scientific reports

Check for updates

OPEN

Investigating the intrinsic top-down dynamics of deep generative models

Lorenzo Tausani[1,2], Alberto Testolin[1,2]✉ & Marco Zorzi[1,3]✉

Hierarchical generative models can produce data samples based on the statistical structure of their training distribution. This capability can be linked to current theories in computational neuroscience, which propose that spontaneous brain activity at rest is the manifestation of top-down dynamics of generative models detached from action-perception cycles. A popular class of hierarchical generative models is that of Deep Belief Networks (DBNs), which are energy-based deep learning architectures that can learn multiple levels of representations in a completely unsupervised way exploiting Hebbian-like learning mechanisms. In this work, we study the generative dynamics of a recent extension of the DBN, the *iterative* DBN (iDBN), which more faithfully simulates neurocognitive development by jointly tuning the connection weights across all layers of the hierarchy. We characterize the number of states visited during top-down sampling and investigate whether the heterogeneity of visited attractors could be increased by initiating the generation process from biased hidden states. To this end, we train iDBN models on well-known datasets containing handwritten digits and pictures of human faces, and show that the ability to generate diverse data prototypes can be enhanced by initializing top-down sampling from "chimera states", which represent high-level features combining multiple abstract representations of the sensory data. Although the models are not always able to transition between all potential target states within a single-generation trajectory, the iDBN shows richer top-down dynamics in comparison to a shallow generative model (a single-layer Restricted Bolzamann Machine). We further show that the generated samples can be used to support continual learning through generative replay mechanisms. Our findings suggest that the top-down dynamics of hierarchical generative models is significantly influenced by the shape of the energy function, which depends both on the depth of the processing architecture and on the statistical structure of the sensory data.

In the past decade, cognitive neuroscience has witnessed a paradigm shift in understanding the brain as a predictive machine[1]. This theoretical framework is rooted in the idea that the brain is perpetually engaged in the construction of internal models of its surrounding environment, leveraging the interplay between sensory data and prior experience to guide adaptive behavior[2]. Therefore, a possible role for recurrent loops in the cerebral cortex could be to integrate bottom-up sensory observations with top-down contextual priors, which are encoded using multiple levels of representation in hierarchical generative models[3,4].

A key issue in modern neuroscience is understanding the properties and role of the so-called *spontaneous brain activity*, which is the neuronal activation recorded at rest when external stimuli are weak or absent[5]. Spontaneous cortical activity has been found to alternate between motifs defined by regional axonal projections, reflecting multiple modes of sensory processing[6]. Four different lines of evidence support the functional importance of this intrinsic activity[7]: (i) it accounts for most of the brain's energy consumption[8]; (ii) it shows sophisticated dynamics that organize into distinct spatio-temporal patterns[9]; (iii) it predicts individual differences in cognitive functions and learning[10]; (iv) it recapitulates task-evoked activity and seems to support off-line learning[11].

A recent theoretical proposal[7] is that spontaneous activity is the manifestation of top-down dynamics occurring in generative models, whose goal is to estimate the latent factors underlying the observed data distribution[2]. This idea entails a strong connection between spontaneous and task-related brain activity: when engaged in a task, the generative model prioritizes the maximization of accuracy in that specific task, while at rest it reproduces plausible task-related activation patterns, which can be used to estimate generic spatio-temporal priors summarizing a wide variety of tasks in a low-dimensional representation[7] or to support continual learning through replay mechanisms[12]. This also implies that during spontaneous activity the model

[1]Department of General Psychology and Padova Neuroscience Center, University of Padova, Padova, Italy. [2]Department of Mathematics, University of Padova, Padova, Italy. [3]IRCCS San Camillo Hospital, Venice, Italy. ✉email: alberto.testolin@unipd.it; marco.zorzi@unipd.it

should explore neuronal states similar to those experienced during task periods, for example by generating configurations analogous to those emerging from sensory perception. The idea that spontaneous brain activity is the manifestation of top-down generative dynamics detached from action-perception cycles[7] is also in agreement with modeling work suggesting that the brain at rest is in a state of maximum metastability[13], where periods of stability are interleaved by periods of instability that allow flexible transition to different stable states[14].

The main objective of the present work is to investigate whether the top-down dynamics emerging from hierarchical generative models implemented in deep neural networks[15] could capture some of the key aspects of spontaneous brain activity observed during resting conditions. In particular, we study whether top-down generation of plausible perceptual states could result in stable attractors and whether deep networks can explore multiple stable states starting from specific (and possibly biased) initial conditions. Our modeling approach is based on a popular class of generative neural networks called Deep Belief Networks (DBN)[16]. DBNs have been extensively used in computational neuroscience and cognitive science as models of brain function[17–22] due to the biological plausibility of their learning and inference schemes, which are based on Hebbian-like mechanisms[4] that can be linked to biophysical models of neuronal dynamics[23,24]. Notably, DBNs have also been recently implemented on memristive devices, demonstrating that this class of generative models can run on low-power neuromorphic hardware[25].

DBNs are usually built by stacking together multiple Restricted Boltzmann Machines (RBMs), which are undirected graphical models formed by two layers of symmetrically connected neurons. Visible neurons encode sensory data (e.g., pixels in an image), while hidden neurons discover latent features through unsupervised generative learning[26]. This class of generative neural networks belongs to the family of energy-based models: plausible configurations of network states can be generated by alternating Gibbs sampling, where the activations of hidden neurons are sampled conditional to the activations of visible neurons, and vice versa[4]. Learning in DBNs has traditionally relied on a greedy, layer-wise training approach: the connection weights of layer n are changed only after layer $n-1$ has been fully trained. Although efficient from a computational perspective, this learning modality is clearly implausible from the perspective of cognitive (neuro)science, because cortical circuits develop holistically during learning. In this work, we therefore exploit a recently introduced iterative learning scheme[27], which allows tuning the entire hierarchy of connections in a DBN using a variant of the original contrastive divergence learning algorithm[26]. In analogy to the fast feed-forward sweep observed in cortical circuits, where neuronal activity is rapidly routed to a large number of visual areas after stimulus presentation[28,29], in the iterative DBN (henceforth iDBN) the sensory input is immediately propagated throughout the entire processing hierarchy. Concurrently with the fast feed-forward sweep, top-down generative connections are locally used to reconstruct data representations at each level of the hierarchy, mimicking the kind of processing supported by recurrent and horizontal connections within cortical areas[28,30].

Despite their popularity as neurocomputational models, DBNs (or the iDBN version used here) have not yet been used to simulate the spontaneous generation of sensory patterns due to the intrinsic difficulty of sampling stable network states in a top-down, unconstrained fashion. In theory, one could randomly initialize the visible neurons and then perform alternating Gibbs sampling over the entire network hierarchy until the process converges toward a steady-state distribution (regardless of the initial starting point). However, in practice, reaching this steady state may take an impractically long time[16]. Possible strategies to mitigate these issues could be to "warm-start" the sampling process by initializing bottom-up the network state by providing an input sample, or to constrain top-down generation by biasing the hidden representations at the deepest layer of the hierarchy using categorical information associated with the data patterns[16,31]. Biased hidden representations have usually been created using two main approaches. One possibility is to add a multimodal RBM to the top of the network hierarchy, which is jointly trained using as input both the internal representation of the input patterns learned by the DBN and the corresponding categorical labels encoded using one-hot vectors[16]. After learning, the label neurons can be clamped to a certain state, and the top RBM settles to equilibrium, thus recovering the internal representation of the given data class. The generative connections of the DBN can then be used to sample an image on the visible layer in a single top-down pass: the generated image can be thought of as the model prototype for the corresponding class representation. A similar result can be achieved more efficiently by directly mapping the internal representation to the class label through a linear projection[31]. These constrained sampling strategies are particularly useful when the distribution being learned by the model is complex and has multiple modes; however, they usually drive the model towards a single stable attractor and do not allow to flexibly explore different states.

Here, we propose a novel way to initialize the biased hidden state, which consists of blending together the internal representations of multiple classes, with the goal of starting the generation process from an intermediate point between different attractor basins: we call this method *chimera biasing*. We trained iDBNs on two popular datasets commonly used in computer vision and cognitive modeling research: the MNIST dataset[32], which contains images of handwritten digits, and the CelebA dataset[33], which contains images of human faces. The learning process was fully unsupervised: the objective was to build a hierarchical generative model that captures the training distribution by discovering increasingly more complex visual features in hidden representations[31]. Following standard practice, for MNIST we adopted a iDBN architecture with three hidden layers[16]. We then replicated a similar architecture for the CelebA dataset (see Methods). We propose an original method to quantify the flexibility and heterogeneity of state space exploration, and we test the different top-down generation schemes to initialize the initial hidden state of the iDBN.

