The Thesis Committee for Margaret C. von Ebers
certifies that this is the approved version of the following thesis:

From Sensory Input to Cognitive Maps: Exploring the Significance of Spatial Representations in Artificial Hippocampal Models

SUPERVISING COMMITTEE:

Xue-Xin Wei, Supervisor

Risto Miikkulainen, Co-supervisor

From Sensory Input to Cognitive Maps: Exploring the Significance of Spatial Representations in Artificial Hippocampal Models

by

Margaret C. von Ebers

Thesis

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science

The University of Texas at Austin

August 2024

Acknowledgments

To my thesis advisors, Dr. Xue-Xin Wei and Dr. Risto Miikkulainen, thank you so much for taking me on as a student in your labs, and for your patient guidance during my time here which has taught me so much.

To my parents Paul and Jill, my sisters Rachel and Tracy, my brothers David and Daniel, to my Babi, and to every other member of my family, thank you for being there for me during the last two years. You celebrated my acceptance to this program with me, and later listened empathetically as I threatened to run away to live on a farm.

Thank you so much to my lab mates Dylan, Eric, Yuezhang, Thomas, Marlan, and Zhongxuan. I did not expect to laugh so much at my desk, and it made the work so much easier. I will not miss that countdown.

And finally, thank you to my lovely and kind friends. I'd like to particularly mention David and Yara, but I am so thankful for my whole community either here in Austin or in my phone.

Abstract

From Sensory Input to Cognitive Maps: Exploring the Significance of Spatial Representations in Artificial Hippocampal Models

Margaret C. von Ebers, MS
The University of Texas at Austin, 2024

SUPERVISORS: Xue-Xin Wei, Risto Miikkulainen

A wide range of computational models have been proposed to explain how the hippocampus supports spatial and non-spatial reasoning. A recent model demonstrated that prediction of observations alone creates representations which contain spatial information and resemble the activity of hippocampal cells. This work explores three key questions. Does the prediction of visual elements in natural scenes induce a true, usable world model, challenging the notion that specialized neural architectures are necessary for spatial cognition? Do the "place cells" identified in this model function in accordance with our current understanding of their biological counterparts? And, are functional cell types such as place cells truly "functional", or are they heretofore mislabeled correlates of sensory information? This investigation reveals that prediction of visual elements in this scheme induces not a cognitive map, but instead local and non-functional features which are easily misidentified as true place cells. The study further proposes that genuine hippocampal features may serve more complex functions than the image-processing artifacts that superficially resemble them.

Table of Contents

List of Figures

Chapter 1: Introduction

Imagine driving a car through the downtown of a city you've never been to before, looking for parking. Their street signs are a different color than your town's, but you're still able to use them for navigation. You see a sports stadium, and understand that there's probably a parking garage nearby. Upon finding a garage, you realize that the front entrance is under construction, but you know to round the block and enter the back instead. You can do all of this because you naturally relate your specific sensory input to useful abstractions and then develop a flexible understanding of how they relate to each other. How exactly is your brain able to do this? If an artificial system existed whose "neurons" responded like our neurons during this task, would the system be able to navigate just the same? Scientists trying to understand and replicate these powerful cognitive capacities have two sources of information: the behavior of biological agents, and the neural response patterns that accompany it. Understanding how these neural patterns give rise to flexible behavior is an open question and one that is crucial for both neuroscience and artificial intelligence research. This thesis investigates the question of whether the appearance of hippocampal activity patterns in artificial models indicates the formation of a cognitive map of the kind that supports such impressive functionality in biological agents.

1.1 Motivation

Ever since O'Keefe and Dostrovsky (O'Keefe and Dostrovsky, 1971) discovered neurons in the rodent hippocampus which display striking preference for certain locations in the environment, neuroscientists have been fascinated with cataloguing the numerous hippocampal cell types and their purported functions. The "cognitive map" said to be generated by these cells initially appeared to support flexible behavior in space, such as planning routes or taking novel shortcuts. However, recent

research has indicated that this spatial map may be one instance of a general coding mechanism that synthesizes observations and gives them a relational structure so that they can be retrieved and utilized flexibly. Striking non-spatial examples include grid cells in the entorhinal cortex responding to stimuli which varies across two abstract dimensions (Constantinescu et al., 2016), and so-called place cells tracking "locations" in other sensory domains such as sound frequencies (Aronov et al., 2017). Given the wide variety of experimental results, and the fact that these cell type expressions are often less amenable to interpretation, understanding how these representations form and what role they take in supporting these functions is crucial for unraveling the brain's fundamental processing principles.

Computational neuroscientists have long been tackling this challenge by developing models that must generate representations resembling place and grid cells in the process of solving (generally) spatial tasks. Concurrently, researchers in reinforcement learning and large language models, despite achieving impressive and emergent advancements in model capabilities, are still striving to create systems that can learn continuously and reason flexibly in novel domains as humans can. Harnessing the power of this biological "GPS" could have far-reaching implications for any field modeling complex real-world interactions, particularly in self-driving cars and robotics. Although language and mathematical reasoning stem from different brain regions, decoding the hippocampus's ability to construct structure from sequences could enhance general-purpose reasoning in AI systems across various domains.

Before this research uncovers fundamental insights about generalization, it holds significant potential for validating experimental data and testing neuroscientific theories in novel contexts. Robust computational models of cognitive functions can be ablated, reconstructed, and applied to scenarios more difficult to study in vivo, serving as valuable tools for validating and extending our understanding of previously gathered experimental data.

1.2 Research Questions

Recent studies have used deep artificial neural networks to not only explain some functional cell types in the hippocampus but also to raise fundamental questions about the way these cells are traditionally classified in biological studies. Two currently prominent works challenge the long-held belief that specialized neural architectures are necessary for spatial cognition. The first, a work by Gornet and Thomson (Gornet and Thomson, 2023), proposes that the prediction of visual input in naturalistic environments induces a cognitive map which supports localization and gives rise to striking place cell-like representations. The second, by Luo et al. (Luo et al., 2024), makes the claim that not even prediction is necessary: spatial knowledge can arise out of systems which create sufficiently complex representations of sensory input, place cell-like representations can be observed in these systems, and these supposed functional cells do not appear to uniquely support this spatial knowledge. Both works raise important questions about the nature of the relationship between sensory inputs and their corresponding downstream representations in the hippocampus.

Gornet and Thomson assert that their predictive coding model forms a cognitive map because it replicates both the form and the function of place cells. In this instance, "form" refers to the model's ability to generate units that preferentially respond to different areas of the environment. "Function" describes how the model's proposed mechanism of action induces a cognitive map which is indicated by increased spatial information in the latent space. This thesis questions if such a simple architecture does induce a true cognitive map, and does so by questioning the form and function. Is the model's purported mechanism actually being carried out, or can its spatial information be attributed to simple local image feature prediction? If the model is not operating as previously hypothesized, is it possible that the novel place cell representations are instead supporting the cognitive map under the classical understanding of place cell function, or can they be attributed to the same effects as demonstrated in Luo et al.? Finally, do these findings suggest that the model's

"place cells" are superfluous, or that our traditional understanding of hippocampal place cells may need revision?

This thesis provides experimental evidence to support arguments that the predictive coding model does not and cannot support a flexible cognitive map of the environment which could be used to support all the neural function attributed to the hippocampal formation. The first two experimental findings support an alternative hypothesis that spatial information in the model is attributable to general function of attention blocks operating on image features, rather than encoded locations within a broader map. First, agents with different movement characteristics are developed to generate training observations for the model. It was found that spatial information encoded in the model is not significantly affected by these changes in the movement, which would not be the case for a model which relies on an encoded location being shifted forward in time accurately. It was also found that when an additional stream of information is integrated that directly tells the model about the movement which corresponds to its observations, the model gets better at predicting the next observation, and its representations contain more spatial information. However, varying the movement information did not result in expected changes in the model's predicted next observation, further supporting the theory that the model is not learning to represent the underlying movement. Next, it is shown that while the predictive coding model displays a wealth of functional cell types including place cells, the ablation of these does not cause performance deficits in localization in a way that would indicate that these features are supporting a cognitive map. Finally, it is shown that the functional cells shown in the image processing models from Luo et al. do not replicate to the larger environment, leading to discussions on the validity of their methods in supporting their theory that functional cell types in the hippocampus are truly superfluous.

1.3 Outline

This paper is divided into several sections. Chapter 2, "Background", gives an overview of relevant previous work regarding cognitive maps, starting with the experimental data and theories supporting different hippocampal cell types in mammals. It also details relevant computational models, spanning efforts in reinforcement and representation learning. This chapter ends with introductions to the works from Gornet and Thomson and Luo et al. which are the focus of this paper. Chapter 3, "The Effects of Transition Statistics on Localization", introduces variations to the way that data is collected for predictive coding models and discusses their effects on performance across different tasks. Chapter 4, "Integration of Vestibular Input", shows how the model makes use of a complementary stream of information during training, and explains why action signals may be critical to cognitive map formation. Chapter 5, "Content of Image Latents", applies methods from Luo et al. to draw comparisons between the activations of image-processing networks and the predictive coder in localization tasks. Chapter 6, "Discussion and Future Work", summarizes the experimental findings found in this thesis, provides additional support from the literature for the claims, and discuss future avenues towards complete cognitive mapping models. Chapter 5, "Conclusion", offers final reflections on the implications of this work.

Chapter 2: Background

This chapter presents a comprehensive overview of the biological and computational foundations underlying this research. It begins by exploring the historical discoveries and recent assessments of hippocampal function in animals. The discussion then examines relevant computational models of cognitive maps, highlighting the distinctions between various approaches and their current limitations. Finally, the chapter introduces the two key studies central to this research, providing a detailed overview of their methodologies that will be utilized in subsequent analyses.

2.1 Neural Substrates of Cognitive Maps

Decades of research in neuroscience have indicated that the hippocampus is a critical region for functions such as spatial reasoning and episodic memory, as well as more abstract social or logical cognition. (Whittington et al., 2022a) Most experimental data in this area focuses specifically on space. The first and most striking results came from rodents navigating mazes and small laboratory environments, where a variety of functional cell type classifications were determined that might support the cognitive map of space. These include place cells (O'Keefe and Dostrovsky, 1971), which respond strongly to one area of the environment and do not generalize to new locations; grid cells (Hafting et al., 2005), which smoothly tile the environment; head-direction cells (Taube et al., 1990), and more.

How these internal representations are formed remains heavily debated. The hippocampus receives highly processed inputs from multiple sensory modalities (e.g. visual, olfactory, or vestibular systems) (Hitier et al., 2014) as well as information about goals from the prefrontal cortex (Eichenbaum, 2017). Striking differences in the experimental realizations of these cells have been observed across species, likely due to differences in the quality, content, and sampling methods of these inputs. In

particular, primates appear to exhibit much more mixed selectivity, with hippocampal rhythmic theta oscillations occurring with eye saccades instead of during locomotion as is the case with rodents. (Piza et al., 2024)

Further complicating the efforts to catalog cell types and understand the mechanism that they support is the fact that similar representations exist in varying degrees across the cortex. The hippocampus takes a role in shaping the representations found in earlier sensory areas (Saleem et al., 2018) and grid-cell-like codes have been found to organize conceptual knowledge in the entorhinal cortex and ventromedial prefrontal cortex (Constantinescu et al., 2016).

2.2 Computational Cognitive Maps

Many computational models have been proposed to explain how hippocampal representations might be formed. Spatial maps in the brain are typically considered to be the result of networks that predict location by integrating the agent's movements from vestibular inputs, a process referred to as "path integration". (McNaughton et al., 2006) This computation can be implemented in a class of models called continuous attractor networks (Zhang, 1996; Samsonovich and McNaughton, 1997; Burak and Fiete, 2009), and recurrent neural networks have been shown to form response patterns resembling grid cells while solving the path integration task (Cueva and Wei, 2018). Similar bodies of work show how models can take advantage of an existing map of an environment, literal or represented as place cells, in order to navigate (Banino et al., 2018).

Another class of cognitive mapping models attempts to remove the focus from mapping Euclidean space, instead focusing on how structure can be built out of more general sequences. While these models are typically explained from a reinforcement learning framework for convenience, they are generally trained without reinforcement and show how prediction can allow for sequences of observations without rewards to become a useful representation. Models such as the Tolman-Eichenbaum Machine

(TEM; Whittington et al., 2020) or Spatial Memory Pipeline (SMP; Uria et al., 2022) use an abstract path integration module which uses the content of sensory observations to create a representation which is made up of both the nonspecific and the specific so that these representations might quickly generalize to new environments. Doing away with path integration entirely, Clone Structured Causal Graphs (CSCG; George et al., 2021; Raju et al., 2024) use a hidden Markov model with cloned latent states to show how arbitrary representations can be used to detangle aliased observations.

The throughline that can be seen across all of these implementations is that they focus primarily on how an explicit history of actions might allow for the construction of a map of space. In this realm, while sensory inputs have been recognized to play an important role in anchoring and re-calibration of the map, action signaling in the way of path integration or grid construction is considered to be the ultimate backbone. However, all of the models mentioned above rely on some simplified input representation which is not accurate to the actual biological system: the path integration models receive direct training signals or instantiations regarding the spatial map, and TEM's abstract path integrator operates on allocentric actions and observations. CSCG makes use of egocentric actions and receives no additional training signal than the history of observations it received, but it abstracts its (discrete) observations away so much that they do not support generalization to new environments.

2.3 A Predictive Coding-Based Model

It is possible that one could learn all that is necessary to know about how space works just from building a representation of sensory information in the right way? Does the prediction task induce neural populations to learn both a flexible and implicit path integration system, as well as a way to generalize observations so that they can be used in the future? The model here under examination by Gornet and Thomson shows how a model might encode information about an agent's

location in an environment simply through the prediction of the next observation. This is informed by the theory of predictive coding, which proposes that the general function of neural processes across the brain is to efficiently encode an expected representation of what sensory observations will appear in the next moment in time, making use of spatial-temporal regularities in the world. Originally introduced to explain the inhibitory response in the retina, Rao and Ballard (1999) further developed this framework for the primary visual cortex. Outside of the realm of neuroscience, Poincaré (2015) discussed how movement across space could generate regular and predictable transformations of observations which could be stitched together to form an understanding of the structure of the environment.

2.3.1 Methods and Results

Gornet and Thomson first motivated the predictive coding model with a mathematical theorem. First-person images $I_0, I_1, \ldots, I_n$ were sampled from positions and associated orientations $(x_0, \theta_0), (x_1, \theta_1), \ldots, (x_n, \theta_n)$. Statistical inference of predicting the next image I_{n+1} initially takes the form:

$$P(I_{k+1} \mid I_0, I_1, \ldots, I_k) = \frac{P(I_0, I_1, \ldots, I_k, I_{k+1})}{P(I_0, I_1, \ldots, I_k)} \tag{2.1}$$

Because the movement of the agent is determined by a variable velocity applied to the position with a fixed time step, the motion of the agent is generated by a Markov process with transition probabilities $P(x_{i+1}|x_i)$. Thus, since $P(I_0, I_1, \ldots, I_k, I_{k+1})$ can be considered to be a function on an implicit set of spatial coordinates, the equation can be rewritten to reflect this as an integral over all possible paths in the environment Ω:

$$\begin{aligned} P(I_{k+1} \mid I_0, I_1, \ldots, I_k) &= \int_\Omega \mathrm{d}x \, P(x_0 \ldots x_k) \frac{P(I_0, I_1, \ldots, I_k \mid x_0 \ldots x_k)}{P(I_0, I_1, \ldots, I_k)} P(x_{k+1} \mid x_k) P(I_{k+1} \mid x_{k+1}) \\ &= \int_\Omega \mathrm{d}x \, \underbrace{P(x_0 \ldots x_k \mid I_0, I_1, \ldots, I_k)}_{\text{encoding (1)}} \quad \underbrace{P(x_{k+1} \mid x_k)}_{\text{spatial transition probability (2)}} \quad \underbrace{P(I_{k+1} \mid x_{k+1})}_{\text{decoding (3)}} \end{aligned} \tag{2.2}$$

Fixed-length sequences of these observations are individually encoded with a ResNet-18 encoder (He et al., 2016) to produce a sequence of latents. This encoder uses a U-Net (Ronneberger et al., 2015) architecture to pass latents into a residual stream for the subsequent three blocks of alternating multi-headed attention (Vaswani, 2017) (with eight heads and a causal mask to enforce temporal structure) and feed-forward layers. The output of these attention blocks is hypothesized to be a set of latent predictions that represent the model's expectation of the agent's next position and thus what the next observation will look like. These latents are then passed through a matching ResNet-18 decoder with transposed convolutions to make the predicted next image. The model is trained with a mean-squared error loss in pixel space. It is hypothesized that the encoder, by producing a set of processed visual feature latents, is encoding an estimate of the probability that an agent has traveled along a path of positions associated with the provided image sequence. The self-attention blocks should learn the transition probabilities between the inferred current position (x_k, θ_k) and the next estimated position (x_{k+1}, θ_{k+1}), and finally the model's decoder computes $P(I_{k+1}|x_{k+1}, \theta_{k+1})$. Model architecture and an example of how agent movement generates training data is illustrated in Figure 2.1.

