Predictive Learning Induces Probabilistic Cognitive Maps

Yeowon Kim
ywxxx0306@kaist.ac.kr

Yul HR Kang
yulkang@kaist.ac.kr

Department of Bio and Brain Engineering
Korea Advanced Institute of Science and Technology (KAIST)
Daejeon, South Korea

Abstract

Navigation requires inferring one's pose (location and heading) in an environment based on noisy and ambiguous egocentric sensory inputs. While place cells in the brain are thought to represent an animal's allocentric location and associated uncertainty, the mechanisms by which these probabilistic representations are learned remain unclear. To address this, we developed a model of an agent that navigates using noisy egocentric visual and self-motion signals. We demonstrate that, when the agent is trained to predict future visual stimuli, its hidden representations closely resemble the posterior belief about pose, as computed by a Bayesian ideal observer. Moreover, these hidden representations, like the posterior beliefs of the ideal observer, also resembled place cell activity both in familiar and unfamiliar environments. This resemblance was significantly weaker when the agent was trained as an autoencoder to reproduce its current visual input. Our findings suggest that learning to predict noisy sensory inputs can give rise to probabilistic cognitive maps—probabilistic representations of latent states such as pose—which are essential for Bayesian inference in the brain.

1 Introduction

Navigation requires localizing oneself in an environment. However, one's pose (location and heading direction) is not directly observable (i.e., is a latent state), and must be inferred given noisy and partial information such as egocentric visual and self-motion signals. Recent findings suggest that uncertainty about the latent state must be considered for optimal localization[1]. Indeed, this uncertainty is considered by individuals during navigation, and is represented by place fields in the brain[1, 2]. However, it remains unclear how this probabilistic representation of the latent state is acquired in the brain.

Previous studies have demonstrated that predicting sensory observations during navigation contributes to the emergence of stable latent space representations [3, 4, 5, 6, 7, 8]. However, these studies have not addressed the formation of probabilistic beliefs, which must represent not only a point estimate of the latent variable but also the associated uncertainty. In this study, we demonstrate that probabilistic representations naturally arise when a neural network is trained to predict future sensory inputs.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

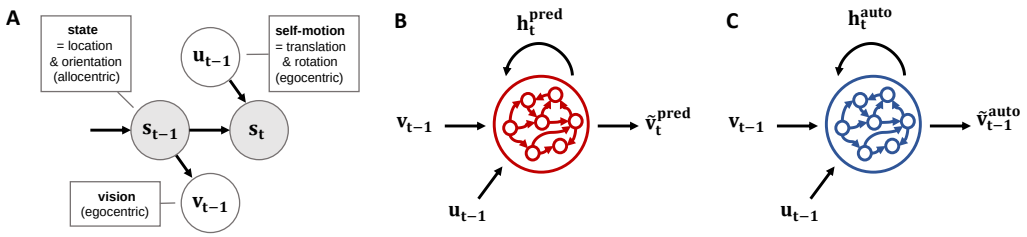

Figure 1: **A**. Generative model of navigation used by the ideal observer. The agent's state (**s**, the combination of its allocentric orientation and location) is updated by egocentric self-motion (**u**) and give rise to visual input (**v**). **B**. Architecture of the pred-RNN. **C**. Architecture of the autoencoder.

We developed a variant of an autoencoder that processes noisy egocentric visual stimuli and self-motion signals from an agent navigating an environment. Our findings show that the hidden state representation resembles the belief of a handcrafted ideal observer about its location. This representation not only reflects the optimal estimate of the location but also the estimate's uncertainty. Remarkably, even when the actual environment differed from the one the agent believed it was in, the hidden state of the model still resembled the ideal observer's belief. Furthermore, the activity of hidden state activity exhibited place-cell like activities, displaying characteristics consistent with known features of place cells. These results suggest that learning to predict upcoming noisy sensory inputs may be a mechanism for learning probabilistic representation in the brain, even in a natural task like spatial navigation where the relationship between the sensory input and the latent state is complex.

2 Results

2.1 Predictive recurrent neural network (pred-RNN) model

To study the probabilistic beliefs of an agent given noisy sensory inputs for localization during navigation, we first constructed a Bayesian ideal observer, which has been shown to explain animals' behavior and neural activity[1, 9, 10, 11]. The ideal observer uses Bayesian filtering to update a posterior distribution ("belief" $\mathbf{p}^{\text{ideal}}$) over its position and heading direction ("state" **s**), which is not directly observable, and hence must be inferred from its noisy egocentric visual and self-motion inputs (**v** & **u**, Fig. 1A). The belief shows varying levels of uncertainty depending on the history of the visual and self-motion inputs (Fig. 2A, top three rows).

