

---

# THE NEURAL CODING FRAMEWORK FOR LEARNING GENERATIVE MODELS

---

**Alexander Ororbia\***

Department of Computer Science  
Rochester Institute of Technology  
Rochester, NY 14623  
ago@cs.rit.edu

**Daniel Kifer**

Department of Computer Science & Engineering  
The Pennsylvania State University  
State College, PA 16801  
duk17@psu.edu

## ABSTRACT

Neural generative models can be used to learn complex probability distributions from data, to sample from them, and to produce probability density estimates. We propose a computational framework for developing neural generative models inspired by the theory of predictive processing in the brain. According to predictive processing theory, the neurons in the brain form a hierarchy in which neurons in one level form expectations about sensory inputs from another level. These neurons update their local models based on differences between their expectations and the observed signals. In a similar way, artificial neurons in our generative models predict what neighboring neurons will do, and adjust their parameters based on how well the predictions matched reality. In this work, we show that the neural generative models learned within our framework perform well in practice across several benchmark datasets and metrics and either remain competitive with or significantly outperform other generative models with similar functionality (such as the variational auto-encoder).

**Keywords** Generative models · biologically plausible learning · predictive processing · credit assignment

## 1 Introduction

One way to understand how the brain adapts to its environment is to view it as a type of generative pattern-creation model [24], one that is engaged in a never-ending process of self-correction, often without external teaching signals (or labels) [62]. Under this perspective, the brain is continuously making predictions about elements of its environment, a process that allows it to infer useful representations of the sensory data that it receives [65] as well as to synthesize novel patterns, which could serve as the potential basis for long-term planning and imagination itself [13]. From the theoretical viewpoint of predictive processing [13], the brain could be likened to a hierarchical model [22] whose levels are implemented by neurons (or clusters of neurons). If levels are likened to regions of the brain, the neurons at one level (region) attempt to predict the state of neurons at another level (region) and adjust/correct their local model synaptic parameters based on how different their predictions were from the observed signal. Furthermore, these neurons utilize various mechanisms to laterally stimulate/suppress each other [47] to facilitate contextual processing (such as grouping/segmenting visual components of objects in a scene). As we will demonstrate in this article, this viewpoint can be turned into a powerful framework for learning generative models.

In machine learning, one central goal is to construct agents that learn distributed representations that extract the underlying structure of data, without the use of explicit supervisory signals such as human-crafted labels, i.e., unsupervised learning. Generative models, or models capable of synthesizing instances of data that resemble a database of collected patterns, that are based on deep artificial neural networks (ANNs), e.g., variational autoencoders [43] or generative adversarial networks [28], have been shown to be one way of acquiring these representations. Once trained, an ANN model is used to “fantasize” patterns by injecting it with noise and propagating this noise through the system until the output nodes are reached.

However, despite the success in deploying ANNs as generative models across a variety of applications, the way that ANNs operate and learn is a far-cry from the neuro-mechanistic story we described earlier [15, 82]. Specifically, ANN generative models are trained with the popular workhorse algorithm known as back-propagation of errors (backprop) [70], which is an elegant mathematical solution to the credit assignment problem in deep networks – synaptic weights are adjusted through the use of teaching signals that are created by propagating an error, which exists exclusively at the output of the ANN, backwards along a feedback pathway [63] (a path created by re-using the same weights that transmitted signals forward [30]). By virtue of this formulation, backprop imposes the constraint that the ANN take the form of a directed feedforward structure (and does not permit the use non-differentiable activation functions and makes integrating other mechanisms such as lateral connectivity difficult). While the neurons in an

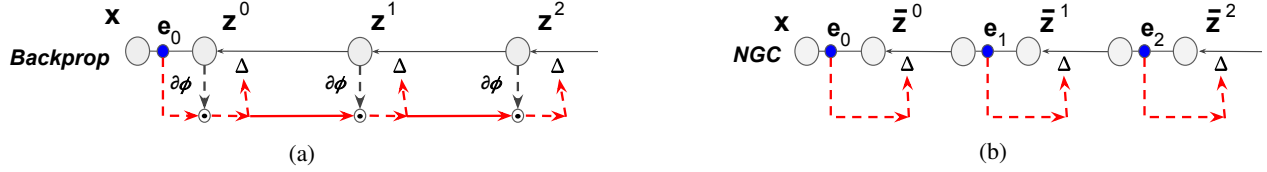


Figure 1: **Backprop in contrast with neural coding:** (a) Credit assignment in backprop requires a strict, global feedback pathway, which requires the completion of the forward pass that carries information upstream (right to left), that carries the error message  $e^0$  at the output layer back along (left to right) the same synapses used in the forward pass to update downstream neurons  $\bar{z}^1$  and  $\bar{z}^2$ . (b) Our proposed neural generative coding (NGC) model sidesteps this neurobiologically implausible requirement by learning with short, local error transmission pathways made possible through recurrent error synapses and stateful neural activities. Credit assignment under NGC operates with local mismatch signals,  $e^1$  and  $e^2$ , that readily communicate this information to their respective layers,  $\bar{z}^1$  and  $\bar{z}^2$ . Black arrows indicate forward propagation while red arrows indicate backwards transmission. Solid lines indicate that a signal is transformed along a synapse(s) while dashed arrows indicate direct copying of information (in other words, it represents the identity function, or that no transformation is applied to the incoming information).  $\partial\phi$  shows communication of the neuron’s first derivative,  $\Delta$  represents the computed change to the synapse (of the forward pass) that will use the (nearby) error signal, and  $\odot$  indicates multiplication of the incoming signals.

ANN are usually arranged hierarchically, they do not make local predictions and they do not laterally affect each other’s activity. Furthermore, synaptic adjustment in backprop-based models is done non-locally, while in neurobiological networks this adjustment is often argued to be done locally [32, 52, 10, 40] (there are far more local connections than long-range connections [83] with the neocortex adhering a local connectivity pattern [82]) – that is, neurons make use of the information immediately available to them (in both time and space) and do not wait on distant regions in order to adjust their synapses (with global information provided through neurotransmitters such as dopamine). In response to the above problems, the statistical learning community has developed a plethora of mechanisms that either modify backprop through a heuristic or additional constraint [66, 39, 54, 4] or, recently, worked on developing learning procedures that embody elements of biological neuronal function and computation while enabling backprop-level learning [37, 46, 5, 48, 73, 63] (see Supplementary Note 2 for a more detailed review). However, while insights from each development have proven valuable in bridging backprop with brain-like computation, many of these ideas only address one or a few of the issues described earlier and tend to focus on simple problems in classification. While the question as to how credit assignment is exactly implemented in the brain is an open one, it would prove useful to machine learning, (computational) neuroscience, and cognitive science to have a framework that demonstrates how a neural system can learn something as complex as a generative model without backprop, using mechanisms and rules that are brain-inspired.

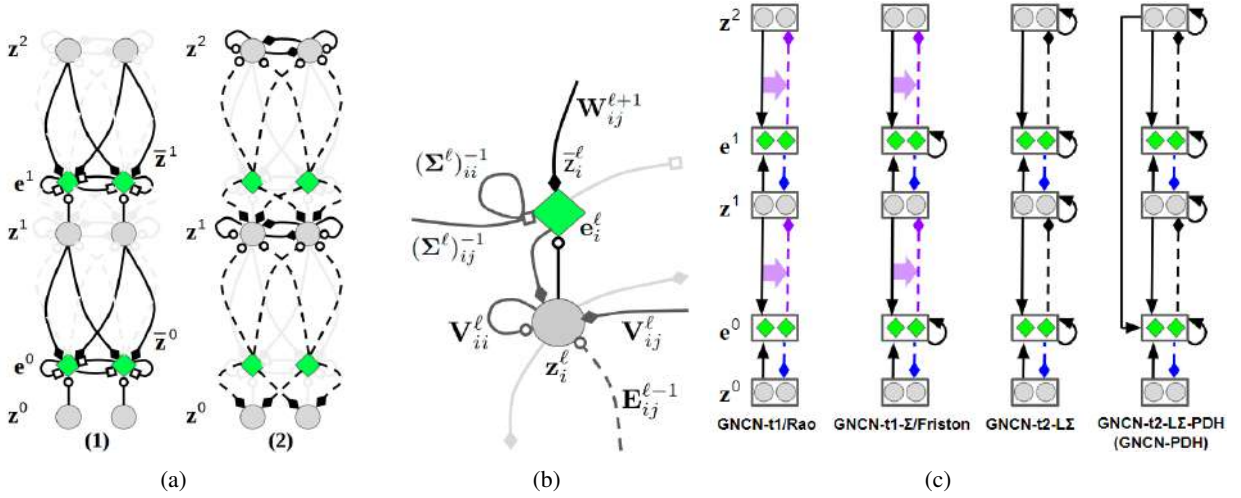
In this work, inspired by predictive processing theory [77, 8, 76, 13] and building on free energy minimization [22, 23], which crucially casts predictive processing formally as variational Bayesian inference, we propose the neural generative coding (NGC) computational framework as a powerful way to learn generative ANNs, resolving several of the key backprop-centric issues described above. Furthermore, we show that certain settings of NGC recover the work proposed in [68] and [22]. We find that NGC models, including [68] and [22], not only remain competitive with backprop-based generative ANNs across several datasets in terms of pattern creation, such as the variational autoencoder [43], but that they also outperform these models on tasks that they are not directly trained for, such as classification and pattern completion. Our results further demonstrate that besides unifying predictive processing models, NGC allows for integration of improvements such as learnable recurrent error synapses and laterally-driven sparsity. As a result, our work presents promising evidence that brain-inspired alterations to traditional deep learning techniques can be a viable source of performance gains.

## 2 Results

### 2.1 Generative Neural Coding Learns Viable Auto-Associative Generative Models

**Notation:** In this paper,  $\odot$  indicates a Hadamard product,  $\oslash$  indicates element-wise division, and  $\cdot$  indicates a matrix/vector multiplication (or dot product if the two objects it is applied to are vectors of the exact same shape) and  $(\mathbf{v})^T$  denotes the transpose. We denote  $\mathbf{v}_i$  (bold font indicates vector/matrix) means that we retrieve the  $i$ th element  $v_i$  (italic indicates single scalar element) in the vector  $\mathbf{v}$  and  $\mathbf{W}_{ij}$  means that we retrieve the element  $W_{ij}$  in the  $i$ th row and  $j$ th column of matrix  $\mathbf{W}$ .

**Problem Setting:** We start with a description of the problem setting – an agent must learn to approximate a probability distribution from a dataset  $\mathbf{X}$  of samples. For notational reasons, this dataset is presented in column-major order, so that each column  $\mathbf{x}$  represents a record (also known as an example or item).  $\mathbf{X}$  has  $D$  rows and  $S$  total columns. The items are assumed independent, so that  $p(\mathbf{X}) = \prod_{\mathbf{x} \in \mathbf{X}} p(\mathbf{x})$  and  $\log p(\mathbf{X}) = \sum_{\mathbf{x} \in \mathbf{X}} \log p(\mathbf{x})$ . We are interested in directed generative models that are capable of producing *explicit* density estimates of the data distribution, i.e., models that estimate a probability density function (PDF) over a sample space, and we will leave the examination of most *implicit* density estimators (i.e., models that do not produce explicit density estimates of the PDF but yield a function that produces samples from the estimated distribution), based on generative adversarial networks [28] for future work.



**Figure 2: Neural generative coding computation & circuitry:** (a) The two key computation steps taken by an entire NGC network (a GNCN-t2-LS) when processing an input ( $z^0 = x$ ): (1) prediction and laterally-weighted error computation, (2) error-correction of neural states. In this diagram, we depict a toy network with 3 layers of 2 state neurons (grey circles), i.e.,  $z^2 = ([z_0^2, z_1^2])^T$ ,  $z^1 = ([z_0^1, z_1^1])^T$ ,  $z^0 = ([z_0^0, z_1^0])^T$ , that are updated iteratively over  $T$  time steps. Two of these layers are linked to error neurons (green diamonds), i.e.,  $e^1 = ([e_0^1, e_1^1])^T$ ,  $e^0 = ([e_0^0, e_1^0])^T$ , which compute mismatch messages that are propagated throughout the system. The bottom layer  $z^0$  receives sensory input, i.e., an image. (b) The basic neural computational unit that an NGC model is composed of, consisting of a single state neuron  $z_i^\ell$  and an error neuron  $e_i^\ell$  at layer  $\ell$ . In the circuit, observe that a state neuron not only receives messages from error neurons (carried by  $E_{ij}^{\ell-1}$  synapses) but also a self-excitation signal and inhibition signals from laterally connected neurons (via  $V_{ij}^\ell$  synapses). The error neuron receives a gain signal (via  $(\Sigma^\ell)^{-1}$  synapses) from laterally connected error neurons. Note that filled diamonds indicate inhibitory signals, non-filled circles indicate excitatory signals, and empty squares indicate multiplicative signals. (c) The different GNCN architectures (under the NGC framework) experimented with in this study. Solid black arrow represents generative weights, dashed black arrow represents error weights, and solid pink arrow represents error weights that are function of a specific set of generative weights (horizontal pink arrow illustrates which generative weights map to which error weights).

**The Typical Deep Learning Approach:** In modern-day deep learning practice, a feedforward ANN, also called a *decoder*, would be constructed to model the desired input distribution. The decoder (NN) takes as input a noise vector or a sampled latent variable  $z$  and maps it to the parameters of a probability distribution, such as a mean and covariance of a multivariate Gaussian, or, as in this paper, the mean of a multivariate Bernoulli distribution, i.e.,  $z^0 = \text{NN}(z)$  where  $x_i \sim B(n=1, z_i^0)$ . This artificial neural network would typically be made up of  $L+1$  layers of neurons ( $L$  layers of hidden neurons and one output layer of neurons), where the state in layer  $\ell$  is represented by a vector  $z^\ell$ . Each layer  $\ell$  is interpreted as a transformation of the layer before it. In essence,  $z^{\ell-1} = \phi^{\ell-1}(\mathbf{W}^\ell \cdot z^\ell)$  where  $z^L$  of layer  $L$  to be the same as the input noise vector  $z$ . The output  $z^0$  of this decoder (Figure 1 a) would be the parameters of a probability distribution, such as the mean of a Bernoulli distribution, or mean and covariance of a Gaussian (see Supplementary Note 7 for descriptions of the backprop-based networks used in this study). One can sample from this distribution to get a sample point  $x$  (or use the mean of the distribution directly). To stabilize and speed up the model’s learning process, an encoder is typically introduced which also takes in as input the sensory input  $x$  to be predicted. The encoder is designed to drive the parameters of a distribution, normally multivariate Gaussian, that shape and control the form of the latent variable  $z$ , i.e.,  $\mu_z, \sigma_z^2 = \text{NN}_e(x)$  where  $\sigma_z^2 = \Sigma_z \odot \mathbf{I}$  (or diagonal covariance).

To fit this model to the data, one would choose the weight parameters  $\mathbf{W}^\ell$  to minimize a loss function  $\psi$  such as the negative log-likelihood, typically using some variant of stochastic gradient descent. Often the backprop algorithm is used to compute the partial derivatives of  $\frac{\partial \psi}{\partial \mathbf{W}^\ell}$  needed for this optimization. Computing the necessary derivatives according to backprop entails first computing an error signal at the output (downstream) layer, or  $e^0 = \frac{\partial \psi}{\partial z^0}$ . This error signal is then transmitted to internal (upstream) neurons by carrying this signal back along the forward synapses that were originally used to transform  $z^L$  (in short, this is done by multiplying the signal with the transpose of the forward weight matrices). Furthermore, knowledge of the derivative of each activation function  $\phi^\ell$  is required during these computations (as shown in Figure 1, a).

**Backprop-Learning versus Brain-Like Learning:** While the backprop algorithm described above has proven to be popular and effective in training ANNs (including generative models [45]), it has certain mechanisms that differ from the current understanding of brain-like learning. For example, in backprop:

1. Synapses that make up the forward information pathway need to directly be used in reverse to communicate teaching signals (the weight transport problem [30]),
2. Neurons need to be able to know about and communicate their own activation function’s first derivative,
3. Neurons must wait for the neurons ahead of them to percolate their error signals way back so they know when and how to adjust their own synapses (the update-locking problem [41]),

4. There is a distinct form of information propagation for error feedback, one that starts from the system's output and works its way back to the input layer (see Figure 1 a), which does not influence neural activity (the global feedback pathway problem [63]), i.e., backprop creates signals that only affect weights but do not (at least directly) affect/improve the network's representations of the environment,
5. The error signals have a one-to-one correspondence with neurons.

The goal of this paper is to present a modeling and learning framework that uses fewer mechanisms that are incompatible with current understanding of brain-like learning. Specifically, we aim to address the first 4 items.