Gornet and Thomson use the Malmo package to build a test environment within Minecraft (Johnson et al., 2016). The environment is a complex scene measuring 40 x 65 blocks, displayed in Figure 2.2. The original model presented is trained for 200 epochs with gradient descent optimization with Nesterov momentum (Sutskever et al., 2013), a weight decay of 5×10^{-6}, and a learning rate of 10^1 adjusted by OneCycle learning rate scheduling (Smith and Topin, 2019), with which it achieves an MSE of 0.094.

Gornet and Thomson first compared the predictive coding model with an autoencoder which has the same ResNet-18 encoder and decoder architecture. This autoencoder reconstructs individual images without prediction and serves as a baseline indicator for how much spatial content can be gleaned from the similarity between

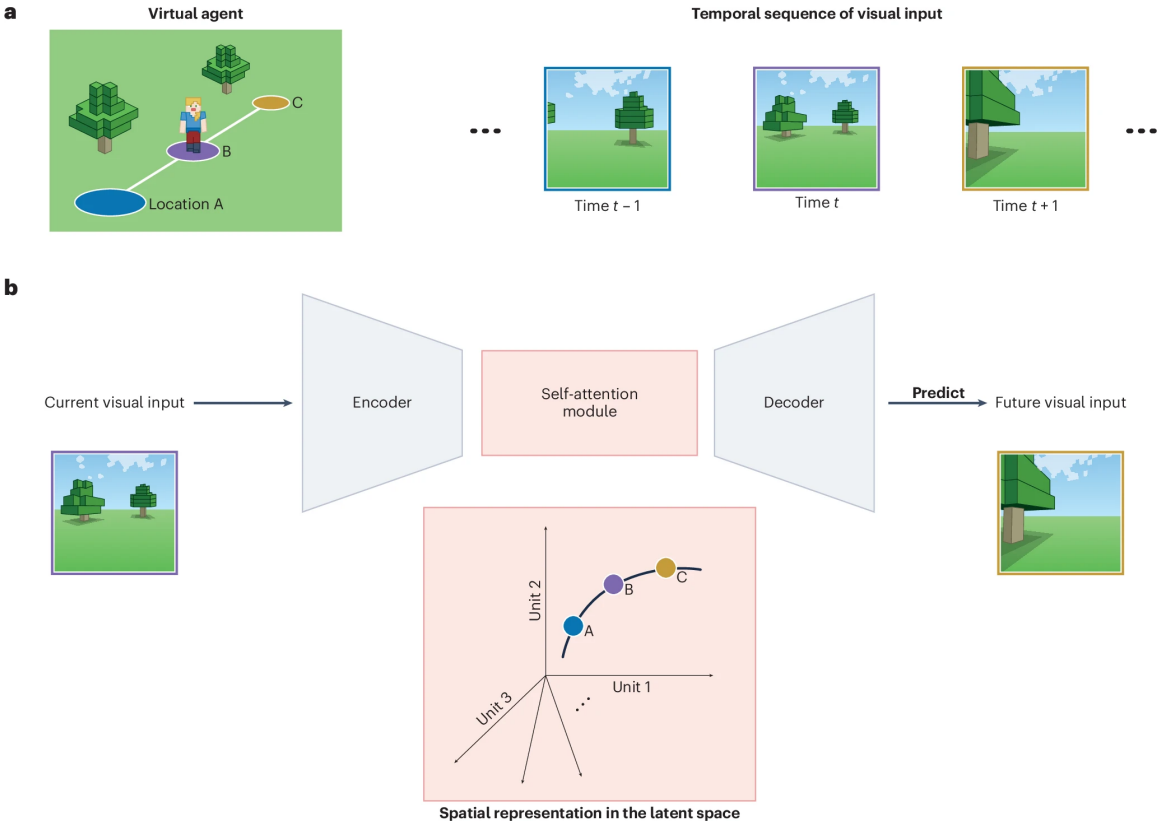

Figure 2.1: **Depiction of the data collection process and model architecture. a,** An agent navigates in a virtual environment and collects visual observations at every step. These image sequences are used to train the model. **b,** The neural network is trained to predict future visual input on the basis of current visual input. A representation of space should emerge from the self-attention module after training the network to solve the visual prediction task. Figure from (von Ebers and Wei, 2024).

Figure 2.2: **A bird's eye view of the test environment.** A bridge restricts the agent's motion, trees provide visual occlusion, and a large cave serves as a global landmark.

images instead of the underlying geometry of the environment. The latents recovered from the predictive coding model outperform those from the autoencoder in two downstream tasks: predicting location and recovering accurate distances between locations. Importantly, the latent population in the predictive coding model before the predictive attention blocks is demonstrated in supplementary information from Gornet and Thomson to not have this additional spatial content, and performs similarly to the autoencoder latents in the localization task.

Additionally, individual activations from just before the decoder in the predictive coding model were treated as potential cells in the hippocampus. Each activation was thresholded at its 90th percentile value, and for each location, the cell was considered to be "activated" if it was above this threshold. It was found that these cells tend to activate at localized regions across the environment. Finally, they show that every location in the environment can be represented by unique combinations of overlapping "place field" regions and that the Hamming distance given by counting the number of overlapping regions can recover the physical distance between locations.

2.3.2 What is Missing?

While Gornet and Thomson's work presents impressive initial results supporting a novel theory of hippocampal function, several critical questions remain. Their model, if accurately described by the proposed mathematical formulation, could induce a cognitive map capable of supporting various navigational behaviors. This formulation, which maps sensory observations to latent states and describes transitions between these states, does indeed align with other cognitive mapping models discussed earlier. However, the presence of spatial content in the model's latents and place cell-like activations, while noteworthy, does not conclusively prove that the model functions as hypothesized. Alternative explanations that do not necessarily result in a flexible map warrant consideration.

Two main factors challenge the notion that the model is encoding images into

estimated locations:

1. Gornet and Thomson demonstrate that latents before the attention heads contain significantly less spatial content than those after. While they interpret this as evidence that prediction is necessary to induce a spatial map, it contradicts their hypothesis that the encoder learns a direct mapping between images and their associated locations and poses.

2. The authors present their model as probabilistic, implying that the encoder and decoder estimate distributions. However, it's important to note that these encoder-decoder architectures are not variational autoencoders, which would directly estimate distributions. Instead, they primarily deal with sets of processed image features. They provide further support for this setup in the supplemental information, but it opens the door for alternative hypotheses.

This thesis proposes that, instead, the attention heads may be learning transformations between processed image features rather than transformations between encoded locations. Experiments in Chapters 3 and 4 aim to demonstrate model behavior more consistent with this alternate formulation.

Under this hypothesis, the increased success in location decoding tasks could be attributed to the fact that image features processed by attention blocks will contain information about previous images in the sequence. This effect might be comparable to providing the downstream model with multiple images for location prediction instead of just one. However, the question remains: do these post-attention image features still represent an implicit code of location within a connected map? Existing research on vanilla transformers suggests they can achieve impressive next-step prediction without a coherent world model (discussed further in Chapter 6).

Gornet and Thomson take care to not attribute too much biological meaning to their place cells, and do not claim that their model's conception of space is functionally supported by these cells, instead simply noting that it is encouraging that their model

of hippocampal function appears to display hippocampal response patterns. However, in the absence of evidence supporting their proposed mechanism of action, it would be useful to their argument if these functional cells were uniquely critical for performance. This would support the model's function under the classical view of place cells, which is that they are meaningful and necessary for localization within space. The next section introduces a work that directly contradicts the idea that these cells are unique or functional in artificial systems, and applies its methods to the predictive coding model in Chapter 5 to exhaust this last potential avenue of support for the cognitive map hypothesis.

2.4 Critically Examining Functional Cell Types

The predictive coding model previously introduced asserts that prediction is an essential component of the function of the hippocampus, and demonstrates that prediction induces place cell-like representations. (Luo et al., 2024) make the further claim that instead of a cognitive map-supporting system being necessary, essentially any sensory processing system with enough complexity contains significant information about space to solve simple localization tasks, and that place cell representations arise from these systems but do not contribute to their function in ways previously ascribed to functional cell types.

This could be interpreted on its face as a caution against confusing the invariant representations formed in artificial image processing with those that support spatial reasoning in the hippocampus. Instead, Luo et al. propose that this is all that real place cells are, or at least that real place cells are similarly inevitable and superfluous. This new theory of hippocampal representations is a salacious one, but not unfounded by the literature. Several ablation studies in the hippocampus have reported inconclusive results, and mixed coding responses have been widely observed in the place of the immediately intuitive place cells, particularly in species like primates that have advanced visual systems. Additionally, there are contemporary movements

in both neuroscience and computer science to move away from examining single neurons as interpretable units of computation.

2.4.1 Methods and Results

Luo et al. propose an experimental setup to demonstrate their claims, utilizing a small Unity environment that resembles a typical laboratory setting used for studying free-moving rodents. They expose several deep image-processing networks to first-person images from this environment. These networks vary in architecture, including convolutional networks like VGG-16 (Simonyan, 2014) as well as the non-convolutional Vision Transformers (ViT; Dosovitskiy et al., 2021), with weights either pretrained on ImageNet-1k (Deng et al., 2009) or randomly initialized. Notably, these models are not further trained on the laboratory setting. The researchers use the activations from these models, when shown the observations, to train linear regression models. These decoders predict three key aspects of spatial cognition: location, head direction, and distance to the nearest wall – all functions hypothesized to be performed by various hippocampal cell types. Importantly, all models achieved high accuracy across all tasks.

Furthermore, Luo et al. classify the model activations according to typical guidelines applied to functional cell types in neuroscience. They discover that all models exhibit large, diverse populations of functional cell types when exposed to this laboratory setting, including significant mixed encoding of head direction and place, which closely resembles primate data. To test the importance of these functional cells, they then sort the model activations by their rankings according to these metrics and progressively remove increasing amounts of the top-ranked units. Surprisingly, they find that removing these functional cells does not result in a significant decrease in the performance indicated by their purported function. For instance, lesioning place cells does not lead to a substantial decline in localization performance, and lesioning border cells do not affect the "distance to nearest wall" task.

2.4.2 What is Missing?

The emergence of place cell-like representations across diverse architectures, without uniquely contributing to localization tasks, offers a potential explanation for the presence of place fields in the predictive coding model. However, to establish a useful baseline for subsequent analyses, it is crucial to consider the limitations of this study. A primary concern lies in the experimental environment's potential biases. The small size of the environment, absence of aliasing or visual occlusions, and the visually-grounding presence of a low wall may contribute to the observed diversity of functional cell types and potentially diminish the significance of the lesioning study results. In this setting, ablating place cells may not significantly impair localization, as the agent's position remains easily determinable from other functional cells. It remains to be seen whether the predictive coding model induces place cell-like activations in larger environments and whether baseline, untrained models exhibit similar behavior. Additionally, is it possible that the larger environment might lead to different outcomes in the lesioning study?

This thesis primarily examines these limitations in the context of the predictive coding model and does not adopt the further conclusion of Luo et al. that real hippocampal cell types are superfluous. Operating under the assumption that hippocampal functional cells really are functional could provide a stronger point against the predictive coding model. However, the analysis is still strong without this conclusion. A full rebuttal of these claims would be out of the scope of this work as this is an open question within neuroscience, though since this thesis does try to use artificial systems to develop an understanding of how sensory information is used to create hippocampal representations, some initial evidence and argumentation casting doubt on the broad claims of Luo et al. is offered in Chapters 5 and 6.

This thesis primarily examines these limitations within the context of the predictive coding model, without adopting Luo et al.'s broader conclusion regarding the superfluity of real hippocampal cell types. Operating under the assumption that hip-

pocampal functional cells serve genuine functions could provide a stronger argument against the predictive coding model. However, the analysis remains robust without this conclusion. While a comprehensive rebuttal of the claims of Luo et al. falls outside the scope of this work, given that it remains an open question in neuroscience, Chapters 5 and 6 present initial evidence and argumentation casting some doubt. This discussion aligns with the thesis's aim to leverage artificial systems in understanding the connection between sensory information and hippocampal representations.

Chapter 3: The Effects of Transition Statistics on Localization

If the contents of the encoded representation are in doubt, one first place to look is at how the attention blocks learn transitions between what's being encoded, whether that is image features or a more explicit code of location and pose. These transitions are determined by the underlying movement of an agent which generates image observations. Supplementary information from Gornet and Thomson indicates that an agent randomly sampling angular velocity at each time step results in significantly reduced spatial content in post-prediction latents. This finding supports their proposed mechanism: the model learns how locations and poses arise from observations, and accurate prediction of location transitions may enable the construction of a comprehensive environmental understanding, leading to high spatial content in the post-prediction latent space.

This section presents a more robust version of this experiment by varying randomness qualities in the trajectories in a more systematic manner by introducing an additional class of agents with more biologically accurate movement. The performance of models trained on this data instead provides evidence for the alternative hypothesis and establishes a foundation for examining these model variants under different testing conditions in subsequent chapters.

3.1 Methods

The first major change made to the agent's movement was to add the option to modulate randomness in the trajectory with a coefficient. Previously, data for the model was generated by an agent selecting two random points in the Minecraft environment and following a path between them determined by the A* algorithm, recording observations at every time step. In the new setup at a low randomness

coefficient, the agent starts at a random location in the environment and samples speed and angular velocity values such that it generally travels in a straight line, with some small addition for both variables from normal distributions. As the randomness coefficient increases, the variance of the two distributions is increased, such that the trajectory eventually becomes overpowered by random movements. The agent will also make some necessary moves to avoid obstacles. 32510 images from each of these coefficients make up datasets for training.

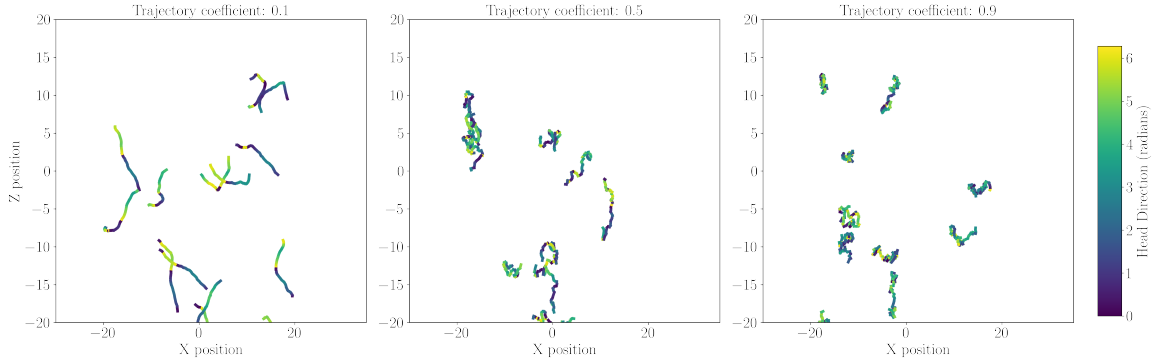

Figure 3.1: **Examples of the effects of increasingly random movements in head-fixed models.** Trajectories resembling these examples will be used to train three separate instances of the predictive coding model.

