# Automated construction of cognitive maps with predictive coding

James A. Gornet[1,2*] Matt Thomson[1,2]
jgornet@caltech.edu mthomson@caltech.edu

[1]California Institute of Technology, Division of Biology and Biological Engineering, Pasadena, CA, USA
[2]California Institute of Technology, Computation and Neural Systems, Pasadena, CA, USA

Humans construct internal cognitive maps of their environment directly from sensory inputs without access to a system of explicit coordinates or distance measurements. While machine learning algorithms like SLAM utilize specialized inference procedures to identify visual features and construct spatial maps from visual and odometry data, the general nature of cognitive maps in the brain suggests a unified mapping algorithmic strategy that can generalize to auditory, tactile, and linguistic inputs. Here, we demonstrate that predictive coding provides a natural and versatile neural network algorithm for constructing spatial maps using sensory data. We introduce a framework in which an agent navigates a virtual environment while engaging in visual predictive coding using a self-attention-equipped convolutional neural network. While learning a next image prediction task, the agent automatically constructs an internal representation of the environment that quantitatively reflects spatial distances. The internal map enables the agent to pinpoint its location relative to landmarks using only visual information.The predictive coding network generates a vectorized encoding of the environment that supports vector navigation where individual latent space units delineate localized, overlapping neighborhoods in the environment. Broadly, our work introduces predictive coding as a unified algorithmic framework for constructing cognitive maps that can naturally extend to the mapping of auditory, sensorimotor, and linguistic inputs.

Space and time are fundamental physical structures in the natural world, and all organisms have evolved strategies for navigating space to forage, mate, and escape predation.[1,2]. In humans and other mammals, the concept of a spatial or cognitive map has been postulated to underlie spatial reasoning tasks[3–5]. A spatial map is an internal, neural representation of an animal's environment that marks the location of landmarks, food, water, shelter, and then can be queried for navigation and planning. The neural algorithms underlying spatial mapping are thought to generalize to other sensory modes to provide cognitive representations of auditory and somatosensory data[6] as well as to construct internal maps of more abstract information including concepts[7,8], tasks [9], semantic information[10–12], and memories[13]. Empirical evidence suggest that the brain uses common cognitive mapping strategies for spatial and non-spatial sensory information so that common mapping algorithms might exist that can map and navigate over not only visual but also semantic information and logical rules inferred from experience[6,7,14]. In such a paradigm reasoning itself could be implemented as a form of navigation within a cognitive map of concepts, facts, and ideas.

Since the notion of a spatial or cognitive map emerged, the question of how environments are represented within the brain and how the maps can be learned from experience has been a central question in neuroscience[15]. Place cells in the hippocampus are neurons that are active when an animal transits through a specific location in an environment[15]. Grid cells in the entorhinal cortex fire in regular spatial intervals and likely track an organism's displacement in the environment[16,17]. Yet with the identification of a substrate for the representation of space, the question of how a spatial map can be learned from sensory data has remained, and the neural algorithms that enable the construction of spatial and other cognitive maps remain poorly understood.

Empirical work in machine learning has demonstrated that deep neural networks can solve spatial navigation tasks as well as perform path prediction and grid cell formation[18,19]. Cueva & Wei[18] and Banino *et al.*[19] demonstrate that neural networks can learn to perform path prediction and that networks generate firing patterns that resemble the firing patterns of grid cells in the entorhinal cortex. However, these studies allow an agent to access environmental coordinates explicitly[18] or initialize a model with place cells that represent specific locations in an arena[19]. In machine learning and autonomous navigation, a variety of algorithms have been developed to perform mapping tasks including SLAM and monocular SLAM algorithms[20–23] as well as neural network implementations[24–26]. Yet, SLAM algorithms contain many specific inference strategies, like visual feature and object detection, that are specifically engineered for map building, wayfinding, and pose estimation based on visual information. A unified theoretical and mathematical framework for understanding

