Research proposal - Tom George

Introduction

Intelligent agents must learn to represent the external world and themselves within it, then dynamically use these representations to solve tasks such as navigation, planning and decision making. Although biological and artificial intelligences do this for different reasons (e.g. survival vs. maximisation of an objective function) they often develop similar representations when trained on comparable tasks [1, 2], strongly indicating we can learn about one by studying the other. My research focuses on three fundamental questions; **how do neural systems learn**, **what representations do neural systems use**, and **how do neural dynamics and representations interact to enable flexible behaviour?** Using theoretical and data-driven approaches I study these in the context of the mammalian spatial-memory system where decades of research have revealed a rich landscape of learning phenomena [3–5], architectures [6], representations [7] and dynamics [4,8–10]. By building computational models I show how biology can be reconciled with statistical and machine learning models of artificial learning systems (NeuroAI [11]) providing novel interpretations for both. Along the way I have **developed popular open-source computational tools** to analyse neural data [12] and to build realistic models of spatial navigation [13].

Past work

(Work 1) Tooling for modelling navigation and neural representations in 2D spaces [13]. Studying the brain's role in spatial navigation often necessitates creating synthetic behaviour and/or neural data. After discovering that no single framework existed for this purpose I developed the RatInABox python toolkit [13] which generates realistic synthetic trajectories for agents exploring one- or two-dimensional environments along with associated neural data (place cells, grids cells etc.). I engaged and led a rapidly emerging open-source developer community whose contributions expanded RatInABox from a simple data generator into a comprehensive toolkit for navigational research, including features such as policy control, sensory inputs, multiagent interactions as well as representation and reinforcement learning capabilities. RatInABox jointly models navigation *and* neurons whilst remaining computationally lightweight. It has received over 40,000 downloads to date, and, in its first year, has already facilitated a promising new wave of hippocampal studies (as shown in [12, 14–24]) including my own on predictive representations [25, 26] (Work 2), hippocampal generative modelling [27] (Work 3) and neural data analysis (Aim 1).

(Work 2) Plausible mechanisms for learning predictive representations [25,26]. Predictive representations, specifically *successor features* (SRs) [28,29], enable agents to rapidly learn and transfer environmental structure. Previous studies have suggested that place cells [30] function like SRs for spatial navigation [31], however temporal difference (TD) learning [32] (the default learning rule) is not compatible with hippocampal biological constraints. I sought an alternative mechanism based on a spiking variant of Hebbian learning called STDP [3] and phase precession, a phenomenon which coordinates neural activity such that cells with potentially quite distant place fields spike in close and coordinated temporal proximity. This "compression" of behaviour down to short timescales enables STDP to bind place cells together in the order they are traversed, rapidly learning an SR-like representation for the current policy. This work provided one of the first biologically plausible mechanisms of how predictive maps can be learned [33,34] and bridged a fundamental mismatch between neural plasticity and behavioural timescales.

In a follow-up study [26], I sought a theoretical explanation for this result, deriving an equivalence between my phase-precession led STDP learning mechanism and TD(λ) [32], an extension of the TD learning algorithm which accelerates learning with memory traces.

(Work 3) A generative model of the hippocampal formation [27]. The mammalian spatial-memory system possesses "generative" functionality [35], exemplified by various forms of *mind travel* [36] (the shift of an internal position encoding away from the true location to generate new trajectories) and *path integration* [37, 38] (where self-location predictions can be generated in the absence of sensory input). Existing models which explain how the hippocampal formation path integrates use deep architectures trained with non-local learning rules [1, 2, 39, 40], leaving open the question of how this is learned in the brain. My NeurIPS paper [27] introduced a biologically plausible model of the hippocampal formation, demonstrating that local learning and message passing was sufficient. One important component was that fast theta-band oscillations (5-10 Hz) control the direction of information flow through a two-layer network, akin to a Helmholtz machine [41] trained using a high-frequency wake-sleep algorithm [42]. Local error-minimising learning rules yielded a ring-attractor in the hidden layer (closely matching that of entorhinal grid cells) capable of robust path integration. My work unified oscillations, Hebbian learning and the hippocampal "loop" under a single normative framework.