Then, we hypothesized that the brain might learn to represent such probabilistic beliefs by predicting the upcoming sensory input[12]. To test this hypothesis, we constructed a predictive recurrent neural network (pred-RNN) model that predicts the next visual input given the history of visual and self-motion inputs (Fig. 1B). The model compresses the incoming noisy visual input ($\mathbf{v}_{t-1}$) and feeds it into a recurrent layer along with self-motion signals. This provides the recurrent layer the information necessary to update the beliefs about the latent state in Bayesian filtering. We call the activity of this recurrent layer as the network's state ($\mathbf{h}_t^{\text{pred}}$). This state is then decompressed to yield a predicted visual input of the *next* time step ($\tilde{\mathbf{v}}_t^{\text{pred}}$) which it did successfully (Fig. 2A, Row 6).

As a control, we also trained another network ("autoencoder") with the identical structure and input as the pred-RNN (Fig. 1C). The only difference was that it was trained to reproduce the visual input at the *current* time step ($\tilde{\mathbf{v}}_{t-1}^{\text{auto}}$), which it also did successfully (Fig. 2A, Row 4). To ensure successful training for both networks, we tried multiple learning rates and selected the best rate for each network. To emphasize, while we refer to this network as an autoencoder, distinct from the pred-RNN described above, the two networks had the same architecture and received the same input at each time step. The only difference was which visual input they were trained to match: the current time step's (autoencoder) or the next time step's (pred-RNN).

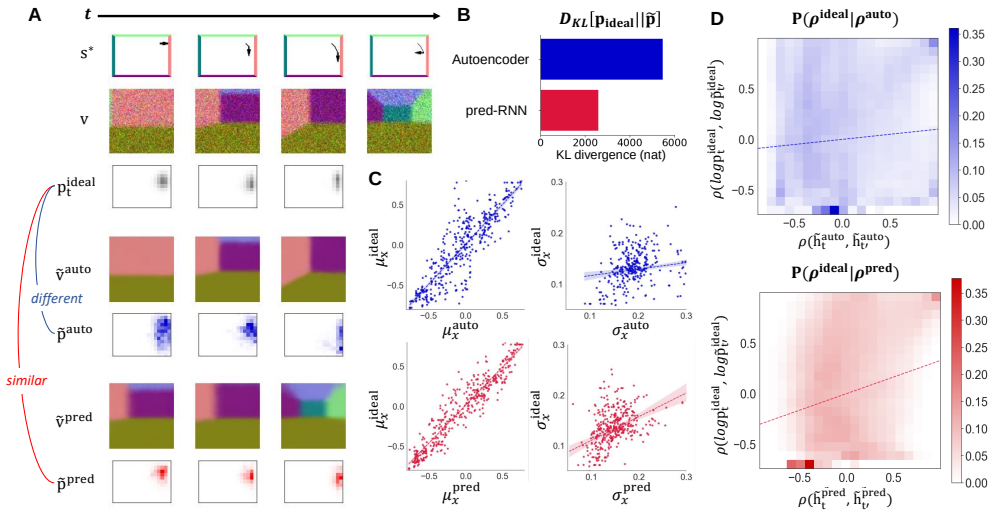

Figure 2: **A**. *Row 1.* Top-down view of an agent navigating an environment. The arrow shows the current location and head direction. Thin curves indicate the past trajectory. *Row 2.* Noisy first-person view of the 3D environment (**v**). *Row 3.* Posterior distribution of the location calculated by the ideal observer. *Row 4.* Reconstruction of the current visual input by the autoencoder. *Row 5.* Posterior distribution decoded from $\mathbf{h}^{\text{auto}}$. *Row 6.* Prediction of the next visual input by pred-RNN. *Row 7.* Posterior distribution decoded from $\mathbf{h}^{\text{pred}}$. **B**. D_{KL} between the ideal observer posterior ($\mathbf{p}^{\text{ideal}}$) and the posterior decoded from the autoencoder ($\tilde{\mathbf{p}}^{\text{auto}}$) vs. that from the pred-RNN ($\tilde{\mathbf{p}}^{\text{pred}}$). **C**. Comparisons of the momentary mean/SDs of $\mathbf{p}^{\text{ideal}}$ and those of $\tilde{\mathbf{p}}^{\text{auto/pred}}$ (*Row 1/2*). **D**. 2D histograms showing correlation between the representational similarities of the ideal observer's posterior ($\log \tilde{\mathbf{p}}^{\text{ideal}}$) and the hidden states of the pred-RNN ($\tilde{\mathbf{h}}^{\text{pred}}$, Top), and the autoencoder ($\tilde{\mathbf{h}}^{\text{auto}}$, Bottom).