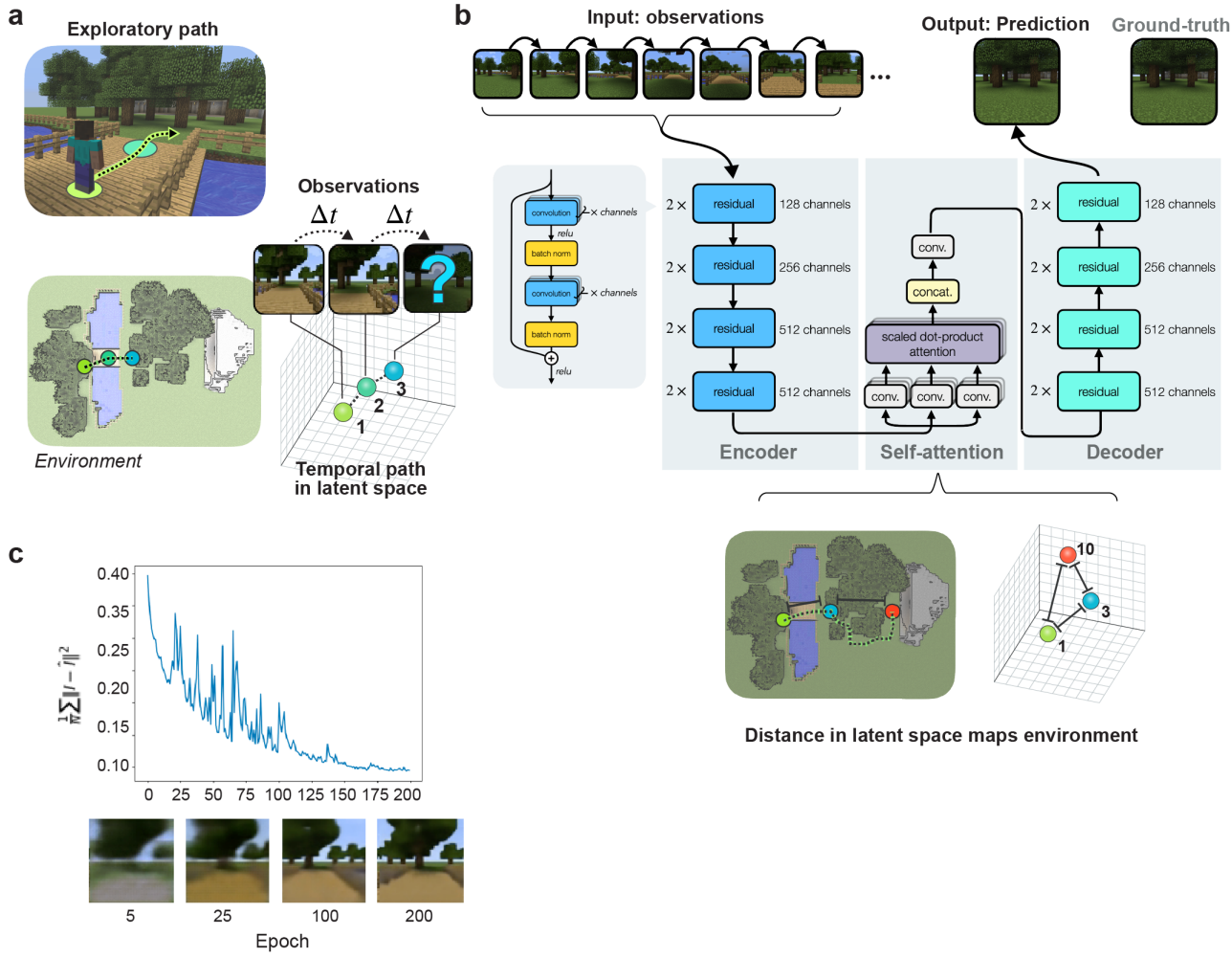


**Figure 1. A predictive coding neural network explores a virtual environment. a,** in predictive coding, a model predicts observations and updates model given the prediction error the latent density using the prediction error. **b,** a self-attention-based encoder-decoder neural network architecture learns to perform predictive coding. A ResNet-18 convolutional neural network acts as an encoder; self-attention is performed with 8 heads, and a corresponding ResNet-18 convolutional neural network performing decoding to the predicted image.**c,** the neural network learns to perform predictive coding effectively—with a mean-squared error of 0.094 between the actual and predicted images.

the mapping of spaces based on sensory information remains incomplete.

Predictive coding has been proposed as a unifying theory of neural function where the fundamental goal of a neural system is to predict future observations given past data[27–29]. When an agent explores a physical environment, temporal correlations in sensory observations reflect the structure of the physical environment. Landmarks nearby one another in space will also be observed in temporal sequence. In this way, predicting observations in a temporal series of sensory observations requires an agent to internalize some implicit information about a spatial domain. Historically, Poincare motivated the possibility of spatial mapping through a predictive coding strategy where an agent assembles a global representation of an environment by gluing together information gathered through local exploration[30,31]. The exploratory paths together contain information that could, in principle, enable the assembly of a spatial map for both flat and curved manifolds. Yet, while the concept of predictive coding for spatial mapping is intuitively attractive, a major challenge is the development of algorithms that can glue together local information gathered by an agent into a global, internally consistent environmental map.

Here, we demonstrate that a neural network trained on a predictive coding task can construct an implicit spatial map of an environment by assembling observations acquired along local exploratory paths into a global representation of a physical space within the network's latent space. The strategy can be implemented by a feed-forward, encoder-decoder network architecture where the encoder network embeds images collected by an agent exploring an environment into an internal representation of space. Within the embedding, the distances between images reflect their relative spatial

position, not object-level similarity between images. During exploratory training, the network implicitly assembles information from local paths into a global representation of space as it performs a next image inference problem. Fundamentally, we connect predictive coding and mapping tasks, demonstrating a computational and mathematical strategy for integrating information from local measurements into a global self-consistent environmental model.

## Mathematical formulation of spatial mapping as predictive coding

First, we formulate a theoretical model of visual predictive coding and demonstrate that the predictive coding problem can naturally be solved by an inference procedure that constructs an implicit representation of an agent's environment. Moreover, the theoretical analysis suggested that the underlying inference problem that can be solved naturally by an encoder-decoder neural network that infers spatial position based upon observed image sequences.

.We consider an agent exploring an environment, $\Omega \subset \mathbb{R}^2$, while acquiring visual information in the form of pixel valued image vectors $I_x \in \mathbb{R}^{m\times n}$ given an $x \in \Omega$. The agent's environment $\Omega$ is a bounded subset or $\mathbb{R}^2$ that could contain obstructions and holes. In general, at any given time, $t$, the agent's state can be characterized by a position $x(t)$ and orientation $\theta(t)$ where $x(t)$ and $\theta(t)$ are coordinates within a global coordinate system unknown to the agent. While both position $x$ and orientation $\theta$ can be accommodated within our statistical inference framework, for expository convenience, we consider an agent that adopts a constant orientation while moving along a series of positions $x(t)$.

The agent's environment comes equipped with a visual scene, and the agent makes observations by acquiring image vectors $I_{x_k} \in \mathbb{R}^{m\times n}$ as it moves along a sequence of points $x_k$. At every position x , the agent acquires an image by effectively sampling from an image the conditional probability distribution $P(I|x_k)$ which encodes the probability of observing a specific image vector $I$ when the agent is positioned at position $x_k$. The distribution $P(I|x)$ has a deterministic and stochastic component where the deterministic component is set by landmarks in the environment while stochastic effects can emerge due to changes in lighting, background, and scene dynamics. Mathematically, we can view $P(I|x)$ as a function on a vector bundle with base space $\Omega$ and total space $\Omega \times I$. The function assigns an observation probability to every possible image vector for an agent positioned at a point $x$.

In the predictive coding problem, the agent moves along a series of points $x_0, x_1, \ldots, x_k$ while acquiring images $I_0, I_1, \ldots I_k$. The motion of the agent in $\Omega$ is generated by a Markov process with transition probabilities $P(x_{i+1}|x_i)$. Note that the agent has access to the image observations $I_i$ but not the spatial coordinates $x_i$. Given the set $\{I_0 \ldots I_k\}$ the agent aims to predict $I_{k+1}$. Mathematically, the image prediction problem can be solved theoretically through statistical inference by (a) inferring the posterior probability distribution $P(I_{k+1}|I_0, I_1 \ldots . I_k)$ from observations. Then, (b) given a specific sequence of observed images $\{I_0 \ldots I_k\}$, the agent can predict the next image $I_{k+1}$ by finding the image $I_{k+1}$ that maximizes the posterior probability distribution $P(I_{k+1}|I_0, I_1 \ldots . I_k)$.