Ongoing and Future work

(Aim 1) I will develop efficient and scalable methods for neural latent discovery, enabling more accurate interpretation of neural tuning curves and dynamics. Latent variable models (LVMs) aim to find low-dimensional factors which explain high-dimensional neural data [43–57]. This contrasts a more traditional, but still widely used, approach where *tuning curves* are plotted to visualise each neurons activity as a function of a particular behavioural variable (e.g. the animal's location) [30, 58–60]. In an ongoing line of work I am developing a novel technique which blends these approaches by recurvsively fitting and optimizing tuning curves through a procedure related to expectation-maximisation [61]. This is fast (10 - 100x faster than comparable methods [55, 62–64]), performant, and conceptually simple but, like other modern LVM approaches, moves beyond the restrictive paradigm of explaining data exclusively in terms of behaviour. Early results are promising; optimized place fields are sharpened by my procedure and biases in their tuning curve shapes (e.g. that place fields close to walls are smaller [65]) are reduced implying space is encoded in a more uniform and precise manner than previously thought. Similar results apply to a motor-task dataset from somatosensory cortex [66]. I expect this tool to have impact in fields where disambiguating between theoretical hypotheses requires highly accurate characterisation of tuning curves and/or latent dynamics. It could also serve as a pre-processing step for neural datasets or for real-time latent decoding.

(Aim 2) I will apply latent variable methods to open neural datasets in order to disentable the effect of synaptic structure and neural dynamics from representations. Latent variable models provide a means to reinterpret neural data in terms of the dynamics (the latent factors) and a mapping from these dynamics to spikes. Thus, there are two ways the brain can perform a task or adapt to new conditions; it can do so *dynamically* through its latent or *structurally* by synaptically adjusting its tuning curves. Consider the SR model of place fields discussed earlier: place fields can show SR-like skewing for structural reasons because their receptive fields are synaptically modified (e.g. see Work 1 [25] and [33, 34]) or for dynamical reasons [67] because the latent variable has a bias for exploring upcoming future locations causing place cells to fire early and thus skew backwards (e.g. due to theta sequences [68]). Disambiguating between these hypotheses is critical if we hope to truly understand how hippocampus supports spatial behaviour and memory. Building on my new latent variable methods (Aim 1) I will analyse new and existing neural dataset to explore which features of neural tuning curves can be attributed to synaptic learning and which to latent dynamics. This technique will be validated against carefully constructed synthetic datasets (Work 3) where the ground truth is known. This project will provide a new perspective on the role of dynamics in neural computation and will be a stepping stone towards a general theory of neural dynamics and structure (Aim 3).

(Aim 3) I will develop a general theory of when dynamics (i.e. "context") and structure ("weights") are used to solve cognitive tasks. The distinction between dynamics and structure is not restricted to spatial cognition. In machine learning, transformers-based architectures [69] are able to solve tasks either "in-weights" or "in-context" [70, 71]. In reinforcement learning, model-based methods find optimal actions at test-time through dynamical rollouts in contrast to model-free methods which learn policies "in weights" [32]. Recently, experimental [72] and theoretical [73] work has demonstrated that hippocampal episodic memory and prefrontal working memory store memories in synapses vs. in neural activity respectively. All in all, I see it as crucial for the field to move towards a unified understanding of why, and when, dynamics vs. structure is used to solve cognitive tasks. My research agenda will seek to answer this question at multiple levels, building upon Aims 1 and 2, while also embracing broader aspects beyond spatial cognition and the brain.

Summary

I have built tools for modelling spatial-memory and used these to study how/what neural systems learn. My works provide biologically plausible explanations of predictive representations and generative models in the hippocampus. My future research will focus on developing methods for neural latent discovery, disentangling latent dynamics from representations and moving towards a general theory of neural dynamics and structure.