2.2 The pred-RNN encodes posterior beliefs

To examine whether the pred-RNN's state ($\mathbf{h}^{\text{pred}}$) and/or that of the autoencoder's ($\mathbf{h}^{\text{auto}}$) represents the ideal observer's uncertainty about its location, we decoded each network's state using a single fully connected (FC) layer, followed by a softmax function to obtain a probability distribution ($\tilde{\mathbf{p}}^{\text{pred}}$ or $\tilde{\mathbf{p}}^{\text{auto}}$). We optimized the FC layer to minimize the Kullback-Leibler Divergence (D_{KL}) of the ideal observer posterior ($\mathbf{p}^{\text{ideal}}$) from the decoded distribution, which we call the decoded posterior ($\tilde{\mathbf{p}}^{\text{pred/auto}}$). Note that this optimization procedure was identical for the two networks. We then fixed the FC layer's weights for each network and decoded $\tilde{\mathbf{p}}^{\text{pred/auto}}$ with a test data set, and compared their match to $\mathbf{p}^{\text{ideal}}$.

We found that $\tilde{\mathbf{p}}^{\text{pred}}$ matched $\mathbf{p}^{\text{ideal}}$ significantly better than $\tilde{\mathbf{p}}^{\text{auto}}$ did ($\Delta D_{\text{KL}} > 2500$; Fig. 2A rows 3, 5, and 7 and Fig. 2B). To determine what aspect of the decoded distributions accounted for this difference, we further compared the correlations of the mean and standard deviation (SD) of $\tilde{\mathbf{p}}^{\text{pred/auto}}$ with those of $\mathbf{p}^{\text{ideal}}$. We found that not only the decoded means but also the decoded SDs of the pred-RNN showed significant Pearson correlation with those of the ideal observer along both x & y axes ($\rho^{\text{pred}}_{\mu_x/\mu_y/\sigma_x/\sigma_y} = 0.94/0.92/0.46/0.59$, all $p < 10^{-4}$), which were significantly higher than those of the autoencoder (all $p \leq 0.001$). Therefore, pred-RNN not only encoded the estimate (mean) of the location better but also the associated uncertainty (SD) better than the autoencoder.

We corroborated this finding by ruling out potential alternative explanations for the correlation. We constructed a multiple regression model that explains $\sigma^{\text{ideal}}_{\text{x/y}}$ with not only $\sigma^{\text{pred}}_{\text{x/y}}$ but also with other variables known to correlate with the ideal observer's uncertainty about its location, such as the agent's true x and y location, distance to the closer of the north and south walls, distance to the closer of the west and east walls, sine and cosine of the heading direction, translation, and rotation[2, 13]. We compared this model ($\mathcal{M}^{\text{x/y}}_1$)'s fit with a reduced model without $\sigma^{\text{pred}}_{\text{x/y}}$ ($\mathcal{M}^{\text{x/y}}_2$) using the Bayesian Information Criterion (BIC). The comparison overwhelmingly supported the

full model ($\mathcal{M}_1^{x/y}$; $\Delta\text{BIC} > 300$ for both x and y), supporting that the pred-RNN's representation of uncertainty cannot be explained by the representation of the other variables.

2.3 The pred-RNN's representation resembles that of an ideal observer

A potential concern is that our method relies on decoding uncertainty using the ideal observer's posterior distribution as the ground truth. This raises questions about whether the internal representations learned by our models inherently align with the ideal observer's representations. To address this, we conducted analyses similar to Representational Similarity Analysis (RSA) to explore the relationship between the representations learned by our two models and those of the ideal observer[5, 14]. We first applied principal component analysis (PCA) to reduce the dimensionality of $\mathbf{h}^{\text{pred}}$ and $\mathbf{h}^{\text{auto}}$ to 10 principal components ($\tilde{\mathbf{h}}^{\text{pred}}$, $\tilde{\mathbf{h}}^{\text{auto}}$). Similarly, we applied PCA to log-transformed $\mathbf{p}^{\text{ideal}}$ to obtain 10 PCs ($\log \tilde{\mathbf{p}}^{\text{ideal}}$). We then computed the Pearson correlation coefficients between pairs of time steps within each of these transformed representations. The correlations within $\tilde{\mathbf{h}}^{\text{pred}}$ exhibited a much stronger alignment with those within $\log \tilde{\mathbf{p}}^{\text{ideal}}$ (Fig. 2D), as reflected on a higher Pearson correlation ($\rho(\rho(\log \tilde{\mathbf{p}}_t^{\text{ideal}}, \log \tilde{\mathbf{p}}_{t'}^{\text{ideal}}), \rho(\tilde{\mathbf{h}}_t^{\text{pred}}, \tilde{\mathbf{h}}_{t'}^{\text{pred}})$=0.32), compared to the autoencoder ($\rho(\rho(\log \tilde{\mathbf{p}}_t^{\text{ideal}}, \log \tilde{\mathbf{p}}_{t'}^{\text{ideal}}), \rho(\tilde{\mathbf{h}}_t^{\text{auto}}, \tilde{\mathbf{h}}_{t'}^{\text{auto}})$=0.12) over 2000 time steps. This result shows that, even without explicit decoding, the ideal observer's representational geometry resembles the pred-RNN's more closely compared to the autoencoder's.