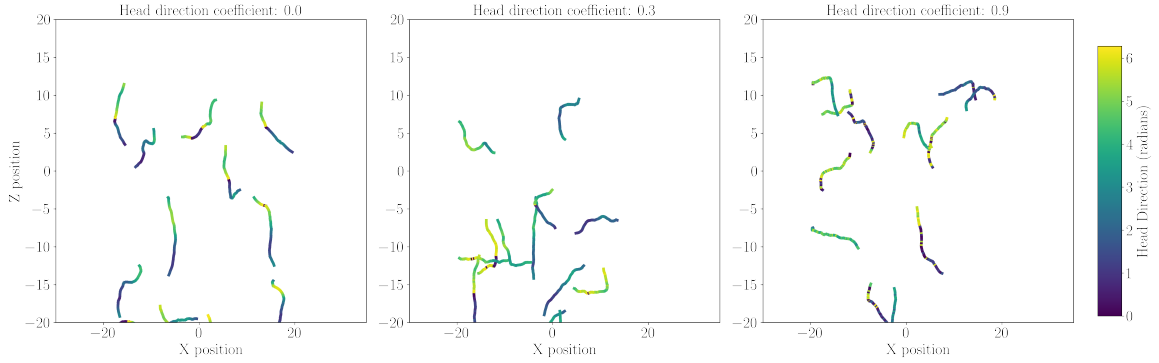

Figure 3.2: **Examples of the effects of increasingly random movements in non-head-fixed models.** Trajectories resembling these examples will be used to train three separate instances of the predictive coding model.

These datasets represent a trade-off: at lower coefficients, the transition statistics are much more regular and presumably easier for the self-attention blocks to

predict. However, a straighter trajectory also implies a final position that is much further away from the initial, which would generally induce more visual changes and less ability to build up evidence. Examples of these trajectories are shown in Figure 3.1.

A second class of agent motion models detangles the head direction and the movement direction. Reflective of sensory input, the exploration strategies of rodents and non-human primates are the primary driver between the differences in hippocampal representations between the two species. Risking oversimplification, Piza et al. (2024) found that rats primarily sample their environment by locomoting to different locations with very small head turns, while primates tend to stop and scan the environment with large, sweeping head movements. Adding a moving head direction tests whether the predictive coding model can accommodate an additional degree of complexity, and adds more biological plausibility to a model that relies on visual input which is closer to the signal strength of a primate instead of a rat. Examples of these trajectories can be seen in Figure 3.2.

The performance impacts of these model data adjustments were measured in two primary ways. First, the model's performance on the validation set for the next-step prediction task is examined. Additionally, the effects of these changes on the spatial content in the post-prediction model activations were measured using both nonlinear and linear decoders. Modified methods from Gornet and Thomson were employed to train a nonlinear decoder in the form of a small convolutional neural network. This decoder receives normalized post-prediction activations as input and is trained to minimize the mean squared error for the predicted 2D location. The decoder distinguishes between multiple views for every location and measures performance on a held-out set of random locations and views from across the environment, representing a more stringent test than that used in the original work. Details of the original architecture and the modifications used in this study can be found in Appendix A. The localization plots include an upper bound in the form of a noise model, which returns the actual position with some additive Gaussian noise, as well as

a lower bound model, where predicted positions were randomly sampled from the occupancy distribution of the training data. As a more rigorous test of spatial content, a simple linear decoder is also introduced, trained on the same data as the nonlinear decoder. This support vector machine reports five-fold cross-validated results.

3.2 Results

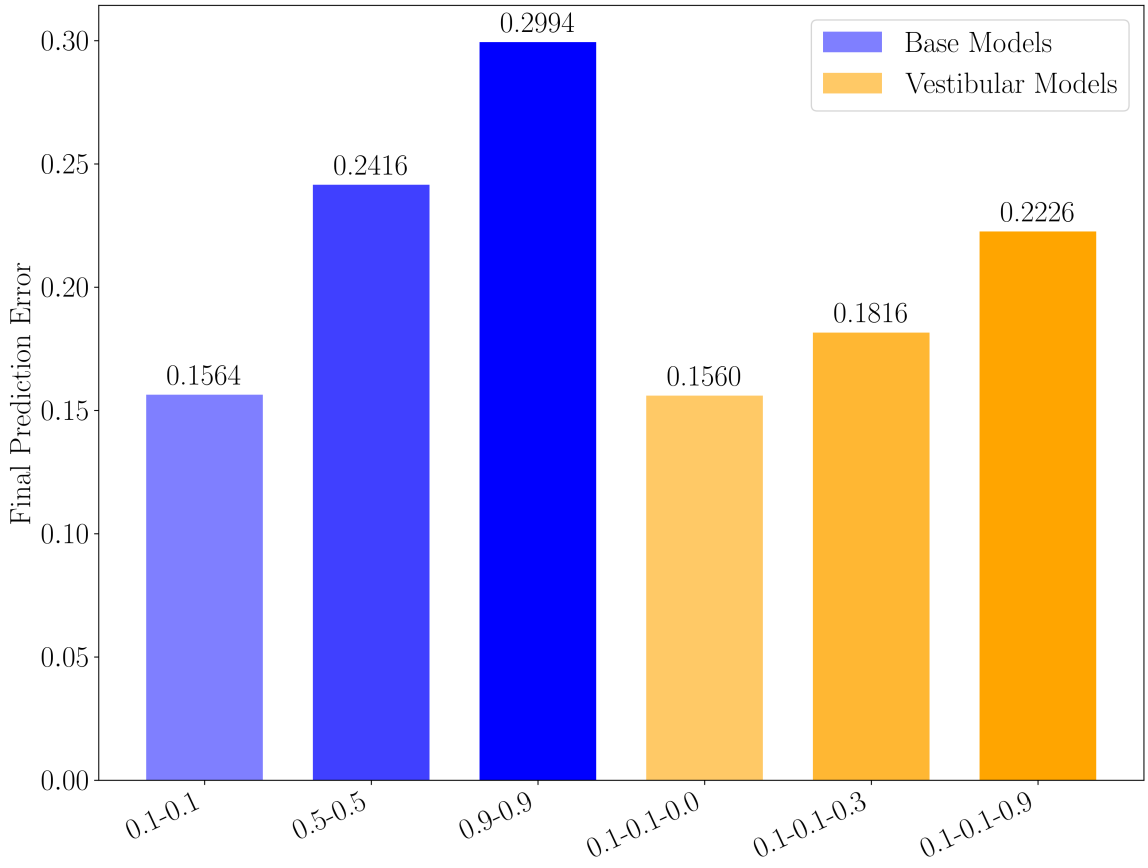

Figure 3.3: **Final mean squared error of predicted images on the validation set.** Head-fixed models were shown in blue and non-head-fixed models were shown in orange. Both classes of models exhibit decreasing performance with additional randomness.

Results for the training objective of mean squared error between images in pixel space were shown in Figure 3.3 for all variations of the predictive coding model.

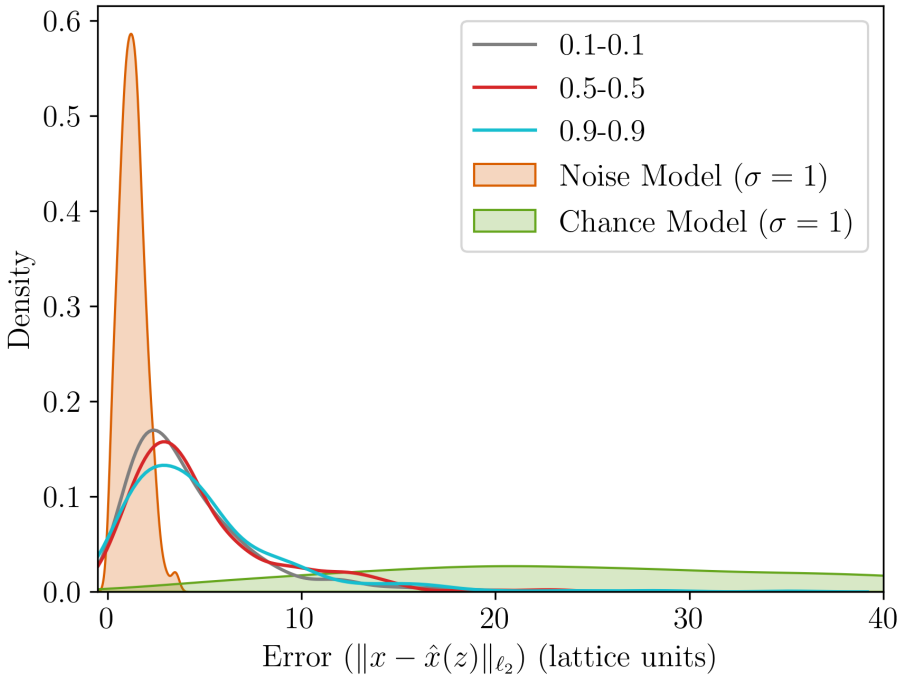

Figure 3.4: **Prediction errors of positions from the predictive coder's latent space, across different head-fixed model trajectory randomness settings.** All models report similar localization performance with a trained nonlinear decoder, indicating that the development of spatial content is not closely tied to transition statistics.

Predictably, increasing randomness in both the head-fixed and non-head-fixed classes of models universally results in decreasing performance for this task. Notably, the least random models of the two classes have almost exact prediction performance, despite the additional complexity of movement, and the models with increased head direction randomness outperform the models with increased movement randomness. This could be explained by the fact that the non-head-fixed models only really induce randomness in one degree of freedom (angular velocity), while the other models have two (angular velocity and speed).

Results for the location decoding task with the nonlinear decoders are presented in Figure 3.6. The performance indicated by the predictive coding model trained on the original movement scheme was successfully replicated. All models,

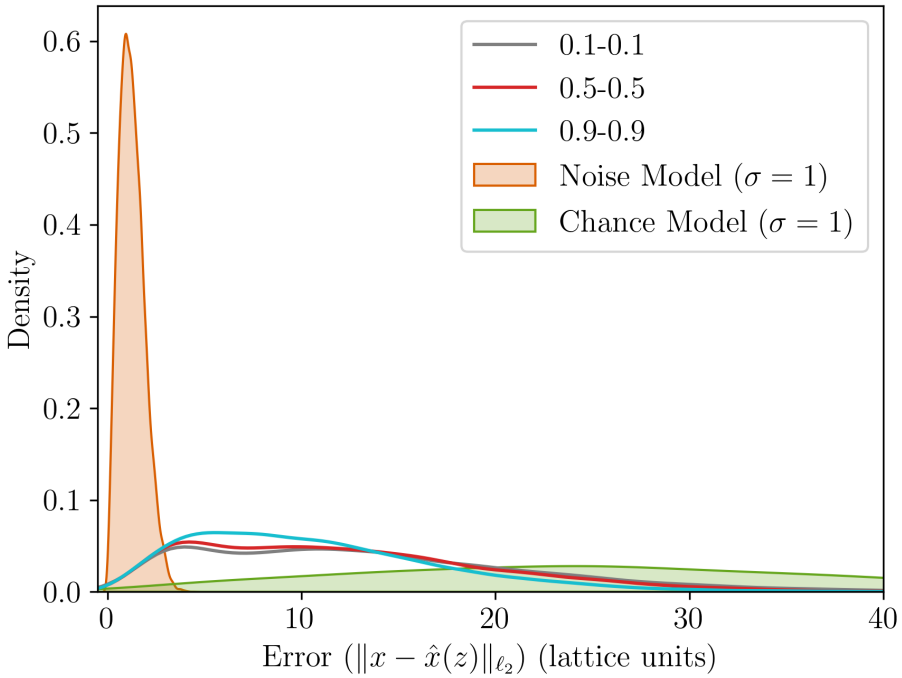

Figure 3.5: **Prediction errors of positions from the predictive coder's latent space, across different model trajectory randomness settings, in head-fixed models.** All models report similar localization performance with linear decoders, indicating that the development of spatial content is not closely tied to transition statistics.

across varying levels of randomness and in both head-fixed and non-head-fixed classes, demonstrate similar and impressive localization performance. This effect is further corroborated by trained linear decoders, as shown in Figures 3.5 and 3.7. Notably, these findings do not align with the results reported by Gornet and Thomson, which suggest that increased randomness in model trajectory should increase the difficulty in constructing a cognitive map from individual locations, leading to decreased spatial content. The most random models in this study exhibit movements that are considerably more extreme and unpredictable than those reported by Gornet and Thomson. Possible explanations for this discrepancy include a more even distribution of visits across the environment in the current study, the provision of more training data, and the implementation of a more robust spatial content test.

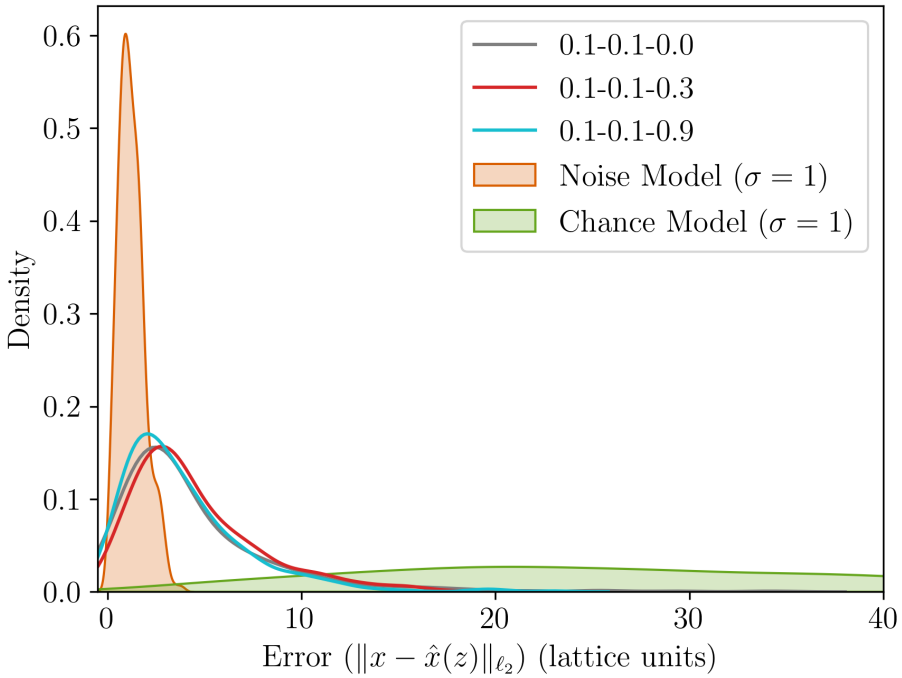

Figure 3.6: **Prediction errors of positions from the predictive coder's latent space, across different non-head-fixed model trajectory randomness settings.** All models report similar localization performance with a trained nonlinear decoder, indicating that the development of spatial content is not closely tied to transition statistics.

Instead, these results support the hypothesis that the model is not constructing representations of location, but rather learning simple local transitions between image features. Under this interpretation, performance in the next-step prediction task would indeed decrease with more random transitions. However, the spatial content contained in the post-prediction representations may be explained as follows: attention performed on image features will produce an output set of image features containing information about previous images in the sequence. While Gornet and Thomson present more striking results in a different environment with severely aliased observations, it remains that under the alternative hypothesis, if the sequence includes information from before the aliased observation, a nonlinear decoder would be able to use that to distinguish locations.

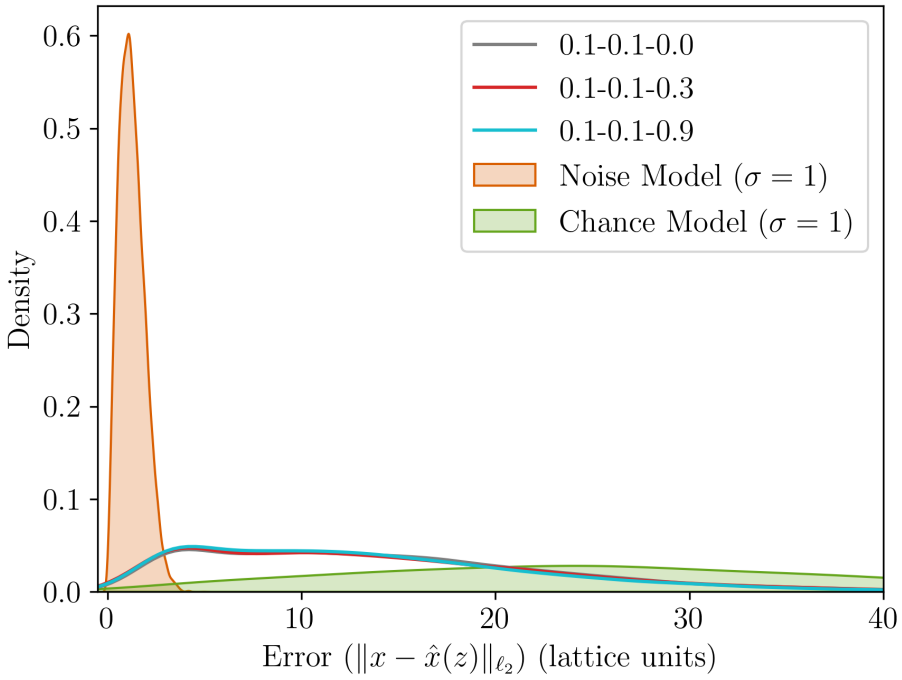

Figure 3.7: **Prediction errors of positions from the predictive coder's latent space, across different model trajectory randomness settings, in non-head-fixed models.** All models report similar localization performance with linear decoders, indicating that the development of spatial content is not closely tied to transition statistics.