**The Neural Generative Coding Approach:** In contrast to the backprop-based way of designing and training ANN generative models, our proposed framework, neural generative coding (NGC), provides one way to emulate the several neuro-biological principles and properties described above by proposing a family of models and their corresponding training procedure. In this framework, any single model is referred to as a generative neural coding network (GNCN, see Supplementary Note 1 for naming convention details). A GNCN model has  $L + 1$  layers of neurons (also called state variables)  $\mathfrak{N}^0, \mathfrak{N}^1, \dots, \mathfrak{N}^L$ , where  $\mathfrak{N}^0$  is the output layer. Each layer  $\mathfrak{N}^\ell$  has  $J^\ell$  neurons and each neuron has a latent state value represented by a single number. The combined latent state of all neurons in layer  $\mathfrak{N}^\ell$  is represented by the vector  $\mathbf{z}^\ell \in \mathcal{R}^{J^\ell \times 1}$  (initially  $\mathbf{z}^\ell = \mathbf{0}$  when a new data point is encountered), with  $\mathbf{z}^0$  being the same as the data  $\mathbf{x}$  (it is typically clamped to the input, i.e.,  $\mathbf{z}^0 = \mathbf{x}$ ). The network is interpreted as a specification of the probability  $P(\mathbf{z}^0 = \mathbf{x}, \mathbf{z}^1, \dots, \mathbf{z}^L) = P(\mathbf{z}^0 | \mathbf{z}^1) \dots P(\mathbf{z}^{L-1} | \mathbf{z}^L) P(\mathbf{z}^L)$ , similar to a Bayesian network, and we use the notation  $\mathbf{Z} = \{\mathbf{z}^1, \dots, \mathbf{z}^L\}$  to refer to the state of all of the *intermediate* neurons (i.e., excluding the output). Thus, layer  $\mathfrak{N}^\ell$  represents the conditional probability  $P(\mathbf{z}^\ell | \mathbf{z}^{\ell+1})$ , with the last layer  $\mathfrak{N}^L$  representing  $P(\mathbf{z}^L)$ . For image data, the distribution  $P(\mathbf{z}^0 | \mathbf{z}^1)$  associated with the output layer is multivariate Bernoulli with mean vector  $\bar{\mathbf{z}}^0$  (which depends on  $\mathbf{z}^1$ ). All of the other distributions  $P(\mathbf{z}^\ell | \mathbf{z}^{\ell+1})$  are multivariate Gaussians with mean vector  $\bar{\mathbf{z}}^\ell$  (which depends on  $\mathbf{z}^{\ell+1}$ ) and covariance matrix  $\Sigma^\ell$ . The mean vector  $\bar{\mathbf{z}}^\ell$  for layer  $\mathfrak{N}^\ell$  is obtained in a feed-forward manner from the latent state of the neighboring layer (biases/offset terms have been omitted for clarity). Specifically, we model this generative process with the following equation:

$$\bar{\mathbf{z}}^\ell \leftarrow g^\ell \left( \overbrace{\mathbf{W}^{\ell+1} \cdot \phi^{\ell+1}(\mathbf{z}^{\ell+1})}^{\text{local, top-down prediction}} + \alpha_m \overbrace{(\mathbf{M}^{\ell+2} \cdot \phi^{\ell+2}(\mathbf{z}^{\ell+2}))}^{\text{local, auxiliary prediction}} \right), \quad (1)$$

where  $\mathbf{W}^\ell$  is a forward/generative weight matrix,  $\alpha_m = 0$ ,  $g^\ell$  and  $\phi^{\ell+1}$  are activation functions, and  $\cdot$  indicates matrix multiplication. Thus each layer  $\mathfrak{N}^\ell$  is specified by two functions  $g^\ell$  and  $\phi^\ell$ , a trainable weight matrix  $\mathbf{W}^\ell$  and a covariance matrix  $\Sigma^\ell$  (the last layer  $\mathfrak{N}^L$  is specified just by the activation function  $\phi^L$ ). Note in Equation 1, the mean vector  $\bar{\mathbf{z}}^\ell$  depends on the sampled realization  $\mathbf{z}^{\ell+1}$  from the previous layer, making this a hierarchical, Gaussian latent variable model. Notice that, optionally, if the binary coefficient  $\alpha_m = 1$ , each layer also incorporates a learnable auxiliary generative matrix  $\mathbf{M}^{\ell+2}$ , which conveys and injects state value information from the layer  $\mathfrak{N}^{\ell+2}$  into the prediction of layer  $\mathfrak{N}^\ell$  through a linear combination. We append the suffix “-PDH” (partially decomposable hierarchy) to a GNCN model name when  $\alpha_m = 1$  (see Figure 2c for a visual depiction).

The goal of any GNCN model is to learn a joint distribution over its  $L + 1$  neural states, i.e.,  $p(\mathbf{z}^0, \mathbf{z}^1, \dots, \mathbf{z}^L)$ , from which the marginal distribution of the data may be obtained via  $p(\mathbf{x}) = \int_{\mathbf{Z}} p(\mathbf{x}, \mathbf{z}^1, \dots, \mathbf{z}^L) d\mathbf{Z} = \int_{\mathbf{Z}} p(\mathbf{z}^0, \mathbf{z}^1, \dots, \mathbf{z}^L) d\mathbf{Z}$ . Given that minimizing  $-\log p(\mathbf{x})$  directly is intractable in general, our approach for training is to approximately minimize the log-likelihood based on the ideas behind the Expectation-Maximization (EM) algorithm. Specifically, we work with the analogue of the *complete-data* likelihood, which adds in the latent variables of the network (it does not marginalize over them) and sets up a 2 step process that adjusts the latent variables (like an E-step) and then updates the network parameters (M step).

**Training the Model:** The complete data log-likelihood  $\psi$  (also referred to as total discrepancy [62]) of the observed data  $\mathbf{x}$  and the latent variables  $\mathbf{z}^1, \dots, \mathbf{z}^L$  is defined formally as follows:

$$\psi = \sum_j \left( \mathbf{x}_j \log \bar{z}_j^0 + (1 - \mathbf{x}_j) \log(1 - \bar{z}_j^0) \right) + \sum_{\ell=1}^L \left( -\frac{1}{2} \log |(\Sigma^\ell)^{-1}| - \frac{1}{2} (\mathbf{z}^\ell - \bar{\mathbf{z}}^\ell)^T \cdot (\Sigma^\ell)^{-1} \cdot (\mathbf{z}^\ell - \bar{\mathbf{z}}^\ell) \right). \quad (2)$$

Importantly, the above complete data log likelihood connects our NGC models (and total discrepancy) to the principle of free energy [24, 22], given that the sum of (weighted) prediction errors defined in Equation 2 can be shown to be a form of variational free energy that is minimized through variational inference. Explicitly characterizing a neural system engaged with optimizing the above objective as a generative model, as we do in this paper, grounds NGC as performing a form of approximate Bayesian inference (much as [22] did for the early predictive coding model of [68]) and connects it to perception as (unconscious) inference [33] and the larger theoretical framework of the Bayesian brain [17, 44].

Since all of the latent variables are continuous, the updates below follow the form of the exact gradient, i.e., backprop (allowing for gradient descent), to optimize the latent variables and the parameters. The log likelihood has the following partial derivatives:

$$\frac{\partial \psi}{\partial (\Sigma^\ell)^{-1}} = \frac{1}{2} \Sigma^\ell - \frac{1}{2} (\mathbf{z}^\ell - \bar{\mathbf{z}}^\ell) \cdot (\mathbf{z}^\ell - \bar{\mathbf{z}}^\ell)^T \quad (3)$$

$$\frac{\partial \psi}{\partial \Sigma^\ell} = -\frac{1}{2} \Sigma^{-1} + \frac{1}{2} \Sigma^{-1} \cdot (\mathbf{z}^\ell - \bar{\mathbf{z}}^\ell) \cdot (\mathbf{z}^\ell - \bar{\mathbf{z}}^\ell)^T \cdot \Sigma^{-1} \quad (4)$$

$$\frac{\partial \psi}{\partial \mathbf{W}^0} = \left( \frac{\partial g^0}{\partial \mathbf{h}^0} \odot (\mathbf{x} \odot \bar{\mathbf{z}}^0 - (\mathbf{1} - \mathbf{x}) \odot (\mathbf{1} - \bar{\mathbf{z}}^0)) \right) \cdot (\phi^1(\mathbf{z}^1))^T \quad (5)$$

where  $\odot$  is element-wise division,  $\odot$  is element-wise product, and  $\mathbf{h}^\ell = \mathbf{W}^\ell \cdot \phi^{\ell+1}(\mathbf{z}^{\ell+1})$

$$\frac{\partial \psi}{\partial \mathbf{W}^\ell} = \frac{\partial \psi}{\partial \mathbf{h}^\ell} \cdot (\phi^{\ell+1}(\mathbf{z}^{\ell+1}))^T \quad (6)$$

$$\frac{\partial \psi}{\partial \mathbf{z}^1} = \frac{\partial}{\partial \mathbf{z}^1} \left( \sum_j \left( \mathbf{x}_j \log \bar{\mathbf{z}}_j^0 + (1 - \mathbf{x}_j) \log(1 - \bar{\mathbf{z}}_j^0) \right) \right) - (\Sigma^\ell)^{-1} \cdot (\mathbf{z}^1 - \bar{\mathbf{z}}^1) \quad (7)$$

$$\frac{\partial \psi}{\partial \mathbf{z}^\ell} = \left( \frac{\partial \bar{\mathbf{z}}^{\ell-1}}{\partial \mathbf{z}^\ell} \cdot \left( (\Sigma^{\ell-1})^{-1} \cdot (\mathbf{z}^{\ell-1} - \bar{\mathbf{z}}^{\ell-1}) \right) \right) - (\Sigma^\ell)^{-1} \cdot (\mathbf{z}^\ell - \bar{\mathbf{z}}^\ell) \quad (8)$$

In this work, we incorporate two key concepts from local representation alignment (LRA) [64, 63]: 1) the use of error synapses to directly resolve the weight-transport problem, and 2) the omission of derivatives of activation functions which yield synapse rules that function like error-Hebbian updates. To incorporate these modifications, we introduce a vector of error neurons  $\mathbf{e}^\ell$  (depicted in Figure 1 b) that are tasked with computing how far off the mean vector is from the relevant nearby state. Note that the error neurons themselves are derived from the likelihood function:

$$\mathbf{e}^0 = \frac{\partial \psi}{\partial \bar{\mathbf{z}}^0} = (\mathbf{x} \odot \bar{\mathbf{z}}^0 - (\mathbf{1} - \mathbf{x}) \odot (\mathbf{1} - \bar{\mathbf{z}}^0)) \quad (9)$$

$$\mathbf{e}^\ell = \frac{\partial \psi}{\partial \bar{\mathbf{z}}^\ell} = \overbrace{(\Sigma^{\ell-1})^{-1}}^{\text{lateral modulation}} \cdot \overbrace{(\mathbf{z}^{\ell-1} - \bar{\mathbf{z}}^{\ell-1})}^{\text{mismatch signal}} \quad (10)$$

and are implemented as separate sets of activities (like the state neurons). In Figure 2 (b), we depict the circuit for a single pair neurons at a layer  $\mathfrak{N}^\ell$ , i.e., a state neuron  $\mathbf{z}_i^\ell$  and an error neuron  $\mathbf{e}_i^\ell$ . Notice that, in Equation 10 above, in order to compute the state of the error neuron  $\mathbf{e}_i^\ell$ , the covariance matrix  $\Sigma^\ell$  acts as a lateral modulation matrix, which is inspired by the neuro-mechanistic concept of precision weighting in predictive processing theory [55]. It allows error neuron  $\mathbf{e}_i^\ell$  to dynamically amplify/reduce the learning signal (i.e.,  $\mathbf{z}_i^\ell - \bar{\mathbf{z}}_i^\ell$ ) of its corresponding state neuron  $\mathbf{z}_i^\ell$ , based on the learning signals of the other state neurons. Empirically, we found these modulatory synapses to improve the crispness of model samples. Note that Equation 10 would be applied to all  $J_\ell$  error neurons for each layer  $\ell = 1, \dots, L-1$ .

**Updating the States and Synapses:** To carry the activities/messages of the error neurons in order to calculate the update to  $\mathbf{z}^\ell$ , we replace the term  $\frac{\partial \mathbf{z}^{\ell-1}}{\partial \mathbf{z}^\ell}$  with a learnable error matrix  $\mathbf{E}^\ell$ . This substitution allows us to rewrite Equations 7 and 8 as follows:

$$\frac{\partial \psi}{\partial \mathbf{z}^1} \approx \Delta \mathbf{z}^1 = \mathbf{E}^1 \cdot \left( \frac{g^0}{\mathbf{h}^0} \odot \mathbf{e}^0 \right) - \mathbf{e}^1 = \mathbf{E}^1 \cdot (\mathbf{z}^1 - \bar{\mathbf{z}}^1) - \mathbf{e}^1 \quad (11)$$

$$\frac{\partial \psi}{\partial \mathbf{z}^\ell} \approx \Delta \mathbf{z}^\ell = \mathbf{E}^\ell \cdot \mathbf{e}^{\ell-1} - \mathbf{e}^\ell. \quad (12)$$

Once the update for any layer  $\mathfrak{N}^\ell$  has been calculated, the corresponding state neurons will proceed to update their state values. By analogy with latent variable models, where the state neurons  $\mathbf{z}^\ell$  correspond to latent variables, this act can be viewed as an attempt to modify the states in a way that improves the complete data log-likelihood (i.e. modifying the  $\mathbf{z}^\ell$  to cause  $p(\mathbf{x}, \mathbf{z}^1, \dots, \mathbf{z}^L)$  to increase). One possible neuroscience-inspired way to perform this update is shown in Equation 13 (here  $\beta$  is a hyperparameter akin to the machine learning concept of a learning rate). Specifically, this update is:

$$\mathbf{z}_i^\ell \leftarrow \mathbf{z}_i^\ell + \beta \left( -\gamma \mathbf{z}_i^\ell + \overbrace{\sum_{j \in J_{\ell-1}} (\mathbf{E}_{ij}^\ell \mathbf{e}_j^{\ell-1})}^{\Delta \mathbf{z}_i} \right). \quad (13)$$

Here  $\mathbf{z}_i^\ell$  is modified through three terms. The first is a decaying pressure caused by the leak term  $-\gamma \mathbf{z}_i^\ell$ , controlled by the strength factor  $\gamma$ . The second term,  $-\mathbf{e}_i^\ell$ , can be interpreted as top-down pressure where  $\mathbf{e}_i^\ell$  is a measure of how much the neuron's state

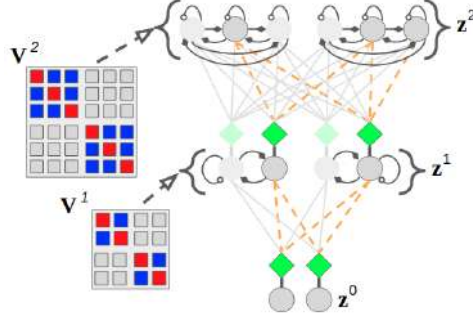


Figure 3: **Lateral dynamics of a generative neural coding network:** Depending on how the lateral connectivity matrices,  $\mathbf{V}^1$  &  $\mathbf{V}^2$ , are designed, different competition patterns emerge among neurons in network. In the system with lateral matrices as shown above, each neuron is driven by its own self-excitation (red-colored blocks) and laterally inhibits (blue-colored blocks) other neurons that inhabit the same group (of 2 or 3 units). In the figure, one possible outcome (or state of the agent) of such a competition is shown (after  $T$  steps). Self-excitation and lateral inhibition strengths are controlled by coefficients set a priori. Dashed orange edges show actively used synapses, filled diamonds indicate inhibitory signals, and non-filled circles indicate excitatory signals.

differs from the predicted state  $\mathbf{z}^\ell$  that is computed by the layer above. Finally, the third term adds in the error message from each error neuron  $\mathbf{e}_j^{\ell-1}$  in the layer below, communicated by special error synapses  $\mathbf{E}^{\ell-1}$ , acting as a form of bottom-up pressure. It is important to mention that, in this paper, we investigated another particular form of Equation 13 as follows:

$$\mathbf{z}_i^\ell \leftarrow \mathbf{z}_i^\ell + \beta \left( -\gamma \mathbf{z}_i^\ell + \frac{\partial \phi(\mathbf{z}_i^\ell)}{\partial \mathbf{z}_i^\ell} \left[ \sum_{j \in J_{\ell-1}} (\mathbf{U}_{ij}^\ell \mathbf{e}_j^{\ell-1}) - \mathbf{e}_i^\ell \right] - \lambda \text{sign}(\mathbf{z}_i) \right). \quad (14)$$

where the last term  $-\lambda \text{sign}(\mathbf{z}_i)$  is a kurtotic prior that can be imposed to encourage most activity values to be closer to zero in a given neural state  $\mathbf{z}^\ell$ . Under the NGC framework, models that use separate, learnable error synapses  $\mathbf{E}^\ell$  will be referred to as “Type 2” (t2) and those that use (non-learnable) error synapses that are a function of the forward weights  $\mathbf{W}^\ell$  will be labeled as “Type 1” (see Figure 2c for visual depictions of these models and Supplementary Note 1 for details on the NGC framework model naming convention). We can then manipulate certain variable values to recover different classical predictive coding models: 1) if  $\gamma = 0$ ,  $\mathbf{U} = (\mathbf{W})^T$ ,  $\Sigma^\ell = \sigma_z^2 \mathbf{I}$ , and  $\phi^\ell(v) = \tanh(v)$ , then we recover the model proposed in [68] – we will refer to this as GNCN-t1/Rao (using  $\gamma > 0$  yields the state equation of [80]), and 2) if  $\gamma = 0$ ,  $\mathbf{U} = -(\mathbf{W})^T$ , and  $\phi^\ell(v) = \max(0, v)$ , then we recover the neural implementation of the graphical model proposed in [22] – we will refer to this as GNCN-t1- $\Sigma$ /Friston (see Figure 2c).