2.4 The pred-RNN encodes posterior beliefs even in deformed environments

Thus far, we have demonstrated that the pred-RNN can learn a probabilistic representation resembling that of an ideal observer. A natural next question is whether this representation also resemble the neural activity observed in the brain.

To address this question, we hypothesized that (1) place cells, traditionally thought to represent the animal's allocentric location [15, 16], may actually represent the animal's beliefs about its location, and (2) these beliefs are represented by the pred-RNN, making its activity similar to that of place cells. To rigorously test these hypotheses, we compared the pred-RNN's activity with place cell activity not only when the animal's beliefs were expected to closely match its true allocentric location, but also when they were expected to diverge. The latter case occurs when the animal is placed in an environment that appears similar to a familiar one, but is unknowingly stretched or compressed, as in experiments by O'Keefe and Burgess (1996)[17]

We first confirmed that the pred-RNN's activity resembles the ideal observer's beliefs, even when the environment is unfamiliar. To achieve this, we had the ideal observer interpret the sensory input from the unfamiliar environment using the map of the familiar environment to form its beliefs. In parallel, we trained the pred-RNN's weights to predict the next visual input in the familiar environment, and froze its weights. We then compared the beliefs decoded from its activity with the ideal observer's beliefs in an unfamiliar environment. Specifically, the agent was trained in a vertical rectangle-shaped environment, and tested in three deformed environments: a small square, a horizontal rectangle, and a large square. Importantly, the same weights were used for both the pred-RNN and its decoding in the unfamiliar environments as those trained in the familiar environment (vertical rectangle). In all three test environments, $\tilde{\mathbf{p}}^{\text{pred}}$ matched $\mathbf{p}^{\text{ideal}}$ significantly better than $\tilde{\mathbf{p}}^{\text{auto}}$ (Fig. 3B), with the D_{KL} between $\mathbf{p}^{\text{ideal}}$ and $\tilde{\mathbf{p}}^{\text{pred}}$ consistently smaller than that for $\tilde{\mathbf{p}}^{\text{auto}}$ (Fig. 3C). These results demonstrate that the pred-RNN's representation closely resembles the ideal observer's beliefs, even when the beliefs are based on an incorrect model of the environment.

2.5 The pred-RNN's activity resembles place cell activity across deformed environments

Having established that the pred-RNN encodes posterior beliefs, we then compared its activity directly with that of place cells. After applying nonlinear transformation (involving exponentiation, a shift, and normalization), multiple units in both $\mathbf{h}^{\text{pred}}$ and $\mathbf{h}^{\text{auto}}$ displayed place cell-like activity in their spatial rate maps (see Supplementary materials). Notably, the rate maps from $\mathbf{h}^{\text{pred}}$ exhibited more distinct and localized areas of activity, closely resembling place fields (Fig. 3D).

We then quantitatively assessed whether the activities of $\mathbf{h}^{\text{pred}}$ and $\mathbf{h}^{\text{auto}}$ match the known properties of place cells in deformed environments. Specifically, it has been reported that the place fields stretch or compress along the direction of environmental expansion or compression [17], aligning