The posterior probability distribution $P(I_{k+1}|I_0, I_1, \ldots . I_k)$ is by definition

$$P(I_{k+1}|I_0, I_1, \ldots., I_k) = \frac{P(I_0, I_1, \ldots, I_k, I_{k+1})}{P(I_0, I_1, \ldots, I_k)}.$$

If we consider $P(I_0, I_1 \ldots I_k, I_{k+1})$ to be a function of an implicit set of spatial coordinates $x_i$ where the $x_i$ provide an internal representation of the spatial environment. Then, we can express the posterior probability $P(I_{k+1}|I_0, I_1 \ldots . I_k)$ in terms of the implicit spatial representation

$$\begin{aligned} P(I_{k+1}|I_0, I_1, \ldots, I_k) &= \int_\Omega \mathbf{dx}\, P(x_0, x_1, \ldots, x_k) \frac{P(I_0, I_1 \cdots I_k|x_0, \ldots, x_k)}{P(I_0, I_1, \ldots., I_k)} P(x_{k+1}|x_k) P(I_{k+1}|x_{k+1}) \\ &= \int_\Omega \mathbf{dx}\, \underbrace{P(x_0, x_1, \ldots, x_k|I_0, I_1 \cdots, I_k)}_{\text{encoding(1)}} \quad \underbrace{P(x_{k+1}|x_k)}_{\text{spatial transition probability (2)}} \quad \underbrace{P(I_{k+1}|x_{k+1})}_{\text{decoding (3)}} \end{aligned} \tag{1}$$

where in 1 the integration is over all possible paths $\{x_0 \ldots x_k\}$ in the domain $\Omega$ and $\mathbf{dx} = dx_i \ldots dx_k$. Equation 1 can be interpreted as a path integral over the domain $\Omega$. The path integral assigns a probability to every possible path in the domain and then computes the probability that the agent will observe a next image $I_k$ given an inferred location $x_{k+1}$. In detail term 1 assigns a probability to every discrete path $\{x_0 \ldots x_k\} \in \Omega$ as the conditional likelihood of the path given the observed sequences of images $\{I_0 \ldots I_k\}$ . Term 2 computes the probability that an agent at a terminal position $x_k$ moves to the position $x_{k+1}$ given the Markov transition function $P(x_{k+1}|x_k)$. Term 3 is

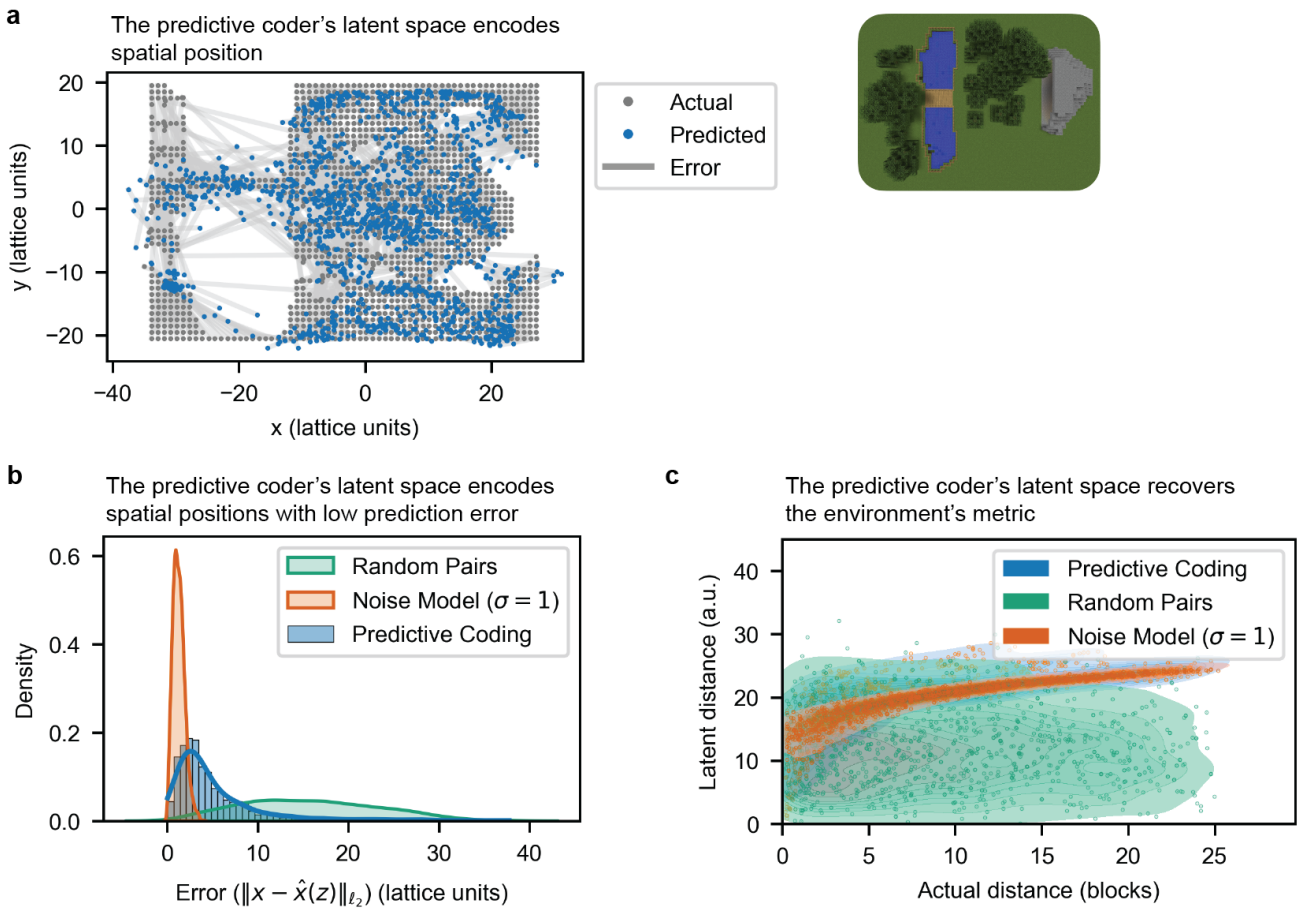


**Figure 2. Predictive coding neural network constructs an implicit spatial map. a-b,** The latent space encodes spatial position. **b,** a neural network predicts the spatial location from the predictive coding's latent space. **b,** predictive coding's latent space encodes spatial position. The histogram of prediction errors of positions from the predictive coder's latent space. The additive noise model and the random shuffle model provide baselines for the minimum and maximum errors, respectively. **c,** predictive coding's latent distances recover the environment's spatial metric. Sequential visual images are mapped to the neural network's latent space, and the latent space distances ($\ell_2$) are plotted with physical distances onto a joint density plot.

the conditional probability that image $I_{k+1}$ is observed given that the agent is at position $x_{k+1}$.

Conceptually, the product of terms solves the next image prediction problem in three steps. First (1), estimating the probability that an agent has traversed a particular sequence of points given the observed images; second (2), estimating the next position of the agent $x_{k+1}$ for each potential path; and third (3), computing the probability of observing a next image $I_{k+1}$ given the inferred terminal location $x_{k+1}$ of the agent. Critically, an algorithm that implements the inference procedure encoded in the equation would construct an internal but implicit representation of the environment as a coordinate system $\mathbf{x}$ that is learned by the agent and used during the next image inference procedure. The coordinate system provides an internal, inferred representation of the agent's environment that is used to estimate future image observation probabilities. Thus, our theoretical framework demonstrates how an agent might construct an implicit representation of its spatial environment by solving the predictive coding problem.