For both datasets, we show that starting the top-down generation from chimera hidden states leads to a better exploration of the attractor landscape, and we also show that deeper architectures implementing hierarchical generative models incorporate a more versatile sampling dynamics, which allows to visit a higher number of perceptual states compared to single-layer models[34]. In a series of continual learning simulations, we further demonstrate that the proposed top-down sampling scheme can be readily used to support interleaved learning[12],

which allows incorporating information from a new distribution into the generative model without disrupting previously acquired knowledge by exploiting generative replay mechanisms[35].

Materials and methods

Datasets for generative tasks

The MNIST dataset consists of 28x28 grayscale images of handwritten digits from 0 to 9. It is subdivided into a training set containing 60000 samples and a testing set of 10000 samples. CelebA contains RGB images of 202599 celebrity faces, each annotated with 40 binary attributes. Following previous work, pre-processing included resizing images to dimensions of 64x64 pixels, converting them to grayscale, and binarizing them using the Sauvola-Pietikainen algorithm[36]. Images were assigned to four mutually exclusive categories evaluating the logical conjunction between the attributes *male/female* and *smiling/not smiling* (i.e., female not smiling, male not smiling, female smiling, male smiling). To avoid class imbalance, each class was under-sampled to ensure that each category had the same number of examples (27256 items per category).

Deep belief networks

Deep Belief Networks (DBNs) are energy-based generative models composed of several layers of stochastic binary neurons (see Fig. 1a). They are built by stacking together Restricted Boltzmann Machines (RBMs), which are undirected graphical models with a bipartite structure that enables efficient probabilistic inference and learning[26]. A RBM consists of two layers, named *visible* and *hidden* layers, respectively. The two layers are fully connected by the weight matrix w. Input patterns (e.g., the pixels of an image) are fed to the model by clamping them to the visible neurons, while hidden neurons act as latent factors that are used to compactly represent the main statistical features of the data distribution. When neurons in one layer are clamped to a particular state (e.g., the visible neurons v are clamped to one training pattern), the activation probability of all neurons in the hidden layer h can be efficiently computed in one parallel step using the sigmoid activation function:

$$P(h|v) = \prod_i P(h_i|v) \tag{1}$$

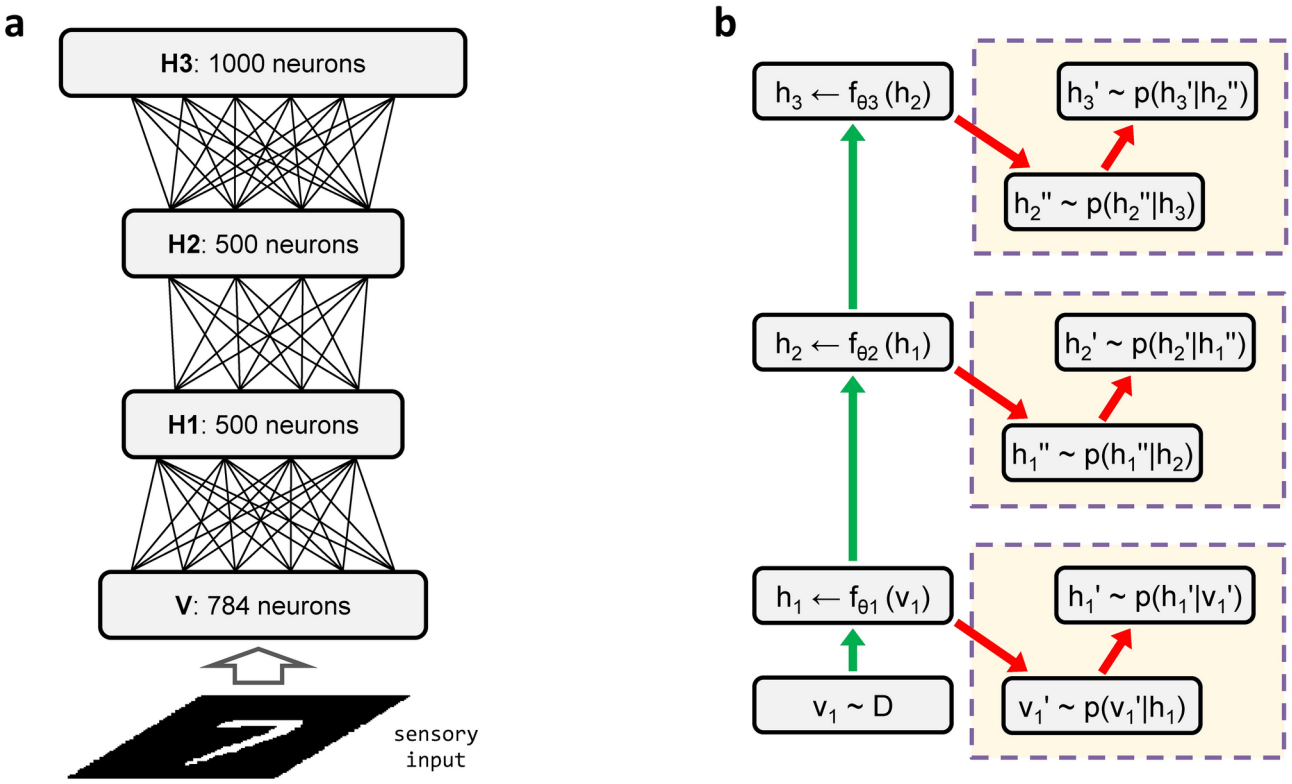

Fig. 1. Graphical representation of the deep belief network and the iterative version of the contrastive divergence learning algorithm used in our study (reprinted from[27]). (**a**) The deep belief network trained on the MNIST dataset is composed by three hidden layers (H1, H2 and H3) and one visible layer (V), which is clamped on the sensory input. (**b**) In the iDBN learning algorithm, green arrows represent bottom-up recognition connections, while red arrows represent top-down generative processing. Yellow boxes enclose local computations. In such iterative learning scheme, input signals sampled from the training dataset D are immediately propagated through the entire deep network, and top-down processing is performed locally at each layer to jointly learn all connection weights..

where:

$$P(h_i = 1|v) = \frac{1}{1 + e^{-\sum_j w_{ij} v_j}} \tag{2}$$

RBMs can be efficiently trained in a completely unsupervised way using the contrastive divergence algorithm[26], which performs gradient descent on the empirical negative log-likelihood of the training data. The derivative of the log-likelihood of a training example with respect to the weight w_{ij} can be analytically defined as:

$$\frac{\partial \log P(v, h; W)}{\partial w_{ij}} = \langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{model} \tag{3}$$

where the angle brackets are used to denote expectations under the distribution specified by the empirical data distribution (positive phase) and the model distribution (negative phase).

DBNs are usually created by training a sequence of RBMs following a greedy layer-wise approach[16]. As discussed in the Introduction, here we exploit a recent and more neurobiologically plausible iterative learning algorithm (iterative DBN or iDBN), which jointly tunes all weights of the network[27]. In the iDBN, the activations of all hidden layers in the deep network are created following each sensory experience, by sequentially propagating the sensory input across the entire processing hierarchy (green arrows in Fig. 1b). This process mimics the fast feed-forward sweep observed in cortical circuits, where neuronal activity is rapidly routed to a large number of visual areas after stimulus presentation[28,29]. Concurrently with the fast feed-forward sweep, top-down generative connections are locally used to reconstruct the data representations at each level of the hierarchy (red arrows in Fig. 1b), mimicking the kind of processing supported by recurrent and horizontal connections within cortical areas[30].

We also trained a single-layer RBM (using the standard contrastive divergence algorithm) as a baseline model to assess whether the iDBN's deeper architecture entails a richer generative dynamics.

Model architecture and learning hyperparameters

Model architecture and learning hyperparameters were chosen following previous work[16,31]. In our simulations, we trained different iDBN architectures for each dataset. Both iDBNs consisted of a visible layer and three hidden layers, with the visible layer comprising a number of neurons equivalent to the number of pixels in each dataset sample (that is, 784 for MNIST and 4096 for CelebA). The MNIST model architecture (see Fig. 1a) consisted of hidden layers with 500, 500, and 1000 neurons, respectively (we only reduced the size of the top layer to speed up training). The architecture of the CelebA model was defined by matching the ratio of hidden to visible neurons for each layer in the MNIST model, resulting in hidden layers of dimensions 2500, 2500, and 5250. The single-layer RBM models used as baselines consisted of a visible layer and a hidden layer which featured the same dimensionality of the last hidden layer of the deeper networks.