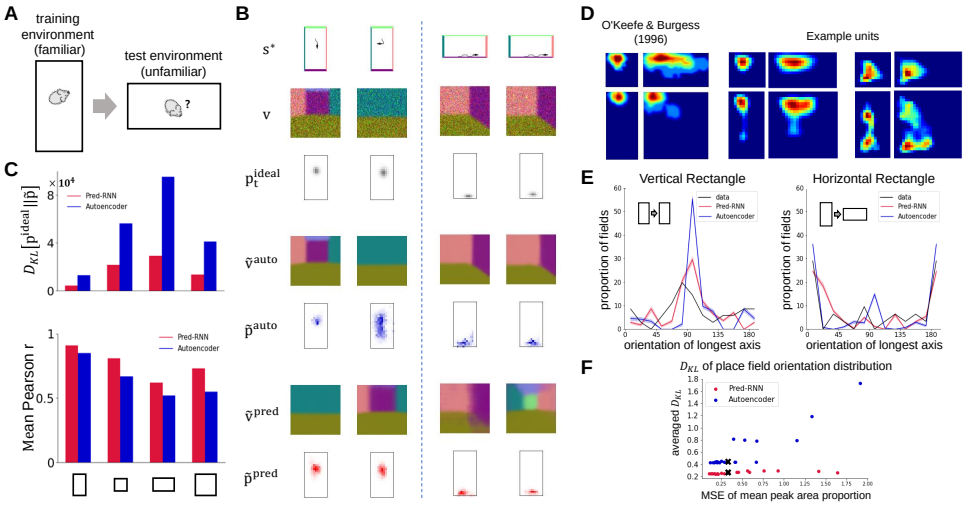

Figure 3: **A**. Agent tested in an unfamiliar environment after being trained in a familiar environment. **B**. Output and (decoded) posterior distributions of each model in the vertical rectangle (familiar environment) and the horizontal rectangle (unfamiliar environment). **C**. D_{KL} between the ideal observer's posterior ($\mathbf{p}^{\mathrm{ideal}}$) and the posteriors decoded from the autoencoder ($\tilde{\mathbf{p}}^{\mathrm{auto}}$) and pred-RNN ($\tilde{\mathbf{p}}^{\mathrm{pred}}$) across different test environments. **D**. (Left) Firing rate map of a place cell when the rat was familiarized with the vertical rectangle environment and was then recorded in all four environments [17]. (Middle) Firing rate map of a unit in $\mathbf{h}^{\mathrm{pred}}$, showing fields stretched along with the environment, similar to the empirical data. (Right) Example of a unit in $\mathbf{h}^{\mathrm{auto}}$, displaying extraneous peaks. **E**. The distribution of place field orientations (the axis of the longest extent of the field) of the units from the animals, $\mathbf{h}^{\mathrm{pred}}$, and $\mathbf{h}^{\mathrm{auto}}$. **F**. D_{KL} of the distribution of orientations between the data from the animals and the models, averaged across four environments. It is plotted against the mean squared error (MSE) of the peak areas of the place fields between the animals and the models. While the model's place field areas depend on the choice of free parameters for the nonlinear transformation, the pred-RNN consistently exhibited a D_{KL} smaller than that of the autoencoder. Black markers correspond to the examples shown above.

the longest axis of the fields ("place field orientation") with the longest axis of the environment. We found that the activity of the pred-RNN exhibited this property significantly more strongly than that of the autoencoder. Across all combinations of free parameters tried for the nonlinear transformation, the average D_{KL} between the empirical place field orientation distribution and that of the pred-RNN was smaller than that of the autoencoder (Fig. 3E). This demonstrates that the pred-RNN's hidden unit activity closely resembles not only the ideal observer's beliefs about location but also place cell activities in both familiar and unfamiliar environments. Since in the latter (unfamiliar) environments the beliefs cannot match a "true" location, this supports our hypothesis that the activity of both place cells and the pred-RNN encodes the animal or agent's beliefs about its location, rather than the true allocentric location *per se*.

3 Conclusion

In this study, we investigated whether probabilistic representations of a latent state can be learned without supervision (i.e., without providing the network with the correct latent state). We found that when a neural network is trained to predict upcoming egocentric visual inputs during navigation, its hidden states encode the ideal observer's beliefs about pose, a latent state. This probabilistic representation not only aligned with the ideal observer's beliefs but also with the place cell activity. Furthermore, this match occurred not only in a familiar environment but also in an unfamiliar one, where the beliefs were expected not to match the true latent state. These findings support our hypotheses that (1) the place fields do not necessarily represent the animal's true allocentric location (which is not directly available to the animal but must be inferred from sensory observations), but

its beliefs about the location, and (2) such probabilistic representations of beliefs can be learned purely by predicting upcoming sensory inputs.

We plan to extend this work in several key directions. First, we aim to develop an analytical explanation for the empirical findings in this study. This will clarify the conditions under which such probabilistic representations can arise. Second, we will train the agent to learn to act toward a reward based on the predictive representation, to test its behavioral significance.

Studies on Marr's computational level often leave unclear how the brain learns to perform such computations. Our study bridges the computational and algorithmic levels, by providing a concrete, realistic mechanism by which a computationally optimal probabilistic representation—resembling that of an ideal observer's beliefs—can emerge without supervision.

References

[1] Yul HR Kang, Daniel M Wolpert, and Máté Lengyel. Spatial uncertainty and environmental geometry in navigation. *bioRxiv*, 2023.

[2] S. Tanni, W. de Cothi, and C. Barry. State transitions in the statistically stable place cell population correspond to rate of perceptual change. *Current Biology*, 2022.

[3] Stefano Recanatesi, Matthew Farrell, Janik Born, Gabriel K. Ocker, and Maile Byron. Predictive learning as a network mechanism for extracting low-dimensional latent space representations. *Nature Communications*, 12(1):21696, 2021.