The three step inference procedure represented in the equation for $P(I_{k+1}|I_0 \ldots I_k)$ can be directly implemented in a neural network architecture. The first term acts as an 'encoder' network that computes the probability that the agent has traversed a path $x_0 \ldots x_k$ given an observed image sequence $I_0, \ldots, I_k$ that has been observed by the network (Figure 1(**b**)). The network can, then, estimate the next position of the agent $x_{k+1}$ given an inferred location $x_k$, and apply a decoding network to compute $P(I_{k+1}|x_{k+1})$ while outputting the prediction $I_{k+1}$ using a decoder. A network trained through visual experience must learn an internal coordinate system and representation $\mathbf{x}$ that not only offers an environmental representation but also establishes a connection between observed images $I_j$ and inferred locations $x_j$.

### A neural network performs accurate predictive coding within a virtual environment

Motivated by the implicit representation of space contained in the predictive coding inference problem, we developed a computational implementation of a predictive coding agent, and studied the representation of space learned by that agent as it explored a virtual environment. We first create an environment with the Malmo environment in Minecraft[32]. The physical environment measures $40 \times 65$ lattice units and encapsulates three aspects of visual scenes: a cave provides a global visual landmark, a forest provides degeneracy between visual scenes, and a river with a bridge constrains how an agent traverses the environment (Figure 1(a)). An agent follows paths, determined by $A^*$ search, between randomly sampled positions and receives visual images along every path.

To perform predictive coding, we construct an encoder-decoder convolutional neural network (CNN) with a ResNet-18 architecture[33] for the encoder and a corresponding ResNet-18 architecture with transposed convolutions in the decoder (Figure 1(**b**)). The encoder-decoder architecture uses the U-Net architecture[34] to pass the encoded latent units into the decoder. Multi-headed attention[35] processes the sequence of encoded latent units to encode the history of past visual observations. The multi-headed attention has $h = 8$ heads. For the encoded latent units with dimension $D = C \times H \times W$, the dimension $d$ of a single head is $d = C \times H \times W/h$.

The predictive coder approximates predictive coding by minimizing the mean-squared error between the actual observation and its predicted observation. The predictive coder trains on $82,630$ samples for 200 epochs with gradient descent optimization with Nesterov momentum[36], a weight decay of $5 \times 10^{-6}$, and a learning rate of $10^{-1}$ adjusted by OneCycle learning rate scheduling[37]. The optimized predictive coder has a mean-squared error between the predicted and actual images of 0.094 and a good visual fidelity (Figure 1(**c**)).

## Predictive coding network constructs an implicit spatial map

We show that the predictive coder creates an implicit spatial map by demonstrating it recovers the environment's spatial position and distance. We encode the image sequences using the predictive coder's encoder to analyze the encoded sequence as the predictive coder's latent units. To measure the positional information in the predictive coder, we train a neural network to predict the agent's position from the predictive coder's latent units (Figure 1(**a**)). The neural network's prediction error

$$E(x, \hat{x}) = \|\hat{x} - x\|_{\ell_2}$$

indirectly measures the predictive coder's positional information. To provide comparative baselines, we construct two different position prediction models to lower bound and upper bound the prediction error. To lower bound the prediction error, we construct a model that gives the agent's actual position with small additive Gaussian noise

$$\hat{x} = x + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma).$$

To upper bound the prediction error, we construct a model that shuffles the agent's actual position *without replacement*. To compare the predictive coder to the baselines, we compare the prediction error histograms (Figure 2(**b**)).

The predictive coder encodes the environment's spatial position to a low prediction error (Figure 2.d). The predictive coder has a mean error of 5.04 lattice units and > 80% of samples have an error < 7.3 lattice units. The additive Gaussian model with $\sigma = 4$ has a mean error of 4.98 lattice units and > 80% of samples with an error < 7.12 lattice units. The shuffle model, on the other hand, has a mean error of 15.87 lattice units and > 80% of samples have an error < 22.24 lattice units.

We show the predictive coder's latent space recovers the local distances between the environment's physical positions. For every path that the agent traverses, we calculate the local pairwise distances in physical space and in the predictive coder's latent space with a neighborhood of 100 time points. To determine whether latent space distances correspond to physical distances, we calculate the joint density between latent space distances and physical distances (Figure 2(**c**)). We model the latent distances by fitting the physical distances with additive Gaussian noise to a logarithmic function

$$d(z, z') = \alpha \log(\|x - x' + \epsilon\|) + \beta, \epsilon \sim \mathcal{N}(0, \sigma).$$

In addition, as a null distribution, we shuffle the physical positions and calculate the latent distances on this shuffled set. The modeled distribution is concentrated with the predictive coder's distribution with a Kullback-Leibler divergence ($\mathbb{D}_{\mathrm{KL}}(p_{\mathrm{PC}} \| p_{\mathrm{model}})$) of 0.429 bits. The null distribution shows a low overlap with the predictive coder's distribution with a $\mathbb{D}_{\mathrm{KL}}(p_{\mathrm{PC}} \| p_{\mathrm{null}})$ of 2.441 bits.

## Predictive coding network learns spatial proximity not image similarity

In the previous section, we show that a neural network that performs predictive coding learns an internal representation of its physical environment within its latent space. Here, we demonstrate that the prediction task itself is essential for spatial mapping . Prediction forces a network to learn spatial proximity and not merely image similarity. Many frameworks including principal components analysis, IsoMap[38], and autoencoder neural networks can collocate images by visual similarity. While similar scenes might be proximate in space, similar scenes can also be spatially divergent. For example, the virtual environment we constructed has two different 'forest' regions that are separated by a lake. Thus, in the two forest environments might generate similar images but are actually each closer to the lake region than to one another (Figure1. )

To demonstrate the central role for prediction in mapping, we compared the latent representation of images