However, instead of directly using this update rule, we may further incorporate another property of the brain – activation sparsity (through lateral inhibition/excitation). Sparsity in real neuronal and artificial systems is often argued to be useful for learning compact representations (most activity values will be at/near zero), allowing for efficient storage and vastly improved energy efficiency. To emulate this type of sparsity, we integrate an explicit mechanism to force neurons to compete for activation (in contrast to the kurtotic prior used in Equation 14), where we take inspiration from the known occurrence of lateral synapses in cortical regions of the brain (which are thought to facilitate contextual processing [75]). To do this, we introduce two terms to Equation 13 that use excitatory/inhibitory synapses stored in a matrix  $\mathbf{V}^\ell$ . This means that state neurons are updated as follows:

$$\mathbf{z}_i^\ell \leftarrow \mathbf{z}_i^\ell + \beta \left( \underbrace{-\gamma \mathbf{z}_i^\ell}_{\text{leak}} + \underbrace{\left( \sum_{j \in J_{\ell-1}} \mathbf{E}_{ij}^\ell \mathbf{e}_j^{\ell-1} \right) - \mathbf{e}_i^\ell}_{\text{bottom-up + top-down pressures}} - \underbrace{\left( \sum_{j \in J_\ell, j \neq i} \mathbf{V}_{ij}^\ell \phi^\ell(\mathbf{z}_j^\ell) \right)}_{\text{lateral inhibition}} + \underbrace{\mathbf{V}_{ii}^\ell \phi^\ell(\mathbf{z}_i^\ell)}_{\text{self-excitation}} \right) \quad (15)$$

Depending on the values set in  $\mathbf{V}^\ell$ , different types of sparsity patterns emerge, creating a flexible means for testing the benefits/drawbacks of different kinds of lateral competition patterns in an interpretable manner. Figure 3 provides a graphical example of the type of interaction pattern we found worked well for the GNCN in this study, i.e., we forced  $J_\ell/K$  groups (or columns) of  $K$  neurons to compete with each other (see Supplementary Note 6). The model employing Equation 15 and  $\alpha_m = 0$  (in its Equation 1) will be referred to as GNCN-t2-L $\Sigma$  and when  $\alpha_m = 1$  (in its Equation 1) the model will be referred to as the GNCN-PDH.

The synaptic matrix updates in Equations 5 and 6 can be written in terms of the error neurons:

$$\frac{\partial \psi}{\partial \mathbf{W}^0} \propto \Delta \mathbf{W}^0 = \mathbf{e}^0 \cdot (\phi^1(\mathbf{z}^1))^T \quad (16)$$

$$\frac{\partial \psi}{\partial \mathbf{W}^\ell} \propto \Delta \mathbf{W}^\ell = \mathbf{e}^\ell \cdot (\phi^{\ell+1}(\mathbf{z}^{\ell+1}))^T \quad (17)$$

and, following in line with LRA, the error synapses can be updated as:

$$\Delta \mathbf{E}^\ell = \lambda \left( \phi^{\ell+1}(\mathbf{z}^{\ell+1}) \cdot (\mathbf{e}^\ell)^T \right) \quad (18)$$

Table 1: **Reconstruction & Likelihood Measurements:** Generative modeling results across datasets. The binary cross entropy (BCE) of the model reconstructions and the marginal log likelihood  $\log p(\mathbf{x})$  (in nats) on the test set are reported (lower BCE is better and  $\log p(\mathbf{x})$  closer to 0.0 is better). We **bold** the best two scores of the models with respect to BCE (first column) as well as with respect to  $\log p(\mathbf{x})$  (second column). (Metrics averaged over 10 trials - we report their mean and standard deviation.)

Model		BCE	$\log p(\mathbf{x})$		BCE	$\log p(\mathbf{x})$
GMM		–	$-185.37 \pm 0.47$		–	$-426.73 \pm 1.03$
RAE		$60.946 \pm 0.58$	$-117.009 \pm 0.181$		$175.597 \pm 0.506$	$-234.411 \pm 1.173$
GVAE-CV		$69.904 \pm 0.304$	$-106.413 \pm 0.057$		$180.812 \pm 1.054$	$-229.124 \pm 0.946$
GVAE		$75.521 \pm 2.194$	$-100.097 \pm 2.257$		$199.374 \pm 0.913$	<b><math>-218.67 \pm 0.43</math></b>
GAN-AE		$73.54 \pm 3.492$	<b><math>-94.418 \pm 0.069</math></b>		$199.677 \pm 3.014$	<b><math>-216.763 \pm 1.156</math></b>
GNCN-t1/Rao		$66.466 \pm 0.065$	$-102.222 \pm 0.016$		$136.492 \pm 0.149$	$-226.371 \pm 0.187$
GNCN-t1- $\Sigma$ /Friston		<b><math>47.246 \pm 0.191</math></b>	$-101.921 \pm 0.242$		<b><math>114.629 \pm 0.353</math></b>	$-229.037 \pm 0.517$
GNCN-t2-L $\Sigma$		$59.658 \pm 0.021$	$-98.271 \pm 0.247$		$135.864 \pm 0.165$	$-220.41 \pm 0.223$
GNCN-PDH		<b><math>44.857 \pm 0.054</math></b>	<b><math>-97.255 \pm 0.084</math></b>		<b><math>128.389 \pm 0.23</math></b>	$-219.82 \pm 0.641$
GMM		–	$-302.96 \pm 0.47$		–	$-88.63 \pm 0.43$
RAE		$102.723 \pm 0.777$	$-138.762 \pm 1.427$		$40.408 \pm 0.755$	$-50.47 \pm 2.073$
GVAE-CV		$111.719 \pm 0.134$	$-131.816 \pm 0.295$		$49.579 \pm 4.588$	$-41.339 \pm 0.84$
GVAE		$121.618 \pm 0.453$	<b><math>-127.644 \pm 0.041</math></b>		$38.349 \pm 0.74$	$-44.067 \pm 0.24$
GAN-AE		$132.943 \pm 4.372$	$-130.872 \pm 1.795$		$40.396 \pm 0.743$	<b><math>-40.684 \pm 0.131</math></b>
GNCN-t1/Rao		$95.803 \pm 0.025$	$-136.632 \pm 0.465$		$34.595 \pm 0.062$	$-44.621 \pm 0.128$
GNCN-t1- $\Sigma$ /Friston		<b><math>81.734 \pm 0.421</math></b>	$-133.506 \pm 0.267$		<b><math>31.169 \pm 0.023</math></b>	$-42.526 \pm 0.059$
GNCN-t2-L $\Sigma$		$97.802 \pm 0.714$	$-132.041 \pm 0.312$		$33.287 \pm 0.011$	$-41.523 \pm 0.058$
GNCN-PDH		<b><math>90.058 \pm 0.375</math></b>	<b><math>-130.031 \pm 0.162</math></b>		<b><math>26.764 \pm 0.02</math></b>	<b><math>-40.821 \pm 0.195</math></b>

where  $\lambda \in [0, 1]$  controls the strength of the error synaptic adjustment (usually set to values  $< 1$ ). Note that no partial derivatives of the activation function  $\phi^\ell$  is required (in Type 2 GNCN models) – we are using an error-Hebbian style update rule, which has been shown to be effective empirically and crucially removes the need for state neurons to be aware of their own point-wise activation derivative. Note that this means  $\phi^\ell$  could easily be a discrete function in this case, such as the signum or Heaviside.

Putting all of the components above together, given a data point(s)  $\mathbf{x}$ , we can characterize the neural dynamics of the NGC system (graphically depicted in Figure 2 a) and train any GNCN following an online alternating maximization approach as follows:

**// Initialization:**

1. Set  $\mathbf{z}^0 = \mathbf{x}$  (clamp data to output neurons), and set  $\mathbf{z}^\ell = \mathbf{0}$ ,  $\forall \ell \geq 1$
2. Compute mean estimates  $\bar{\mathbf{z}}^\ell$ ,  $\forall \ell \geq 0$  (Equation 1) & use Equation 9 and Equation 10 to initialize  $\mathbf{e}^\ell$ ,  $\forall \ell \geq 0$

**// Latent Update Step:** Search for most probable value of  $\mathbf{z}^\ell$ ,  $\forall \ell$

3. Use Equation 11 to get  $\Delta \mathbf{z}^1$  & Equation 12 to get  $\Delta \mathbf{z}^\ell$ ,  $\forall \ell > 1$ . States are modified via (vectorized form of Equation 15):

$$\begin{aligned}\mathbf{z}^1 &= \mathbf{z}^1 + \beta \left( \Delta \mathbf{z}^1 - \mathbf{V}^1 \cdot \mathbf{z}^1 - \gamma \mathbf{z}^1 \right) \\ \mathbf{z}^\ell &= \mathbf{z}^\ell + \beta \left( \Delta \mathbf{z}^\ell - \mathbf{V}^\ell \cdot \mathbf{z}^\ell - \gamma \mathbf{z}^\ell \right), \forall \ell > 1\end{aligned}$$

4. Run Equation 9 and 10 to obtain updated error neurons, i.e.,  $\mathbf{e}^\ell$ ,  $\forall \ell \geq 0$
5. Repeat Steps 3 and 4 for  $T$  iterations

**// Parameter Update Step:** Update parameters given estimated latent states

6. Update synaptic matrices using Equations 16, 17, 3, and 18 (and normalize current matrices).

At test time, to reconstruct  $\mathbf{x}$ , one should only use Steps 1-5 of the recipe above.  $\gamma$  controls the strength of the decay/leak applied to  $\mathbf{z}^\ell$  and corresponds to placing an additional  $\mathcal{N}(\mu = 0, \Sigma = \lambda \mathbf{I})$  prior over the latent states. In essence, the above complete process depicts that the GNCN adjusts its synaptic weight matrices once activity values for all  $\mathbf{z}^\ell$  and  $\mathbf{e}^\ell$  have been found after a  $T$ -step episode. The synaptic matrix updates are simply the products between relevant state activities and error neurons. Finally, after each weight update has been made, a GNCN's weight matrices are normalized such that the Euclidean norms of their columns are 1.0 (this step, which requires non-local information to perform, could possibly be induced biologically through the use of external neuromodulatory signals). In essence, the steps described above and shown in Figure 2 (a) illustrate that the NGC framework embodies the idea that neural state and synaptic weight adjustment are the result of a process of generate-then-correct, or *continual error correction*, in response to samples of the agent's environment.

**Generative Modeling Results:** The model framework that we have described so far would be immediately useful for creating an auto-associative memory of sensory input, i.e., upon receiving a particular sensory input, the model would be able to recall seeing it by accurately reconstructing it. Furthermore, since the model is learning an estimator of the input distribution  $p(\mathbf{x})$ , we may sample from it (see Supplementary Methods) to synthesize or “hallucinate” data patterns, as we will demonstrate later in this section.



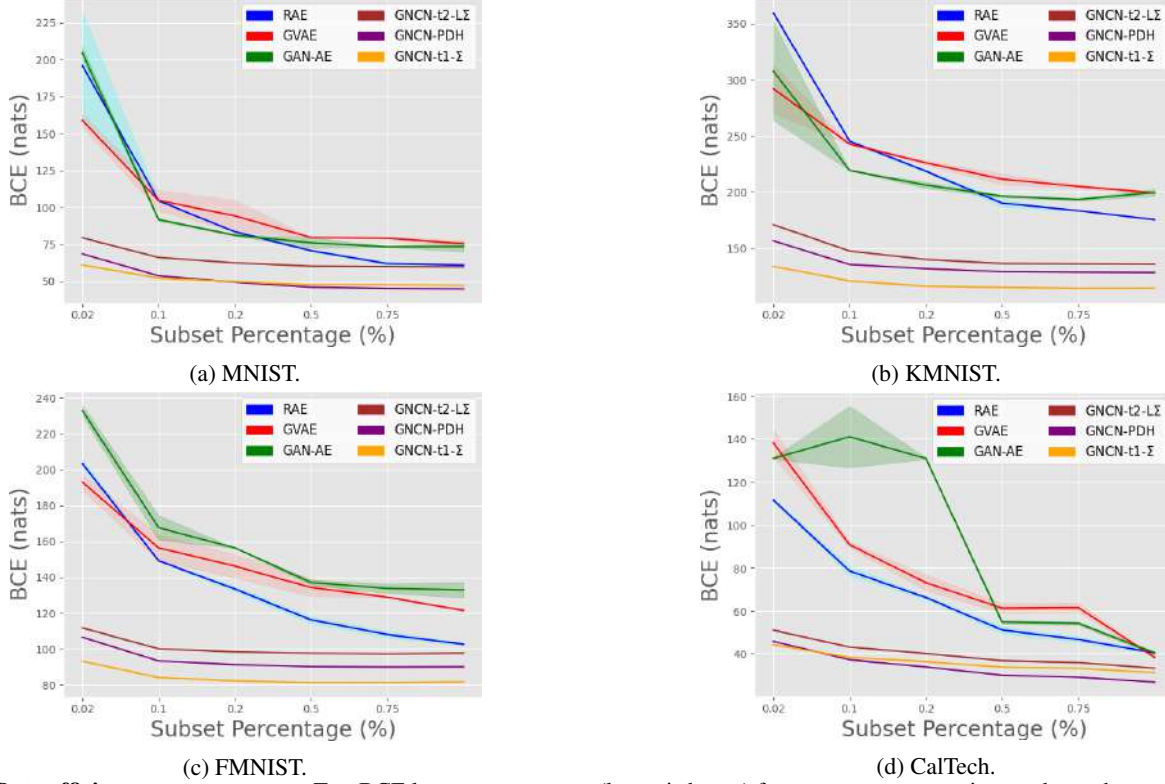


Figure 4: **Data efficiency measurements:** Test BCE loss measurements (lower is better) for pattern reconstruction on data subsets of increasing size (averaged over 10 trials). Curves depict the mean (solid line) and the standard deviation (lighter-colored envelope) over 10 trials. (Note that GNCN-t1- $\Sigma$  is also referred to as GNCN-t1- $\Sigma$ /Friston.)

To evaluate our framework (see Supplementary Methods), nine approaches were compared across four image datasets (Table 1), each of which contained a training subset (from which we created an additional validation subset) and a test subset (see Supplementary Methods). One is a Gaussian mixture model (GMM) and eight are neural models – four of these are backprop-based (see Supplementary Methods and Supplementary Note 6) and four are NGC models, i.e., a GNCN-t1/Rao (which is equivalent to the model of [68] and [80]), a GNCN-t1- $\Sigma$ /Friston (which is equivalent to the model of [22]), a GNCN-t2-L $\Sigma$ , and a GNCN-PDH. All neural models were constrained to have their top-most layer to contain 20 processing elements (the GNCN-t2-L $\Sigma$  and GNCN-PDH were restricted to 20 neural columns) and all had approximately the same total number of synapses. With respect to NGC models,  $T = 50$  (see Supplementary Note 4). The regularized auto-encoder (RAE), the auto-associative network least equipped to serve as a data synthesizer, reaches lower reconstruction error, in terms of binary cross entropy (BCE), compared to the other backprop-based generative models, i.e., the Gaussian variational autoencoder (GVAE), the constant-variance GVAE (CV-GVAE), and the adversarial autoencoder (GAN-AE). However, while the CV-GVAE, GVAE, and GAN-AE models yield worse reconstruction than the RAE, they obtain much better log likelihoods, especially compared to the GMM baseline, indicating that they strong data samplers. It makes sense that these models obtain better likelihood at the expense of BCE given that their optimization objective imposes a strong pressure to craft a proper (Gaussian) distribution over latent variables in addition to reconstructing data samples (in the case of the GAN-AE, the pressure comes from forcing the discriminator to distinguish between fake and real latent variables). Interestingly enough, we see that the GNCN-t2-L $\Sigma$  and GNCN-PDH obtains competitive log likelihood with the CV-GVAE, GVAE, and GAN-AE with the GNCN-PDH yielding the best log likelihood out of all GNCN models for all four datasets. Notably, the GNCN models result in the best reconstruction across all four datasets, with the GNCN-t1- $\Sigma$ /Friston and GNCN-PDH offering the lowest BCE. On MNIST, we note that the GNCN-PDH outperforms some other related prior models [9] that used our same experimental setup, e.g., a restricted Boltzmann machine with  $\log p(\mathbf{x}) = -112$  nats, a denoising autoencoder with  $\log p(\mathbf{x}) = -142$  nats (trained via backprop) and  $\log p(\mathbf{x}) = -116$  nats (using the walk-back algorithm). As indicated by our results, an NGC model’s sampling and reconstruction ability is very promising. Worthy of note, however, is that we find that backprop-based autoencoders appear to do better at matching the data’s class frequency than the GNCNs studied in this article (see Supplementary Note 3).