It remains possible that the predictive coding model could be encoding space in a manner both explicit, as hypothesized by Gornet and Thomson, and robust to the findings presented in this section. "Localization" alone may not be a sufficiently stringent test of representations in a system that should be capable of using these representations for flexible navigation. The next chapter continues to add supporting information to the conclusions of this thesis by further examining the content of the transitions learned by the predictive portion of the network.

Chapter 4: Integration of Vestibular Input

One of the original strengths of the predictive coder model is its reliance solely on sensory data, without direct training signals regarding its location within the environment. If the model functions as originally intended, additional information appears to be unnecessary. However, integrating a stream of information that describes observational changes in a manner reflecting the underlying environment and maintaining an egocentric perspective offers several advantages to this investigation without compromising the biological plausibility of the predictive coding model. This chapter details the effects of including vestibular signaling in the form of information about the agent's speed and angular velocity (and, when applicable, the separate head direction angular velocity).

The hippocampal formation is known to receive and be affected by vestibular signaling (Hitier et al., 2014; Stackman et al., 2002), which carries information about how the agent's body position changes in response to actions taken. These signals are hypothesized to stabilize representations of space when sensory information is insufficient or confusing. For example, in primates, representations of angular head velocity serve to modulate principal hippocampal cells during rapid head-gaze movements. (Piza et al., 2024) This also can be illustrated by a common-sense example: imagine closing one's eyes in an unfamiliar environment and taking a few steps. The vestibular signal is noisy, but it allows for the maintenance of a mental map of the local environment and an understanding of how short-term future actions will alter it. However, the exact mechanisms by which sensory and vestibular signals collaborate to perform these functions remain an open problem in the field.

For the purposes of this study, adding action signals represents another test of what the predictive coding model is actually learning transitions between. While transformations between image features, as indicated in the alternate hypothesis, will to some degree reflect metrics such as speed and angular velocity that shift the

egocentric frame of the environment, it is expected that this relationship will be less direct than one in which locations are directly encoded and the attention blocks must learn to represent speed and angular velocity in order to model the transitions between them.

4.1 Method

The success of this integration is examined in three ways: performance on the next-image-prediction training objective, latent localization performance, and steering ability. In the steering test, images were taken from the localization dataset such that every sequence comprised observations from the same position with a small amount of noise in the position. These were fed to a trained model, and a new action signal was provided with an angular velocity set to 0 for each time step and a constant speed, which was varied. This speed variation should produce predictable changes in the model's image outputs as well as the estimated position outputted by trained nonlinear and linear location decoders from the previous chapter.

Several methods of integrating vestibular inputs were attempted. Initially, code from Gornet and Thomson, not included in their final work, was utilized. Here, the model was given not only a sequence of images but also a sequence of corresponding speed and angular velocity pairs. These action signals were scaled up with a feedforward network to match the size and shape of the image latents. They were then concatenated onto the encoded image data and convolved together, resulting in the same shape as the original image latents. These latents were then passed through the rest of the network as normal. Similarly, image latents were scaled down with a convolutional layer to 10 dimensions, added to scaled-up action latents, and scaled up to the original size of the pre-attention latents before being passed through the rest of the network. Under the assumption that the encoder is transforming input images into representations of location and view, these methods can be understood as adding "intentionality" before prediction.

Ultimately, neither of these two methods was sufficiently performant in any metric to display results. In both cases, performance actually declined slightly in the next-step prediction and localization tasks, and when examining steering, an applied speed at every step produced no significant change in the decoded images or in the predicted location. Inspired by relative positional encodings in large language models, an attempt was made to remove any convolution that might destroy information in the two streams, instead either simply adding the transformed action latents before prediction or concatenating them. The last method doubles the size of the latents which pass through the self-attention blocks and into the decoder from 128x8x8 to 256x8x8, and the network architecture was scaled to match. The results of this last method will be discussed in the next section.

4.2 Results

As shown in Figures 4.1 and 4.2, the vestibular signal as a relative positional encoding which is concatenated to the image latents created models with significantly improved performance on the next-image prediction task in both head-fixed and non-head-fixed models, as well as across all types of trajectory. However, it was not the case that this entirely stabilized the performance of the more random trajectory models, even though the model is now receiving and evidently making use of noiseless information that should almost entirely determine the content of the next image. Adding vestibular information also improves the quality of the localization in post-prediction latents, as seen in Figures 4.3 and 4.4.

The steering capability of the vestibular predictive coding model was tested using two datasets: one with the heading direction consistently set to 0 degrees, and another at 90 degrees. Given the environment's layout, it was anticipated that a speed input greater than one and an angular velocity of zero applied to image inputs facing 0 degrees should result in an increase in the z coordinate, while for images facing 90 degrees, a decrease in the x coordinate was expected. Both directions were

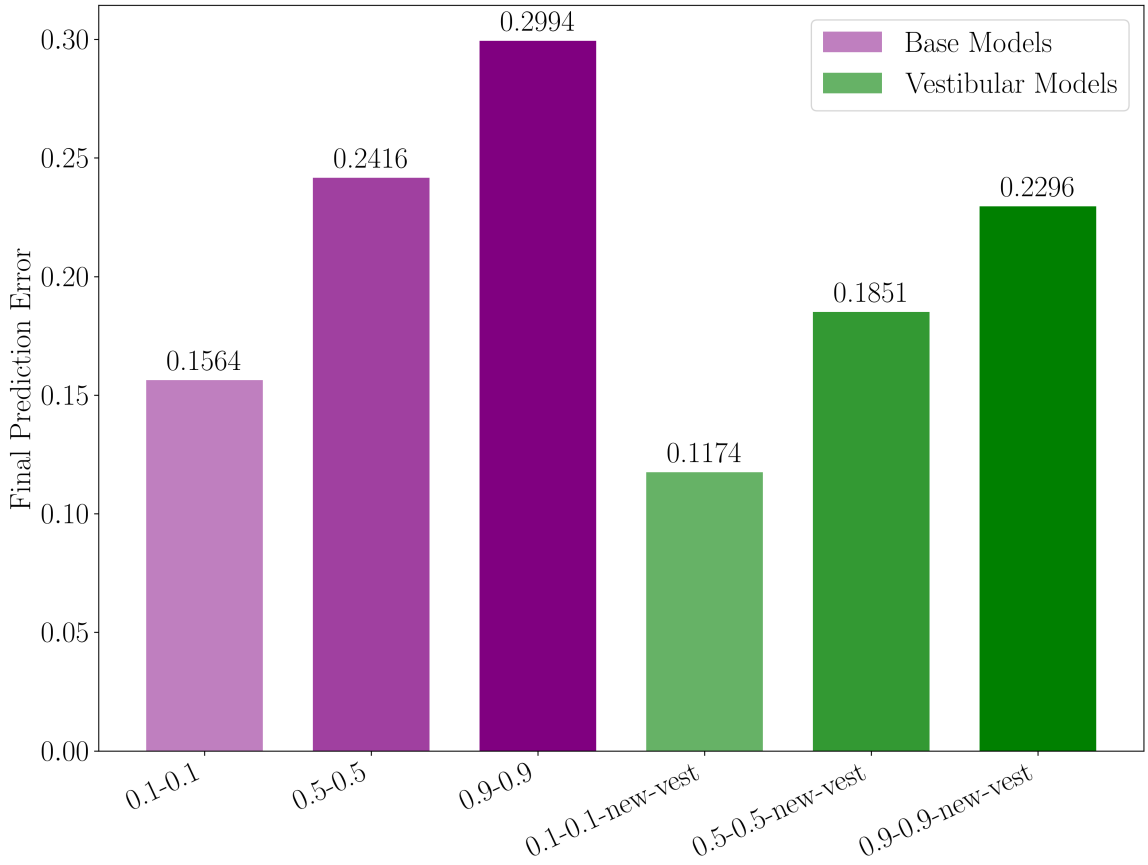

Figure 4.1: **Final image prediction mean squared error on the validation set, for head-fixed agents.** Models with vestibular signaling display overall better performance, but increasingly random trajectories still affect model performance.

studied to eliminate potential biases induced by the environment's shape. The effect was examined using both the last predicted latent in the sequence (after 20 speed applications) and the first predicted latent (where no previous image inputs could overpower the speed input). It was hypothesized that models with more random trajectories might make greater use of the vestibular signal. Figure 4.5 displays the results for the 0.9 head-fixed predictive coding model. However, none of the head turn or non-head turn models across all randomness levels demonstrated the expected effect.

These results paint a complicated picture of function in the predictive coding

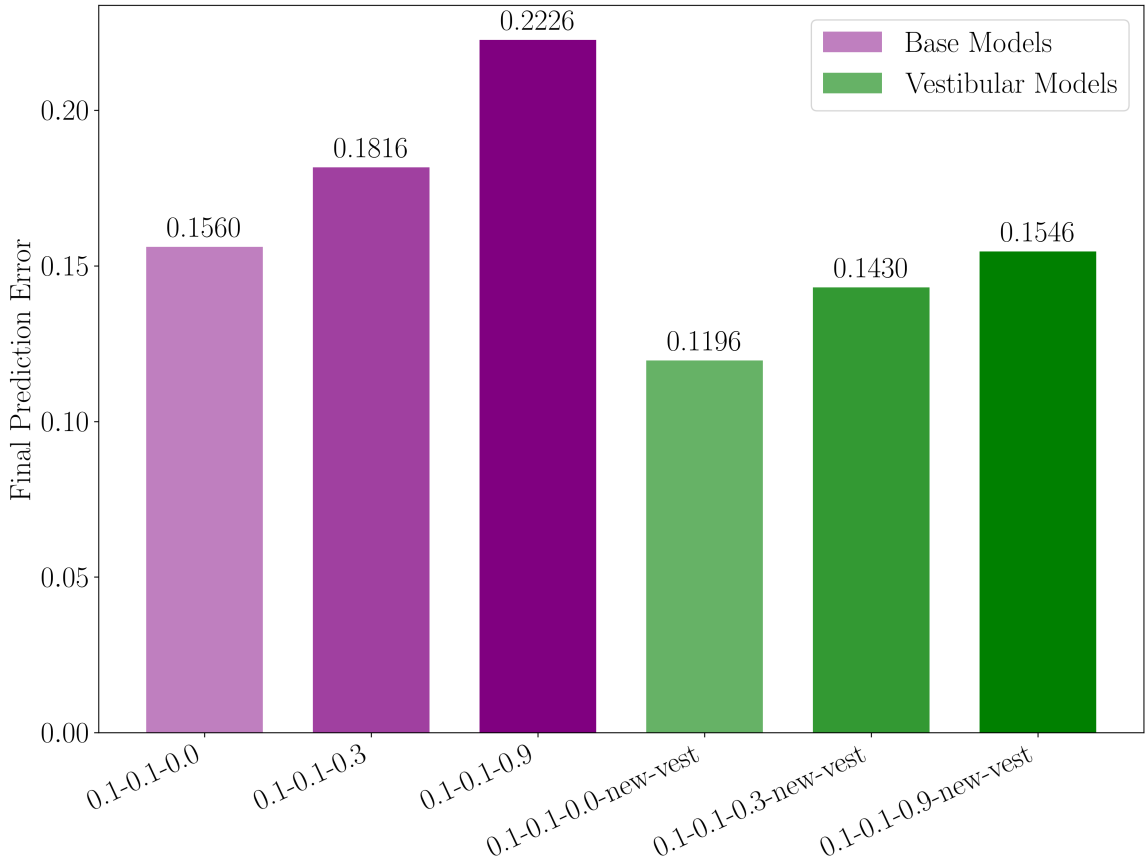

Figure 4.2: **Final image prediction mean squared error on the validation set, for agents with variable head direction.** The non-head-fixed agents display similar performance improvements with the addition of vestibular signaling to the head-fixed agents.

model. In the last section, it was reported that invariance of spatial content in the face of input information supported the alternative hypothesis, and yet here, results show that an additional stream of information does improve the spatial content. While these results might initially appear to support the conclusion that location is being encoded and transitions between locations are being learned, two factors suggest otherwise.

Firstly, the models continue to exhibit significant differences in next-step prediction across randomness settings. Given that the provided speed and angular ve-

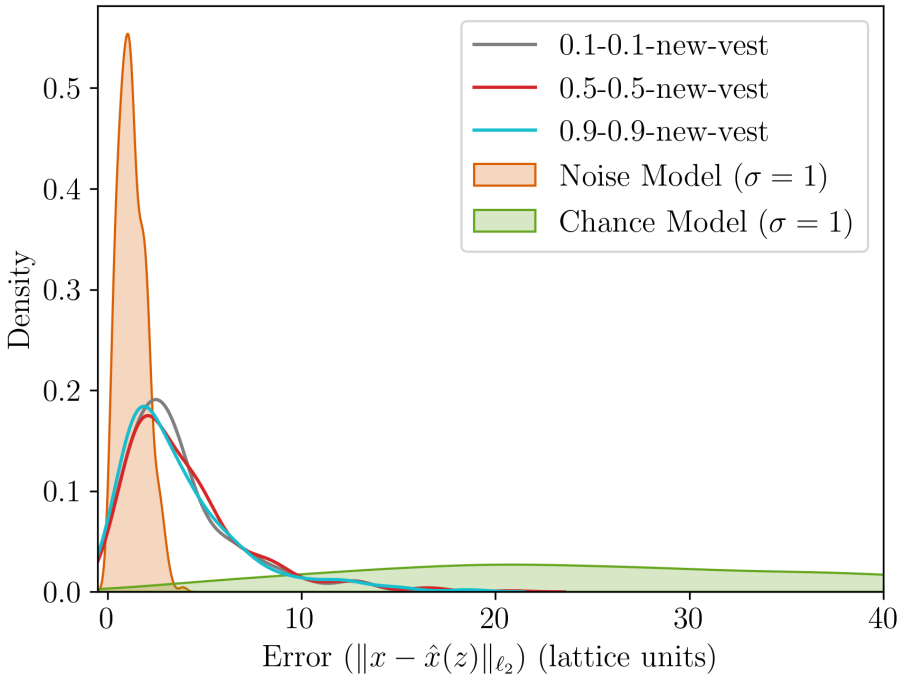

Figure 4.3: **Prediction error as reported by a nonlinear decoder on the post-prediction latents in the head-fixed predictive coding models with vestibular signalling.** Vestibular signalling increases the amount of spatial content found in the predictive coding model's post-prediction latents.

locity signals are noiseless, it is puzzling that the model, over an extended training period with ample data, fails to produce consistent results for this task regardless of the quality of trajectory. Secondly, the observed lack of steering ability is concerning. If this occurred without any performance improvement, it could be attributed to the model learning to prioritize the denser image signal over information regarding movement statistics, potentially indicating a need for reconsideration of the methods applied here. However, the fact that this signal is utilized by the model to realize performance gains, yet fails to produce expected results when varied suggests that the attention heads are not learning a single egocentric transition reflected across multiple modalities. Instead, this gives more credence to the idea that local image feature transitions are being learned, and vestibular signaling, while related to these transitions, cannot be employed to produce equivalent egocentric transformations in

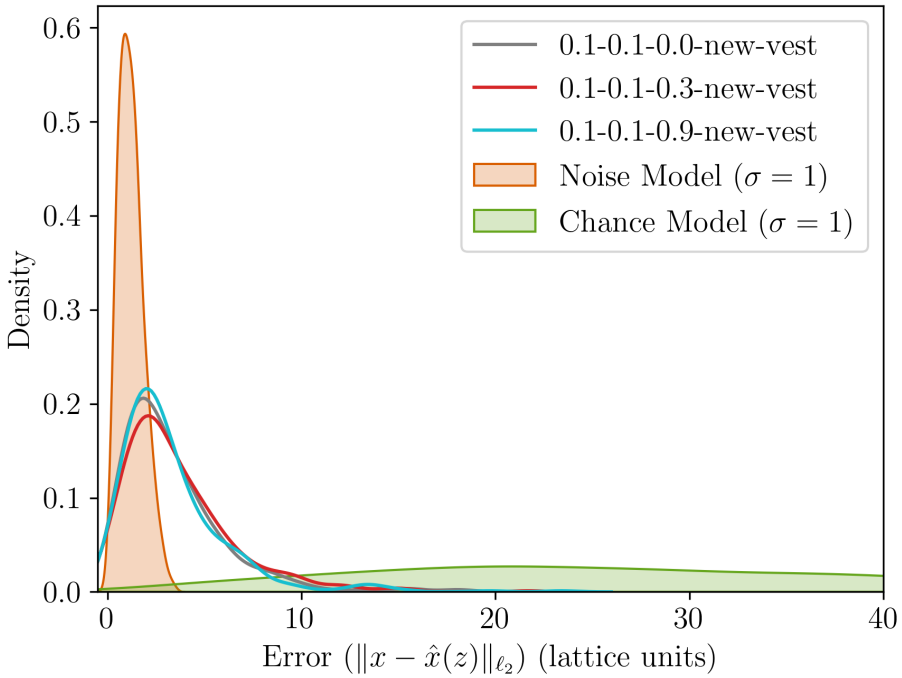

Figure 4.4: **Prediction error as reported by a nonlinear decoder on the post-prediction latents in the non-head-fixed predictive coding models with vestibular signalling.** Vestibular signalling increases the amount of spatial content found in the predictive coding model's post-prediction latents.

the observations.