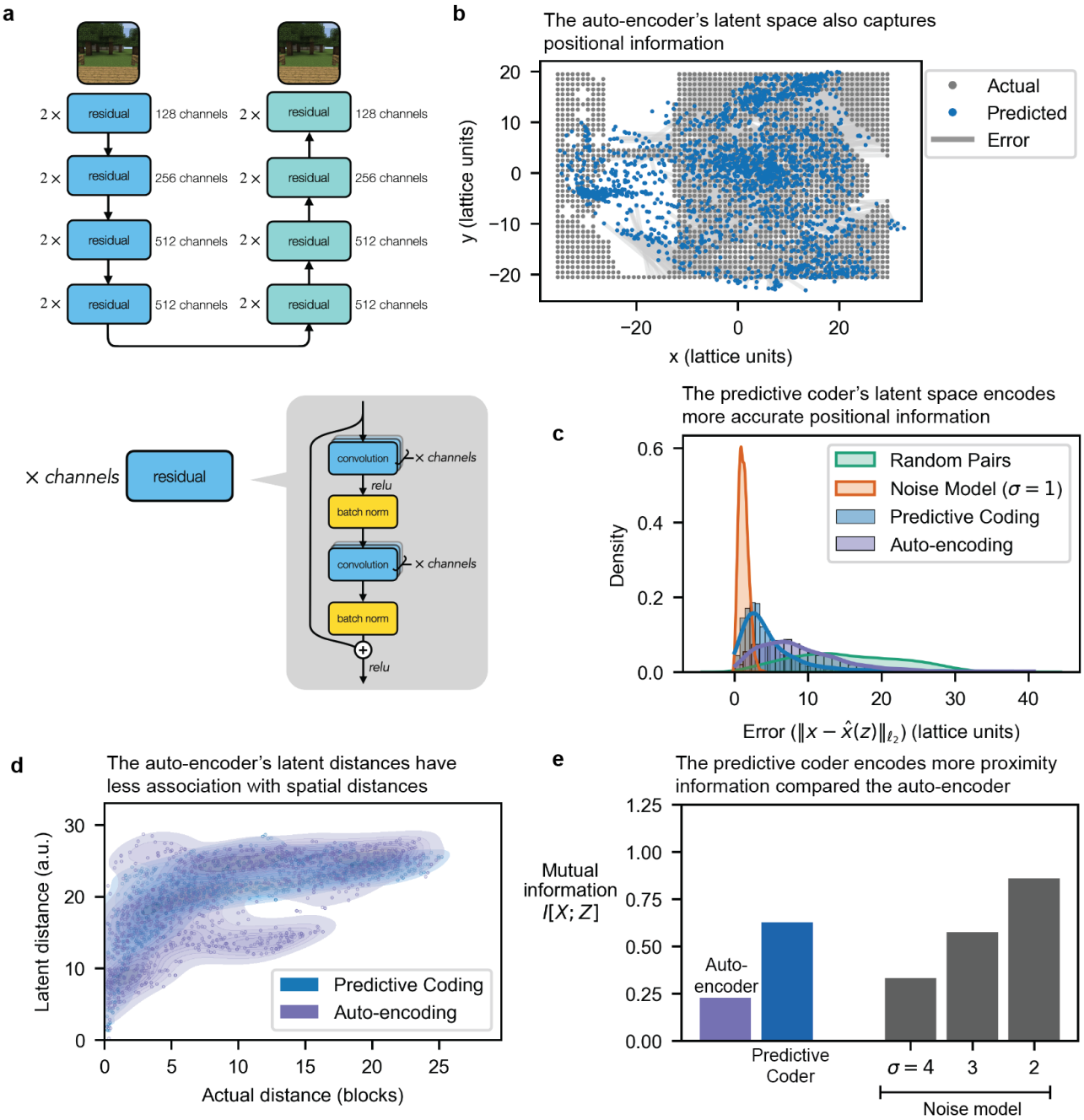


**Figure 3. Predictive coding network learns spatial proximity not image similarity a,** an autoencoding neural network compresses visual images into a low-dimensional latent vector and reconstructs the image from the latent space. Auto-encoder trains on visual images from the environment *without any sequential order*. **b-c,** auto-encoding encodes lower resolution in positional information. **b,** a neural network predicts the spatial location from the auto-encoding's latent space. **c,** auto-encoding captures less positional information compared to predictive coding. The histogram shows the prediction errors of positions from the latent space of both the auto-encoder and the predictive coder. **d,** latent distances, however, show a weaker relationship with physical distances, as the joint histogram between physical and latent distances is less concentrated. **e,** predictive coding's latent units communicate more fine-grained spatial distances whereas auto-encoding communicates broad spatial regions. Joint density plots show the association between latent distances and physical distances for both predictive coding and auto-encoding. Predictive coding's latent distances increase with spatial distances, with a higher concentration compared to auto-encoding.

generated by the predictive coding network to a representation learned by an autoencoder. The auto-encoder network has a similar architecture to the predictive encoder but encodes a *single* image observation in a latent space, and decodes the same observations. As the auto-encoder only operates on a single image—rather than a sequence, the auto-encoder learns an embedding based on image proximity not underlying spatial relationships. As with the predictive coder, the auto-encoder (Figure 3(**a**)) trains to minimize the mean-squared error between the actual image and the predicted image on 82, 630 samples for 200 epochs with gradient descent optimization with Nesterov momentum, a weight decay of $5 \times 10^{-6}$, and a learning rate of $10^{-1}$ adjusted by the OneCycle learning rate scheduler. The auto-encoder has mean-squared error of 0.039 and a high visual fidelity.

The predictive coder encodes more positional information in its latent space than the auto-encoder. As with the predictive coder, we train an auxiliary neural network to predict the agent's position from the auto-encoder's latent units (Figure 3(**b**)). The neural network's prediction error indirectly measures the auto-encoder positional information. The auto-encoder has greater than 80% of its points and has a prediction error of less than 13.1 lattice units, as compared to the predictive coder that has > 80% of its samples have a prediction error of 7.3 lattice units (Figure 3(**c**)).

We also show that the predictive coder recovers the environment's spatial distances with finer resolution compared to the auto-encoder. As with the predictive coder, we calculate the local pairwise distances in physical space and in the auto-encoder's latent space, and we generate the joint density between the physical and latent distances (Figure 3**d**). Compared to the predictive coder's joint density, the auto-encoder's latent distances increase with the agent's physical distance. The auto-encoder's joint density shows a larger dispersion compared to the predictive coder's joint density, indicating that the auto-encoder encodes spatial distances with higher uncertainty.

We can quantitatively measure the dispersion in the auto-encoder's joint density by calculating mutual information of the joint density (Figure 3(**e**))

$$I[X;Z] = \mathbb{E}_{p(X,Z)}\left[\log \frac{p(X,Z)}{p(X)p(Z)}\right].$$

The auto-encoder has a mutual information of 0.227 bits while the predictive coder has a mutual information of 0.627 bits. As a comparison, positions with additive Gaussian noise having a standard deviation $\sigma$ of 2 lattice units has a mutual information of 0.911 bits. The predictive coder encodes 0.400 additional bits of distance information to the auto-encoder. The predictive coder's additional distance information of 0.4 bits exceeds the auto-encoder's distance information of 0.227 bits, which indicates the temporal dependencies encoded by the predictive coder capture more spatial information compared to visual similarity.