As shown in Figure 4, we measured the data efficiency of the GNCN-t1- $\Sigma$ /Friston, GNCN-t2-L $\Sigma$ , and GNCN-PDH as well as several representative backprop-models (RAE, GVAE, and GAN-AE). Specifically, we train each model (to convergence) using a subset of the training data of varying size (we created six versions of each dataset’s training subset using either 2%, 10%, 20%, 50%, 75%, or 100% of the original sample), and plot final test BCE at convergence (versus percentage of original data used on the x-axis), for each of the models. Desirably, the result demonstrates that the predictive processing NGC models generalize better given varying



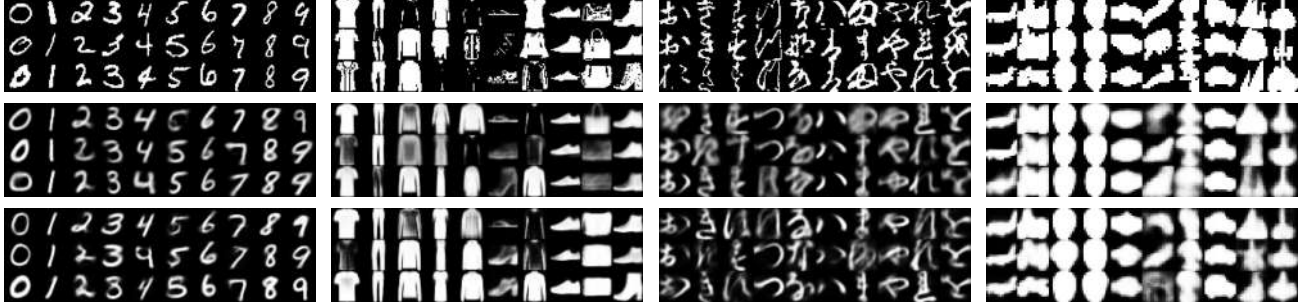


Figure 5: **Model sample visualization:** Shown are samples from models with overall best BCE-log  $p(\mathbf{x})$  balance (each column corresponds to one class). Top row shows dataset samples, middle row shows adversarial autoencoder (GAN-AE) samples, and bottom row shows generative neural coding network (GNCN-t2-LΣ) samples. Columns are arranged, left to right, as: a) MNIST, b) KMNIST, c) FMNIST, and d) CalTech.

Table 2: **Downstream performance across datasets:** (a) The classification error (Err) of a log-linear model fit to each model’s latent codes and the masked mean squared error (M-MSE) of the model’s pattern completion ability on the test set are reported (lower is better – two best scores are in **bold** with respect to each metric/column). (Metrics averaged over 10 trials - we report their mean and standard deviation.) (b) Masked patterns in first row and pattern completions from the GNCN-PDH in the second row.

(a) Model	Err (%)	M-MSE	Err (%)	M-MSE
DSRN	<b>1.93 ± 0.04</b>	–	<b>10.02 ± 0.08</b>	–
RAE	11.99 ± 0.47	22.10 ± 1.44	39.32 ± 0.95	30.69 ± 0.45
GVAE-CV	17.25 ± 0.50	20.37 ± 0.15	40.22 ± 0.05	23.69 ± 1.61
GVAE	9.13 ± 0.18	20.06 ± 0.57	48.09 ± 2.32	31.04 ± 0.26
GAN-AE	13.03 ± 1.68	15.87 ± 0.24	38.66 ± 0.87	22.97 ± 0.94
GNCN-t1/Rao	8.57 ± 0.03	4.54 ± 0.01	30.75 ± 0.06	9.16 ± 0.022
GNCN-t1-Σ/Friston	5.40 ± 0.01	<b>3.11 ± 0.02</b>	22.89 ± 0.59	<b>7.64 ± 0.08</b>
GNCN-t2-LΣ	2.38 ± 0.04	4.15 ± 0.02	<b>15.45 ± 0.43</b>	9.23 ± 0.02
GNCN-PDH	<b>2.28 ± 0.04</b>	<b>3.04 ± 0.02</b>	17.01 ± 0.32	<b>8.43 ± 0.04</b>

DSRN	<b>9.75 ± 0.09</b>	–	<b>31.56 ± 0.01</b>	–
RAE	25.14 ± 0.24	56.83 ± 0.73	38.10 ± 0.49	23.56 ± 0.91
GVAE-CV	24.19 ± 0.24	53.46 ± 2.39	38.64 ± 1.69	27.22 ± 1.43
GVAE	30.27 ± 0.57	50.81 ± 4.63	40.36 ± 0.43	24.08 ± 2.31
GAN-AE	24.67 ± 0.35	42.31 ± 3.82	37.34 ± 1.01	19.47 ± 0.88
GNCN-t1/Rao	20.28 ± 0.01	6.81 ± 0.01	34.23 ± 0.01	2.11 ± 0.03
GNCN-t1-Σ/Friston	17.78 ± 0.12	<b>6.01 ± 0.04</b>	30.60 ± 0.28	<b>1.97 ± 0.01</b>
GNCN-t2-LΣ	<b>16.85 ± 0.04</b>	7.46 ± 0.05	<b>29.11 ± 0.37</b>	2.23 ± 0.04
GNCN-PDH	17.11 ± 0.26	<b>6.01 ± 0.02</b>	29.54 ± 0.18	<b>1.82 ± 0.01</b>

<b>(b) MNIST</b>	<b>KMNIST</b>

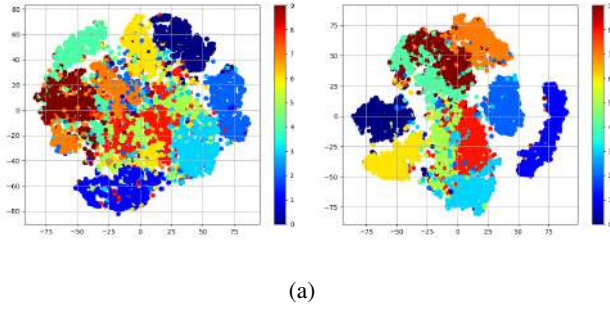
  

<b>FMNIST</b>	<b>CalTech</b>

amounts of data, even with less data compared to the autoencoder models. This benefit is useful given that it offsets the higher per-sample processing cost of the predictive processing models. Furthermore, in Figure 5, we present random samples from each class obtained from either: 1) the original dataset, 2) ancestrally sampling the GAN-AE, or 3) ancestrally sampling the GNCN-t2-LΣ (see Supplementary Table 1 for additional visualization of nearest-neighbor samples that match an original data point for each class). Note that we trained well-regularized MLP classifier trained on the original dataset and then used it to automatically annotate the samples produced by each model. Observe that both the GNCN-t2-LΣ and GAN-AE yield reasonably good-looking sample images for all four datasets and the differences in perceptual quality between the models’ sets of samples is marginal. This is encouraging, given that our goal was to demonstrate that an NGC model could be competitive with backprop-based generative models.

## 2.2 Neural Generative Coding Yields Strong Downstream Pattern Classifiers

All of the generative models we have experimented with in this paper are unsupervised in nature, meaning that by attempting to learn a density estimator of the data’s underlying distribution, the representations acquired by each might prove useful for downstream applications, such as image categorization. To evaluate each how useful each model’s latent representations might be when attempting to discriminate between samples, we evaluate the performance of a simple log-linear classifier, i.e., maximum entropy, that is fit to each model’s topmost latent variable using the labels accompanying each dataset. For all models, we measure the classification error (**Err**), as a percentage, on the test set of each benchmark in Table 2, where the closer a model is to 0%, the better. In addition, we provide the results for simple, purely discriminative baseline for context (the DSRN), which is simply a backprop-trained sparse recitfier network [27] that is constrained to have the same number of synapses as the generative models. As



	GNCN-t2-LΣ			GNCN-t1-Σ/Friston		
	$\rho_a^1$	$\rho_a^2$	$\rho_a^3$	$\rho_a^1$	$\rho_a^2$	$\rho_a^3$
<b>MNIST</b>	0.21	0.18	0.16	0.74	0.60	0.63
<b>KMNIST</b>	0.24	0.21	0.18	0.35	0.50	0.67
<b>FMNIST</b>	0.21	0.20	0.18	0.41	0.44	0.68
<b>CalTech</b>	0.21	0.19	0.17	0.50	0.68	0.69

(b)

Figure 6: **Examination of model latent codes:** (a) t-SNE plot of latent codes of the GAN-AE and GNCN-t2-LΣ on the MNIST database. (b) Sparsity levels in the GNCN-t2-LΣ versus the GNCN-t1-Σ/Friston [22],  $\rho_a^\ell$  indicates sparsity for layer  $\ell$ . (Analysis performed on training set.)

we see in Table 2, in terms of test error, the NGC models (GNCN-t1/Rao, GNCN-t1-Σ/Friston, GNCN-t2-LΣ, and GNCN-PDH) are competitive with the purely discriminatively-trained DSRN and outperform all of the other generative models (even outperforming the DSRN in one out of the four cases).

In Figure 6, we provide qualitative evidence that the latent representations of an NGC model (a GNCN-t2-LΣ) appear to yield a stronger, natural separation of the test data points into seemingly class-respective clusters as compared to the GAN-AE (one of the best performing backprop-based models with respect to both log likelihood and reconstruction error). Again, we emphasize that the GNCN-t2-LΣ acquired these representations without labeled information, meaning that the class-based relationships have emerged as a result of its very sparse neural activities. This offers some promising evidence that an NGC model’s representations offer benefits beyond the original density estimation task, allowing reuse of the same system for downstream tasks like categorization.

We hypothesize that the GNCN-t2-LΣ’s latent codes make it easier for a linear classifier to separate out patterns by category as a result of the fact that they are sparse. Crucially, this model’s lateral structure creates columns of neural processing units that fight for the right to explain the input, meaning that only a few within a group would be active when processing particular patterns. To quantify this sparsity in the GNCN-t2-LΣ, we measure and report in the table of Figure 6 (b) the (mean) proportion  $\rho_a^\ell$  of neurons at layer  $\ell$  that were active (we counted whether each unit for a given pattern satisfied  $z_i^\ell > \epsilon$ , where  $\epsilon = 1e^{-6}$ ) in each model on each dataset. We compare this model to the GNCN-t1-Σ/Friston [22], one of the best performing GNCNs that used a kurtotic prior to induce sparsity, and observe that the lateral inhibition/excitation greatly reduces the amount of neural firing while still obtaining top performance. Note that the activities of the GNCN-t2-LΣ are quite sparse (the highest/worst-case value was on KMNIST in layer 1 with a sparsity of 21%) and, furthermore, the number of active neurons is lower in deeper layers, with 11-15% sparsity for all datasets in the top layer  $z^3$ . We believe that the GNCN-t2-LΣ’s (and GNCN-PDH’s) structured form of sparsity aids it in modeling input, i.e., yielding strong likelihood and reconstruction error, facilitating better separation of nonlinear data, and opening the door to designing models that more directly adhere to homeostatic constraints similar to those imposed by the brain. Though models like the GAN-AE, GVAE, and CV-GVAE reach good log likelihoods too, it is possible that since their learning process focuses on shaping their latent spaces to dense, multivariate Gaussian distributions, there is little chance for useful sparsity to emerge as a by-product, reducing the possibility that the latent space might result in beneficial side-effects.

From a neurobiological perspective, it is well-known that lateral synapses often facilitate contextual processing [16] and offer a natural form of “activity sharpening” [2]. We believe that this is an important structural prior that biases the an NGC model like the GNCN-t2-LΣ or GNCN-PDH towards acquiring more economical representations [6], where progressively fewer neurons at layers progressively farther away from the input stimulus work to encode information (or are non-zero). Much like the more recent incarnations of spike-and-slab coding [29], our form of structurally-enforced sparsity has a much stronger regularization effect on the model and ensures that the generative distribution of the neural activities is truly sparse. This is unlike more traditional approaches where sparsity is weakly enforced through the use of factorial kurtotic priors applied during inference [59, 68]. Our NGC framework not only yields naturally sparse codes but also offers flexibility in exploring other types of lateral connectivity patterns beyond the choice made in this study.

### 2.3 Neural Generative Coding Can Conduct Pattern Completion

Another interesting ability attributed to auto-associative memory models is their ability to complete partially-corrupted or incomplete patterns. In the real world, this type of scenario would often occur in the form of object occlusion, where the view of an object might be partially obstructed from the agent’s view. Being able to “imagine” the rest of the object might prove useful when planning to grasp it or manipulate it in some fashion. To test each model’s ability to complete patterns, we conducted an experiment where the right half of each image in each dataset was masked and each model was tasked with predicting the deleted portions. In Table 2, we report the masked mean squared error (M-MSE, see Supplementary Methods) of each model on each dataset’s test set.

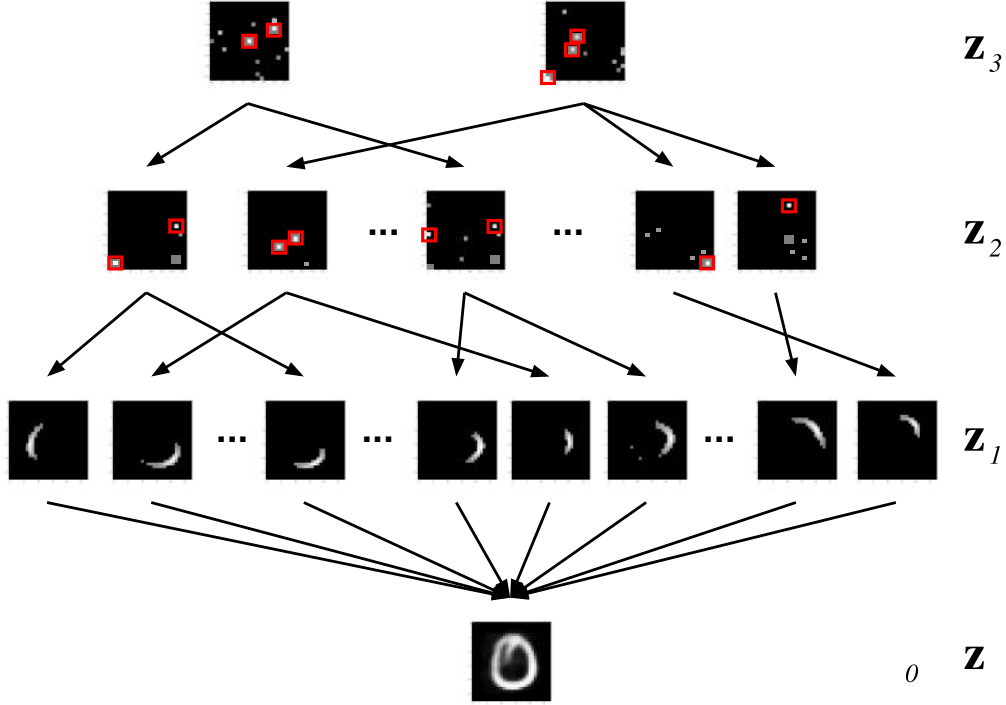


Figure 7: **The feature composition process:** Illustration of an NGC network composing the digit zero that was randomly sampled from the MNIST database (all the feature maps shown in this image have been extracted from one of the GNCN-t2-L $\Sigma$  models trained on MNIST). Note that layers 2 and 3 ( $z_2$  and  $z_3$ ) control a hierarchical neuron selection pathway which leads to excitation (turning “on”) or suppression (turning “off”) of certain low-level neurons in layer 1 ( $z_1$ ) that contain visual features that are super-imposed to generate a final output image. The image maps shown corresponding to  $z_2$  and  $z_3$  are visualized synaptic weight vectors that appear to serve as blueprints for choosing units (white indicate neurons to excite and black indicates neurons to suppress, after normalization was applied to the original maps and values lower than a threshold of 0.5 were set to black). Red squares have been added to indicate the neurons with some of the highest activity values that select the most important visual features that will be used to compose a final output predicted image.

Interestingly enough, we see that the NGC models outperform the other baselines in terms of pattern completion (with GNCN-t1- $\Sigma$ /Friston and GNCN-PDH offering the most competitive performance, with respect to measured M-MSE. We furthermore provide some examples of original data patterns that were masked in Table 2 (in the first and third rows) contrasted with the GNCN-PDH’s completed patterns (in the second and fourth rows). We hypothesize that the reason an NGC model is able to conduct pattern completion better than the autoencoder models is due to its iterative inference process.

### 3 Discussion

Generative models based on artificial neural networks (ANNs) have yielded promising tools for estimating and sampling from complicated probability data distributions [43, 28]. By re-considering how these models might operate and learn, drawing inspiration from one promising neuro-mechanistic account of how the brain interacts and adapts to its environment, i.e., predictive processing [13], we have shown that learning a viable generative model is possible. Specifically, we propose the neural generative coding (NGC) computational framework for learning neural probabilistic models of data, implementing several concrete instantiations of NGC which we called generative neural coding networks (GNCNs). In our experiments, we observe that the GNCNs are not only competitive with several powerful, modern-day backprop-based models on the task of estimating the marginal distribution of the data but they can generalize beyond the task they were originally trained to do. Specifically, we investigated the performance of all of the models on downstream tasks such as pattern completion and pattern categorization and discovered that the unsupervised NGC models outperformed all of the examined backprop-based baselines and were even competitive with an ANN that was directly trained to specialize for classification. As a result, for systems as complex as probabilistic generative models, we have demonstrated that crafting a more fundamentally brain-inspired approach to information processing and credit assignment can yield artificial neural systems that extract rich representations of input data in an unsupervised fashion. Our results demonstrate, on four datasets, that even though extra computation is needed to process each input for a fixed (yet small) stimulus presentation time, the NGC models converge far sooner than comparable backprop-based ones, generalizing well earlier.