[4] Benigno Uria, Borja Ibarz, Andrea Banino, Vinicius Zambaldi, Dharshan Kumaran, Demis Hassabis, Caswell Barry, and Charles Blundell. A model of egocentric to allocentric understanding in mammalian brains. *bioRxiv*, 2020.

[5] J. Gornet and M. Thomson. Automated construction of cognitive maps with visual predictive coding. *Nature Machine Intelligence*, 6:820–833, 2024.

[6] M.C. von Ebers and X.X. Wei. Cognitive maps from predictive vision. *Nature Machine Intelligence*, 6:850–851, 2024.

[7] James C.R. Whittington, Timothy H. Muller, Shirley Mark, Guifen Chen, Caswell Barry, Neil Burgess, and Timothy E.J. Behrens. The tolman-eichenbaum machine: Unifying space and relational memory through generalization in the hippocampal formation. *Cell*, 183(5):1249–1263.e23, 2020.

[8] Christopher J. Cueva and Xue-Xin Wei. Emergence of grid-like representations by training recurrent neural networks to perform spatial localization, 2018.

[9] Florian Kessler, Jan Frankenstein, and Constantin A. Rothkopf. Human navigation strategies and their errors result from dynamic interactions of spatial uncertainties. *Nature Communications*, 15:5677, 2024.

[10] A Castegnaro, Z Ji, K Rudzka, D Chan, and N Burgess. Overestimation in angular path integration precedes alzheimer's dementia. *Curr Biol.*, 33(21):4650–4661.e7, 2023.

[11] K.J. Lakshminarasimhan, E. Avila, X. Pitkow, et al. Dynamical latent state computation in the male macaque posterior parietal cortex. *Nature Communications*, 14:1832, 2023.

[12] James Aitchison. Goodness of prediction fit. *Biometrika*, 62(3):547–554, 1975.

[13] Christian F. Doeller and Neil Burgess. Distinct error-correcting and incidental learning of location relative to landmarks and boundaries. *Proceedings of the National Academy of Sciences*, 105(15):5909–5914, 2008.

[14] Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis–connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008.

[15] J O'Keefe and J Dostrovsky. The hippocampus as a spatial map. preliminary evidence from unit activity in the freely-moving rat. *Brain Res.*, 34(1):171–175, 1971.

[16] John O'Keefe. Place units in the hippocampus of the freely moving rat. *Experimental Neurology*, 51:78–109, 1976.

[17] J O'Keefe and N Burgess. Geometric determinants of the place fields of hippocampal neurons. *Nature*, 381(6581):425–428, 1996.

[18] T Hartley, N Burgess, C Lever, F Cacucci, and J O'Keefe. Modeling place fields in terms of the cortical inputs to the hippocampus. *Hippocampus*, 10(4):369–379, 2000.

Supplementary materials

A. Ideal observer

The ideal observer was constructed as described in [1]. Briefly speaking, the agent receives noisy egocentric visual and self-motion inputs and recurrently its beliefs about its pose (location and heading direction) using Bayesian filtering. See [1] for details.

B. Pred-RNN architecture and training

The architecture of the pred-RNN model mirrors the steps of Bayesian filtering (Fig. 4). The pred-RNN compresses the incoming noisy visual input ($\mathbf{v}_{t-1}$) using three convolutional layers and two fully connected (FC) layers. This compressed representation is then fed into a recurrent FC layer, referred to as the network's state ($\mathbf{h}_{t-1}$), analogous to the measurement step of Bayesian filtering. In this step, $\mathbf{h}_{t-1}$ can be updated based on the new observation, similar to how the ideal observer's posterior distribution ($\mathbf{P}_{t-1}$) is updated using the incoming sensory input ($\mathbf{v}_{t-1}$).

Once the state $\mathbf{h}_{t-1}$ is updated, it is decoded using three convolutional layers and two FC layers to generate the predicted visual input ($\tilde{\mathbf{v}}_{t-1}$). In parallel with the prediction step in Bayesian filtering, where the posterior distribution is projected forward using a control signal ($\mathbf{u}_t$), a control signal containing information about the agent's intended movement is fed into an FC layer to produce $\mathbf{h}_t^{pred}$. This predicted hidden state ($\mathbf{h}_t^{pred}$) is used to generate the predicted visual input for the next time step ($\tilde{\mathbf{v}}_t$), analogous to the predicted posterior ($\mathbf{P}_t^{pred}$) in Bayesian filtering.