Each model was trained using 1-step contrastive divergence with a learning rate $\eta = 0.01$. A momentum term γ was included to accelerate learning: γ was set to 0.5 until the fifth iteration of training and was then updated to 0.9. To regularize the training procedure, a weight decay term equal to 0.0001 was also included. The initial connection weights were sampled from a normal distribution with zero mean and standard deviation equal to 0.01. Training was implemented using a mini-batch scheme with batch size equal to 128. The RBM models were trained using 1-step contrastive divergence with similar learning parameters, which are described in[34].

Top-down sampling from iDBNs

During top-down generation, hidden neurons were binarized via Bernoulli sampling, while real values were retained in the visible layer to generate smoother images. Sample generation was carried out according to two possible schemes: label biasing[31] and chimera methods. The latter implement a novel generation procedure that we propose to maximize state exploration by initiating the process from intermediate states between different data classes. We also implemented a random initialization method as a control condition for the chimera scheme (see below). In all generation schemes, after initializing the top hidden layer, activation is propagated through the hierarchy in a top-down fashion, until the visible layer is reached. This constitutes one *generation step*, and the resulting visible state corresponds to one *generated sample*. The hidden state of the next generation step is instantiated by propagating the generated sample from the previous step in a bottom-up fashion (see panel a in Fig. 2 for a graphical representation). In the generative experiments, 100 samples per each class / class combination were produced, and generativity was quantified as the average number of states explored within a single generation round of 100 steps (see Fig. S1).

Label biasing. The label biasing method aims to map the one-hot encoded class label to a corresponding internal representation through a direct linear projection (see panel b in Fig. 2). This is accomplished by inverting a linear classifier that maps hidden representations to class labels. Given a set of *n* data patterns $D = \{D_1, D_2, \ldots, D_n\}$ and the corresponding one-hot encoded labels $L = \{L_1, L_2, \ldots, L_n\}$, the weight matrix W_{Lc} of the inverted classifier is obtained analytically with the equation 4:

$$W_{Lc} = R_{H_n} L^{+} \tag{4}$$

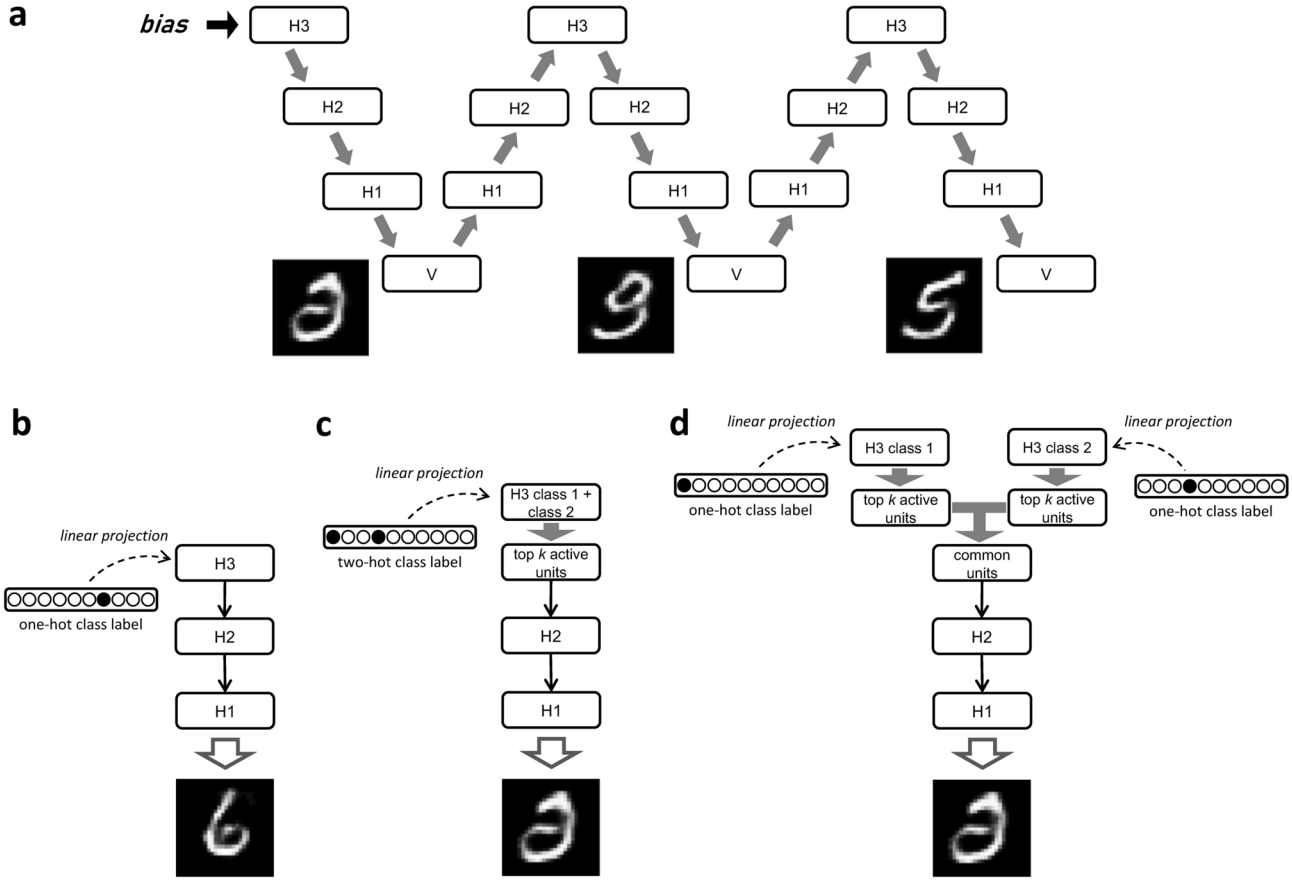

Fig. 2. Schematic representation of the proposed top-down generation framework. (**a**) Top-down generation is implemented as an iterative sampling process: generation unfolds over a sequence of time steps, and an initial *bias* on the top-level hidden state (H3) is imposed using one of the methods described in the remaining panels. (**b**) The label biasing method projects a one-hot class label into the corresponding iDBN hidden state by inverting a linear classifier. (**c**) The double label chimera biasing method creates a mixed hidden representation by using a two-hot class label, and then binarizes the hidden state by selecting only the *k* most active neurons. (**d**) The intersection chimera biasing method creates a mixed hidden representation by first creating two separate representations using label biasing, which are then binarized by selecting only the *k* most active neurons, and finally merged into a single hidden state by considering the set of most active neurons in common.

where L^+ is the Moore-Penrose pseudoinverse of the one-hot encoded label matrix and R_{H_n} are the corresponding hidden representations in the last layer of the iDBN. R_{H_n} is obtained by feed-forward propagation of each of the *n* data patterns D_i. The top-down generation can then be initiated starting from a hidden vector H_n^{LB} that is obtained by multiplying the inverted weight matrix W_{Lc} with the one-hot encoded label of interest L_i:

$$H_n^{LB} = W_{Lc} L_i \tag{5}$$

H_n^{LB} is propagated top-down to produce the first generated sample V^1, which in turn is propagated bottom-up to the last layer H_n, thus forming a new hidden state H_n^2. The same alternation of top-down and bottom-up propagation is performed for each step of generation.

Chimera states. Chimera state methods aim to initiate the generation from hidden states away from attractor basins, making them more prone to flexibly explore the state space. This could be obtained by initializing the generation process from a hidden state that mixes the representations of two different classes. We characterize two possible variants of chimera biasing: the *double label biasing* method (C_{2LB}) and the *intersection* method(C_{int}). Both methods are based on the observation that the distribution of hidden neuron activation in label biasing states for each class is right-skewed, with a small number of highly active neurons that are characteristic of that state (see Fig. S2). In the *double label biasing* method (see panel c in Fig. 2), the label biasing procedure is performed with a label vector L_i with both class indexes active (i.e., two neurons are set to 1, thus forming a "two-hot" class label representation). The resulting hidden representations $H_n^{C_{2LB}}$ in the top iDBN layer are then binarized according to their activation rank: only the top *k* most active neurons are set to 1, while the

others are set to 0. Given that the percentage of active hidden neurons remained fairly stable throughout the label biasing generation process, we decided to set k equal to the average number of active hidden neurons in the first step of generation. The top-down generation then proceeds in the same way as for the simple label biasing method. In the *intersection* method (see panel d in Fig. 2), chimera states between two distinct classes are obtained by first creating two distinct hidden representations, each corresponding to a different class label, which are binarized by selecting the top k most active neurons. The chimera hidden representation $H_n^{C_{int}}$ is then obtained by setting to one only the common set of active neurons shared by the two representations. The chimera hidden state is finally used to initiate the generation process, which proceeds in the same way as in the other biasing methods.