To offer additional insight into what an NGC model’s intermediate representations might look like, we examined, using the MNIST dataset, the generative synaptic weights for each layer of a trained model, of which the visualization results for the bottom layer












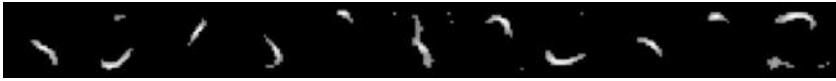


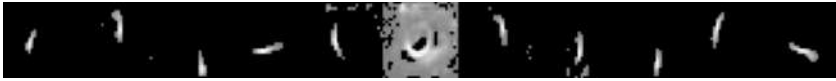















Class	Data	Output	Top Selected Features
"0"			
"1"			
"2"			
"3"			
"4"			
"5"			
"6"			
"7"			
"8"			
"9"			

Table 3: **Feature Composition Analysis of the generative neural coding network:** The "Class" shows the class label of a sample from MNIST, the "Data" column depicts the original pattern, and the "Output" column presents the weighted super-position of the 30 layer 1 synaptic weight vectors (we present the top 11 in the "Top Selected Features" column) associated with the 30 state neurons with highest activity when reconstructing the original datapoint. Note that the feature maps above were extracted from a GNCN-t2-L $\Sigma$  trained on MNIST and were first normalized and values lower than a threshold of 0.5 were set to black in order to improve the clearness of the specifically extracted feature.

(where  $z^1$  predicts  $z^0$ ) are shown in Table 3 (see Supplementary Note 8 for details of the analysis). Upon examination, we observe that a GNCN-t2-L $\Sigma$  appears to learn a latent command structure in its upper layers (layers  $\mathfrak{N}^2$  and  $\mathfrak{N}^3$  – these appear to provide maps for turning off or on state neurons in the level below) that work in tandem to drive a dynamic composition of low-level visual features in layer  $\mathfrak{N}^1$ . In Figure 7, we illustrate how these higher-level maps interact with the layer  $\mathfrak{N}^1$  features, ultimately composing a final output image by super-imposing an intensity-weighted set of low-level features (see the "Output" column of Table 3).

How the brain conducts credit assignment is a fundamental and open question in both (computational) neuroscience and cognitive science. There are many theories that posit how this might happen and our proposed NGC framework represents a scalable, computational instantiation of only one of them, the theory of predictive processing, suggesting that cortical regions in the brain

communicate prediction errors and/or predictions across different regions through a hierarchical message passing scheme [79]. Observe that the direction that this study takes is one that starts from cognitive neuroscientific concepts and ends in the development of a statistical learning algorithm. Specifically, we have shown that one way of emulating some neuro-biological principles yield agents that learn more general-purpose representations, as our results on downstream classification and pattern completion provide evidence for. As for implications for computational neuroscience itself, while an NGC model can embody concepts such as lateral competition-driven sparsity, hierarchical message passing, and local Hebbian-like synaptic adjustment, our framework lacks many important details that might allow it to serve as a means to make falsifiable claims that can be proven or disproven by neurobiological experiments. In addition, an NGC model suffers from other criticisms of many predictive processing models – for example, it requires a one-to-one pairing of error neurons with state neurons [11] (though this can potentially be resolved by decoupling the error neurons and introducing an extra synaptic weight matrix that connects a pool of error neurons of a one size to a pool of state neurons of a different size). However, if the NGC framework is modified significantly to more faithfully model neurobiological details, e.g., synapses are constrained to take on only non-negative values and neurons communicate via spike trains (for example, the NGC could work with leaky integrate-and-fire neurons, similar to that of [61]), it could serve as the means to facilitate refinement of predictive brain theories such as predictive processing [13] and principles such as free energy [24]. Doing so could facilitate stronger synergy between neural computational modeling and the design of agents that solve complex tasks examined in statistical learning.

Given the difficulty in imagining how backprop, in the form it is implemented when training deep ANNs today, might occur in the brain [30, 15], there is value in not only developing approximations of it that might be more brain-like (moving machine learning a bit closer to computational neuroscience) but also in exploring alternatives that start from neuro-cognitive principles, theories, and mechanisms, creating new algorithms that embody particular ideas in cognitive neuroscience at the outset. Taking the second of the last two directions, which is in the spirit of this work and several others [31, 71], might allow us to more easily shed the constraints imposed by backprop in the effort to construct general-purpose learning agents capable of emulating more complex human cognitive function. Doing so might also allow the machine learning community to make further progress on problems even harder than generative modeling, such the problem of active inference [25] and continual temporal prediction [62].

## References

- [1] ACKLEY, D. H., HINTON, G. E., AND SEJNOWSKI, T. J. A learning algorithm for boltzmann machines. *Cognitive science* 9, 1 (1985), 147–169.
- [2] ADESNIK, H., AND SCANZIANI, M. Lateral competition for cortical space by layer-specific horizontal circuits. *Nature* 464, 7292 (2010), 1155–1160.
- [3] ANDERSEN, P., ECCLES, J., AND LØYNING, Y. Hippocampus of the brain: Recurrent inhibition in the hippocampus with identification of the inhibitory cell and its synapses. *Nature* 198, 4880 (1963), 540–542.
- [4] BA, J. L., KIROS, J. R., AND HINTON, G. E. Layer normalization. *arXiv preprint arXiv:1607.06450* (2016).
- [5] BALDI, P., SADOWSKI, P., AND LU, Z. Learning in the machine: Random backpropagation and the learning channel. *arXiv preprint arXiv:1612.02734* (2016).
- [6] BARLOW, H. B. Single units and sensation: a neuron doctrine for perceptual psychology? *Perception* 1, 4 (1972), 371–394.
- [7] BARTUNOV, S., SANTORO, A., RICHARDS, B., MARRIS, L., HINTON, G. E., AND LILLICRAP, T. Assessing the scalability of biologically-motivated deep learning algorithms and architectures. In *Advances in Neural Information Processing Systems* (2018), pp. 9390–9400.
- [8] BASTOS, A. M., USREY, W. M., ADAMS, R. A., MANGUN, G. R., FRIES, P., AND FRISTON, K. J. Canonical microcircuits for predictive coding. *Neuron* 76, 4 (2012), 695–711.
- [9] BENGIO, Y., YAO, L., ALAIN, G., AND VINCENT, P. Generalized denoising auto-encoders as generative models. In *Advances in neural information processing systems* (2013), pp. 899–907.
- [10] BI, G.-Q., AND POO, M.-M. Synaptic modification by correlated activity: Hebb’s postulate revisited. *Annual review of neuroscience* 24, 1 (2001), 139–166.
- [11] BOGACZ, R. A tutorial on the free-energy framework for modelling perception and learning. *Journal of mathematical psychology* 76 (2017), 198–211.
- [12] CLANUWAT, T., BOBER-IRIZAR, M., KITAMOTO, A., LAMB, A., YAMAMOTO, K., AND HA, D. Deep learning for classical japanese literature, 2018.
- [13] CLARK, A. *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford University Press, 2015.
- [14] CRAFTON, B., WEST, M., BASNET, P., VOGEL, E., AND RAYCHOWDHURY, A. Local learning in rram neural networks with sparse direct feedback alignment. In *2019 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)* (2019), IEEE, pp. 1–6.

- [15] CRICK, F. The recent excitement about neural networks. *Nature* 337, 6203 (1989), 129–132.
- [16] CUI, Y., AHMAD, S., AND HAWKINS, J. The htm spatial pooler—a neocortical algorithm for online sparse distributed coding. *Frontiers in Computational Neuroscience* 11 (2017), 111.
- [17] DENEVE, S. Bayesian inference in spiking neurons. *Advances in neural information processing systems* 17 (2005), 353–360.
- [18] DOUGLAS, R. J., KOCH, C., MAHOWALD, M., MARTIN, K., AND SUAREZ, H. H. Recurrent excitation in neocortical circuits. *Science* 269, 5226 (1995), 981–985.
- [19] DUMOULIN, V., GOODFELLOW, I. J., COURVILLE, A., AND BENGIO, Y. On the challenges of physical implementations of rbms. *arXiv preprint arXiv:1312.5258* (2013).
- [20] EL-BOUSTANI, S., IP, J. P., BRETON-PROVENCHER, V., KNOTT, G. W., OKUNO, H., BITO, H., AND SUR, M. Locally coordinated synaptic plasticity of visual cortex neurons in vivo. *Science* 360, 6395 (2018), 1349–1354.
- [21] FRENKEL, C., LEFEBVRE, M., AND BOL, D. Learning without feedback: Direct random target projection as a feedback-alignment algorithm with layerwise feedforward training. *arXiv preprint arXiv:1909.01311* (2019).
- [22] FRISTON, K. Hierarchical models in the brain. *PLoS computational biology* 4, 11 (2008), e1000211.
- [23] FRISTON, K. The free-energy principle: a rough guide to the brain? *Trends in cognitive sciences* 13, 7 (2009), 293–301.
- [24] FRISTON, K., KILNER, J., AND HARRISON, L. A free energy principle for the brain. *Journal of Physiology-Paris* 100, 1-3 (2006), 70–87.
- [25] FRISTON, K., MATTOU, J., AND KILNER, J. Action understanding and active inference. *Biological cybernetics* 104, 1-2 (2011), 137–160.
- [26] GHOSH, P., SAJJADI, M. S., VERGARI, A., BLACK, M., AND SCHÖLKOPF, B. From variational to deterministic autoencoders. In *International Conference on Learning Representations* (2020).
- [27] GLOROT, X., BORDES, A., AND BENGIO, Y. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics* (2011), pp. 315–323.
- [28] GOODFELLOW, I., POUGET-ABADIE, J., MIRZA, M., XU, B., WARDE-FARLEY, D., OZAI, S., COURVILLE, A., AND BENGIO, Y. Generative adversarial nets. In *Advances in neural information processing systems* (2014), pp. 2672–2680.
- [29] GOODFELLOW, I. J., COURVILLE, A. C., AND BENGIO, Y. Scaling up spike-and-slab models for unsupervised feature learning. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 8 (2013), 1902–1914.
- [30] GROSSBERG, S. Competitive learning: From interactive activation to adaptive resonance. *Cognitive science* 11, 1 (1987), 23–63.
- [31] GUERGUIEV, J., LILLICRAP, T. P., AND RICHARDS, B. A. Towards deep learning with segregated dendrites. *ELife* 6 (2017), e22901.
- [32] HEBB, D. O., ET AL. The organization of behavior, 1949.
- [33] HELMHOLTZ, H. v. Concerning the perceptions in general. *Treatise on physiological optics*, (1866).
- [34] HINTON, G. E. Training products of experts by minimizing contrastive divergence. *Neural computation* 14, 8 (2002), 1771–1800.
- [35] HINTON, G. E. To recognize shapes, first learn to generate images. *Progress in brain research* 165 (2007), 535–547.
- [36] HINTON, G. E., ET AL. What kind of graphical model is the brain? In *IJCAI* (2005), vol. 5, pp. 1765–1775.
- [37] HINTON, G. E., AND MCCLELLAND, J. L. Learning representations by recirculation. In *Neural information processing systems* (1988), pp. 358–366.
- [38] HOPFIELD, J. J. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences* 79, 8 (1982), 2554–2558.
- [39] IOFFE, S., AND SZEGEDY, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning* (2015), PMLR, pp. 448–456.
- [40] ISOMURA, T., AND TOYOIZUMI, T. Error-gated hebbian rule: A local learning rule for principal and independent component analysis. *Scientific reports* 8, 1 (2018), 1–11.
- [41] JADERBERG, M., CZARNECKI, W. M., OSINDERO, S., VINYALS, O., GRAVES, A., SILVER, D., AND KAVUKCUOGLU, K. Decoupled neural interfaces using synthetic gradients. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70* (2017), JMLR. org, pp. 1627–1635.
- [42] KINGMA, D. P., AND BA, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
- [43] KINGMA, D. P., AND WELLING, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).



- [44] KNILL, D. C., AND POUGET, A. The bayesian brain: the role of uncertainty in neural coding and computation. *TRENDS in Neurosciences* 27, 12 (2004), 712–719.
- [45] LECUN, Y., BENGIO, Y., AND HINTON, G. Deep learning. *Nature* 521, 7553 (2015), 436–444.
- [46] LEE, D.-H., ZHANG, S., FISCHER, A., AND BENGIO, Y. Difference target propagation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (2015), Springer, pp. 498–515.
- [47] LIANG, H., GONG, X., CHEN, M., YAN, Y., LI, W., AND GILBERT, C. D. Interactions between feedback and lateral connections in the primary visual cortex. *Proceedings of the National Academy of Sciences* 114, 32 (2017), 8637–8642.
- [48] LILLICRAP, T. P., COWNDEN, D., TWEED, D. B., AND AKERMAN, C. J. Random synaptic feedback weights support error backpropagation for deep learning. *Nature communications* 7 (2016), 13276.
- [49] LITTLE, W. A. The existence of persistent states in the brain. *Mathematical biosciences* 19, 1-2 (1974), 101–120.
- [50] LONDON, M., ROTH, A., BEEREN, L., HÄUSSER, M., AND LATHAM, P. E. Sensitivity to perturbations in vivo implies high noise and suggests rate coding in cortex. *Nature* 466, 7302 (2010), 123–127.
- [51] LONG, N. M., LEE, H., AND KUHL, B. A. Hippocampal mismatch signals are modulated by the strength of neural predictions and their similarity to outcomes. *Journal of Neuroscience* 36, 50 (2016), 12677–12687.
- [52] MAGEE, J. C., AND JOHNSTON, D. A synaptically controlled, associative signal for hebbian plasticity in hippocampal neurons. *Science* 275, 5297 (1997), 209–213.
- [53] MAKHZANI, A., SHLENS, J., JAITLEY, N., AND GOODFELLOW, I. Adversarial autoencoders. In *International Conference on Learning Representations* (2016).
- [54] MISHKIN, D., AND MATAS, J. All you need is a good init. *CoRR abs/1511.06422* (2015).
- [55] MORAN, R. J., CAMPO, P., SYMMONDS, M., STEPHAN, K. E., DOLAN, R. J., AND FRISTON, K. J. Free energy, precision and learning: the role of cholinergic neuromodulation. *Journal of Neuroscience* 33, 19 (2013), 8227–8236.
- [56] MOSTAFA, H., RAMESH, V., AND CAUWENBERGHS, G. Deep supervised learning using local errors. *Frontiers in neuroscience* 12 (2018), 608.
- [57] MOVELLAN, J. R. Contrastive hebbian learning in the continuous hopfield model. In *Connectionist Models*. Elsevier, 1991, pp. 10–17.
- [58] NØKLAND, A. Direct feedback alignment provides learning in deep neural networks. In *Advances in Neural Information Processing Systems* (2016), pp. 1037–1045.
- [59] OLSHAUSEN, B. A., AND FIELD, D. J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 381, 6583 (1996), 607–609.
- [60] O'REILLY, R. C. Biologically plausible error-driven learning using local activation differences: The generalized recirculation algorithm. *Neural computation* 8, 5 (1996), 895–938.
- [61] ORORBIA, A. Spiking neural predictive coding for continual learning from data streams. *arXiv preprint arXiv:1908.08655* (2019).
- [62] ORORBIA, A., MALI, A., GILES, C. L., AND KIFER, D. Continual learning of recurrent neural networks by locally aligning distributed representations. *IEEE Transactions on Neural Networks and Learning Systems* (2020).
- [63] ORORBIA, A. G., AND MALI, A. Biologically motivated algorithms for propagating local target representations. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2019), vol. 33, pp. 4651–4658.
- [64] ORORBIA, A. G., MALI, A., KIFER, D., AND GILES, C. L. Deep credit assignment by aligning local representations. *arXiv preprint arXiv:1803.01834* (2018).
- [65] PARR, T., AND FRISTON, K. J. The anatomy of inference: generative models and brain structure. *Frontiers in computational neuroscience* 12 (2018), 90.
- [66] PASCANU, R., MIKOLOV, T., AND BENGIO, Y. On the difficulty of training recurrent neural networks. In *International conference on machine learning* (2013), pp. 1310–1318.
- [67] RAMAKRISHNAN, K., SCHOLTE, H. S., GROEN, I. I., SMEULDERS, A. W., AND GHEBREAB, S. Visual dictionaries as intermediate features in the human brain. *Frontiers in computational neuroscience* 8 (2015), 168.
- [68] RAO, R. P., AND BALLARD, D. H. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience* 2, 1 (1999).
- [69] RICHARDSON, E., AND WEISS, Y. On gans and gmms. *arXiv preprint arXiv:1805.12462* (2018).
- [70] RUMELHART, D. E., HINTON, G. E., AND WILLIAMS, R. J. Learning representations by back-propagating errors. *Nature* 323, 6088 (1986), 533–536.