Networks are optimized by minimizing the sum of the measurement loss (MSE between decoded output of $\tilde{\mathbf{h}}_{t-1}$ and $\mathbf{v}_{t-1}$) and the prediction loss (MSE between decoded output of $\mathbf{h}_t^{pred}$ and $\mathbf{v}_t$). Note that the same decoder network is used for decoding the hidden state into visual input at each time step.

Note that, while we highlighted parallels between the pred-RNN architecture and Bayesian filtering to aid understanding, we did not train the pred-RNN to explicitly mimic Bayesian filtering computations. Instead, it was trained solely to predict the current and upcoming sensory input.

C. Autoencoder architecture and training

The architecture and input of the autoencoder are identical to those of the pred-RNN. The key distinction lies in the training objective: the autoencoder is trained to reconstruct the visual input at the current time step, $\mathbf{v}_{t-1}$. Consequently, the autoencoder's networks are optimized by minimizing the sum of the MSE between decoded output of $\mathbf{h}_{t-1}$ and $\mathbf{v}_{t-1}$, and the MSE between $\mathbf{h}_t^{auto}$ and $\mathbf{v}_{t-1}$.

D. Decoding posterior distribution

Regardless of whether the believed environment $\tilde{\mathcal{E}}$ is the same as or different from the true environment $\mathcal{E}^*$, the ideal observer posterior $\mathbf{p}_t^*$ at time step t is of length $|\tilde{\mathcal{E}}|$. The decoded posterior belief $\tilde{\mathbf{p}}_t$ has the same length, and is always decoded from the hidden state $\mathbf{h}_t$ with $\mathbf{W}_{\tilde{\mathcal{E}}}^{\text{post}}$ with dimensions $|\tilde{\mathcal{E}}| \times |\mathbf{h}|$:

$$\tilde{\mathbf{p}}_t = \text{softmax}\left(\mathbf{W}_{\tilde{\mathcal{E}}}^{\text{post}}\mathbf{h}_t\right) \tag{1}$$

Note that the belief of being at a particular state $\tilde{\mathbf{s}}$ is just the k($\tilde{\mathbf{s}}$)-th element of this vector (where k($\cdot$) is the index of a state), denoted simply as $\tilde{\mathbf{p}}_t(\tilde{\mathbf{s}})$.

E. Ideal observer posterior and decoded distributions over time

Fig. 5 illustrates $\tilde{\mathbf{p}}^{\text{pred}}$ (red) and $\tilde{\mathbf{p}}^{\text{auto}}$ (blue), alongside $\mathbf{p}^{\text{ideal}}$ (grey), across 10 consecutive time steps. $\tilde{\mathbf{p}}^{\text{pred}}$ consistently shows a closer alignment with $\mathbf{p}^{\text{ideal}}$ compared to $\tilde{\mathbf{p}}^{\text{auto}}$. This is particularly evident in situations where the agent has access to limited visual information, such as when only a single wall is visible. Under these conditions, the autoencoder often produces an excessively broad posterior, indicating an overestimation of uncertainty and a failure to accurately represent

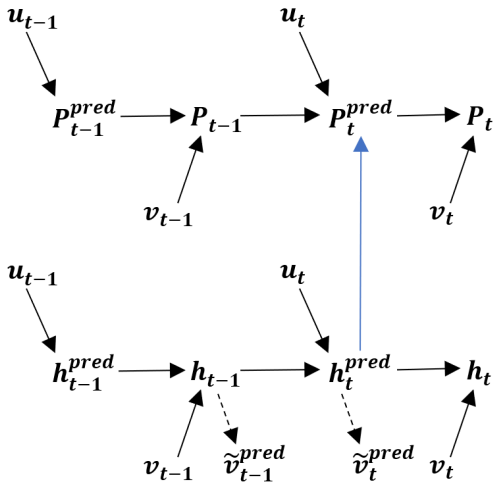

Figure 4: Diagram of pred-RNN's architecture, and its alignment with Bayesian filtering steps, showing validity of regressing $\mathbf{h}_t^{pred}$ to $\mathbf{P}_t^{pred}$.

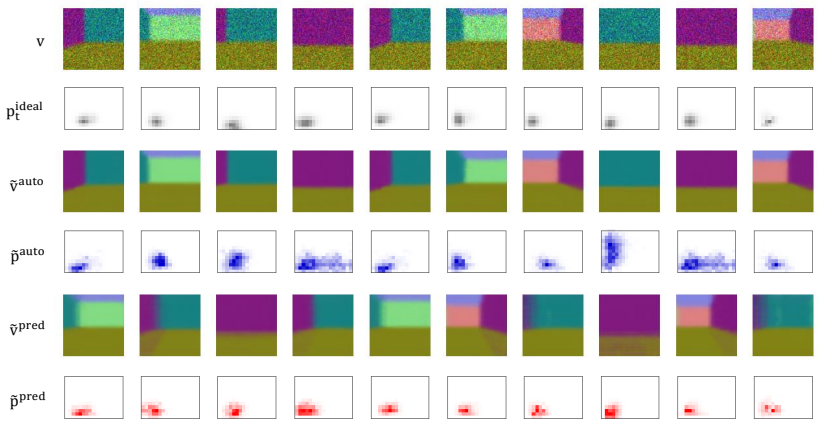

Figure 5: Decoded posterior distribution of pred-RNN and autoencoder across 10 time steps

the agent's location. This discrepancy underscores the pred-RNN's capability to form more precise and reliable representations by considering the history of the visual input.