## Predictive coding generates units with localized receptive fields that support vector navigation

In the previous section, we demonstrate that the predictive coding neural network captures spatial relationships within an environment containing more internal spatial information than can be captured by an auto-encoder network that encodes image similarity. Here, we analyze the structure of the spatial code learned by the predictive coding network. We demonstrate that each unit in the neural network's latent space activates at distinct, localized regions—akin to place fields in the mammalian brain—in the environment's physical space (Figure 4(**a**)). These place fields overlap and their aggregate covers the entire physical space. Each physical location, is represented by a unique combination of overlapping regions encoded by the latent units. This combination of overlapping regions recovers the agent's current physical position. Furthermore, given two physical locations, there now exist two distinct combinations of overlapping regions in latent space. The differences in these two combinations, the Hamming distance, provides the distance between the two physical locations (Figure 4(**b**)). By comparing the combinations of overlapping regions at different positions, the neural network can perform vector navigation given its place fields.

To support this proposed mechanism, we first demonstrate the neural network generates place fields. In other words, units from the neural network's latent space produce localized regions in physical space. To determine whether a latent unit is active, we threshold the continuous value with its 90th-percentile value. To measure a latent unit's localization in physical space, we fit each latent unit distribution, with respect to physical space, to a two-dimensional Gaussian distribution (Figure 4(**c**), top)

$$p = \frac{1}{2\pi|\Sigma|}\exp\left[-\frac{1}{2}(x-\mu)^{\intercal}\Sigma^{-1}(x-\mu)\right]$$

We measure the area of the ellipsoid given by the Gaussian approximation where $p \geq 0.0005$ (Figure 4(**c**), bottom). The area of the latent unit approximation measures how localized a unit is compared to the environment's area, which measures $40 \times 65 = 2,600$ lattice units. The latent unit approximations have a mean area of 254.6 lattice units and a 80% of areas are < 352.6 lattice units, which cover 9.79% and 13.6% of the environment, respectively.

The units in the neural network's latent space provide a unique combinatorial code for each spatial position.The aggregate of latent units covers the environment's entire physical space. At each lattice block in the environment, we calculate the number of active latent units (Figure 4(**d**), left). The number of active latent units is different in 87.6% of the lattice blocks. Every lattice block has at least one active latent unit, which indicates the aggregate of the latent units cover the environment's physical space.

Lastly, we demonstrate that the neural network can measure physical distances and could perform vector