- [71] SACRAMENTO, J., PONTE COSTA, R., BENGIO, Y., AND SENN, W. Dendritic cortical microcircuits approximate the backpropagation algorithm. *Advances in neural information processing systems* 31 (2018), 8721–8732.
- [72] SANTURKAR, S., SCHMIDT, L., AND MADRY, A. A classification-based study of covariate shift in gan distributions. In *International Conference on Machine Learning* (2018), PMLR, pp. 4480–4489.
- [73] SCELLIER, B., AND BENGIO, Y. Equilibrium propagation: Bridging the gap between energy-based models and backpropagation. *Frontiers in computational neuroscience* 11 (2017), 24.
- [74] SCELLIER, B., GOYAL, A., BINAS, J., MESNARD, T., AND BENGIO, Y. Generalization of equilibrium propagation to vector field dynamics. *arXiv preprint arXiv:1808.04873* (2018).
- [75] SPRATLING, M., AND JOHNSON, M. Dendritic inhibition enhances neural coding properties. *Cerebral Cortex* 11, 12 (2001), 1144–1149.
- [76] SWANSON, L. R. The predictive processing paradigm has roots in kant. *Frontiers in Systems Neuroscience* 10 (2016), 79.
- [77] VON HELMHOLTZ, H. *Ueber das Sehen des Menschen ein Populär Wissenschaftlicher Vortrag...* Leopold Voss, 1855.
- [78] WACONGNE, C., CHANGEUX, J.-P., AND DEHAENE, S. A neuronal model of predictive coding accounting for the mismatch negativity. *Journal of Neuroscience* 32, 11 (2012), 3665–3678.
- [79] WACONGNE, C., LABYT, E., VAN WASSENHOVE, V., BEKINSCHTEIN, T., NACCACHE, L., AND DEHAENE, S. Evidence for a hierarchy of predictions and prediction errors in human cortex. *Proceedings of the National Academy of Sciences* 108, 51 (2011), 20754–20759.
- [80] WHITTINGTON, J. C., AND BOGACZ, R. An approximation of the error backpropagation algorithm in a predictive coding network with local hebbian synaptic plasticity. *Neural computation* 29, 5 (2017), 1229–1262.
- [81] XIAO, H., RASUL, K., AND VOLLGRAF, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747* (2017).
- [82] ZADOR, A. M. A critique of pure learning and what artificial neural networks can learn from animal brains. *Nature communications* 10, 1 (2019), 1–7.
- [83] ZHANG, K., AND SEJNOWSKI, T. J. A universal scaling law between gray matter and white matter of cerebral cortex. *Proceedings of the National Academy of Sciences* 97, 10 (2000), 5621–5626.

## Appendix

### Datasets

All of the datasets used in this paper, except for CalTech 101, which already contained binary images, contained gray-scale pixel feature values in the range of  $[0, 255]$ . The images in these databases were first pre-processed by normalizing the pixels to the range of  $[0, 1]$  by dividing them by 255 and finally converted to binary values by thresholding at 0.5 as in [9].

The MNIST dataset contains  $28 \times 28$  images containing handwritten digits across 10 categories. Fashion MNIST (FMNIST) [81], which was proposed as a challenging drop-in replacement for MNIST, contains  $28 \times 28$  grey-scale images depicting clothing items (out of 10 item classes). Kuzushiji-MNIST (KMNIST) is another  $28 \times 28$  image dataset containing hand-drawn Japanese Kanji characters [12]. The Caltech 101 Silhouettes database contains  $16 \times 16$  binary pixel images across 100 different categories. Each training subset had 60000 samples and the testing subset had 10000 (the standard test split of each dataset was used in this paper), with the exception of Caltech 101, which had 8596 training samples and 2302 test samples. A validation subset of 2000 samples was drawn from each training set to be used for tuning model meta-parameters (Caltech had 2257 samples).

### Baseline Model Descriptions

The baseline backprop-based models implemented for this article included a regularized auto-associative (autoencoding) network (RAE), a Gaussian variational autoencoder with fixed (spherical) variance (GVAE-CV) [26], a Gaussian variational autoencoder (GVAE) [43], and an adversarial autoencoder (GAN-AE) [53]. All models were constrained to have four layers like the NGC models. For all backprop-based models, the sizes of the layers in between the latent variable and the input layer  $\tilde{\mathbf{z}}^0$  were chosen such that the total synaptic weight count of the model was approximately equal to the GNCNs, the linear rectifier was used for the internal activation function, i.e.,  $\phi^\ell(v) = \max(0, v)$ , and weight values (for any  $\mathbf{W}^\ell$  and  $\mathbf{E}^\ell$ ) were initialized from a centered Gaussian distribution with a standard deviation  $\sigma$  that was tuned on held-out validation data. To further improve generalization ability, the decoder weights of all autoencoders were regularized and some autoencoder models had specific meta-parameters that were tuned using validation data. Notably, the GAN-AE was the only model that required a specialized gradient descent rule, i.e., Adam [42], in order to obtain good log likelihood and to stabilize training (stability is a known issue related to GANs [28]). Finally, we implemented an optimized Gaussian mixture model (GMM) that was fit to the training data via expectation-maximization with the number of mixture components chosen based on preliminary experiments that yielded best performance. The only GMM implementation detail worthy of note was that we clipped the images sampled from the mixture model to lie in the range  $[0, 1]$  (improving likelihood slightly).

### Experimental Setup and Task Design

**Training Setup:** The parameters of all models, whether they were updated via backprop or by the NGC learning process, were all optimized using stochastic gradient descent using mini-batches (or subsets, randomly sampled without replacement) of 200 samples for 50 passes through the data (epochs). For the backprop-based models, we re-scaled the gradients [66] by re-projecting them to a Gaussian ball with radius of 5.0, which we found ensured stable performance across trials. For each model, upon completion of training, we fit a Gaussian mixture model (GMM) to the topmost neural activity layer (to serve as the model prior). This density estimator was trained with expectation-maximization and contained  $K = 75$  components, where each component  $k \in K$  defines a multivariate Gaussian with mean  $\mu_k$  and covariance  $\Sigma_k$  parameters as well as a mixing coefficient  $\pi_k$ .

**The Density Modeling Task:** Given a dataset  $\mathbf{X} \in \{0, 1\}^{D \times S}$ , where  $S$  is the number of vector pattern samples and  $D$  is the dimensionality of any given pattern, the goal is to learn a density model of  $p(\mathbf{X})$ , or the probability of sensory input data, where subset of  $B$  vectors is denoted as  $\mathbf{x} \in \{0, 1\}^{D \times B}$ . We parameterize the probability distribution  $p(\mathbf{X})$  via  $p_\Theta(\mathbf{X})$  by introducing learnable parameters  $\Theta$ . Since computing the marginal log likelihood  $\log p_\Theta(\mathbf{X})$  directly is intractable for all the models in this paper, we estimate it by calculating a Monte Carlo estimate using 5000 samples according to the recipe: 1) sample the GMM prior, 2) ancestrally sample directed neural generative model (whether a baseline or a GNCN) given the samples of the prior.

The reconstruction metric, binary cross-entropy (BCE), also known as the negative Bernoulli log likelihood, was computed as follows:  $BCE(\mathbf{X}, \hat{\mathbf{X}}) = -\frac{1}{S} \sum_{s=1}^S \sum_{d=1}^D \left( \mathbf{X} \log(\hat{\mathbf{X}}) + (1 - \mathbf{X}) \log(1 - \hat{\mathbf{X}}) \right) [d, s]$  (measured in nats).  $\hat{\mathbf{X}} \in \{0, 1\}^{D \times S}$  is the predicted sensory input (matrix) produced from a model under evaluation.

**The Pattern Completion Task:** For this task, we test each model’s ability to complete patterns where the images of each dataset were partially masked and each model was tasked with completing the masked images. Specifically, half of each image  $\mathbf{x}$  (containing  $\sqrt{D}$  columns of  $\sqrt{D}$  pixels) was masked according to a binary mask  $\mathbf{m}$  where  $\sqrt{D}/2$  columns were set to 1 and the rest 0, i.e.,  $\mathbf{x}_m = \mathbf{x} \odot \mathbf{m}$ . We report the masked mean squared error (M-MSE) of each model on each dataset’s test set, which is computed per

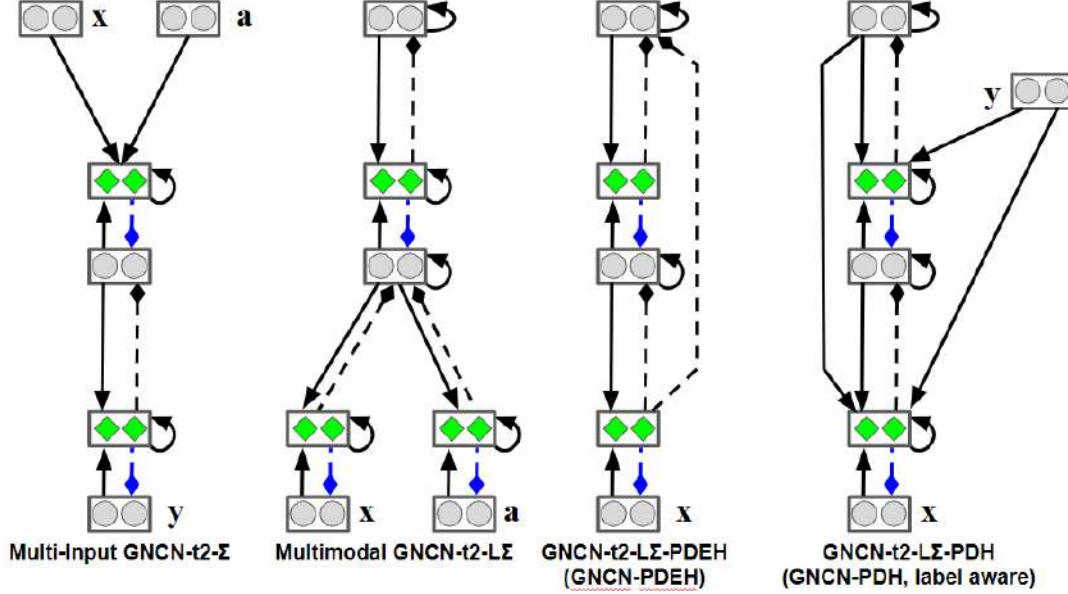


Figure 8: **Variant NGC Architectures:** Ordered left to right, possible architectures in NGC include: 1) a multi-input GNCN-t2- $\Sigma$ , 2) a multimodal GNCN-t2-L $\Sigma$ , 3) a GNCN-t2-L with an asymmetric error transmission pathway (a GNCN-PDEH), and 4) a label-aware GNCN-PDH (or a multi-input GNCN-t2-L $\Sigma$ ) with an asymmetric generative pathway.

image as follows:  $M\text{-MSE}(\mathbf{X}, \hat{\mathbf{X}}, \mathbf{M}) = ((\hat{\mathbf{X}} - \mathbf{X}) \odot (1 - \mathbf{M}))^T \cdot ((\hat{\mathbf{X}} - \mathbf{X}) \odot (1 - \mathbf{M}))$  (measured in nats), where  $\mathbf{M} \in \{0, 1\}^{D \times S}$  is a masking matrix  $\mathbf{m}$  where each column contains one mask vector per image vector  $\mathbf{x}$ .

**The Classification Task:** In this task, each model’s prediction error is measured on the test sample’s label set. This is possible given that each dataset  $\mathbf{X}$  also comes with a set of target annotations that be encoded into a  $C$ -dimensional space yielding the label matrix  $\mathbf{Y} \in \mathcal{R}^{C \times D}$  (one label vector per sample), where  $C$  is the number of unique categories labeled a priori. We measure the classification error (**Err**) (as a percentage %) on the test set, as follows:

$$\text{Err}(\mathbf{Y}, \hat{\mathbf{Y}}) = \left( 1 - \frac{1}{S} \sum_{s=1}^S \begin{cases} 1 & \hat{y}_s = y_s \\ 0 & \hat{y}_s \neq y_s \end{cases} \right)$$

where  $\hat{y}_s = \arg \max_d(\hat{\mathbf{Y}}[:, s])$ , the class index chosen by the model, and  $y_s = \arg \max_d(\mathbf{Y}[:, s])$ , the index of the actual class label. Note that  $\hat{\mathbf{Y}} \in \mathcal{R}^{C \times D}$  is the collected set of predictions from the linear classifier fit to a model’s latent space.

## Neural Generative Coding Model Procedures

In this section, we detail the sampling and image completion procedures for the GNCN models.

**Sampling from the Model:** Since the optimization procedure, whether via gradient descent or another method is not guaranteed to find globally optimal parameter settings (since the objective function is not convex), the distribution of the latent state  $\mathbf{z}^L$  of the final layer will not be Gaussian. To estimate it, after training on the data, we obtain the corresponding  $\mathbf{z}_i^L$  value for each data point  $\mathbf{x}_i$ . We fit a Gaussian mixture model to this collection of values  $\mathbf{z}_1^L, \dots, \mathbf{z}_N^L$  (where  $N$  is the number of training points). Then, in order to reduce variance during sampling, we take the following approach. First, we sample  $\mathbf{z}^L$  from the Gaussian mixture model, recursively set  $\bar{\mathbf{z}}^{\ell-1} \leftarrow g^{\ell-1}(\mathbf{W}^{\ell-1} \phi^\ell(\mathbf{z}^\ell))$  and output  $\mathbf{z}^0$ . This is similar to how variational auto-encoders are sampled in practice, where the input is a Gaussian and the output is the mean vector.

**Image Completion:** In the event that incomplete sensory input  $\mathbf{x}$  is provided to the GNCN, i.e., portions of  $\mathbf{x}$  are masked out by the variable  $\mathbf{m} \in \{0, 1\}^{J_0 \times 1}$ , we may infer the remaining portions of  $\mathbf{x}$  by utilizing the output error neurons of the GNCN and treating the bottom sensory layer  $\mathbf{z}^0$  as a partial latent state. Specifically, we update the missing portions, i.e.,  $1 - \mathbf{m}$ , of  $\mathbf{z}^0$  as follows:

$$\mathbf{z}^0 = \mathbf{x} \odot \mathbf{m} + \left( \mathbf{z}^0 + \beta \left( -\frac{\partial \psi}{\partial \bar{\mathbf{z}}^0} \right) \right) \odot (1 - \mathbf{m}) = \mathbf{x} \odot \mathbf{m} + \left( \mathbf{z}^0 - \beta \mathbf{e}^0 \right) \odot (1 - \mathbf{m}). \quad (19)$$

## Supplementary Note 1: On The Neural Generative Coding Framework

**NGC Model Structure and Naming Convention:** Naming GNCN models under the NGC framework entails appending a suffix to the end of the model sub-name. The hyphenated suffix “-t1” refers to a GNCN model that uses “Type 1” error synapses, or, specifically, error synapses that are a function of the GNCN’s forward generative weights, i.e., they are not physically separate/distinct synapses. The suffix “-t2” refers to a GNCN model with “Type 2” error synapses, or separate learnable synaptic parameters that focus on solely transmitting error messages through the model. An additional suffix is added to the model depending on whether on it contains lateral synapses in its state variables, i.e., “-L”, and whether or not it contains lateral precision weights in its error neurons, i.e., “- $\Sigma$ ”. As a complete example, a GNCN that contains Type 2 error synapses and lateral synapses in both its state and error neurons would be designated as GNCN-t2-L $\Sigma$ .

NGC Model Name	Error Type	Precision Wghts	Lateral Wghts	$\alpha_m$ Value	Uses $\partial\phi$
GNCN-t1/Rao	Type 1	No	No	0	Yes
GNCN-t1- $\Sigma$ /Friston	Type 1	Yes	No	0	Yes
GNCN-t2-L $\Sigma$	Type 2	Yes	Yes	0	No
GNCN-t2-L $\Sigma$ -PDH (GNCN-PDH)	Type 2	Yes	Yes	1	No

While this study explores four variant GNCN models that can be derived from the general NGC computational framework we put forth, there are many other possible architectures that can handle other styles of problems. This flexibility is owed to the fact that our NGC framework supports asymmetry with respect to the structure of the forward generative pathway and that of the error transmission pathway. Additionally, if an NGC model utilizes a non-hierarchical structure in its generative/prediction neural structure, as in the case of one of this paper’s main models, it receives one final suffix “-PDH” (for partially decomposable hierarchy). Note that in the case of the GNCN-t2-L $\Sigma$ -PDH, we, in this paper, for further convenience, abbreviate it to GNCN-PDH. Alternatively, if an NGC model contains a non-hierarchical error structure (given that its forward and error transmission pathways are not strictly required to be symmetric), it receives the final suffix “-PDEH” (for partially decomposable error hierarchy). Note that special cases of our framework have been provided special name modifications, e.g., GNCN-t1 is also referred to as GNCN-t1/Rao [68] and GNCN-t1- $\Sigma$  is also referred to as GNCN-t1- $\Sigma$ /Friston [22].

Figure 8 depicts four other alternative neural circuit structures that would tackle, respectively, 1) clamped, multiple input to generated/predicted outputs (as in the case of tasks such as direct classification), 2) multi-modal generative modeling (for example, crafting a circuit that jointly learns to synthesize an image and discrete one-hot encoding of a word/character at time step  $t$  within a sequence), 3) a generative model where upper layers receive error messages from layers other than its immediately connected one (a GNCN-PDEH, where PDEH means partially decomposable error hierarchy), i.e., layer  $\ell = 2$  receives error messages from layer  $\ell = 1$  and  $\ell = 0$ , and 4) a label-aware generative model that forms a partially decomposable hierarchy in its forward generative structure (GNCN-PDH driven by labels as input).