Random initialization. As a baseline condition for the chimera states we also tested a *random initialization* scheme that controlled for the potential effect of sparsity of the initial hidden representation. In this case the generation process was initialized from a hidden state with k randomly selected active neurons (i.e., $h_i = 1$, where i belongs to a set of k indexes sampled uniformly from the set of indexes of the final layer $I = \{1, 2, .., n\}$). k was derived from the chimera method. All other neurons of the hidden representation were set to 0.

Characterization of generative dynamics

State classifiers

To determine whether the generated samples corresponded to well-formed perceptual states, we exploited as a state classifier two convolutional neural networks trained on the original datasets to recognize digit/face classes in a supervised way. We used VGG-16[37] for MNIST and ResNet18[38] for CelebA. The VGG-16 architecture consisted of 3 fully connected layers on top of 4 VGG block units, with a softmax layer for the final classification. The ResNet18 architecture consisted of an initial convolutional and max-pooling layer, followed by 4 sets of 2 residual blocks each. At the top of the residual blocks, a fully connected layer was followed by a softmax layer, which allowed for image classification. The ResNet model was pre-trained using the ImageNet dataset[39] and fine-tuned on CelebA. Both models were trained for 20 epochs with mini-batches of 64 examples, monitoring generalization performance on a separate validation set. VGG-16 parameters were optimized using stochastic gradient descent (learning rate $\eta = 0.1$), while Adam was adopted for the ResNet model ($\eta = 0.001$). At the end of the training phase, both classifiers achieved very high validation accuracy (VGG-16: 99.3%, ResNet18: 97.8%).

Since the iDBN trained on the MNIST images often generated patterns that were not recognizable as any digit, the VGG model was trained to also identify poorly formed patterns as *non-digits*. For this reason, its training set also included ∼60000 examples of non-digit patterns. Of these, 10% were scrambled digit images, while the remaining patterns consisted of training set images with a random number of adjacent active pixels (i.e. not black, intensity > 0) set to 0. We chose to use this method of producing non-digits based on empirical observations of cases where iDBN generation resulted in shapes that were unidentifiable as digits by humans. A class associated with unidentifiable patterns was not implemented for the ResNet model, as the iDBN trained on CelebA almost never generated unrecognizable patterns.

Generativity metrics

We characterized the generative dynamics of the models using quantitative metrics that aim to capture the diversity and stability of state exploration during top-down generation. The *number of unique states* visited in a single generation trajectory of 100 steps was used to quantify the heterogeneity of state exploration. This metric did not include non-digit states identified for the MNIST model. The *number of transitions* occurring in each generation process was calculated to quantify the dynamism of state exploration. We defined the transition as a change in sample classification from one generation step to the following one. With the same aim, a transition matrix was inferred from all state transitions in the generation procedure (i.e. considering all samples and all their generation steps). Each entry of the matrix indicates the probability of transitioning from one class to another one, and was computed by counting all transitions from that state to the other, divided by the total number of transitions from that state. Each measure is reported with the associated standard error of the mean.

Spatio-temporal dynamics of hidden states during generation

We used a population analysis[40] of neural activity to analyze the spatio-temporal dynamics of network states across multiple generation runs triggered by each of the different initialization methods in the MNIST iDBN model. We recorded the activity of hidden neurons for 100 generation steps obtained from label biasing ($n = 10$, all digits), chimera states obtained with the intersection method ($n = 45$, all chimeras), and random initializations ($n = 10$) with k set to be equal to the number of active hidden neurons of each of the 10 most effective chimeras (i.e., yielding the largest number of visited states). For each initialization condition and method (e.g., label "3"), we collected 100 runs and computed the average hidden activations (i.e., an average trial). We then used Principal Component Analysis (PCA) to reduce the dimensionality of data and project them into the first two (or three) principal components (PCs). This allowed us to trace the spatio-temporal trajectories of network states in a low dimensional space. In order to assess how non-digit states clustered in the latent space, we also computed the mode of the state visited during each generation step.

Continual learning simulations

In the continual learning experiments, the iDBN was initially trained on the MNIST dataset, and subsequently re-trained on two different datasets: EMNIST, a dataset composed of handwritten letters, and Fashion-MNIST (fMNIST), which contains grayscale images of 10 categories of fashion items (t-shirts/tops, trousers, pullovers, dresses, coats, sandals, shirts, sneakers, bags, and ankle boots). To maintain the number of output classes, for the EMNIST dataset we selected the examples from the 10 central uppercase letters of the English alphabet (i.e.,

from H to Q). This subset of letters was chosen because, among the subsets tested, it was the one that resulted in the highest drop in read-out accuracy on the MNIST dataset during continual learning.

In both cases, the retraining datasets were composed of 156 batches of 128 elements each, for a total of 19968 examples. In the *sequential learning* condition, the retraining dataset only contained samples from the new datasets (i.e., EMNIST or fMNIST), while in the *interleaved learning* condition the retraining was performed using a dataset composed of half new data (EMNIST/fMNIST) and half MNIST data (*experience replay*) or samples produced by the iDBN model (*generative replay*). The generated samples were obtained by performing either chimera biasing or label biasing on the iDBN trained on MNIST data for 100 generation steps. We generated chimera biasing samples using the intersection method, as it systematically led to a higher number of visited states during sample generation. Samples were sorted from 100 generation chains for label biasing (10 for each digit) and 3 generation chains for each of the 45 unique chimera digit combinations for chimera biasing. The quality of continual learning was assessed by training linear read-outs on the deepest hidden layer of the iDBN, with the goal of decoding the new data (EMNIST/FMINST) at every epoch of retraining, and by testing the read-out classifier trained on the MNIST data at the end of the initial training phase.

To investigate the impact of the biasing method on the results of the continual learning experiment, we also performed generative replay using images generated with the random initialization of the hidden state. In this case, images were sampled from 1000 generation chains with random initialization. The value of the hyperparameter *k* was set to be equal to the average number of active neurons between the chimera states ($k = 63$, $n = 45$) rounded to the nearest integer. The neurons active were randomly sampled for each generation chain.

Results

As a preliminary analysis, we evaluated whether iDBNs can be used in a bottom-up sensory-driven modality to support recognition of sensory patterns. To this end, we trained a linear read-out classifier to classify the top-level internal representations of the iDBN into the corresponding categories. The classifier achieved a high decoding accuracy for both datasets (MNIST: 96.8%; CelebA: 79.8%), which is in line with previous studies[27,31] and with the notion that deep networks learn increasingly more disentangled representations of sensory manifolds[41].

Generative dynamics induced by the sampling initialization schemes

We applied to the iDBN the different top-down generative sampling schemes. In all cases, the generation starts from a *biased* hidden state in the top layer of the hierarchy, which sequentially drives the activation of the layers below until the visible layer is reached, with the goal of forming a plausible image pattern in the sensory space. The activation of visible neurons is then propagated up through the hierarchy of hidden layers, and down again. This alternation of top-down and bottom-up passes is repeated multiple times (see Fig. 2). In the first step of top-down generation, all sampling methods result in well-formed visual representations in the perceptual layer (top row samples in Fig. 3). The images produced by the iDBN resemble class prototypes associated with the category imposed by the biasing mechanism: in the case of the label biasing method, the initial states correspond to the chosen category, while in the case of the chimera methods (columns highlighted in red), the images resemble a mixture of the two activated categories. Some categories (i.e., digits 0 and 1) are associated with stronger attractors, and thus the visible pattern remains fairly stable throughout the generation sequence (Fig. S3). Other categories tend to fade away and collapse into null states, while others allow to more flexibly explore plausible visible configurations by dynamically switching between different perceptual categories. Interestingly,

Fig. 3. Examples of sample generation from MNIST and CelebA over representative sampling steps. For both datasets, generations from intermediate chimera states (red boxes) evidence a richer generation trajectory with respect to label biasing (other columns). For label biasing, a generated sample is shown for each class of the datasets (MNIST: digits from 0 to 9; CelebA: female not smiling, male not smiling, female smiling, male smiling). For chimera biasing, the generated MNIST samples represent the intersections $\{0, 3\}$ (left, obtained with C_{2LB} method) and $\{2, 9\}$ (right, obtained with C_{int} method), while for CelebA both samples represent the intersection {male smiling, male not smiling} (left: obtained with C_{2LB} method, right: obtained with C_{int} method).

the chimera biasing scheme, and in particular the intersection chimera method, seems to bias the hidden representations toward more metastable states, allowing for a more heterogeneous exploration of plausible perceptual configurations.