It is conceivable that an alternative method of integrating vestibular signals could be more effective and might predict the intended effect. This possibility could imply a model in which action signals do induce a cognitive map, a concept explored further in Chapter 6. Nevertheless, these findings, in conjunction with arguments and results from previous chapters, provide substantial evidence to suggest that the predictive coding model is not stitching together locations into a map through prediction.

One last potential mechanism for a cognitive map in the predictive coding model remains through its display of post-prediction place cell-like units. These representations, which have been recharacterized as simple image latents so far in this

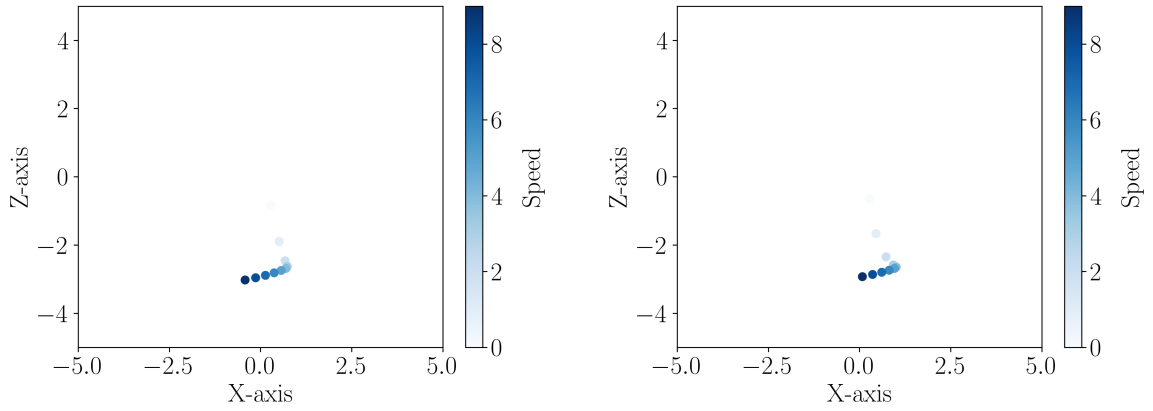

Figure 4.5: **The result of speed inputs on predicted location.** Increasing amounts of speed were applied to the model, and predicted position was decoded using a trained nonlinear decoder. The model facing 0 degrees should produce an increase in position along the z-axis **(left)**. This effect is not observed. The model facing 90 degrees should produce a decrease in position along the x-axis **(right)**. This effect is not observed.

work, do exhibit striking similarities to place cells. This observation raises the possibility that prediction transforms visual features into a representation resembling place cells both superficially and functionally. Such a finding would suggest that the results of this chapter and the previous one do not disprove the predictive coding model, but rather indicate intriguing conclusions about how cognitive maps are computed through prediction.

Chapter 5: Content of Image Latents

While the appearance of place cells in the predictive coding model is a novel and encouraging sign for a model of the hippocampus, Gornet and Thomson are cautious not to place undue importance or meaning on their place cell-like representations. They state: "The appearance of place cell-like firing is common in simpler networks that perform spatial navigation such as (Treves et al., 1992) and (Franzius et al., 2007). It is currently unclear whether artificial place cell-like behavior corresponds to biological place cells. Artificial place cell-like behavior could be an artifact of the simplified inputs or spatial coordinates." It is possible that the place cell-like behavior in their work is indeed just an artifact of simplified visual inputs, a conclusion which would lead to the end of the search for mechanisms supporting a cognitive map. If this is the case, a further discussion arises regarding the claims of Luo et al.: are real hippocampal cells explainable through the same mechanism?

This chapter replicates the cataloging and lesioning experiments from (Luo et al., 2024), first described in Chapter 2, to the larger environment which has been used to test the predictive coding model thus far. It also newly applies these experiments to the predictive coding model. The primary aim remains to test functional cell types in the predictive coding model. However, the implications of replicating results with the baseline models in the larger environment are also discussed.

5.1 Functional Cell Classification

5.1.1 Methods

Images from different grid locations and viewing angles were collected within a small square environment, which is bordered by a short wall and is generally reminiscent of a typical laboratory environment. Individual images were processed by image-processing deep neural network architectures including VGG-16 (Simonyan,

2014), ResNet-50 (He et al., 2016), and ViT-16 (Dosovitskiy et al., 2021), and each architecture was additionally varied by either being pretrained on the image classification dataset ImageNet-1k (Deng et al., 2009), or by having their weights totally randomized. None of the models were further trained with images from the laboratory setting.

Activations from these models after they were shown views from the environment were classified "based on standard criteria for place, direction, and border cells (Banino et al., 2018) (Tanni et al., 2022)". Place fields can be identified by characteristics such as size, number of place fields, or maximal activations within each field, and so several are examined – any unit that has at least one qualifying place field is classified as a place cell. Luo et al. use adapted methods from the cited works to define a qualified place field as one that contains at least 10 pixels but less than half the size of the environment. This is linearly scaled to 30 pixels for the Minecraft recreation of the smaller environment to adapt for its slightly increased size, and to 90 pixels for the larger environment. Furthermore, some of the stricter barriers to the classification of place cells such as stability of the field across multiple visits and iterative thresholding (to ensure that the field arises smoothly from the environment) were originally left out of Luo et al. and have been added back into this analysis. Head direction and border cell classification was left unaltered.

Finally, the setting has been changed from Unity to Minecraft, and the size of the environment from 17x17 to 24x24 (to accommodate for scaling differences between the two platforms), as well as the number of evenly-spaced head directions samples per location from 25 to 17 (to accommodate for computational demands). In accordance with the Unity environment, the Minecraft environment is situated with several different types of trees in the background of the setting in order to decrease aliasing. These changes did not have a significant effect when replicating results with any of their stated models. An example of an observation in the Minecraft recreation can be seen in Figure 5.1.

Figure 5.1: **The small world testing environment.** The Luo et al. environment was replicated in Minecraft for the experiments in this work. Different types of trees were added outside of the testing area to de-alias observations as in the original Unity testing environment.

5.1.2 Results

The results from Luo et al. were replicated in this work for ResNet-50 (pre-trained) due to the similarity to the predictive coding model's encoder architecture of ResNet-18. This analysis revealed several findings of interest. Before making any change to the place cell classification, the previous ResNet-50 results were replicated in the smaller Minecraft environment. These were overwhelmed with cells of mixed place and view encoding, which is noted by the authors to mirror many findings in primates. However, two changes were made in this thesis that explain the differences in results reported here. Firstly, the stricter place cell classification used here led to

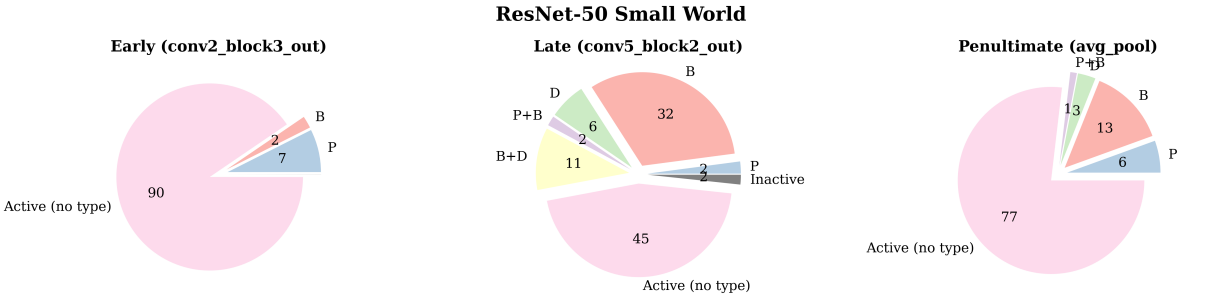

Figure 5.2: **Cell type breakdowns for the predictive coding model in the large world.** In the small environment, the ResNet-50 model produces a diverse population of cell types across multiple layers.

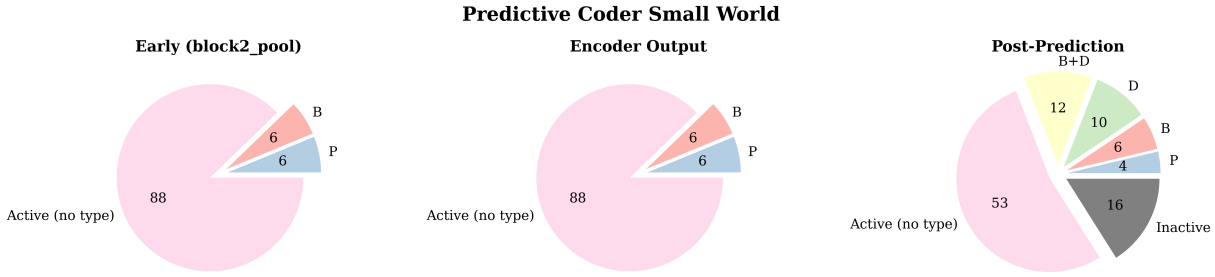

Figure 5.3: **Cell type breakdowns for the predictive coding model in the small world.** In the small environment, the head-fixed predictive coding model produces a diverse population of cell types in the post-attention latents.

most cells not fitting the criteria for any cell type. Secondly, an error was found in the open source code released by Luo et al. which mistakenly was using the mean of the activation map to classify direction rather than the method detailed in Appendix B. Results are shown in Figure 5.2. Cell types vanished even further in the more complex environment across all layers, though most strikingly in the final layers, where more invariant visual features would typically be expected. Results for the bigger environment are shown in Figure 5.5

Next, the same examination was applied to the predictive coding model. An early convolutional block in the encoder was chosen, as well as the pre-prediction

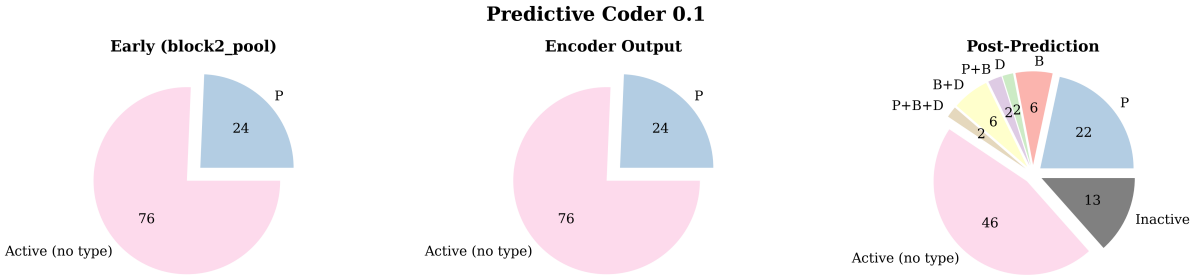

Figure 5.4: **Cell type breakdowns for the predictive coding model in the large world.** In the complex environment, the head-fixed predictive coding model produces a diverse population of cell types in the post-attention latents.

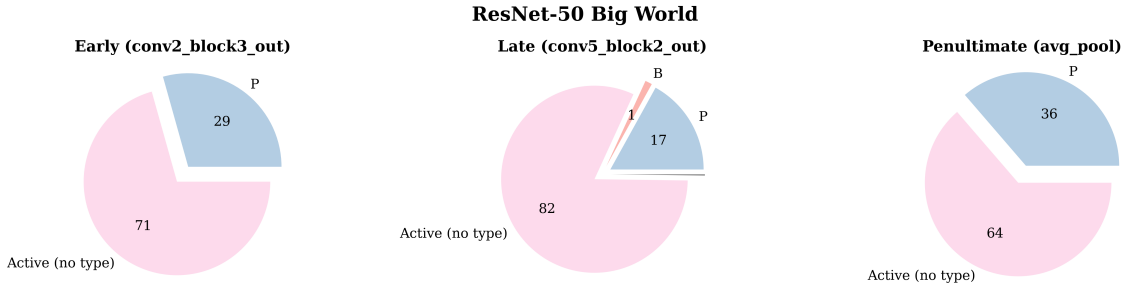

Figure 5.5: **Cell type breakdowns for the ResNet-50 model in the large world.** In the complex environment, the ResNet-50 model displays almost only place cells.

latents and the post-prediction latents. One predictive coding model was trained on the small environment for further comparison, and another low randomness head-fixed predictive coding model was used for investigating in the big environment. The pre-prediction results for the predictive coding model in both environments approximately mirror the ResNet-50 results, but the post-prediction latents appear to splinter off into a much more diverse population of different cell types in both environments. Example activation maps for classified place cells for the ResNet-50 model and the predictive coding model in the large world can be found in Appendix B.

It is particularly intriguing that the post-prediction latents in the predictive

coding model appear to support many more types of functional cells in the larger environment, even including some of the mixed encoding that was originally reported in Luo et al. under the looser standards. One potential explanation for this is that the predictive coding model really is inducing a cognitive map and that it uses these diverse types to navigate in the complex environment. Another explanation follows from the observation that these results have a striking similarity to the reported results of the Vision Transformer (ViT) in Luo et al., not only in the diversity of the cell types observed, but also in the appearance of a sizable population of inactive units. Vision transformers of course imbue image representations with contextual information in the same way that this thesis proposes the predictive coding model works. If it is to be proven that the predictive coding model is inducing these cell types to support localization, rather than as a consequence of their architecture, it must next be examined how their removal affects localization.

5.2 Functional Cell Performance and Lesioning

5.2.1 Methods

Luo et al. also examined the performance of the model activations in predicting location, head direction, and distance to the closest wall: tasks meant to target the respective responsibilities of place, head direction, and border cells. The examination here will focus on the location task. Linear regression models were trained on representations from selected layers to predict these targets, with 70% of the locations and views held out for a rigorous test of generalization across the environment. Chance decoding results were calculated in two ways: in one, locations (or head direction) were randomly shuffled, in the other, the center location/direction is always predicted. The decision was made to allow the linear decoder to train on 80% of the locations in the large world, in order to produce a better baseline for the lesioning studies. Distance error for these figures are represented as Minecraft units, normalized to the size of the environment.