## Supplementary Note 2: Related Work on Biologically-Inspired Models & Learning Algorithms

As mentioned in the introduction, some of the more notable criticisms of backprop include:

1. Synapses that make up the forward information pathway need to directly be used in reverse to communicate teaching signals (the weight transport problem),
2. Neurons need to be able to communicate their own activation function’s first derivative,
3. Neurons must wait for the neurons ahead of them to percolate their error signals way back before adjusting their own synapses (the update-locking problem),
4. There is a distinct form of information propagation through a long, global error feedback pathway that only affect weights but does not (at least directly) affect the network’s internal representations,
5. The error signals have a one-to-one correspondence with neurons.

The properties above are inherent to backprop and do not conform to known biological feedback mechanisms underlying neural communication in the brain [15]. The brain, in contrast, is heavily recurrently connected [3, 18], allowing for complementary pathways to form that would allow for percolation of error/mismatch information [51, 78]. Biological neurons communicate binary spike signals, making it unlikely that they also sport specialized circuitry to communicate the derivative of a loss function with respect to their activities [35] (though some recent studies have suggested that real neurons might communicate with rate codes [50]). Furthermore, it is more commonly accepted that neurons in the brain learn “locally” [32, 20], modulated globally by signals provided through neuromodulators such as dopamine, i.e., they operate with only immediately available information (such as their own activity and that of nearby neurons that they are connected to), making it unlikely that a global feedback pathway drives synaptic weight adjustments. In addition, several of the problems above result in practical problems – the weight transport problem has been shown to create memory access pattern issues in hardware implementations [14] and the global feedback pathway itself is one

key source behind the well-known exploding and vanishing gradient problems [66] in deep ANNs, yielding unstable or ineffective learning unless specific heuristics are employed.

In recent research, developing learning procedures that enable backprop-level learning while embodying elements of actual neuronal function has seen increasing interest in the machine learning community. However, while insights provided by each development have proven valuable, increasing the evidence that shows how a backprop-free form of adaptation can be consistent with some aspects real networks of neurons, many of these ideas only address one or a few of the issues described earlier. Random feedback alignment algorithms [5, 48] address the weight transport problem, and to varying degrees, the update locking problem [58, 56, 21], but are fundamentally emulating backprop’s differentiable global feedback pathway to create teaching signals (and pay a reduction in generalization ability the farther away they deviate from backprop [7, 21]). In addition, experimentally, the success of these approaches depends on how well the feedback weights are chosen a priori (instead of learning them). Other procedures, like local representation alignment [63] and target propagation [37, 46], which also resolve the weight transport problem and eschew the need for differentiable activations, fail to address the update locking problem, since the various incarnations of these require a full forward pass to initiate inference. Procedures have been proposed to address the update locking problem, such as the method of synthetic gradients [41], but still require backprop to compute local gradients. It remains to be seen how these algorithms could be adapted to learn generative models. Other systems, such as those related to contrastive Hebbian learning (CHL) [57, 60, 74], are much more biologically-plausible but often require symmetry between the forward and backward synaptic pathways, i.e., failing to address weight transport. More importantly, CHL requires long settling phases in order to compute activities and teaching signals, resulting in long computational simulation times. However, while some of these algorithms have shown some success in ANN training [46, 48, 63], they focus on classification, which is purely supervised and arguably a simpler problem than generative modeling.

Boltzmann machines [1], which are generalizations of Hopfield networks [49, 38] to incorporate latent variables, are a type of generative network that also contains lateral connections between neurons much as the models in our NGC framework do. While training for the original model was slow, a simplification was later made to omit lateral synapses, yielding a bipartite graphical model referred to as a harmonium, trained by contrastive divergence [34], a local contrastive Hebbian learning rule. However, while powerful, the harmonium could only synthesize reasonable-looking samples with many iterations of block Gibbs sampling and the training algorithm suffered from mixing problems (leading low sample diversity among other issues) [19]. Another kind of generative model [36] can be trained with the wake-sleep algorithm (or the up-down algorithm in the case of deep belief networks [36]), where an inference (upward) network and a generative (downward) network are jointly trained to invert each other. Unfortunately, wake-sleep suffers from instability, struggling to produce good samples of data most of the time, due to the difficulty both networks have in inverting each other due to layer-wise distributional shift. Motivated by the deficiencies in models learned by contrastive divergence/wake-sleep, algorithms have been created for auto-encoder-based models [9] but most efforts today rely on backprop.

### Supplementary Note 3: On Evaluating Mode Capture

Given that a common problem in training generative models is mode collapse, we measure some distributional properties of several of our baseline models and NGC models. Note that it is difficult to directly and automatically determine the labels of the samples produced by the unsupervised models investigated in this paper (aside from manual qualitative inspection) and it is an open research area/problem to develop better measurements for evaluating mode-capturing ability of generative models in general. Among the myriad of current experimental approaches proposed for measuring the degree of modal capture, we opted to implement and measure the number of statistically different bins (NBD) from [69], which has been argued to be a potentially useful metric for evaluating the degree of mode collapse that a given generative model might have experienced (values closer to 0 mean that the model has likely captured most of the modes of the data’s underlying distribution). To complement this metric, we also measure the Kullback-Leibler divergence (KL-D) between a pool of samples generated by our model (equal to the size of the original dataset) and the original dataset samples – we estimate the empirical mean and covariance matrices of each and calculate the closed-form multivariate Gaussian KL-D, to be specific.

To further utilize the labels, we also present results using another approach proposed in [72]. Specifically, we train a well-regularized multilayer perceptron (MLP) classifier (as mentioned above) on the full original dataset and use it to automatically annotate the samples produced by a given generative model. Once the samples have been annotated, we compute the frequencies of each label class (and furthermore, normalize these values to lie in the range  $[0, 1]$ ) and plot these class probabilities in Figure 9 (we also plot the ground truth distribution for reference). Also note that, since even a powerful MLP classifier such as the one we trained incurs error (it naturally cannot reach 100% test error), the frequency measurements should be taken with annotator error in mind.

Based on the results presented in Figure 9 and Table 4, we see that while it does not appear that the predictive processing NGC models suffer from any severe form of mode collapse, they do not appear to capture the frequency distribution of the classes quite as well as the GAN-AE. With respect to the G-KL and NBD metrics, it does appear that the GNCN-PDH and GNCN-t1- $\Sigma$  do well (often performing among the top scores across all datasets), though the GAN-AE and GVAE appear to perform the best currently with respect to these mode measurements. Future work will entail uncovering the potential reason why the NGC models do not, at least upon first examination, match the class frequency of the data as well as the autoencoders.



Model		G-KL	NDB		G-KL	NDB
Baseline		1229.3871	$1.00 \pm 0.0$		1988.3081	$1.00 \pm 0.0$
RAE		$180.621 \pm 5.966$	$0.72 \pm 0.014$		$668.821 \pm 36.12$	$0.72 \pm 0.014$
GVAE-CV		$254.976 \pm 3.036$	$0.63 \pm 0.042$		$704.495 \pm 29.394$	$0.645 \pm 0.049$
GVAE		$295.973 \pm 23.266$	$0.71 \pm 0.028$		$804.943 \pm 18.041$	$0.685 \pm 0.021$
GAN-AE		$540.408 \pm 25.943$	$0.665 \pm 0.035$		$1325.641 \pm 39.115$	$0.74 \pm 0.014$
GNCN-t1/Rao		$682.454 \pm 1.732$	$0.77 \pm 0.042$		$1316.362 \pm 2.463$	$0.755 \pm 0.049$
GNCN-t1- $\Sigma$ /Friston		$531.206 \pm 3.735$	$0.775 \pm 0.021$		$1205.361 \pm 6.5$	$0.79 \pm 0.014$
GNCN-t2-L $\Sigma$		$596.67 \pm 0.106$	$0.78 \pm 0.014$		$1218.252 \pm 2.761$	$0.795 \pm 0.12$
GNCN-PDH		$446.053 \pm 3.369$	$0.69 \pm 0.002$		$1150.434 \pm 2.68$	$0.745 \pm 0.007$
Baseline		1811.2068	$1.0 \pm 0.0$		704.3340	$1.0 \pm 0.0$
RAE		$551.805 \pm 14.6$	$0.74 \pm 0.028$		$57.235 \pm 4.925$	$0.34 \pm 0.042$
GVAE-CV		$661.615 \pm 4.182$	$0.725 \pm 0.007$		$216.271 \pm 8.387$	$0.32 \pm 0.014$
GVAE		$797.931 \pm 1.731$	$0.685 \pm 0.021$		$205.223 \pm 1.316$	$0.27 \pm 0.002$
GAN-AE		$1279.573 \pm 43.131$	$0.805 \pm 0.035$		$266.874 \pm 8.692$	$0.365 \pm 0.092$
GNCN-t1/Rao		$1173.705 \pm 8.71$	$0.79 \pm 0.014$		$314.684 \pm 1.871$	$0.325 \pm 0.021$
GNCN-t1- $\Sigma$ /Friston		$1040.528 \pm 1.019$	$0.805 \pm 0.021$		$300.303 \pm 0.659$	$0.4 \pm 0.028$
GNCN-t2-L $\Sigma$		$1180.764 \pm 7.35$	$0.86 \pm 0.002$		$290.617 \pm 0.299$	$0.395 \pm 0.049$
GNCN-PDH		$1091.876 \pm 8.8$	$0.76 \pm 0.014$		$204.032 \pm 1.067$	$0.31 \pm 0.099$

Table 4: **Distributional measurements:** Model ability to match the implicit modes underlying each dataset, i.e., MNIST, KMNIST, FMNIST, and CalTech. Metrics reported include the Gaussian Kullback-Leibler divergence (G-KL) and number of statistically different bins (NDB). (Metrics averaged over 10 trials - we report their mean and standard deviation.)

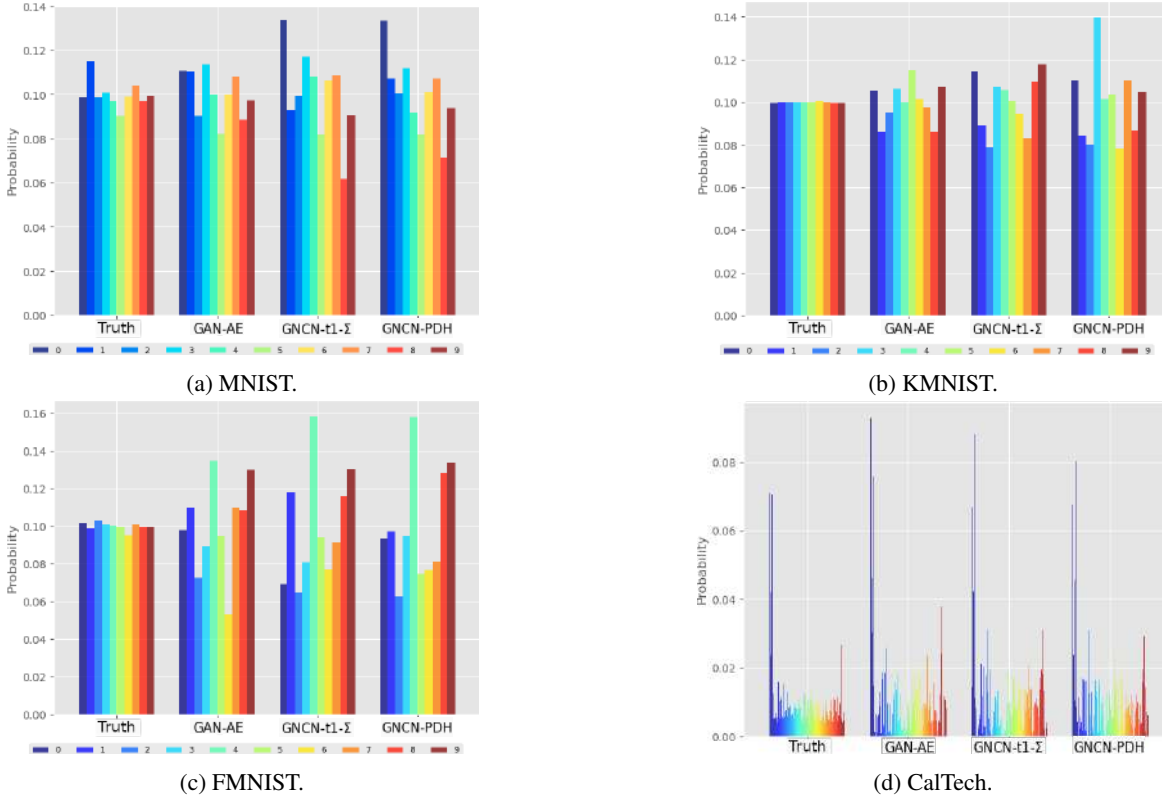


Figure 9: **Class distribution visualization:** Approximate label distributions (averaged over 10 trials) produced by each model – samples from each model were automatically annotated by a regularized MLP discriminator separately trained on original ground truth data. (a) MNIST class distribution, (b) KMNIST class distribution, (c) FMNIST class distribution, and (d) CalTech class distribution. (Note that GNCN-t1- $\Sigma$  is also referred to as GNCN-t1- $\Sigma$ /Friston.)

## Supplementary Note 4: On Model Complexity of Neural Generative Coding

### On Setting Parameter Complexity

Model complexity was selected based on consideration of the hardware resources available and preliminary experimentation with the validation set of each dataset, i.e., each benchmark had a validation subset that was randomly sampled without replacement

Model	MNIST Samples	KMNIST Samples
Truth		
RAE		
GAN-AE		
GNCN-t1- $\Sigma$		
GNCN-t2-L $\Sigma$		
	FMNIST Samples	CalTech Samples
Truth		
RAE		
GAN-AE		
GNCN-t1- $\Sigma$		
GNCN-t2-L $\Sigma$		

Table 5: **Nearest neighbor sample matches:** For each dataset, i.e., MNIST, KMNIST, FMNIST, and CalTech, we depict the nearest neighbor samples for each class from: ground truth (Truth), the regularized autoencoder (RAE), the adversarial autoencoder (GAN-AE), the GNCN-t1- $\Sigma$  (also referred to as GNCN-t1- $\Sigma$ /Friston), and the GNCN-t2-L $\Sigma$ . Specifically, we present the top 10 nearest neighbor matches to a randomly sampled digit for each class from the original dataset.

(per class) from the training set, as described in the paper. Backprop-based variational autoencoders are typically designed with low-dimensional latent spaces in mind so we investigated different latent code sizes in the range of [5, 30] and found that 20 worked best. Then, we constrained the GNCNs that had lateral synapses to only have 20 neural columns (since each column roughly functions as a latent variable) in their topmost layers and explored values for the lower levels in the range of [100, 400] and found that 360 was a good choice (GNCN-t1 [68] and GNCN-t1- $\Sigma$  [22] were set to use 360 neurons in the top layer). For the autoencoder models, the hidden layer sizes were set to be equal and a coarse grid search was conducted through the range between 100 and the maximum number possible for equal-sized layers such that the model could only have a maximum number of parameters equal to the total of the NGC models (meaning they could have fewer synapses if that yielded better performance on the validation set), to ensure fair comparison.

The optimal dimensionality of  $\mathbf{z}^\ell$  was also tuned through preliminary experimentation using held-out validation data. We chose the linear rectifier activation function for the NGC models because we desired strictly positive activity values (which work well with the formulation of lateral inhibition we present in this work). For the GNCN-t1- $\Sigma$  [22], we found that the linear rectifier worked best while the hyperbolic tangent worked best for GNCN-t1 [68] (for both of these models, the coefficient weighting the kurtotic prior was tuned on validation data).

### On Model Run-time Complexity

Note that an NGC model, require multiple steps of processing to obtain their latent activities for a given input. Naturally, per sample, this means that all of the predictive processing models we have explored would be slower than the feedforward autoencoder models (which conduct inference with a single feedforward pass). In short, while an autoencoder (with  $L + 1$  layers) would roughly only require  $2 * L$  matrix multiplications (the most expensive operation in the neural systems we investigate), any NGC predictive processing model would require at least  $2 * L * T$  multiplications. Note that, as observed in our data efficiency plots in the main paper, we note that this more expensive per-sample cost is desirably offset by convergence with fewer samples in comparison to backprop models. Furthermore, specialized hardware could take advantage of the NGC's inherent parallelism to speed up the process.

One key to reducing the inherent cost of an NGC model's iterative processing include designing alternative state update equations (whereas the ones explored in this paper embody a form of Euler integration, one could design higher-order integration steps, such as those based on the midpoint method or Runge-Kutta). Another solution could be to craft an amortized inference process, where another neural model (much akin to the encoder in a variational autoencoder) learns to infer the value of the state variables at the end

of an expensive iterative inference process so as to ultimately reduce the amount of processing steps per sample required. We leave investigation of these remedies for future work.