We quantified the number of plausible visible states visited during each generation sequence by training a supervised convolutional network to categorize the images produced by the model (see Methods). This allowed us to study whether the proposed biasing methods are characterized by different generative dynamics and whether changes in model architecture could lead to notable differences in state exploration. It turns out that the sampling dynamics of hierarchical generative models is indeed richer compared to that of their shallow counterpart, implemented as single-layer Restricted Boltzmann Machine (RBM). As shown in Fig. 4a, deep networks (solid lines) explored a significantly higher number of states compared to shallow networks (dashed lines), and this difference was clearly observable for all biasing methods. Furthermore, these quantitative analyses confirm that the chimera methods are more effective in driving the dynamics toward a more heterogeneous exploration of plausible sensory states, especially for the MNIST dataset (blue lines). The two chimera methods yield similar state exploration in the CelebA dataset, whereas in MNIST the chimera intersection method achieves superior results.

The richer generative dynamic of deep networks with respect to single-layer RBMs can be further explored by comparing the probabilities of state transitions, as reported in the transition matrices of Fig. 4b,c. Both panels refer to label biasing generation on the MNIST dataset. The RBM transition matrix (panel b) shows higher probabilities of same-state transitions (percentages on the diagonal) compared to the iDBN, being closer to the static condition where only same-state transitions occur (i.e., a diagonal transition matrix with $P(S_i \rightarrow S_i) = 1$; Euclidean distance $Static - RBM = 0.526$, Euclidean distance $Static - iDBN = 0.811$). Furthermore, transitions between different digits have on average higher probabilities for the iDBN compared to the RBM (average between-digit transition probability for the iDBN: 0.016 ± 0.002, average between-digit transition probability for the RBM: 0.010 ± 0.002).

Population analysis of hidden neurons' activity during generation

We further investigated the generative dynamics of iDBNs using a population analysis of the neural activity of hidden neurons, as typically used in neurophysiological studies[40]. This allowed us to trace the spatio-temporal trajectory of network states in a low dimensional latent space (following PCA, see Method) as a function of the different initialization schemes. The first two principal components accounted for a large portion of the total variance (69.76%). The resulting trajectories considering the activity of all hidden neurons are illustrated in Fig. 5. Qualitatively similar results were obtained when considering each hidden layer separately (see Fig. S5) or when also considering the third principal component to track the trajectories in a 3D latent space (Fig. S6, variance explained = 84.02%). As can be seen in Fig. 5a, the trajectories of label biasing trials are highly heterogeneous, with some digits spanning a large portion of the latent space (e.g. label bias "8"), while others (in particular, "0" and "1") are confined in a small subspace. Short trajectories are of interest because they can be linked to strong attractors, as evidenced by the transition matrix (Fig. 4c). Dark shading in Fig. 5 indicates generation steps associated with non-digit modes. For label biasing, non-digit states cluster in the lower left corner of the latent space, corresponding to a region where multiple trajectories (e.g. label bias "4" and "5") converge. In general, the trajectories for all digits except "0" converge towards negative values of both PC1 and PC2, where the non-digit states are also located. Figure 5b illustrates the behavior of the 10 chimera states (intersection method) that had the highest average number of states visited (see Fig. 4a). Non-digit states of these trials also cluster in the same region of the latent space as for label biasing, supporting the idea that non-digit states act as attractors. However, the trajectories of chimera trials are longer and initiate in a region with extreme positive values of PC1. This likely results in richer generation dynamics that avoid non-digit states. Figure 5c illustrates the trajectories of random initialization trials, with each trial matching one of the 10 chimera states in terms of number of active neurons. All trajectories remain confined in the lower left quadrant of the latent space and rapidly converge to

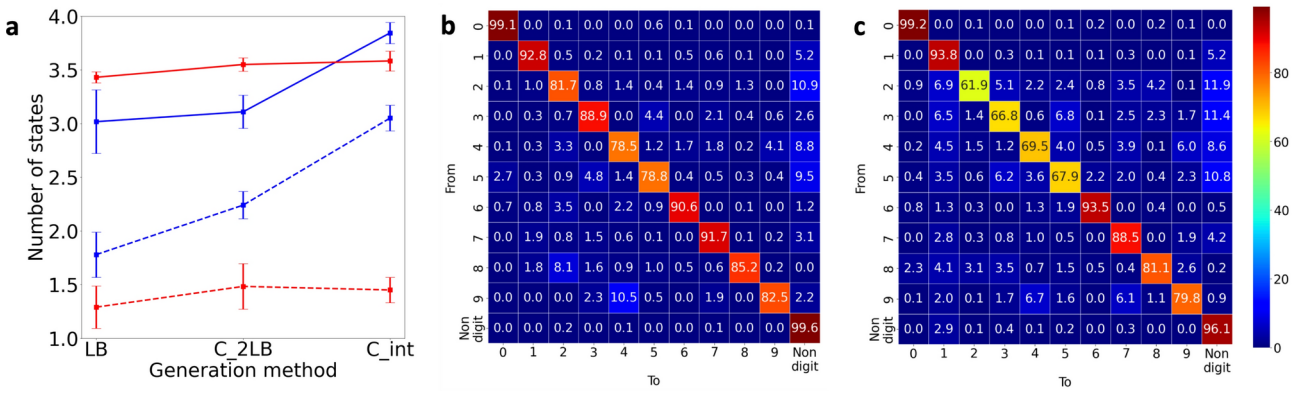

Fig. 4. Generativity metrics. (**a**) Number of visited states by iDBNs (solid lines) and RBMs (dashed lines) for the MINST (blue) and CelebA (red) datasets. (**b**) RBM transition matrix for the MNIST dataset. (**c**) iDBN transition matrix for the MNIST dataset.

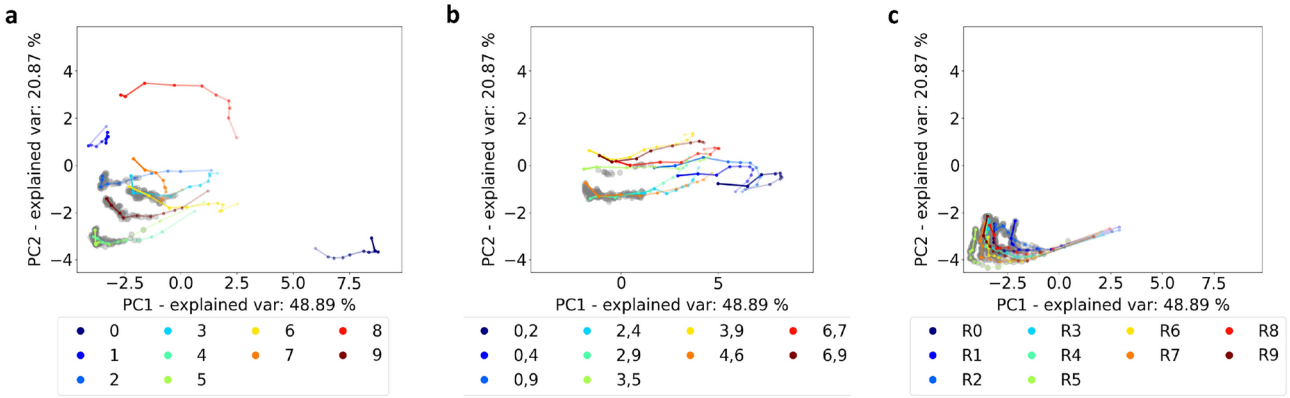

Fig. 5. Population analysis of hidden neurons' activity in the iDBN trained on MNIST data (considering all hidden neurons in the network). Spatio-temporal trajectories of hidden states are projected onto the first two principal components. The different initialization conditions are indicated with the color code reported in the legend. Each trajectory is displayed using 10 evenly spaced time-points in logarithmic scale (across the 100 generation steps), with time represented by increasing color saturation. Dark shading around points indicates steps that correspond to non-digit states. (**a**) Label biasing trajectories for all 10 digits. (**b**) Chimera biasing trajectories for the best 10 chimeras. (**c**) Trajectories of 10 random initializations with sparisty matching the chimera states..

