	0.001
η_b	Inference parameter learning rate	0.001
$[\beta_1, \beta_2]$	Adam optimizer parameters	[0.9, 0.999]

Table S1: **MNIST multicompartment network hyperparameters**

Parameter	Function	Value
$\{N^{(l)}\}$	Layer Widths	$[128, 128, 96]$
σ_b	Inference standard dev.	0.01
σ_p	Generative standard dev.	0.01
N_d	Number of dendritic branches	16
L	Epoch number	50
T_w	Length of Wake phase	200
T_s	Length of Sleep phase	200
K	Batch number	256
η_p	Generative parameter learning rate	0.001
η_b	Inference parameter learning rate	0.001
$[\beta_1, \beta_2]$	Adam optimizer parameters	[0.9, 0.999]

Table S2: **CIFAR10 multicompartment network hyperparameters**

Parameter	Function	Value
$\{N^{(l)}\}$	Layer Widths	$[512, 256, 256, 128, 64, 32, 16]$
σ_b	Denoised inference standard dev.	0.01
$\bar{\sigma}_b$	Noisy inference standard dev.	0.01
σ_p	Denoised generative standard dev.	0.01
$\bar{\sigma}_p$	Noisy generative standard dev.	0.01
L	Epoch number	400
T_w	Length of Wake phase	200
T_s	Length of Sleep phase	50
K	Batch number	512
η_p	Generative parameter learning rate	0.001
η_b	Inference parameter learning rate	0.0001
$[\beta_1, \beta_2]$	Adam optimizer parameters	[0.9, 0.999]

Table S3: **Recurrent network hyperparameters**

Algorithm 1 Wake-Sleep Pseudocode

$e \leftarrow 0$
$\mathcal{L}_{prev} \leftarrow \infty$
while $e < L$ **do** ▷ Repeat for L data epochs
  for t in range$(0, T_w)$ **do** ▷ Wake Phase
    for k in range$(0, K)$ **do** ▷ Sample network states
      $\mathbf{r}^{(k)} \sim b(\mathbf{r}, \theta_b)$
      if epoch is complete **then** $e \leftarrow e + 1$
      end if
    end for
    $\mathcal{L}(\theta_p) \leftarrow \frac{1}{K} \sum_{k=0}^{K} -\log p(\mathbf{r}^{(k)}, \theta_p)$ ▷ Update generative parameters θ_p
    $\Delta\theta_p \leftarrow \eta_p \nabla_{\theta_p} \mathcal{L}(\theta_p)$
    $\theta_p \leftarrow \text{Adam}(\theta_p, \Delta\theta_p)$
  end for
  if $\mathcal{L}(\theta_p) < \mathcal{L}_{prev}$ **then**
    $\mathcal{L}_{prev} \leftarrow \mathcal{L}(\theta_p)$
    for t in range$(0, T_s)$ **do** ▷ Sleep Phase
      for k in range$(0, K)$ **do** ▷ Sample network states
        $\mathbf{r}^{(k)} \sim p(\mathbf{r}, \theta_p)$
      end for
      $\mathcal{L}(\theta_b) \leftarrow \frac{1}{K} \sum_{k=0}^{K} -\log b(\mathbf{r}^{(k)}, \theta_b)$ ▷ Update inference parameters θ_b
      $\Delta\theta_b \leftarrow \eta_b \nabla_{\theta_b} \mathcal{L}(\theta_b)$
      $\theta_b \leftarrow \text{Adam}(\theta_b, \Delta\theta_b)$
    end for
  end if
end while