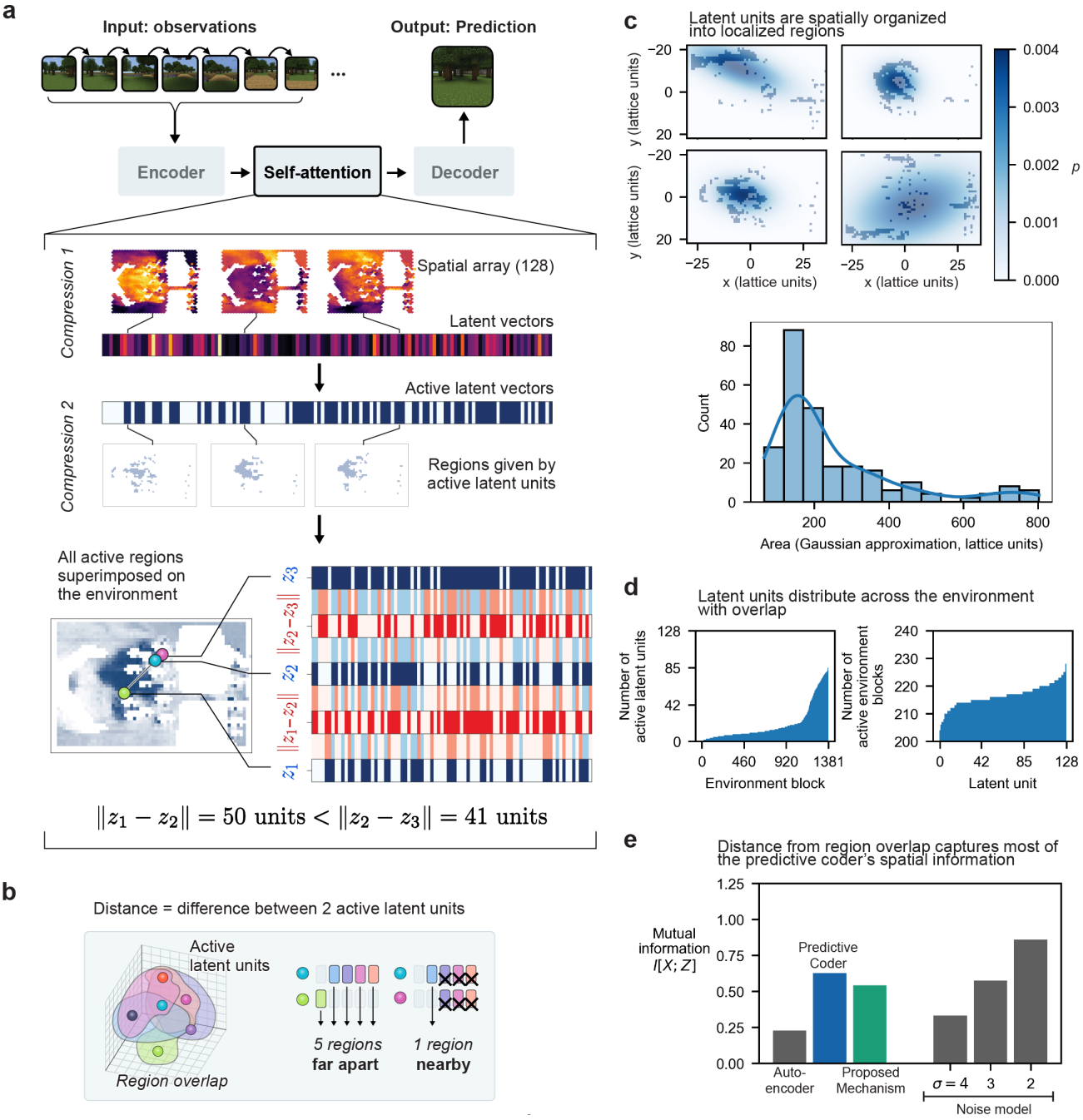


**Figure 4. The predictive coding network generates place fields that support vector based distance calculations** **a,** when encoding past images for predictive coding, the self-attention module generates latent vectors. Each continuous unit in these latent vectors activates in concentrated, localized regions in physical space. These continuous units can be thresholded to generate a binary vector determining whether each unit is active. Each latent unit covers a unique region, and each physical location gives a unique combination of these overlapping regions. As an agent moves away from its original location, the combination of overlapping regions gradually deviates from its original combinations. This deviation, as measured by Hamming distance, correlates with physical distance. **b,** distance is given by the difference in the latent units' overlapping regions. Two nearby locations have small deviations in overlap (right) while two distant locations have large deviations (middle). **c**, latent units are spatially organized into localized regions. The active latent units are approximated by a two-dimensional Gaussian distribution to measure the latent unit's localization (top). The latent units' Gaussian approximations are highly localized with a mean area of $254.6$ for densities $p \geq 0.0005$. **d,** latent units distributed across the environment. The number of latent units was calculated as each lattice block in the environment (left), and the number of lattice blocks were calculated for each active unit (right). The latent units provide a unique combination for $87.6\%$ of the environment, and their aggregate covers the entire environment. **e,** distance from the region overlap captures most of the predictive coder's spatial information. We calculate the distance for every pair of active latent vectors and their respective physical Euclidean distances as a joint distribution. The proposed mechanism captures a majority of the predictive coder's spatial information—as the proposed mechanism's mutual information (0.542 bits) compares to the predictive coder's mutual information (0.627 bits)

navigation by comparing the combinations of overlapping regions in its latent space. We first determine the active latent units by thresholding each continuous value by its 90th-percentile value. At each position, we have a 128-dimensional binary vector that gives the overlap of 128 latent units. At every two positions, we calculate the Hamming distance between each binary latent vector as well as the physical Euclidean distance (Figure 4(**a**), bottom). Similar to the Sections  and , we compute the joint densities of the binary vectors' Hamming distances and the physical positions' Euclidean distances. We then calculate their mutual information to measure how much spatial information the Hamming distance captures. The proposed mechanism for the neural network's distance measurement—the binary vector's Hamming distance—gives a mutual information of 0.542 bits, compared to the predictive coder's mutual information of 0.627 bits and the auto-encoder's mutual information of 0.227 bits (Figure 4(**e**)). Compared to the auto-encoder, the vector based distance calculations capture a majority amount of the predictive coder's spatial information.

## Discussion

Mapping is a general mechanism for generating an internal representation of sensory information. While spatial maps facilitate navigation and planning within an environment, mapping is a ubiquitous neural function that extends to representations beyond visual-spatial mapping. The primary sensory cortex (S1), for example, maps tactile events topographically. Physical touches that occur in proximity are mapped in proximity for both the neural representations and the anatomical brain regions45s[39]. Similarly, the cortex maps natural speech by tiling regions with different words and their relationships, which shows that topographic maps in the brain extend to higher-order cognition. Similarly, the cortex maps natural speech by tiling regions with different words and their relationships, which shows that topographic maps in the brain extend to higher-order cognition. The similar representation of non-spatial and spatial maps in the brain suggests a common mechanism for charting cognitive[40]. However, it is unclear how a single mechanism can generate both spatial and non-spatial maps.

Here, we show that predictive coding provides a basic, general mechanism for charting spatial maps by predicting sensory data from past sensory experiences. Our theoretical framework applies to any vector valued sensory data and could be extended to auditory data, tactile data, or tokenized representations of language. We demonstrate a neural network that performs predictive coding can construct an implicit spatial map of an environment by assembling information from local paths into a global frame within the neural network's latent space. The implicit spatial map depends specifically on the sequential task of predicting future visual images . Neural networks trained as auto-encoders do not reconstruct a faithful geometric representation in the presence of physically distant yet visually similar landmarks.

Moreover, we study the predictive coding neural network's representation in latent space. Each unit in the network's latent space activates at distinct, localized regions—called place fields—with respect to physical space. At each physical location, there exists a unique combination of overlapping place fields. At two locations, the differences in the combinations of overlapping place fields provides the distance between the two physical locations. The existence of place fields in both the neural network and the hippocampus7[15] suggest that predictive coding is a universal mechanism for mapping. In addition, vector navigation emerges naturally from predictive coding by computing distances from overlap- ping place field units. Predictive coding may provide a model for understanding how place cells emerge, change, and function.

Predictive coding can be performed over any sensory modality that has some temporal sequence. As natural speech forms a cognitive map, predictive coding may underlie the geometry of human language. Intriguingly, large language models train on causal word prediction, a form of predictive coding, build internal maps that support generalized reasoning, answer questions, and mimic other forms of higher order reasoning[41]. Similarities in spatial and non-spatial maps in the brain suggest that large language models organize language into a cognitive map and chart concepts geometrically. These results all suggest that predictive coding might provide a unified theory for building representations of information—connecting disparate theories including place cell formation in the hippocampus, somatosensory maps in the cortex, and human language.

## Acknowledgements

We deeply appreciate Inna Strazhnik for her exceptional contributions to the scientific visualizations and figure illustrations. Her expertise in translating our research into clear visuals has significantly elevated the clarity and impact of our paper. We express our heartfelt gratitude to Thanos Siapas, Evgueniy Lubenov, Dean Mobbs, and Matthew Rosenberg for their invaluable and insightful discussions which profoundly enriched our work. Their expertise and feedback have been instrumental in the development and realization of this research. Additionally, we appreciate the insights provided by Lixiang Xu, Meng Wang, and Jieyu Zheng, which played a crucial role in refining various aspects of our study. The dedication and collaborative spirit of this collective group have truly elevated our research, and for that, we are deeply thankful.

## References

1. Epstein, R. A., Patai, E. Z., Julian, J. B. & Spiers, H. J. The Cognitive Map in Humans: Spatial Navigation and Beyond. *Nature Neuroscience* **20.** 1504–1513. (2023) (Nov. 2017).
2. Wang, Z. J. & Thomson, M. Localization of signaling receptors maximizes cellular information acquisition in spatially structured natural environments. *Cell Systems* **13.** 530–546 (2022).