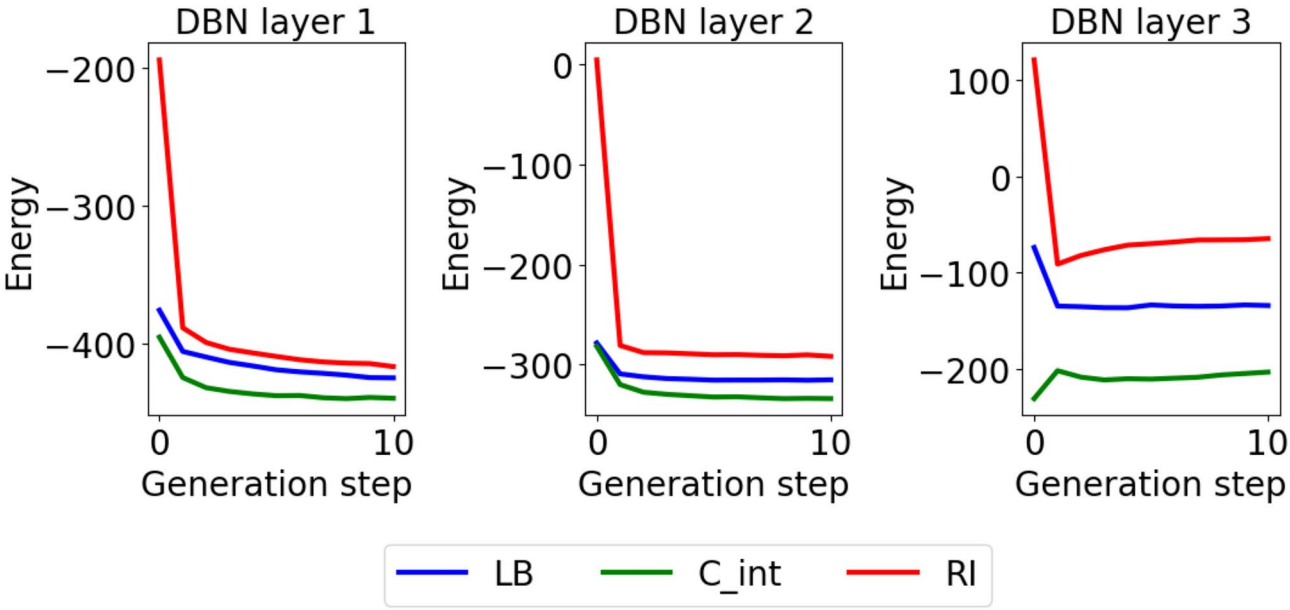

Fig. 6. Network energy during top-down generation. The energy during the first 10 generation steps is plotted separately for each hidden layer and initialization method (LB = label biasing; C_int = chimera intersection method); RI = random initialization; ..

non-digit states. This demonstrates that effective exploration of the network state space obtained with chimera initialization is not due to the cardinality of active neurons.

Overall, our analysis of the spatio-temporal trajectories of network states suggests that both label and chimera biasing are more effective than random initialization, and that chimera is superior to label biasing in inducing richer generative dynamics that tends to avoid getting trapped into a specific attractor or degenerate into non-digit states. This contention is supported by an analysis of how the network energy changes over timesteps during generation (Fig. 6). As expected, the average energy decreases over time in all conditions, but it starts and remains much higher for random initialization. Conversely, chimera initialization is the condition that guarantees the lowest energy across all hidden layers throughout the generation run. The energy difference is particularly marked in the third hidden layer, which is the layer that is directly affected by the biasing methods.

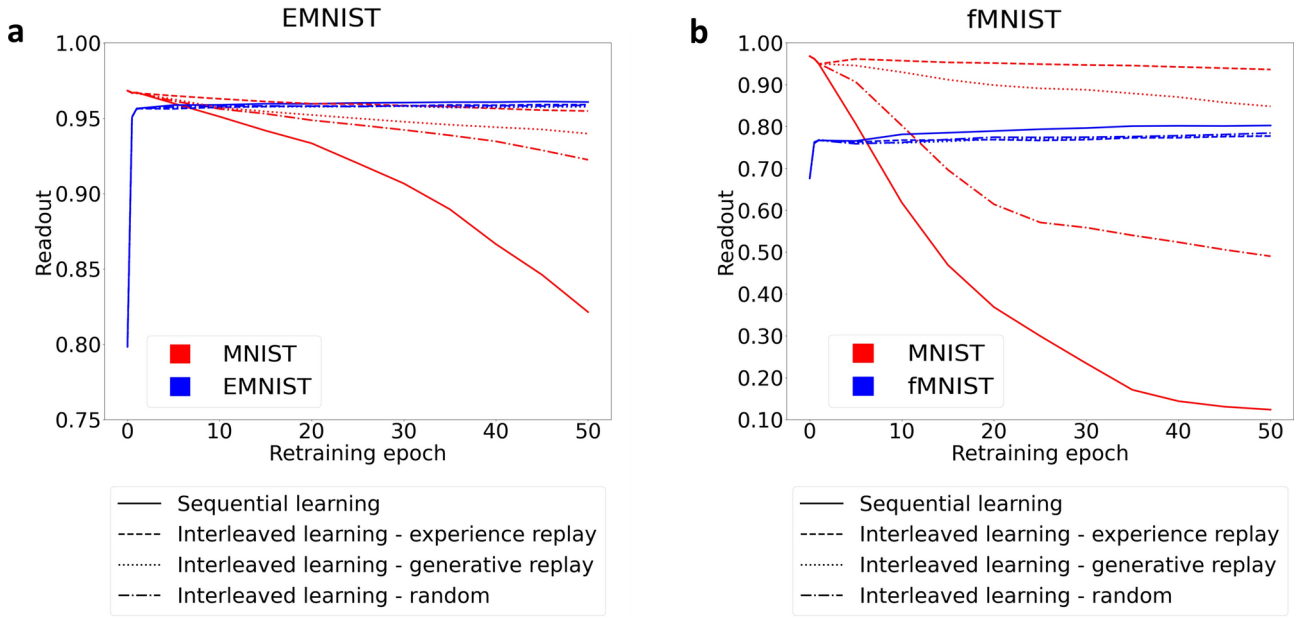

Fig. 7. Read-out accuracy in the continual learning scenarios implemented for the EMNIST (**a**) and fMNIST (**b**) datasets using the chimera biasing method. In both cases, the sequential learning regimen is strongly affected by catastrophic forgetting, as highlighted by the drop in read-out accuracy observed for the original MNIST distribution (solid red curves). Interleaved learning allows to incorporate information from the new distribution (EMNIST or fMNIST) while preserving previous knowledge, as highlighted by the preserved read-out accuracy for the original MNIST distribution. The generative replay method (dotted curves) allows to achieve high decoding accuracy on the new distribution, at the same time preserving almost the same accuracy as the experience replay method (dashed curves) in the original distribution. Interleaved learning is not effective when using samples generated from random initialization (dashed-dotted curves)..

Top-down sampling for continual learning through generative replay

Finally, we investigated whether our top-down sampling schemes could be effectively used to support continual learning. Indeed, artificial neural networks are known to be susceptible to a phenomenon called "catastrophic forgetting", in which knowledge gathered during initial learning phases is progressively lost when learning subsequent datasets and/or tasks[42]. This is in contrast to biological learning, since the acquisition of new information does not necessarily entail the loss of previously acquired knowledge. It has been proposed that continual learning in the brain could be partially supported by interleaved learning, allowing the integration of new information using replay mechanisms[12,35]. Previous work has shown that interleaved learning can be simulated in iDBNs using data samples from the previous distribution[27]: this would correspond to an *experience replay* mechanism, which requires to store (somewhere) the entire set of training patterns from the previous distribution. Here, we extend this framework by implementing a straightforward and general method that has been proposed to instantiate a *generative replay* mechanism[43], which consists of interleaving the new data samples with the data samples generated according to the previously learned generative model, thus eliminating the requirement of storing all previous data patterns.