Performance in these tasks was additionally measured under several methods of lesioning model units. In the original analysis, Luo et al. lesions increasingly large populations of model units in order to determine which population has the most impact on performance in location prediction tasks. They show that across all models, lesioning the cells which show the highest "placeness", "directionness", or "borderness" by the previously-specified metrics does not have more of an effect on performance in any task than randomly shuffling the units and lesioning the same amounts that way. Because metrics such as the number of place fields, max value across fields, place field size, etc. all contribute to an understanding of the "placeness" of a unit, these were considered separately. These results were compared to lesioning of the model units which report the highest coefficients in the linear regression model trained for each task, which shows a much larger effect than lesioning by functional cell type. This effect is expected to be true for the predictive coding model as well (as lesioning the units with high coefficients in a linear decoding model has a foregone conclusion in any study, and serves primarily as a baseline) and focus instead on the lesioning of cell types. This thesis will display results on the lesioning task for the number of place fields and the maximal activations within those place fields, as well as for directionness, although future work should also examine the lesioning effects of cells with the most spatial information content.

5.2.2 Results

Lesioning performance was calculated across the two models in the small environment. The results of lesioning the measures of the size and number of place fields. The ResNet-50 model is able to achieve significant location decoding performance, in line with the results previously shown for the predictive coding models. The ResNet-50 model displays the same lesioning behavior in the Minecraft version of the small environemnt as was previously reported in the small Unity environment, shown in Figures **??**, 5.7, and 5.8.

It should be noted here that the random lesioning is applied to any active unit

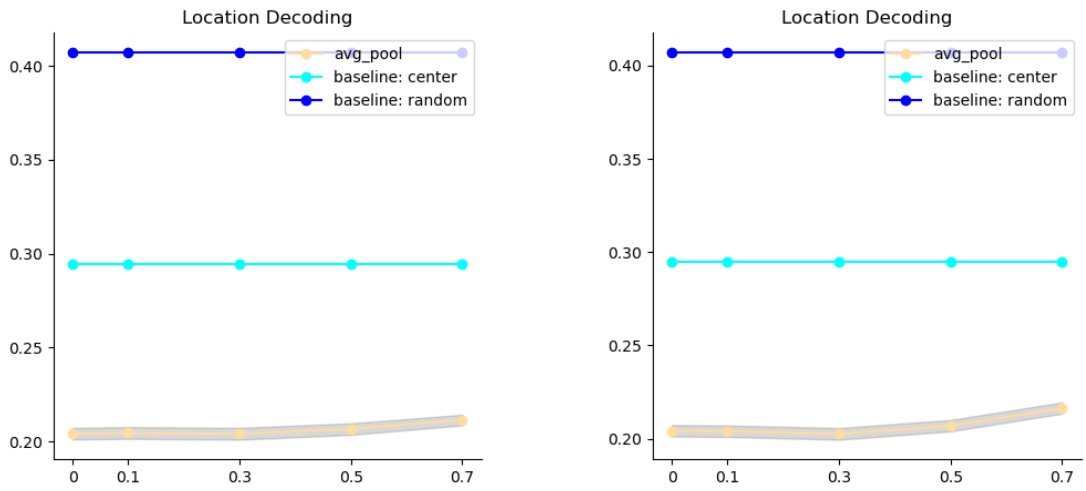

Figure 5.6: **Results on localization task for lesioning by "directionness" on the ResNet-50 model in the small environment.** The ResNet-50 model does not exhibit significantly worse than chance performance when units which display increased directional preference are increasingly removed from the population. Lesioning randomly **(left)**, lesioning top units **(right)**.

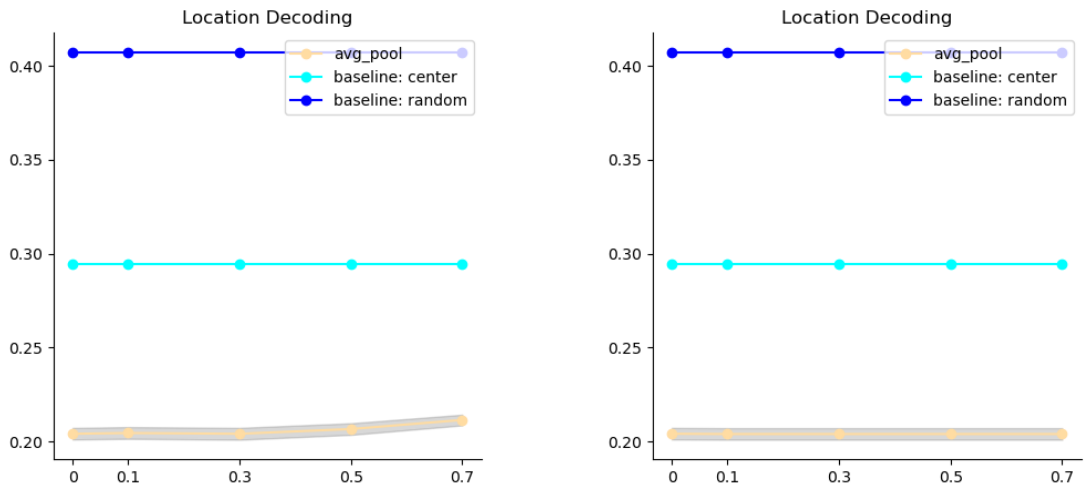

Figure 5.7: **Results on localization task for lesioning by the number of place fields on the ResNet-50 model in the small environment.** The ResNet-50 model does not exhibit worse than chance performance when units which display increased amounts of place fields are increasingly removed from the population. Lesioning randomly **(left)**, lesioning top units **(right)**.

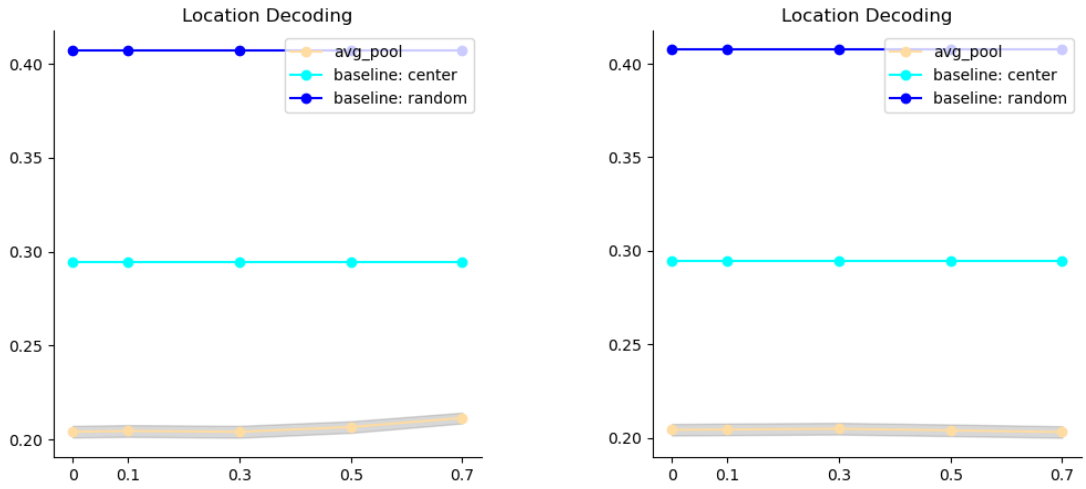

Figure 5.8: **Results on localization task for lesioning by the maximum activation value across the place fields on the ResNet-50 model in the small environment.** The ResNet-50 model does not exhibit worse than chance performance when units that display place fields with high activations are increasingly removed from the population. Lesioning randomly **(left)**, lesioning top units **(right)**.

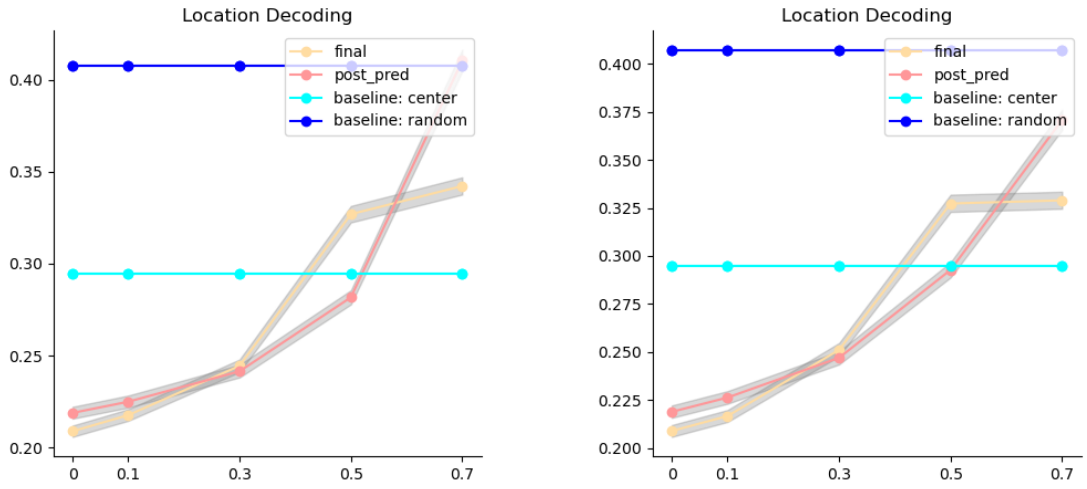

Figure 5.9: **Results on localization task for lesioning by "directionness" on the predictive coding model in the small environment.** The predictive coding model does not exhibit worse than chance performance when units which display increased directional preference are increasingly removed from the population. Lesioning randomly **(left)**, lesioning top units **(right)**.

in the population, whereas the top-most lesioning can only be applied, in the case of the two place field analyses, to units which display any place field at all. Results for directionness do not have this constraint. The baseline location decoding performance for the predictive coding model is very similar to the results of the linear decoder in previous sections. The predictive coding model displays very similar lesioning performance as the ResNet-50 model, results of which are displayed in Figures 5.9, 5.10 and 5.11. No type of sorted lesioning induced any above-chance performance deficits, including the more rigorously defined place fields. The predictive coding model's lesioning induced such large drops in performance due to the small number of units available, however, the lesioning of the top units is still not more significant than the lesioning of random units, indicating no relationship.

One argument that was made earlier is that the localization test may be too easy and overdefined to produce meaningful results in the ablation study. Future

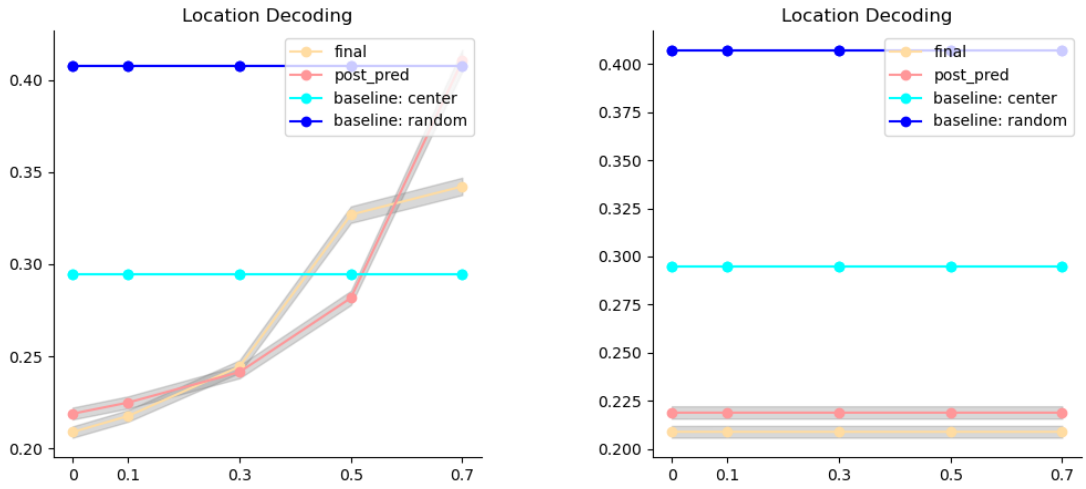

Figure 5.10: **Results on localization task for lesioning by the number of place fields on the predictive coding model in the small environment.** The predictive coding model does not exhibit worse than chance performance when units which display increased amounts of place fields are increasingly removed from the population. Lesioning randomly **(left)**, lesioning top units **(right)**.

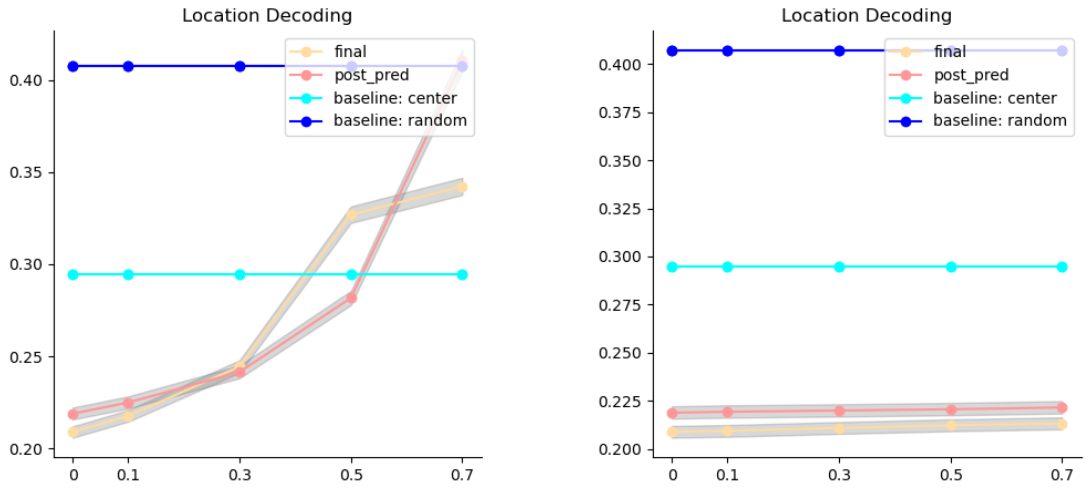

Figure 5.11: **Results on localization task for lesioning by the maximum activation value across the place fields in the predictive coding model in the small environment.** The predictive coding model does not exhibit worse than chance performance when units that display place fields with high activations are increasingly removed from the population. Lesioning randomly **(left)**, lesioning top units **(right)**.

work should examine this effect in the larger and more complex environment.

These results paint the compelling story that the image features of the predictive coding model are not unique: their activation patterns can be explained through other networks that embody the alternative proposed mechanism of action as suggested by this thesis. Additionally, they are not functional: when trained to minimize prediction error, these latent units do not take on interpretable roles to achieve this goal. This work has now, across the last three chapters, completed its argument that the cognitive map is not being supported. Additionally, results in this chapter, while not a conclusive argument against the strong conclusions of Luo et al., suggest a further line of experiments that could be used to explore this issue. The following chapter uses these results to draw final conclusions about the predictive coding model and further discusses the implications of the Luo et al. claims.

Chapter 6: Discussion and Future Work

This chapter includes a wealth of information supporting the conclusions of the thesis and expanding on its implications. First, it summarizes the results found experimentally in this thesis. Next, it provides further discussion on the sweeping claims made in Luo et al. that were not directly addressed in the experimental findings of this thesis. Further support is next provided for the claim that local image features in a vanilla transformer model cannot form a broad cognitive map, that has been proposed earlier in this thesis but not yet substantiated. Finally, while the predictive coding model is insufficient in its current form, it represents an exciting opportunity to advance the field of artificial cognitive maps. This chapter discusses potential improvements to the current formulation of the predictive coding model that could enable it to form a cognitive map from its simple and compelling design.

6.1 Summary of Results

This thesis has provided significant evidence challenging the purported function of the predictive coding model. The model was originally proposed by Gornet and Thomson to implement a mechanism that would generate a mathematically-motivated and robust map of space. Supporting evidence for their mechanistic claim included the spatial content induced by the predictive mechanism and the emergence of place cell-like representations. The alternative theory supported in this thesis proposed that image features, rather than explicit location and pose information, are being encoded and decoded by the model. This other proposal would indicate that the transition statistics learned by the attention heads reflect only surface-level feature changes, albeit with impressive performance.

Chapter 3 supported this hypothesis by demonstrating that difficult-to-learn transition statistics do not affect the encoding of location information in the way that

they should under the old understanding of the model. Chapter 4 further corroborated this claim, producing evidence that the transitions learned by the predictive portion of the network only superficially relate to egocentric movement signals. If integrated correctly under the originally proposed method, these signals would provide the model with noiseless information to directly solve the prediction task under all levels of trajectory randomness. However, this effect was not observed.