F. Spatial rate maps and method for analysis

In the spatial rate maps shown in Fig. 6, clear distinctions emerge between the representations learned by the pred-RNN and the autoencoder. Prior to applying a nonlinear transformation, both models exhibit spatial tuning, but the activity is not as localized compared to the reported mean peak areas in empirical data (Top). To address this discrepancy, we applied nonlinear transformation (Bottom), and used the transformed activity for analysis. We applied the following nonlinear

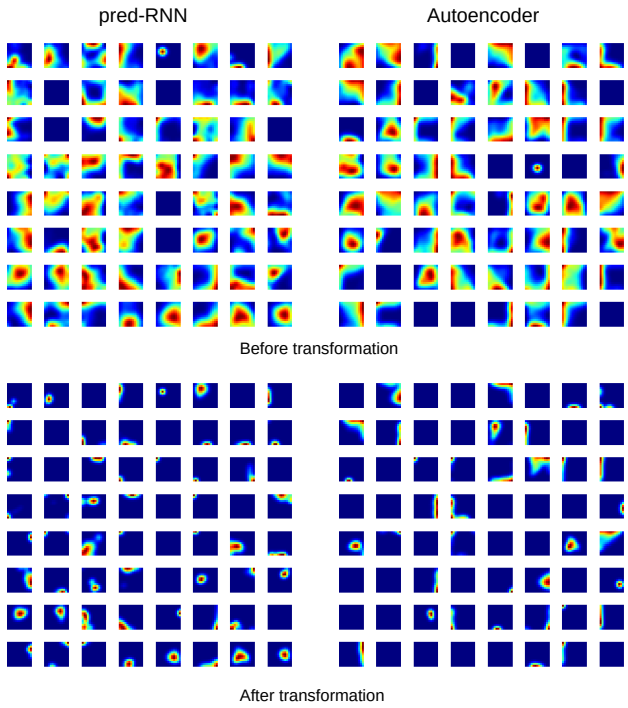

Figure 6: Firing rate maps of units in $\mathbf{h}^{pred}$ and $\mathbf{h}^{auto}$ before and after nonlinear transformation, in small square environment.

transformation to each unit in $\mathbf{h}^{\text{pred}}$ and $\mathbf{h}^{\text{auto}}$:

$$r_{\text{trans}} = \left(\frac{r}{r_{\text{max}}}\right)^{\alpha} - \beta \tag{2}$$

where α represents a free exponent, and β represents a free bias. We performed the analyses across all combinations of α and β, with α ranging from 2 to 5 and β ranging from 0 to 0.3 in increments of 0.05. We selected a pair of free parameters from the set where D_{KL} remained relatively constant across different combinations of α and β.

The autoencoder's rate maps (right) exhibit less distinct place fields, with several units demonstrating activity strictly along the walls of the environment. This suggests a more constrained and potentially less flexible encoding of space, which likely contributes to the disproportionately high place field orientation ratios near 90 degrees. Additionally, the autoencoder's rate maps reveal sparse activation patterns, with many units displaying little to no activity.

In contrast, the pred-RNN's rate maps (left) exhibit more distinct and localized place fields, with sharper and more uniformly distributed activity fields across the environment. These observations suggest that the pred-RNN is better suited for encoding spatial information.

To quantitatively compare the activity patterns of the models with those of place cells, we replicated the analysis procedure used in [17, 18]. Specifically, place cell-like units were identified as those with a region enclosed by a contour at half-maximum firing rate covering more than 1/30 of the bins in the state space. For the analysis of population statistics, we matched the size and composition of the animal data, where 28 units were analyzed—21 units pre-trained in a vertical rectangle and 7 in a horizontal rectangle. We matched them by randomly selecting 21 units from place cell-like units, and 7 additional units after a 90° rotation. (In our setup, all units were trained in a vertical rectangle, so rotating a subset of the data allowed us to match the composition of the units from the animal experiment.) We repeated the analysis with 20 populations selected this way using different random seeds. To obtain the distribution of place field orientation, the axis of greatest extent was measured to the nearest 7.5°, and plotted in 15° bins.