In the continual learning simulations, the iDBN was initially trained on the MNIST dataset, followed by subsequent retraining on the EMNIST[44] or Fashion-MNIST (fMNIST)[45] datasets. We chose these two alternative datasets to investigate whether the distance of the original dataset from the retraining dataset could influence the preservation of internal representations during continual learning. Indeed, it is reasonable to assume that performing continual learning on a similar dataset (for example, EMNIST) allows for the "recycling" of features previously extracted by the model, resulting in a slower rate of forgetting compared to a dataset that shares fewer of these features (like fMNIST). Figure 7 illustrates how the decoding accuracy of the original dataset (MNIST, in red) changes during retraining on the new datasets (EMNIST/fMNIST, in blue). The graphs show that for both the retraining datasets the use of generative replay allows to significantly mitigate catastrophic forgetting, which instead occurs in the sequential learning regimen. The use of original data (experience replay) guarantees better preservation of information from the previous training distribution, but this comes at the cost of having to store all previous training patterns. The use of data patterns sampled from the iDBN (generative replay) still allows to mitigate catastrophic forgetting, both using chimera biasing (Fig. 7) and, to a lesser extent, also using label biasing (Fig. S4). As expected, catastrophic forgetting is more pronounced for the fMNIST dataset, whose features are less related to MNIST, compared to the EMNIST dataset. However, the relative decrease in read-out accuracy obtained with generative replay at the end of the retraining process is similar for the two datasets, since the ratio between the accuracy drop for generative replay with respect to sequential learning is almost the same (0.11 for EMNIST, 0.10 for fMNIST). Generative replay with random initialization is also shown as a baseline condition, with the sparsity of the initial activation (i.e., bias) matched to that of the chimera

intersection method. Note that initializing with a sparse representation might be beneficial if the different categories (or tasks) are internally represented through partially non-overlapping patterns of activation. Indeed, previous work on continual learning has shown that optimal partitioning of the network neurons, either through non-adaptive (random) or adaptive (learned) gating signals[46,47], can alleviate catastrophic forgetting. In our simulations, sparse random initialization does not prevent catastrophic forgetting, though it is still better than purely sequential learning. This suggests that decreasing the representational overlap through the addition of a contextual gating signal would further improve the efficiency of generative replay, as already proposed in the context of hippocampal replay in the complementary learning systems theory[12]. Overall, these results show that the quality of generated samples is crucial to support interleaved learning through generative replay and corroborates the effectiveness of chimera biasing as initialization strategy for iDBNs.

Discussion

In this work, we investigated the generative dynamics of energy-based neural networks by proposing novel top-down sampling schemes that constrain the generation of sensory data by initializing the sampling procedure with biased hidden states. The choice of the iDBN (as opposed to other state-of-the-art generative models used in artificial intelligence research[48]) as a framework for our investigation is supported by the appeal of this type of energy-based neural networks from a cognitive neuroscience perspective, in light of the biological plausibility and the sound probabilistic interpretation of learning and inference schemes[4,31]. Indeed, DBNs have been successfully used for modeling perceptual and cognitive phenomena[18,19,21] and its recent iterative version (iDBN) provides an ideal platform for modeling learning and development from a cognitive science perspective[27,49].

We showed that the initial biased state triggers a sequence of cycles through the entire iDBN hierarchy, alternating top-down and bottom-up passes. Importantly, our sampling schemes allowed transitioning into different stable states and this processing mode leads to the generation of heterogeneous (but well-formed) data patterns in the visible layer. The dynamics of top-down processing was richer in deep networks compared to shallow models, suggesting that hierarchical architectures support a more flexible exploration of plausible sensory states. It is worth emphasizing that this processing mode is fully detached from external stimuli. Indeed, the iDBN operated in an intrinsic mode initiated from internal states at the top of the representational hierarchy, which is consistent with the proposal that spontaneous brain activity at rest is the manifestation of top-down dynamics that occur in generative models[7]. Although the functional role of spontaneous brain activity emerging without explicit external input is still debated, it is clear that its spatio-temporal structure distinguishes it from mere random noise[8,9,50]. Accordingly, the iDBN "at rest" explores neuronal states that are similar to those experienced during learning and inference, and these states reflect low-dimensional abstract representations of the sensory environment[7]. The fact that top-down generative dynamics in these energy-based models recapitulate sensory-driven activation is well-aligned with recent findings in the neuroscience literature, which have shown that sequential firing patterns present in recordings of spontaneous activity closely match those recorded during sensory stimulation[51].

Intrinsic activity initiated by top-down processing could be the basis for endogenous attention[52] as well as for the integration of bottom-up sensory observations with top-down contextual priors[3,4]. Spontaneous brain activity is also though to be crucially linked to learning and memory[7,12] to prevent catastrophic forgetting[42]. In this perspective, we showed that the samples generated by the iDBN can be used in continuous learning scenarios to support the integration of new information into the network weights through generative replay mechanisms.

Previous studies on shallow models such as the Restricted Boltzmann Machine (RBM) have shown that top-down sampling could be challenging due to the presence of large free energy barriers, which prevent flexible exploration of network states[53]. Here we have shown that deeper architectures allow to partially overcome these limitations, although it would be interesting to investigate whether our sampling schemes might allow to force the generation trajectory through specific attractors, as can be done in RBMs variants that learn disentangled representations[36].

In general, we believe that the proposed framework could stimulate further research to bridge the gap between deep learning models and computational neuroscience[4,54–56]. For example, studying the intrinsic top-down dynamics in generative neural networks could help clarify the functional role of spontaneous brain activity in the maintenance and consolidation of sensory-motor information[57], or the role of post-exposure spontaneous activity in refining stimulus encoding and persistence[58]. The approach presented in this work also provides a principled framework to simulate contextual effects resulting from hidden state biasing, which can be used to study how top-down expectations can shape perception[59]. Furthermore, by analyzing the learning trajectories of the iDBN one could investigate how internal models of the environment that support top-down spontaneous cortical activity might get gradually refined throughout development[60]. Research on spontaneous brain activity has also investigated the temporal organization of characteristic resting-state functional connectivity patterns in humans and animals, revealing that the transitions are not random but rather follow specific sequential orders[61]. Although the present approach does not allow to directly model these temporal characteristics of brain signals, we believe this could be a viable direction for future research based on this framework, for example by taking advantage of sequential versions of deep generative models[62].

Interestingly, shallow generative models have recently been used to simulate the activation dynamics of neural assemblies emerging from neuronal recordings in animal models[63], and we believe that computational investigations based on deeper architectures could constitute an important step forward toward a more comprehensive understanding of the endogenous neural dynamics emerging from assembly organization. Another interesting analogy that has not yet been adequately explored is that spontaneous brain activity is usually observed during rest or dreaming, and learning and inference processes in deep generative models have also been characterized in terms of "wake and sleep" cycles[64].

Data Availability

The complete source code to reproduce our simulations can be found here: https://github.com/CCNL-UniPD/top-down

Received: 5 December 2023; Accepted: 26 December 2024
Published online: 22 January 2025