### Supplementary Note 5: On the Omission of Activation Derivatives

Removing the activation derivative as we did in our GNCN-t2-L $\Sigma$  and GNCN-PDH models could be argued to lead to possible value fluctuations that lead to unstable dynamics. Nonetheless, experimentally, we did not observe any strong weight fluctuations in our simulations and we believe that such fluctuations (much like the growing weight value problem inherent to many Hebbian update rules) are unlikely given that our model weight (columns) are constrained to have unit norms. Furthermore, the step size  $\beta$  is usually kept within the range of  $[0.02, 0.1]$  and the leak variable  $-\gamma \mathbf{z}^\ell$  helps to smooth out the values and prevent the occurrence of large values/magnitudes (serving as sort of an L2 penalty over the latent state activities).

In [64, 63], it was argued that so long as the activation function was monotonically increasing (with a similar condition imposed for stochastic activation functions), then the learning process would be stable and the benefit that the point-wise derivative offered would be absorbed by the error synaptic weights introduced to carry error signals. However, note that an approximation for the derivative in the form of a prefactor (derivative) term could be designed/introduced to further safeguard against potential fluctuations (and this will be the subject of future work).

### Supplementary Note 6: Generating the Lateral Competition Matrices

In the main paper, we introduced a lateral competition matrix  $\mathbf{V}^\ell$  that directly affects the latent state  $\mathbf{z}^\ell$ . It is created to contain the self-excitation weights and lateral inhibition weights by using the following matrix equation:  $\mathbf{V}^\ell = \alpha_h(\mathbf{M}^\ell) \odot (1 - I) - \alpha_e(I)$ , where  $I$  is the identity matrix and the masking matrix  $\mathbf{M}^\ell \in \{0, 1\}^{J_\ell \times J_\ell}$  is set by the experimenter (placing ones in the slots where it is desired for neuron pairs to laterally inhibit one another). In this study, we set  $\alpha_e = 0.13$  (the self-excitation strength) and  $\alpha_h = 0.125$  (the lateral inhibition strength). Our mask matrix  $\mathbf{M}^\ell$ , which emphasized a type of group or neural-column form of competition, was generated by the following process:

1. create  $J_\ell/K$  matrices of shape  $J_\ell \times K$  of zeros, i.e.,  $\{\mathbf{S}_1, \mathbf{S}_2, \dots, \mathbf{S}_k, \dots, \mathbf{S}_C\}$  (where  $C = J_\ell/K$ )
2. in each matrix  $\mathbf{S}_k$  insert ones at all combinations of coordinates  $c = \{1, \dots, k, \dots, K\}$  (column index) and  $r = \{1 + K * (k - 1), \dots, k + K * (k - 1), \dots, K + K * (k - 1)\}$  (row index)
3. concatenate the  $J_\ell/K$  matrices along the horizontal axis, i.e.,  $\mathbf{M}^\ell = \langle \mathbf{S}_1, \mathbf{S}_2, \dots, \mathbf{S}_C \rangle$ .

Note that for our proposed integration of the lateral synapses in the state neuron layers, we start out with a probabilistic model and then modify it by introducing sparsity driven by lateral synaptic weights (a neuroscience-inspired idea), which directly modify the values of the state neurons (serving as sort of filter). While we cannot justify them in the probabilistic model, our experiments in the Results section of the main paper show that they improve over those that do not employ them, such as GNCN-t1- $\Sigma$  [22] and GNCN-t1 [68] (which both impose a Laplace distribution over their state neurons to encourage sparsity). Part of our future work will be to derive a probabilistic interpretation of our particular extensions to the NGC model.

With respect to backprop-based neural systems, one could also introduce stateful neurons with similar connectivity to our lateral and precision synapses above by introducing recurrence (as is done in recurrent neural networks). However, to update the weight parameters, one would have to resort to backprop through time (BPTT) and unroll the model over  $T$  steps in time. This would require creating a very deep computational graph and storing the activities and gradients at each time step before a final update to each synaptic weight matrix could be calculated, creating a very expensive memory footprint. An NGC model, in contrast, does not require unrolling and the large memory footprint associated with BPTT-trained recurrent networks.

### Supplementary Note 7: Autoencoder Baseline Model Descriptions

To make learning the decoder function (NN) described in the main paper tractable, it is common practice in the deep learning literature to introduce a supporting function known as the *encoder* [43]. The encoder ( $\text{NN}_e$ ), parameterized by a feedforward network, takes in the input stimulus  $\mathbf{x}$  and maps it to  $\mathbf{z}$  or to a distribution over  $\mathbf{z}$ . Depending on the choice of encoder, one can recover one of the four main baselines we experimented with in this paper.

For all backprop-based baseline models in this paper, the decoder of each was regularized with an additional L2 penalty. Specifically, this meant that their data log likelihood objectives always took the form:  $\psi_{reg} = \psi + \Omega(\Theta_{\text{NN}})$ , where  $\Theta_{\text{NN}} = \{\mathbf{W}^L, \dots, \mathbf{W}^\ell, \dots, \mathbf{W}^1\}$  contains all of the weight matrix parameters of the decoder NN.  $\Omega_{\text{NN}}$  is the regularization function applied to the decoder, i.e.,  $\Omega_{\text{NN}} = -\lambda \sum_{\mathbf{W}^\ell \in \Theta} \|\mathbf{W}^\ell\|_2^2$  where  $\|\mathbf{W}^\ell\|_2$  denotes computing the Frobenius norm of

$\mathbf{W}^\ell$ . During training/optimization with gradient ascent, we do not constrain the column norms of any of the weight matrices for any of the baseline models (as we do for the GNCNs) as we found that doing so worsened their generalization ability.

Furthermore, the number of total layers in the decoder for any model was set to be four – one output and one input layer with two hidden layers in between. The encoder was constrained to be the same – one input and one output layer with layers in between (in the case of the GVAE, CV-GVAE, and GAN-AE, the encoder’s output is technically split into two blocks, as described later). The sizes of the hidden layers were set such that the total number of learnable model weights were approximately equal across all baselines and GNCNs (maximum was 1,400,000 synapses), which means that all models were forced to have the same parameter complexity to avoid any unfair advantages that might come from over-parameterization.

**Regularized Auto-encoder (RAE):** The encoder  $\text{NN}_e$  is designed to be a feedforward network of  $L$  layers of neurons. Each layer is a nonlinear transformation of the one before it, where  $\hat{\mathbf{z}}^\ell = \phi^\ell(\mathbf{E}^\ell \cdot \hat{\mathbf{z}}^{\ell-1})$ . Like in the decoder,  $\phi^\ell$  is an activation function and  $\mathbf{E}^\ell$  is a set of tunable weights. In this paper, we chose  $\phi^\ell$  to be the linear rectifier, i.e.,  $\phi^\ell(v) = \max(0, v)$ . The bottom layer activation was chosen to be the logistic link, i.e.,  $\phi^0(\mathbf{z}) = 1/(1 + \exp(-\mathbf{z}))$ .

Note that in the RAE, the input to the decoder is now  $\mathbf{z} = \hat{\mathbf{z}}^L$ , i.e., the noise sample vector is set equal to top-most layer of neural activities of the encoder. The data log likelihood that the RAE optimizes is:

$$\psi = \sum_j \left( \mathbf{x}[j] \log \mathbf{z}^0[j] + (1 - \mathbf{x}[j]) \log(1 - \mathbf{z}^0[j]) \right) \quad (20)$$

where updates to each weight matrix  $\mathbf{E}^\ell$  (of the encoder) and  $\mathbf{W}^\ell$  (of the decoder) are updated by computing the relevant gradients  $\frac{\partial \psi}{\partial \mathbf{E}}$  and  $\frac{\partial \psi}{\partial \mathbf{W}}$ , respectively. The weight gradients are then used to update model parameters via gradient ascent.

**Gaussian Variational Auto-encoder (GVAE):** Instead of using an encoder to only produce a single value for  $\mathbf{z}$ , we could instead modify this network to produce the parameters of a distribution over  $\mathbf{z}$  instead. If we assume that this distribution is a multivariate Gaussian with a mean  $\mu_z$  and a diagonal covariance  $\sigma_z^2 = \Sigma_z \odot \mathbf{I}$ , we can then modify the RAE’s encoder function to instead be:  $(\mu_z, \sigma_z^2) = \text{NN}_e(\mathbf{x})$ . Specifically, the top-most layer of  $\text{NN}_e(\mathbf{x})$  is actually split into two separate output layers as follows:  $\mu_z = \mathbf{E}_\mu^L \cdot \mathbf{z}^{L-1}$  and  $\sigma_z^2 = \exp(\mathbf{E}_\sigma^L \cdot \mathbf{z}^{L-1})$  (this is also known as the variational autoencoder, or VAE [43]).  $\mathbf{E}_\mu^L$  is the tunable weight matrix for the mean and  $\mathbf{E}_\sigma^L$  is the tunable weight matrix for the variance. The data log likelihood for the GVAE is as follows:

$$\psi = \sum_j \left( \mathbf{x}[j] \log \mathbf{z}^0[j] + (1 - \mathbf{x}[j]) \log(1 - \mathbf{z}^0[j]) \right) - D_{KL} \left( q(\mathbf{z}|\mathbf{x}) || p(\mathbf{z}) \right) \quad (21)$$

where  $q(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mu_z, \sigma_z^2)$  (the Gaussian parameters produced by the encoder  $\text{NN}_e(\mathbf{x})$ ) and  $p(\mathbf{z}) = \mathcal{N}(\mu_p, \sigma_p^2)$  where  $\mu_p = 0$  and  $\sigma_p^2 = 1$  (an assumed unit Gaussian prior over  $\mathbf{z}$ ). The second term in the above objective is the Kullback-Leibler divergence  $D_{KL}$  between the distribution defined by the encoder and the assumed prior distribution. This term is meant to encourage the output distribution of the encoder to match a chosen prior distribution, acting as a powerful probabilistic regularizer over the model’s latent space. Note that this divergence term serves as a top-down pressure on the top-most layer of the encoder while the gradients that flow from the encoder (via the chain rule of calculus) act as a sort of bottom-up pressure. Note that since  $\text{NN}_e(\mathbf{x})$  is a distribution, the input to the decoder is, unlike the RAE, a sample of the encoder-controlled Gaussian, i.e.,  $\mathbf{z} = \mu_z + \sqrt{\sigma_z^2} \odot \epsilon$  where  $\epsilon \sim \mathcal{N}(0, 1)$ .

Gradients of the likelihood in Equation 21 are then taken with respect to all of the encoder and decoder parameters, including the new mean and variance encoder weights  $\mathbf{E}_\mu^L$  and  $\mathbf{E}_\sigma^L$ , which are subsequently updated using gradient ascent. All the other activation functions of the GVAE are set to be the linear rectifier, except for the output function  $\phi^0$  of the decoder, which, like the RAE, is set to be the logistic sigmoid.

**Constant-Variance Gaussian Variational Auto-encoder (CV-GVAE):** This model [26] is identical to the GVAE except that the variance parameters  $\sigma_z^2$  of the encoder are omitted and a fixed (non-learnable) value is chosen instead for the variance (meaning that the diagonal covariance is collapsed further to a single scalar). The exact value for this variance meta-parameter, for each benchmark, was chosen from the range  $[0.025, 1.0]$  by tuning performance to a held-out set of image samples.

**Generative Adversarial Network Autoencoder (GAN-AE):** This model, also referred to as an adversarial autoencoder [53], largely adheres to the architecture of the GVAE except that the second term, i.e., the Kullback-Leibler divergence term, in the data log likelihood is replaced with the adversarial objective normally used to train implicit density estimators like the generative adversarial network (GAN) [28]. As a result, we integrate a third feedforward network, i.e.,  $p_r = \text{NN}_d(\mathbf{z})$ , into the generative model (this module is also referred to as the discriminator). The discriminator is tasked with distinguishing whether an input vector comes from the desired prior distribution  $p(\mathbf{z})$  (set to be a unit Gaussian as in the GVAE) or comes from the encoder network  $\text{NN}_e(\mathbf{z})$  distribution. This task is posed as a binary classification problem, where a sample from the encoder  $\mathbf{z}_f \sim \mathcal{N}(\mu_z, \sigma_z^2)$  is assigned the label of  $c = 0$  (fake sample) and a sample drawn from the prior  $\mathbf{z}_r \sim \mathcal{N}(0, 1)$  is assigned a label of  $c = 1$  (real sample). These fake and real samples are fed through the discriminator which returns a scalar value for each, representing the probability

$p_r = p(c = 1|\mathbf{z})$ . This leads to the modified data log likelihood objective below:

$$\psi = \sum_j \left( \mathbf{x}[j] \log \mathbf{z}^0[j] + (1 - \mathbf{x}[j]) \log(1 - \mathbf{z}^0[j]) \right) + \left( \log(\text{NN}_d(\mathbf{z}_r)) + (1 - \log(\text{NN}_d(\mathbf{z}_f))) \right). \quad (22)$$

However, to update the weights of the GAN-AE, we do not compute partial derivatives of Equation 22 directly. Instead, following in line with the typical multi-step optimization of [28, 53], upon presentation of a sample or mini-batch of samples, we compute gradients with respect to  $\text{NN}_e(\mathbf{x})$ ,  $\text{NN}(\mathbf{z})$ , and  $\text{NN}_d(\mathbf{z})$  separately. Specifically, if we group all of the encoder weights under  $\Theta_{\text{NN}_e}$ , all of the decoder weights under  $\Theta_{\text{NN}}$ , and all of the discriminator weights  $\Theta_{\text{NN}_d}$ , then the gradients of the objective are computed in three separate but successive steps shown below:

$$\Delta_{\text{auto}} = \frac{\partial \sum_j \left( \mathbf{x}[j] \log \mathbf{z}^0[j] + (1 - \mathbf{x}[j]) \log(1 - \mathbf{z}^0[j]) \right)}{\partial (\Theta_{\text{NN}_e} \cup \Theta_{\text{NN}})} \quad (\text{Autoencoder gradients}) \quad (23)$$

$$\Delta_{\text{gen}} = \frac{\partial \left( 1 - \log(\text{NN}_d(\mathbf{z}_f)) \right)}{\partial \Theta_{\text{NN}}} \quad (\text{Generator gradients}) \quad (24)$$

$$\Delta_{\text{disc}} = \frac{\partial \left( \log(\text{NN}_d(\mathbf{z}_r)) + (1 - \log(\text{NN}_d(\mathbf{z}_f))) \right)}{\partial \Theta_{\text{NN}_d}} \quad (\text{Discriminator gradients}) \quad (25)$$

where the above gradient calculations are each followed by a separate gradient ascent update to their relevant target parameters, i.e.,  $\Delta_{\text{auto}}$  is used to update  $\Theta_{\text{NN}_e} \cup \Theta_{\text{NN}}$ ,  $\Delta_{\text{gen}}$  is used to update  $\Theta_{\text{NN}_e}$ , and  $\Delta_{\text{disc}}$  is used to update  $\Theta_{\text{NN}_e} \cup \Theta_{\text{NN}_d}$ .

The number of synaptic weights associated with the discriminator were included in the model’s total parameter count and had two hidden layers of linear rectifier units. Again, like the GVAE, the hidden layer functions  $\phi^\ell$  of the encoder and decoder were chosen to be the linear rectifier and the output  $\phi^0$  of the decoder was set to be the logistic sigmoid.

## Supplementary Note 8: Feature Analysis of Neural Generative Coding

The analysis we conducted on the GNCN-t2-L $\Sigma$ ’s intermediate representations involved, using the MNIST dataset, examining the generative synaptic weights for each layer of a trained GNCN-t2-L $\Sigma$  model. Specifically, when we viewed the weight vectors that conveyed predictions of state neurons in layer 1 to layer 0 (the output), we found that the features resembled rough strokes and digit components (of different orientations/translations). When we viewed the weight vectors relating neurons in layer 2 to layer 1 and layer 3 to layer 2, we found that they rather resembled neural selection “blueprints” or maps that seem to be used to select or trigger lower-level state neurons.

In Figure 7 of the main paper, we illustrate how these higher-level maps seem to interact with the low-level stroke visual features/dictionary. Based on our simple analysis, it appears that the GNCN-t2-L $\Sigma$  learns a type of multi-level command structure, where neurons in one level learn to turn off and turn on neurons in the neighboring level below them, further scaling those they choose to activate by an intensity coefficient. When we reach layer 1, the state neurons chosen from the command structure of layers 2 and 3, as well as their final produced intensity coefficients (ranging from  $[0, \infty)$  due to the fact that any layer of the GNCN-t2-L $\Sigma$  in this paper uses the linear rectifier activation function) resemble and work to produce a composition of low-level features that ultimately produce a complete object or digit. This means that the GNCN-t2-L $\Sigma$  (as well as the other predictive processing models like those of [22], i.e., the GNCN-t1- $\Sigma$ , and [68], i.e., the GNCN-t1, since they process sensory inputs in the same way as the GNCN-t2-L $\Sigma$ ) learns to compose and produce a weighted summation of low-level features akin to the results of sparse coding [59] driven by a complex, higher-level neural latent structure. In the “Output” column of Table 3 of the main paper, we empirically confirm this by summing up the top most highly-activated state neurons in layer  $\mathfrak{N}^1$  (multiplying each by its activation coefficient) – this simple super-position visually yields digits quite similar to the original one presented to the GNCN-t2-L $\Sigma$  model.

While this simple feature analysis is promising, we remark that future work should involve developing methodology to map an NGC model’s layerwise activities to those of actually brain activity (using fMRI data) or to a biological model such as HMAX, such as using the method proposed in [67].