The place cell-like representations observed by Gornet and Thomson were not fully explained by the alternative hypothesis proposed in this thesis. Therefore, further experiments were conducted to explore whether these cell types were uniquely produced by the prediction task and if they could potentially support a cognitive map. Such a result would lend support to both the predictive coding model and traditional understandings of place cell functionality. Chapter 5 applied methods from Luo et al. to demonstrate that these representations are not unique, as they have been observed in a broader class of image processing models. This chapter tied the observed functional cell types back to the alternative hypothesis and showed that these representations do not uniquely support the model's function. Consequently, this final notion of support for Gornet and Thomson's original claims cannot be relied upon.

6.2 Does Experimental Setup Induce Superfluousness?

The further analysis of results from Luo et al. suggested a more complex "functional cell type story" than initially reported. This thesis revealed that much of the previously shown mixed encoding which mirrored primate data relied on either mistakes or generous cataloguing methods. Additionally, many cell types seemed to disappear in larger, more complex environments. However, this does not represent a comprehensive refutation of the ideas presented in their work.

Several preliminary arguments and further tests could potentially address these claims. While the diversity and quantity of place cells in spatial systems are note-

worthy, the lack of significant deficits in localization tasks following lesioning of these functional cell types is difficult to ignore. However, this test may not be sufficiently stringent. It is unsurprising that in a simple environment with minimal visual aliasing and smooth and predictable features such as a wall, a linear decoder might demonstrate good localization performance when given image-processing model activations from an untrained model. In such a setting, even linear decoders trained with pixel activations alone might accomplish this task. This is supported by the surprising performance and functional cell population of the earlier layers of the models examined in Luo et al. Given these conditions, it is reasonable to expect that location is redundantly encoded by various cell types, and that the linear decoder would ultimately rely on a combination of the most convincing ones to decode location.

Luo et al. primarily reinforce the notion that place cells and similar structures should not be strictly tied to Euclidean spatial concepts, a point that is not contested in this thesis. However, it is premature to conclude that functional cell types lack meaningful roles and functions. An alternative hypothesis is that functional cell types arise from the processing of sensory information into a representation which is invariant to selective types of transformations. This hypothesis would explain the potential confusion with activations induced by deep image processing networks, as the inductive bias of the model architecture produces invariant representations of image features (via convolution and max pool operations, or even attention blocks). Under this framework, hippocampal representations would be both reflective of sensory input and genuinely meaningful, though this meaning may rely on a generalized space contextualizing sensory input with information such as internal beliefs and goals.

The question of what hippocampal activations actually "mean", and what role they play, will likely be a hotly debated question in neuroscience for some time. The primary argument in this thesis is that ablation studies should respect the complexity and redundancy of the brain, and be conducted using tasks that are sufficiently difficult and specific in order to produce a deficit when certain cell populations are

targeted. The analysis provided by Luo et al. offers another reason to avoid interpreting functional cells solely under their most literal and Euclidean space-based definitions.

6.3 Limitations of Autoregressive Transformers in Cognitive Map Construction

The experimental findings presented here suggest that the encoder and decoder networks in the predictive coding model do not appear to be learning any special mapping between images and location as indicated by Gornet and Thomson in their probabilistic formulation. Results in Chapter 5 in tandem with previous arguments even support that the model could still achieve good performance with an untrained encoder. Therefore, this model does not appear to take any special performance from being in a continuous domain and could be describable in similar terms as a vanilla transformer, where existing literature on the subject can illuminate issues that this architecture has with cognitive map building.

Gornet and Thomson mention in the last section of their paper that "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 (Brown, 2020). 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." This is true, but neglects to consider the large and quickly developing body of work which indicates that there might be large gaps in the reasoning ability of large language models. Momennejad et al. (2023) directly tested several large language models using their custom cognitive mapping test suite, CogEval, revealing severe deficits across all architectures. In an evaluation modeled on discrete

finite automata, transformers given turn-by-turn sequences of textual descriptions of Manhattan taxi rides displayed near-perfect results in predicting the next turn and encoded current location in their state representations (Vafa et al., 2024). However, the recovered world model deviated significantly from the ground truth, and the model struggled with downstream tasks when environmental modifications were introduced.

Despite these limitations, the transformer may still serve as a valuable starting point for predictive-coding-based models of the hippocampus. Transformers form sequentially invariant representations and exhibit impressive in-context learning capabilities, both desirable traits for architectures supporting general-purpose reasoning. As language models grow in scale and complexity, these inherent strengths may facilitate the development of increasingly sophisticated internal representations resembling cognitive maps. Several other promising architectures incorporate transformer blocks, including TEM-t (Whittington et al., 2022b), Transformer with Discrete Bottleneck (TDB; Dedieu et al., 2024), and various versions of Joint Embedding Predictive Architecture (JEPA) models such as Vision JEPA (V-JEPA; Bardes et al., 2024), which employs vision transformers. TEM-t, in particular, presents an interesting variant of the previous cognitive mapping model, the Tolman-Eichenbaum Machine, leveraging the fact that attention can be computed in modern Hopfield networks (Ramsauer et al., 2020). These models will be further discussed in subsequent sections.

6.4 Possible future improvements on the predictive coding model

The struggle to use action signaling in a biologically plausible way is considered a major missing component of this predictive coding model. However, the model's inability to integrate these signals may stem from its training loss rather than its architecture. From a biological perspective, the hippocampus typically receives highly processed sensory input from areas associated with late-stage visual processing and other sensory areas. A loss function operating in pixel space is akin to asking the

hippocampus to reconstruct every early activation in the primary visual cortex. This could contribute to the model's apparent failure to build a better conceptual understanding of the environment's underlying geometry, as evidenced by the issues in steering efforts. Luo et al. note the presence of hippocampal-like activations in other cortical areas. Rather than this fact being used as a way to explain how hippocampal representation is not unique, this may indicate that the hippocampus extensively utilizes the invariance and organization of these representations, or even plays a role in shaping them.

Initially, a predictive coding model that takes visual inputs and predicts simply the presence of represented objects in a scene (either wholly or partially) was considered for use in this thesis and implementation began. While this approach would diverge from modeling primate hippocampal function (with their foveated and detailed visual input), it could potentially allow for more general and accurate functionality. A more flexible realization of this objective might be found in current exciting work on world models in reinforcement learning, where action consequences are predicted in latent space (Hu et al., 2023; Hafner et al., 2023). Cognitive mapping models are often framed in reinforcement learning terms, although they are primarily distinguished by the quest to find accurate hippocampal representations and build structure from sequences without rewards (despite much progress in reinforcement learning world models being demonstrated in settings with sparse rewards).

Learning to predict the impact of actions in latent space can be further detangled from rewards by taking a look at the very closely related JEPA class of models, which takes in the latents of a transformed image and uses information about the transformation to predict the inverse of this transformation, with the loss entirely in the latent space. (Garrido et al., 2024) provides further motivation for the use of action signaling in the context of a possible future predictive coding model: "If one does not condition the predictor on the transformation parameters, then the best we can hope for is learning representations that are invariant to the data transformations".

Action conditioning allows the JEPA class of models to instead learn equivariant representations of their data, which are much more useful across a variety of downstream tasks. This connection could allow the framing of the cognitive mapping problem as one in which the hippocampus creates equivariant representations of sensory input through action-informed prediction.

6.5 Context within other cognitive mapping architectures

Recontextualizing the predictive coder network allows for its consideration within the broader realm of cognitive mapping models. This section will primarily focus on two that have been previously introduced in Chapter 2. The first particularly relevant model to this discussion is the Tolman-Eichenbaum Machine, or TEM, which views hippocampal representations as a conjunction between sensory stimuli (from the lateral entorhinal cortex) and structure information (from the medial entorhinal cortex, i.e., grid cells). An action signal can lead to the next remembered observation, or the remembrance of an observation can inform which action signals are available, enabling generalization between environments. However, TEM cannot be the complete explanation in its current form as it requires a global coordinate system and operates using allocentric actions.

The clone structured cognitive map (CSCG; Raju et al., 2024) also creates a cognitive map by modeling how observations can be formulated as latent states and the actions connecting these states. However, CSCG does not use observation content for any purpose other than distinguishing them from their surroundings (even in the continuous setting case, where images are condensed with a vector quantizer into discrete states). Consequently, it cannot use this information to generalize and relies on external algorithms to match the appropriate map to the situation. These two models have been speculated to represent two forms of hippocampal function: map as CSCG, and memory as TEM (Whittington et al., 2022a). Bridging the gap between these models in a single framework would be an exciting development. One avenue for

exploration involves variants of both models built with the transformer architecture. TEM-Transformer (TEM-t) performs the same computation as TEM more efficiently by replacing its Hopfield network and RNN with causally-masked attention blocks. Locations in the environment are represented as observation-location pairs; in the attention computation, the value is fixed to the observation (a one-hot vector) while keys and queries rely entirely on the location.

On the other hand, the authors of CSCG propose TDB, or Transformers with Discrete Bottlenecks, as a solution to path-planning in partially observed environments. This causal transformer accepts a sequence of alternating observations and actions $(x_1, a_1, x_2, a_2, \ldots, x_n, a_n)$. Processed observations T_n, now imbued with contextual information, are compressed with a vector quantizer to form a discrete code $d(T_n)$. The corresponding action a_n is then added to this code to produce x_{n+1}. Due to the sequence invariance of the transformer architecture, each code forms a representation of state that includes both context and content. Interpretable graphs can be derived from these codes and used for downstream tasks such as path planning. However, George et al. note that this model relies on multiple sets of discrete codes which are then coalesced, and these discrete codes do not form a detangled latent space reflecting the environment.

The first application of sensory information in this setting is that, as acknowledged by Dedieu et al., high dimensionality could help distinguish the codes and allow for a detangled latent space. However, it would be additionally advantageous to fold in continuous sensory information, as the codes could then be generalized to new situations based on any useful aspect of the observation. One might also imagine a version of TDB that operates primarily in latent space, similar to joint-embedding models. While TDB does include one additional loss term of this nature, taking after (Guo et al., 2022), their loss refers to the dictionary codes and is thus less flexible than what is envisioned here.

Chapter 7: Conclusion

The findings and analysis presented here highlight the complexities and limitations of current approaches in modeling the form and function of cognitive maps. The predictive coding model, while promising, needs additional constraints to piece together observations into a coherent and flexible map of the environment. The predictive coding model's shortcomings in the cognitive mapping task align with broader challenges observed in transformer-based models and large language models, which, despite their impressive capabilities, struggle with flexible reasoning in out-of-distribution domains.

Moving forward, it is proposed that future research should focus on developing models that can better integrate sensory information with action signals, possibly by predicting consequences in latent space rather than pixel space, aligning with current conceptions of the inputs to the hippocampal formation. This approach, inspired by recent work in world models and joint embedding predictive architectures (JEPA), could potentially bridge the gap between map-like (CSCG) and memory-like (TEM) models of hippocampal function. By combining the strengths of these approaches with the insights gained from the predictive coding experiments, it may be possible to create more robust and generalizable models of spatial cognition that more accurately reflect the flexibility and adaptability of biological systems.

Appendix A: Predictive Coding Methods

The code published by Gornet and Thomson was heavily utilized with some edits. The original codebase can be found at https://github.com/jgornet/predictive-coding-recovers-maps. Thank you so much to the original authors for providing this open source code.

A.1 Location Decoders

Gornet and Thomson trained a neural network to predict the agent's position from the predictive coder's latent units. The prediction error of mean-squared error between locations indirectly measures the predictive coder's positional information. Their simple neural network architecture consisted of a convolutional layer, a max-pooling layer, two linear layers, and a ReLU.

Activations were gathered from the predictive coding model for this down-stream task by giving the model sequences of input images. At every non-obstructed grid location in the environment, 10 image observations were collected which were approximately in the same location. The yaw of the agent was fixed to 0 degrees for every image in this dataset. The predictive coding model takes in these 10 image observations and produces a set of 10 predicted latents – only the last latent set in the sequence is used to produce an observation. This means that every location was seen once, with one viewing angle. Additionally, the localization model was trained on the entire dataset with no locations held out. Generalization was measured by adjusting the input image normalization by a constant factor.

Several changes were made to the localization model. First, the dataset was changed such that every location has 50 evenly-spaced heading direction examples. Several sequences of length 20 (as the model in this work uses a longer sequence length than in Gornet and Thomson) are randomly generated from this set of 50

views for each location. Instead of changing the image normalization, a test set consisting of 20% of the view-position combinations randomly selected from across the environment is now held out. The nonlinear location decoder is additionally adjusted to add another convolutional layer, as well as dropout between the two linear layers.

A linear position decoding model was also introduced in order to provide a more robust test of positional content. This model is a linear support vector regressor, trained with 5-fold cross validation to predict normalized locations on the same dataset as the nonlinear model.

Appendix B: Luo et al. Methods

This thesis uses the open source code from Luo et al., which can be found at https://github.com/don-tpanic/Space. Thank you so much to the original authors for providing this open source code.

B.1 Place Field Classification

Activations for every model unit across all locations in each environment were collected to form activation maps. Place, head-direction, and border cells were identified based on specific firing characteristics from (Tanni et al., 2022; Banino et al., 2018). For place cells, the number and size of place fields were examined, defining a qualified field as spanning 10 pixels to half the environment size. Head-direction cells were assessed using the resultant vector length of the directional activity map. For each direction, vectors representing angle and average intensity were created:

$$\mathbf{r}_i = [\beta_i \cos \alpha_i, \beta_i \sin \alpha_i]^T \tag{B.1}$$

where α_i and β_i are the angle and average intensity for direction i. The mean resultant vector was calculated as:

$$\tilde{\mathbf{r}} = \frac{\sum_{n=1}^{N} r_i}{\sum_{n=1}^{N} \beta_i} \tag{B.2}$$

using 24 uniformly spaced angular directions. Border cells were identified using a border score, comparing wall activation to center activation:

$$b_s = \max_{i \in 1,2,3,4} \frac{b_i - c}{b_i + c} \tag{B.3}$$

where b_i is the mean activation within 3 bins of wall i, and c is the average activity beyond this threshold. Units with a border score > 0.5 were considered border-like.

The original work from Luo et al. uses a simplified method of calculating place cells from the cited Tanni et al. work. This thesis uses the full method, which

is distinct due to its higher demand on the continuity and "smoothness" of the place field which is achieved by iterative thresholding of the activations, as well as the requirement that place fields be stable across multiple visits. Since activations were recorded at 17 different heading directions for every location, these were split into two populations for use in the stability analysis.

B.2 Lesioning Analysis

To analyze how different deep neural network model units contribute to the decoding of spatial content in the network, two unit exclusion analyses are employed. In the first, units are excluded based on their spatial profiles, ranking them according to specific spatial measures like field count. The top n% of units with the strongest spatial characteristics are progressively excluded, assessing the impact on decoding performance for each unit type and associated task as the exclusion rate increases. A control group with random unit exclusion is included for comparison.

The second analysis focuses on task-relevance, using the magnitude of learned decoder coefficients to determine unit importance. Units are gradually excluded based on these coefficient magnitudes, linear decoders are retrained with the remaining units, and their performance on spatial tasks is evaluated. Again, a control analysis with random unit exclusion is conducted. For both analyses, spatial decoders are retrained on the remaining model units following the same procedure as in the original decoding process. This approach allows for quantification of the relative importance of different unit types and their specific contributions to spatial information processing within the model. This thesis does not make sure of the coefficient examination.

B.3 Examples of Functional Cell Types

This section contains examples of the functional cell types collected by the adjusted methods in the large testing environments.