References

1. Clark, A. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.* **36**, 181–204 (2013).
2. Parr, T. & Friston, K. J. The anatomy of inference: Generative models and brain structure. *Front. Comput. Neurosci.* **12**, 90 (2018).
3. Friston, K. Hierarchical models in the brain. *PLoS Comput. Biol.* **4**, e1000211 (2008).
4. Testolin, A. & Zorzi, M. Probabilistic models and generative neural networks: Towards an unified framework for modeling normal and impaired neurocognitive functions. *Front. Comput. Neurosci.* **10**, 73 (2016).
5. Raichle, M. E. The brain's dark energy. *Science* **314**, 1249–1250 (2006).
6. Mohajerani, M. H. et al. Spontaneous cortical activity alternates between motifs defined by regional axonal projections. *Nat. Neurosci.* **16**, 1426–1435 (2013).
7. Pezzulo, G., Zorzi, M. & Corbetta, M. The secret life of predictive brains: What?s spontaneous activity for?. *Trends Cognit. Sci.* **25**, 730–743 (2021).
8. Raichle, M. E. & Mintun, M. A. Brain work and brain imaging. *Annu. Rev. Neurosci.* **29**, 449–476 (2006).
9. Fox, M. D. & Raichle, M. E. Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nat. Rev. Neurosci.* **8**, 700–711 (2007).
10. Lewis, C., Baldassarre, A., Committeri, G., Romani, G. & Corbetta, M. Learning sculpts the spontaneous activity of the resting human brain. *Proc. Natl. Acad. Sci. USA* **106**, 17558–17563 (2009).
11. Liu, Y., Dolan, R. J., Kurth-Nelson, Z. & Behrens, T. E. Human replay spontaneously reorganizes experience. *Cell* **178**(3), 640–652 (2009).
12. Kumaran, D., Hassabis, D. & McClelland, J. L. What learning systems do intelligent agents need? complementary learning systems theory updated. *Trends Cognit. Sci.* **20**, 512–534 (2016).
13. Deco, G., Kringelbach, M. L., Jirsa, V. K. & Ritter, P. The dynamics of resting fluctuations in the brain: Metastability and its dynamical cortical core. *Sci. Rep.* **7**, 3095 (2017).
14. Tognoli, E. & Kelso, J. S. The metastable brain. *Neuron* **81**, 35–48 (2014).
15. Goodfellow, I., Bengio, Y. & Courville, A. *Deep L:earning* (MIT Press, 2016).
16. Hinton, G. E., Osindero, S. & Teh, Y.-W. A fast learning algorithm for deep belief nets. *Neural Comput.* **18**, 1527–1554 (2006).
17. Lee, H., Ekanadham, C. & Ng, A. Sparse deep belief net model for visual area v2. *Adv. Neural Inf. Process. Syst.* **20** (2007).
18. Stoianov, I. & Zorzi, M. Emergence of a'visual number sense'in hierarchical generative models. *Nat. Neurosci.* **15**, 194–196 (2012).
19. Testolin, A., Stoianov, I. & Zorzi, M. Letter perception emerges from unsupervised deep learning and recycling of natural image features. *Nat. Hum. Behav.* **1**, 657–664 (2017).
20. Zorzi, M. & Testolin, A. An emergentist perspective on the origin of number sense. *Philos. Trans. R. Soc. B* **373**, 20170043 (2018).
21. Testolin, A., Dolfi, S., Rochus, M. & Zorzi, M. Visual sense of number vs. sense of magnitude in humans and machines. *Sci. Rep.* **10**, 1–13 (2020).
22. Kleinbub, J. R., Testolin, A., Palmieri, A. & Salvatore, S. The phase space of meaning model of psychopathology: A computer simulation modelling study. *PLoS ONE* **16**, e0249320 (2021).
23. Buesing, L., Bill, J., Nessler, B. & Maass, W. Neural dynamics as sampling: A model for stochastic computation in recurrent networks of spiking neurons. *PLoS Comput. Biol.* **7**, e1002211 (2011).
24. Pecevski, D., Buesing, L. & Maass, W. Probabilistic inference in general graphical models through sampling in stochastic networks of spiking neurons. *PLoS Comput. Biol.* **7**, e1002294 (2011).
25. Wang, W. et al. A memristive deep belief neural network based on silicon synapses. *Nat. Electron.* **5**, 870–880 (2022).
26. Hinton, G. E. Training products of experts by minimizing contrastive divergence. *Neural Comput.* **14**, 1771–1800 (2002).
27. Zambra, M., Testolin, A. & Zorzi, M. A developmental approach for training deep belief networks. *Cogn. Comput.* **15**, 103–120 (2023).
28. Lamme, V. A. & Roelfsema, P. R. The distinct modes of vision offered by feedforward and recurrent processing. *Trends Neurosci.* **23**, 571–579 (2000).
29. VanRullen, R. The power of the feed-forward sweep. *Adv. Cogn. Psychol.* **3**, 167 (2007).
30. Kreiman, G. & Serre, T. Beyond the feedforward sweep: Feedback computations in the visual cortex. *Ann. N. Y. Acad. Sci.* **1464**, 222 (2020).
31. Zorzi, M., Testolin, A. & Stoianov, I. P. Modeling language and cognition with deep unsupervised learning: A tutorial overview. *Front. Psychol.* **4**, 515 (2013).
32. LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **86**, 2278–2324 (1998).
33. Liu, Z., Luo, P., Wang, X. & Tang, X. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision* 3730–3738 (2015).
34. Tausani, L., Testolin, A. & Zorzi, M. Investigating the generative dynamics of energy-based neural networks. *arXiv preprint*[SPACE]arXiv:2305.06745 (2023).
35. Wittkuhn, L., Chien, S., Hall-McMaster, S. & Schuck, N. W. Replay in minds and machines. *Neurosci. Biobehav. Rev.* **129**, 367–388 (2021).
36. Fernandez-de Cossio-Diaz, J., Cocco, S. & Monasson, R. Disentangling representations in restricted Boltzmann machines without adversaries. *Phys. Rev. X* **13**, 021003 (2023).
37. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint*[SPACE]arXiv:1409.1556 (2014).
38. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 770–778 (2016).
39. Deng, J. *et al.* Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* 248–255 (Ieee, 2009).
40. Cunningham, J. P. & Yu, B. M. Dimensionality reduction for large-scale neural recordings. *Nat. Neurosci.* **17**, 1500–1509 (2014).
41. DiCarlo, J. J., Zoccolan, D. & Rust, N. C. How does the brain solve visual object recognition?. *Neuron* **73**, 415–434 (2012).
42. French, R. M. Catastrophic forgetting in connectionist networks. *Trends Cogn. Sci.* **3**, 128–135 (1999).
43. Shin, H., Lee, J. K., Kim, J. & Kim, J. Continual learning with deep generative replay. *Adv. Neural Inf. Process. Syst.* **30** (2017).
44. Cohen, G., Afshar, S., Tapson, J. & Van Schaik, A. Emnist: Extending mnist to handwritten letters. In *2017 International Joint Conference on Neural Networks (IJCNN)* 2921–2926 (IEEE, 2017).

45. Xiao, H., Rasul, K. & Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. arXiv: csLG/1708.07747 (2017).
46. Masse, N. Y., Grant, G. D. & Freedman, D. J. Alleviating catastrophic forgetting using context-dependent gating and synaptic stabilization. *Proc. Natl. Acad. Sci. USA* **115**, E10467–E10475 (2018).
47. Verbeke, P. & Verguts, T. Using top-down modulation to optimally balance shared versus separated task representations. *Neural Netw.* **146**, 256–271 (2022).
48. Ho, J., & Jain, A. Denoising diffusion probabilistic models. *Proc. Int. Conf. Neural Inf. Process. Syst.* 6840–6851 (2020).
49. Dolfi, S. *et al.* Weaker number sense accounts for impaired numerosity perception in dyscalculia: Behavioral and computational evidence. *Dev. Sci.* e13538 (2024).
50. He, Y. et al. Uncovering intrinsic modular organization of spontaneous brain activity in humans. *PLoS ONE* **4**, e5226 (2009).
51. Carrillo-Reid, L., Miller, J.-E.K., Hamm, J. P., Jackson, J. & Yuste, R. Endogenous sequential cortical activity evoked by visual stimuli. *J. Neurosci.* **35**, 8813–8828 (2015).
52. Casarotti, M., Lisi, M., Umiltà, C. & Zorzi, M. Paying attention through eye movements: A computational investigation of the premotor theory of spatial attention. *J. Cogn. Neurosci.* **24**, 1519–1531 (2012).
53. Roussel, C., Cocco, S. & Monasson, R. Barriers and dynamical paths in alternating Gibbs sampling of restricted Boltzmann machines. *Phys. Rev. E* **104**, 034109 (2021).
54. De Schutter, E. Deep learning and computational neuroscience. *Neuroinformatics* **16**, 1–2 (2018).
55. Richards, B. A. et al. A deep learning framework for neuroscience. *Nat. Neurosci.* **22**, 1761–1770 (2019).
56. Saxe, A., Nelli, S. & Summerfield, C. If deep learning is the answer, what is the question?. *Nat. Rev. Neurosci.* **22**, 55–67 (2021).
57. Zhang, L., Pini, L., Kim, D., Shulman, G. L. & Corbetta, M. Spontaneous activity patterns in human attention networks code for hand movements. *J. Neurosci.* **43**, 1976–1986 (2023).
58. Lazar, A., Lewis, C., Fries, P., Singer, W. & Nikolic, D. Visual exposure enhances stimulus encoding and persistence in primary cortex. *Proc. Natl. Acad. Sci. USA* **118**, e2105276118 (2021).
59. De Lange, F. P., Heilbron, M. & Kok, P. How do expectations shape perception?. *Trends Cogn. Sci.* **22**, 764–779 (2018).
60. Berkes, P., Orbán, G., Lengyel, M. & Fiser, J. Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. *Science* **331**, 83–87 (2011).
61. Ma, Z. & Zhang, N. Temporal transitions of spontaneous brain activity. *Elife* **7**, e33562 (2018).
62. Gan, Z., Li, C., Henao, R., Carlson, D. E. & Carin, L. Deep temporal sigmoid belief networks for sequence modeling. *Adv. Neural Inf. Process. Syst.* **28** (2015).
63. van der Plas, T. L. et al. Neural assemblies uncovered by generative modeling explain whole-brain activity statistics and reflect structural connectivity. *eLife* **12**, e83139 (2023).
64. Hinton, G. E., Dayan, P., Frey, B. J. & Neal, R. M. The, "wake-sleep" algorithm for unsupervised neural networks. *Science* **268**, 1158–1161 (1995).

Acknowledgements

This work was supported by the European Union - NextGenerationEU as part of the National Recovery and Resilience Plan (PNRR), M4C2 PE0000013 "Future Artificial Intelligence Research - FAIR", CUP J93C24000320007 to M.Z.

Author contributions

A.T. and M.Z. conceived the experiments, L.T. conducted the experiments and analyzed the results. L.T. and A.T. wrote the first draft and M.Z. revised the manuscript. All authors reviewed the final version of the manuscript.

Declarations

Competing interests

The authors declare that they have no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-85055-y.

Correspondence and requests for materials should be addressed to A.T. or M.Z.