3. Anderson, J. *Cognitive Psychology and Its Implications* Ninth edition (Worth Publishers, New York City, Jan. 2020).
4. Rescorla, M. Cognitive maps and the language of thought. *The British Journal for the Philosophy of Science* (2009).
5. Whittington, J. C., McCaffary, D., Bakermans, J. J. & Behrens, T. E. How to build a cognitive map. *Nature neuroscience* **25.** 1257–1272 (2022).
6. Aronov, D., Nevers, R. & Tank, D. W. Mapping of a Non-Spatial Dimension by the Hippocampal–Entorhinal Circuit. *Nature* **543.** 719–722. (2022) (Mar. 2017).
7. Nieh, E. H. *et al.* Geometry of Abstract Learned Knowledge in the Hippocampus. *Nature* **595.** 80–84. (2022) (July 2021).
8. Whittington, J. C. *et al.* The Tolman-Eichenbaum machine: unifying space and relational memory through generalization in the hippocampal formation. *Cell* **183.** 1249–1263 (2020).
9. Wilson, R. C., Takahashi, Y. K., Schoenbaum, G. & Niv, Y. Orbitofrontal cortex as a cognitive map of task space. *Neuron* **81.** 267–279 (2014).
10. Constantinescu, A. O., O'Reilly, J. X. & Behrens, T. E. J. Organizing Conceptual Knowledge in Humans with a Gridlike Code. *Science* **352.** 1464–1468. (2022) (June 2016).
11. Garvert, M. M., Dolan, R. J. & Behrens, T. E. A Map of Abstract Relational Knowledge in the Human Hippocampal–Entorhinal Cortex. *eLife* **6** (ed Davachi, L.) e17086. (2023) (Apr. 2017).
12. Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E. & Gallant, J. L. Natural Speech Reveals the Semantic Maps That Tile Human Cerebral Cortex. *Nature* **532.** 453–458. (2023) (Apr. 2016).
13. Corkin, S. Lasting Consequences of Bilateral Medial Temporal Lobectomy: Clinical Course and Experimental Findings in H.M. *Seminars in Neurology* **4.** 249–259. (2023) (June 1984).
14. Behrens, T. E. *et al.* What is a cognitive map? Organizing knowledge for flexible behavior. *Neuron* **100.** 490–509 (2018).
15. O'Keefe, J. Place Units in the Hippocampus of the Freely Moving Rat. *Experimental Neurology* **51.** 78–109. (2021) (Jan. 1976).
16. Hafting, T., Fyhn, M., Molden, S., Moser, M.-B. & Moser, E. I. Microstructure of a Spatial Map in the Entorhinal Cortex. *Nature* **436.** 801–806. (2020) (Aug. 2005).
17. Amaral, D. G., Ishizuka, N. & Claiborne, B. Neurons, Numbers and the Hippocampal Network. *Progress in Brain Research* **83.** 1–11 (1990).
18. Cueva, C. J. & Wei, X.-X. Emergence of Grid-like Representations by Training Recurrent Neural Networks to Perform Spatial Localization. *arXiv:1803.07770 [cs, q-bio, stat].* arXiv: `1803.07770 [cs, q-bio, stat]`. (2020) (Mar. 2018).
19. Banino, A. *et al.* Vector-Based Navigation Using Grid-like Representations in Artificial Agents. *Nature* **557.** 429–433. (2020) (May 2018).
20. Thrun, S. & Montemerlo, M. The Graph SLAM Algorithm with Applications to Large-Scale Mapping of Urban Structures. *The International Journal of Robotics Research* **25.** 403–429. (2023) (May 2006).
21. Mur-Artal, R. & Tardós, J. D. Visual-Inertial Monocular SLAM With Map Reuse. *IEEE Robotics and Automation Letters* **2** (Apr. 2017).
22. Mourikis, A. I. & Roumeliotis, S. I. *A Multi-State Constraint Kalman Filter for Vision-aided Inertial Navigation* in *Proceedings 2007 IEEE International Conference on Robotics and Automation* (Apr. 2007), 3565–3572.
23. Lynen, S. *et al. Get Out of My Lab: Large-scale, Real-Time Visual-Inertial Localization* (July 2015).
24. Gupta, S. *et al. Cognitive Mapping and Planning for Visual Navigation* Feb. 2019. arXiv: `1702.03920 [cs]`. (2022).
25. Mirowski, P. *et al. Learning to Navigate in Cities Without a Map* in *Advances in Neural Information Processing Systems* **31** (Curran Associates, Inc., 2018). (2022).
26. Duan, Y. *et al.* $RL^2$: Fast Reinforcement Learning via Slow Reinforcement Learning. (2022) (Nov. 2016).
27. Lee, T. S. & Mumford, D. Hierarchical Bayesian inference in the visual cortex. *JOSA A* **20.** 1434–1448 (2003).
28. Mumford, D. *Pattern theory: a unifying perspective* in *First European Congress of Mathematics: Paris, July 6-10, 1992 Volume I Invited Lectures (Part 1)* (1994), 187–224.
29. Rao, R. P. N. & Ballard, D. H. Predictive Coding in the Visual Cortex: A Functional Interpretation of Some Extra-Classical Receptive-Field Effects. *Nature Neuroscience* **2.** 79–87. (2023) (Jan. 1999).
30. Poincaré, H. *The Foundations of Science: Science and Hypothesis, the Value of Science, Science and Method* trans. by Halsted, G. B. (Cambridge University Press, Cambridge, 2015).
31. O'Keefe, J. & Nadel, L. *The Hippocampus as a Cognitive Map* (Clarendon Press ; Oxford University Press, Oxford : New York, 1978).
32. Johnson, M., Hofmann, K., Hutton, T. & Bignell, D. *The Malmo Platform for Artificial Intelligence Experimentation* in *International Joint Conference on Artificial Intelligence* (2016).
33. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. *arXiv:1512.03385 [cs].* arXiv: `1512.03385 [cs]`. (2019) (Dec. 2015).
34. Ronneberger, O., Fischer, P. & Brox, T. *U-Net: Convolutional Networks for Biomedical Image Segmentation* May 2015. arXiv: `1505.04597 [cs]`. (2023).
35. Vaswani, A. *et al. Attention Is All You Need* Aug. 2023. arXiv: `1706.03762 [cs]`. (2023).
36. Sutskever, I., Martens, J., Dahl, G. & Hinton, G. On the Importance of Initialization and Momentum in Deep Learning.
37. Smith, L. N. & Topin, N. *Super-Convergence: Very Fast Training of Neural Networks Using Large Learning Rates* May 2018. arXiv: `1708.07120 [cs, stat]`. (2023).
38. Tenenbaum, J. B., de Silva, V. & Langford, J. C. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science* **290.** 2319–2323. (2023) (Dec. 2000).
39. Rosenthal, I. A. *et al.* S1 Represents Multisensory Contexts and Somatotopic Locations within and Outside the Bounds of the Cortical Homunculus. *Cell Reports* **42.** 112312. (2023) (Apr. 2023).
40. Behrens, T. E. J. *et al.* What Is a Cognitive Map? Organizing Knowledge for Flexible Behavior. *Neuron* **100.** 490–509. (2023) (Oct. 2018).
41. Brown, T. B. *et al.* Language Models Are Few-Shot Learners. *arXiv:2005.14165 [cs].* arXiv: `2005.14165 [cs]`. (2020) (July 2020).