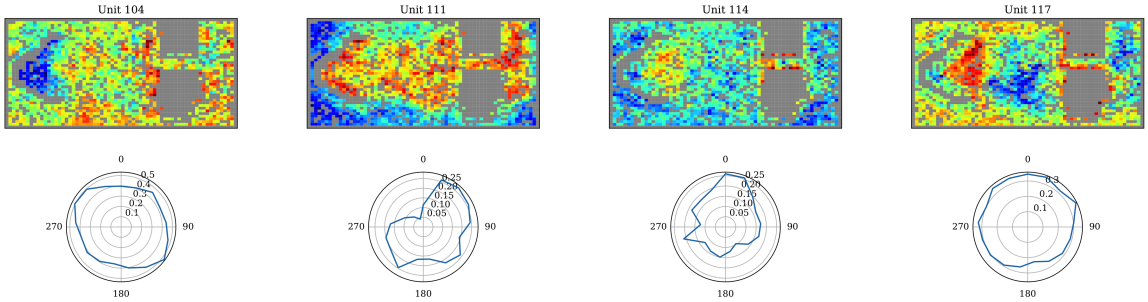

Figure B.1: **Activation maps for units classified as place cells from the post-prediction predictive coding model in the big world.**

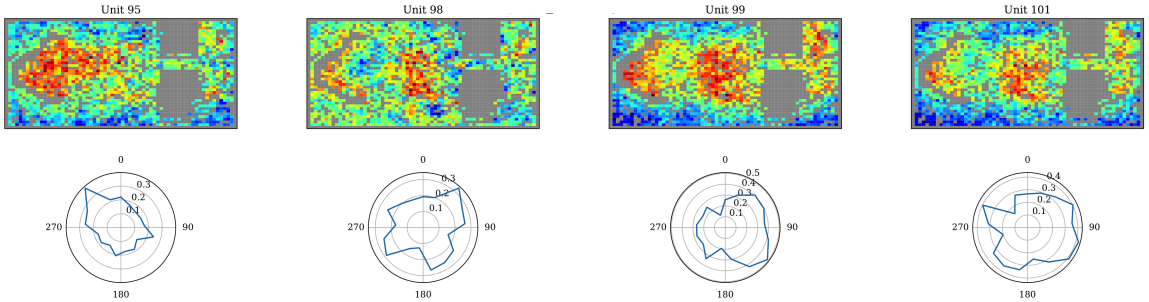

Figure B.2: **Additional activation maps for units classified as place cells from the post-prediction predictive coding model in the big world.**

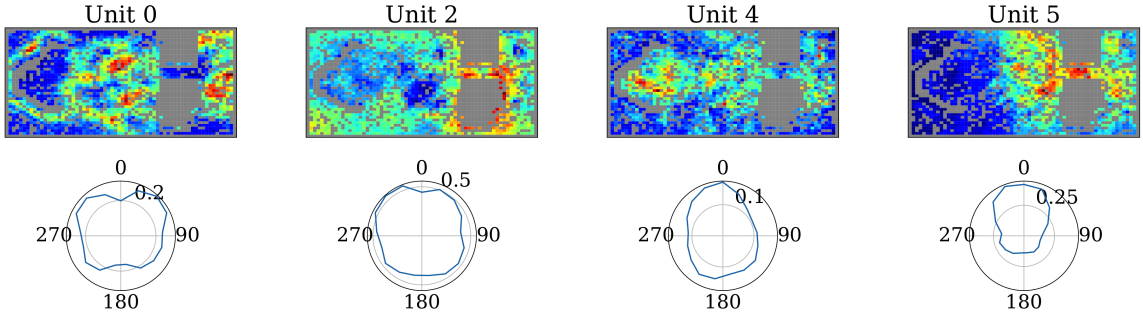

Figure B.3: **Activation maps for units classified as place cells from the last layer of the ResNet-50 model in the big world.**

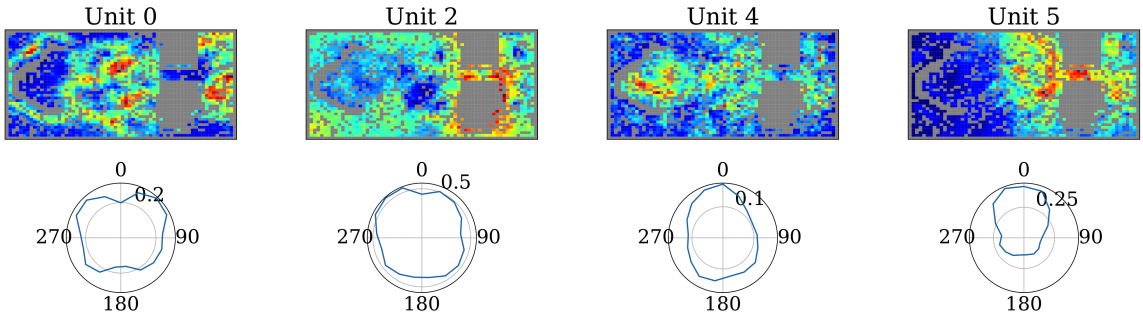

Figure B.4: **Additional activation maps for units classified as place cells from the last layer of the ResNet-50 model in the big world.**

Works Cited

(PDF) Evaluating the World Model Implicit in a Generative Model, a. URL https://www.researchgate.net/publication/381226930_Evaluating_the_World_Model_Implicit_in_a_Generative_Model.

Revisiting Feature Prediction for Learning Visual Representations from Video | Research - AI at Meta, b. URL https://ai.meta.com/research/publications/revisiting-feature-prediction-for-learning-visual-representations-from-video/.

Space is a latent sequence: A theory of the hippocampus, c. URL https://www.science.org/doi/10.1126/sciadv.adm8470.

Dmitriy Aronov, Rhino Nevers, and David W Tank. Mapping of a non-spatial dimension by the hippocampal–entorhinal circuit. *Nature*, 543(7647):719–722, 2017.

Andrea Banino, Caswell Barry, Benigno Uria, Charles Blundell, Timothy Lillicrap, Piotr Mirowski, Alexander Pritzel, Martin J. Chadwick, Thomas Degris, Joseph Modayil, Greg Wayne, Hubert Soyer, Fabio Viola, Brian Zhang, Ross Goroshin, Neil Rabinowitz, Razvan Pascanu, Charlie Beattie, Stig Petersen, Amir Sadik, Stephen Gaffney, Helen King, Koray Kavukcuoglu, Demis Hassabis, Raia Hadsell, and Dharshan Kumaran. Vector-based navigation using grid-like representations in artificial agents. *Nature*, 557(7705):429–433, May 2018. ISSN 1476-4687. doi: 10.1038/s41586-018-0102-6. URL https://www.nature.com/articles/s41586-018-0102-6. Publisher: Nature Publishing Group.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature predic-

tion for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024.

Tom B Brown. Language models are few-shot learners. *arXiv preprint ArXiv:2005.14165*, 2020.

Yoram Burak and Ila R. Fiete. Accurate Path Integration in Continuous Attractor Network Models of Grid Cells. *PLOS Computational Biology*, 5 (2):e1000291, February 2009. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1000291. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000291. Publisher: Public Library of Science.

Xinlei Chen and Kaiming He. Exploring Simple Siamese Representation Learning, November 2020. URL http://arxiv.org/abs/2011.10566. arXiv:2011.10566 [cs].

Alexandra O Constantinescu, Jill X O'Reilly, and Timothy EJ Behrens. Organizing conceptual knowledge in humans with a gridlike code. *Science*, 352 (6292):1464–1468, 2016.

Christopher J. Cueva and Xue-Xin Wei. Emergence of grid-like representations by training recurrent neural networks to perform spatial localization, March 2018. URL http://arxiv.org/abs/1803.07770. arXiv:1803.07770 [cs, q-bio, stat].

Antoine Dedieu, Wolfgang Lehrach, Guangyao Zhou, Dileep George, and Miguel Lázaro-Gredilla. Learning Cognitive Maps from Transformer Representations for Efficient Planning in Partially Observed Environments, January 2024. URL http://arxiv.org/abs/2401.05946. arXiv:2401.05946 [cs].

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, June 2021. URL http://arxiv.org/abs/2010.11929. arXiv:2010.11929 [cs].

Howard Eichenbaum. Prefrontal–hippocampal interactions in episodic memory. *Nature Reviews Neuroscience*, 18(9):547–558, 2017.

Mathias Franzius, Henning Sprekeler, and Laurenz Wiskott. Slowness and sparseness lead to place, head-direction, and spatial-view cells. *PLoS computational biology*, 3(8):e166, 2007.

Quentin Garrido, Mahmoud Assran, Nicolas Ballas, Adrien Bardes, Laurent Najman, and Yann LeCun. Learning and Leveraging World Models in Visual Representation Learning, March 2024. URL http://arxiv.org/abs/2403.00504. arXiv:2403.00504 [cs].

Dileep George, Rajeev V. Rikhye, Nishad Gothoskar, J. Swaroop Guntupalli, Antoine Dedieu, and Miguel Lázaro-Gredilla. Clone-structured graph representations enable flexible learning and vicarious evaluation of cognitive maps. *Nature Communications*, 12(1):2392, April 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-22559-5. URL https://www.nature.com/articles/s41467-021-22559-5.

James Gornet and Matthew Thomson. Automated mapping of virtual environments with visual predictive coding, August 2023. URL http://arxiv.org/abs/2308.10913. arXiv:2308.10913 [cs, eess, q-bio].

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to

self-supervised Learning, September 2020. URL http://arxiv.org/abs/2006.07733. arXiv:2006.07733 [cs, stat].

Zhaohan Guo, Shantanu Thakoor, Miruna Pîslar, Bernardo Avila Pires, Florent Altché, Corentin Tallec, Alaa Saade, Daniele Calandriello, Jean-Bastien Grill, Yunhao Tang, et al. Byol-explore: Exploration by bootstrapped prediction. *Advances in neural information processing systems*, 35:31855–31870, 2022.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Torkel Hafting, Marianne Fyhn, Sturla Molden, May-Britt Moser, and Edvard I. Moser. Microstructure of a spatial map in the entorhinal cortex. *Nature*, 436 (7052):801–806, August 2005. ISSN 1476-4687. doi: 10.1038/nature03721. URL https://doi.org/10.1038/nature03721.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Martin Hitier, Stephane Besnard, and Paul F Smith. Vestibular pathways involved in cognition. *Frontiers in Integrative Neuroscience*, 8:59, 2014. doi: 10.3389/fnint.2014.00059. URL https://doi.org/10.3389/fnint.2014.00059.

Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. GAIA-1: A Generative World Model for Autonomous Driving, September 2023. URL http://arxiv.org/abs/2309.17080. arXiv:2309.17080 [cs].

Matthew Johnson, Katja Hofmann, Tim Hutton, and David Bignell. The malmo platform for artificial intelligence experimentation. In *Ijcai*, volume 16, pages 4246–4247, 2016.

Xiaoliang Luo, Robert M. Mok, and Bradley C. Love. The inevitability and superfluousness of cell types in spatial cognition, January 2024. URL https://www.biorxiv.org/content/10.1101/2024.01.10.575026v1. Pages: 2024.01.10.575026 Section: New Results.

Bruce L McNaughton, Francesco P Battaglia, Ole Jensen, Edvard I Moser, and May-Britt Moser. Path integration and the neural basis of the'cognitive map'. *Nature Reviews Neuroscience*, 7(8):663–678, 2006.

Ida Momennejad, Hosein Hasanbeig, Felipe Vieira, Hiteshi Sharma, Robert Osazuwa Ness, Nebojsa Jojic, Hamid Palangi, and Jonathan Larson. Evaluating Cognitive Maps and Planning in Large Language Models with CogEval, September 2023. URL https://arxiv.org/abs/2309.15129v1.

J. O'Keefe and J. Dostrovsky. The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Research*, 34(1): 171–175, November 1971. doi: 10.1016/0006-8993(71)90358-1.

Diego B. Piza, Benjamin W. Corrigan, Roberto A. Gulli, Sonia Do Carmo, A. Claudio Cuello, Lyle Muller, and Julio Martinez-Trujillo. Primacy of vision shapes behavioral strategies and neural substrates of spatial navigation in marmoset hippocampus. *Nature Communications*, 15(1):4053, 5 2024. doi: 10.1038/s41467-024-48374-2.

Henri Poincaré. *The Foundations of Science: Science and Hypothesis, the Value of Science, Science and Method*. Cambridge University Press, Cambridge, 2015.

Alison R. Preston and Howard Eichenbaum. Interplay of Hippocampus and Prefrontal Cortex in Memory. *Current Biology*, 23(17):R764–R773, September 2013. ISSN 09609822. doi: 10.1016/j.cub.2013.05.041. URL https://linkinghub.elsevier.com/retrieve/pii/S0960982213006362.

Rajkumar Vasudeva Raju, J Swaroop Guntupalli, Guangyao Zhou, Carter Wendelken, Miguel Lázaro-Gredilla, and Dileep George. Space is a latent sequence: A theory of the hippocampus. *Science Advances*, 10(31):eadm8470, 2024.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, et al. Hopfield networks is all you need. *arXiv preprint arXiv:2008.02217*, 2020.

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

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

Aman B. Saleem, E. Mika Diamanti, Julien Fournier, Kenneth D. Harris, and Matteo Carandini. Coherent encoding of subjective spatial position in visual cortex and hippocampus. *Nature*, 562(7725):124–127, October 2018. ISSN 1476-4687. doi: 10.1038/s41586-018-0516-1. URL `https://www.nature.com/articles/s41586-018-0516-1`. Publisher: Nature Publishing Group.

Alexei Samsonovich and Bruce L. McNaughton. Path Integration and Cognitive Mapping in a Continuous Attractor Neural Network Model. *Journal of Neuroscience*, 17(15):5900–5920, August 1997. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.17-15-05900.1997. URL `https://www.jneurosci.org/content/17/15/5900`. Publisher: Society for Neuroscience Section: Articles.

Karen Simonyan. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Leslie N Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, volume 11006, pages 369–386. SPIE, 2019.

Robert W. Stackman, Ann S. Clark, and Jeffrey S. Taube. Hippocampal Spatial Representations Require Vestibular Input. *Hippocampus*, 12(3):291–303, 2002. ISSN 1050-9631. doi: 10.1002/hipo.1112. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1823522/`.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147. PMLR, 2013.

Sander Tanni, William De Cothi, and Caswell Barry. State transitions in the statistically stable place cell population correspond to rate of perceptual change. *Current Biology*, 32(16):3505–3514, 2022.

Jeffrey S Taube, Robert U Muller, and James B Ranck. Head-direction cells recorded from the postsubiculum in freely moving rats. i. description and quantitative analysis. *Journal of Neuroscience*, 10(2):420–435, 1990.

Alessandro Treves, Orazio Miglino, and Domenico Parisi. Rats, nets, maps, and the emergence of place cells. *Psychobiology*, 20(1):1–8, 1992.

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, March 2022. URL `https://www.biorxiv.org/content/10.1101/2020.11.11.378141v2`. Pages: 2020.11.11.378141 Section: New Results.

Keyon Vafa, Justin Y Chen, Jon Kleinberg, Sendhil Mullainathan, and Ashesh Rambachan. Evaluating the world model implicit in a generative model. *arXiv preprint arXiv:2406.03689*, 2024.

Ashish Vaswani. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

Margaret C von Ebers and Xue-Xin Wei. Cognitive maps from predictive vision. *Nature Machine Intelligence*, pages 1–2, 2024.

James C. R. Whittington, David McCaffary, Jacob J. W. Bakermans, and Timothy E. J. Behrens. How to build a cognitive map. *Nature Neuroscience*, 25(10):1257–1272, October 2022a. ISSN 1097-6256, 1546-1726. doi: 10.1038/s41593-022-01153-y. URL `https://www.nature.com/articles/s41593-022-01153-y`.

James C R Whittington, Joseph Warren, and Timothy E J Behrens. Relating Transformers to Models and Neural Representations of the Hippocampal Formation. 2022b.

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, November 2020. ISSN 00928674. doi: 10.1016/j.cell.2020.10.024. URL `https://linkinghub.elsevier.com/retrieve/pii/S009286742031388X`.

K. Zhang. Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: a theory. *Journal of Neuroscience*, 16(6):2112–2126, March 1996. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.16-06-02112.1996. URL `https://www.jneurosci.org/content/16/6/2112`. Publisher: Society for Neuroscience Section: Articles.

Vita

Margaret C. von Ebers was born in Des Moines, Iowa on October 17th 1997, the daughter of Paul and Jill von Ebers. She received her Bachelor of Science degree from Texas A&M University (Gig 'Em Ags) in the spring of 2020. She was accepted into the Computer Science MS program at The University of Texas at Austin in 2022.

Address: 2102 East 8th Street, Austin, TX 78702

This thesis was typeset with LaTeX[†] by the author.

[†]LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.