References

[1] Andrea Banino, Caswell Barry, Benigno Uria, Charles Blundell, Timothy Lillicrap, Piotr Mirowski, Alexander Pritzel, Martin J Chadwick, Thomas Degris, Joseph Modayil, et al. Vector-based navigation using grid-like representations in artificial agents. *Nature*, 557(7705):429–433, 2018.

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

[3] Guo-qiang Bi and Mu-ming Poo. Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. *Journal of neuroscience*, 18(24):10464–10472, 1998.

[4] Margaret F Carr, Shantanu P Jadhav, and Loren M Frank. Hippocampal replay in the awake state: a potential substrate for memory consolidation and retrieval. *Nature neuroscience*, 14(2):147–153, 2011.

[5] Katie C Bittner, Aaron D Milstein, Christine Grienberger, Sandro Romani, and Jeffrey C Magee. Behavioral time scale synaptic plasticity underlies ca1 place fields. *Science*, 357(6355):1033–1036, 2017.

[6] Daniel Bush, Caswell Barry, and Neil Burgess. What do grid cells contribute to place cell firing? *Trends in neurosciences*, 37(3):136–145, 2014.

[7] Edvard I Moser, May-Britt Moser, and Bruce L McNaughton. Spatial representation in the hippocampal formation: a history. *Nature neuroscience*, 20(11):1448–1464, 2017.

[8] William E Skaggs and Bruce L McNaughton. Replay of neuronal firing sequences in rat hippocampus during sleep following spatial experience. *Science*, 271(5257):1870–1873, 1996.

[9] Laure Buhry, Amir H Azizi, and Sen Cheng. Reactivation, replay, and preplay: how it might all fit together. *Neural plasticity*, 2011(1):203462, 2011.

[10] Abraham Z Vollan, Richard J Gardner, May-Britt Moser, and Edvard I Moser. Left-right-alternating theta sweeps in the entorhinal-hippocampal spatial map. *bioRxiv*, pages 2024–05, 2024.

[11] Anthony Zador, Sean Escola, Blake Richards, Bence Ölveczky, Yoshua Bengio, Kwabena Boahen, Matthew Botvinick, Dmitri Chklovskii, Anne Churchland, Claudia Clopath, et al. Catalyzing next-generation artificial intelligence through neuroai. *Nature communications*, 14(1):1597, 2023.

[12] Tom M George, Pierre Glaser, Kimberly Stachenfeld, Caswell Barry, and Claudia Clopath. Simpl: Scalable and hassle-free optimization of neural representations from behaviour. *bioRxiv*, 2024.

[13] Tom M George, Mehul Rastogi, William de Cothi, Claudia Clopath, Kimberly Stachenfeld, and Caswell Barry. Ratinabox, a toolkit for modelling locomotion and neuronal activity in continuous environments. *Elife*, 13:e85274, 2024.

[14] Daniel Levenstein, Aleksei Efremov, Roy Henha Eyono, Adrien Peyrache, and Blake Richards. Sequential predictive learning is a unifying theory for hippocampal representation and replay. *bioRxiv*, pages 2024–04, 2024.

[15] Guillaume Etter, Suzanne van der Veldt, Coralie-Anne Mosser, Michael E Hasselmo, and Sylvain Williams. Idiothetic representations are modulated by availability of sensory inputs and task demands in the hippocampal-septal circuit. *Cell Reports*, 43(11), 2024.

[16] J Quinn Lee, Alexandra T Keinath, Erica Cianfarano, and Mark P Brandon. Identifying representational structure in ca1 to benchmark theoretical models of cognitive mapping. *bioRxiv*, pages 2023–10, 2023.

[17] Sihao Liu, Augustine N Mavor-Parker, and Caswell Barry. Functional connectome: Approximating brain networks with artificial neural networks. *arXiv preprint arXiv:2211.12935*, 2022.

[18] Hugo Chateau-Laurent. *Computational modeling of the interactions between episodic memory and cognitive control*. PhD thesis, Université de Bordeaux, 2024.

[19] Janis Samuel Keck, Caswell Barry, Christian F Doeller, and Jürgen Jost. Symmetry and generalization in local learning of predictive representations. *bioRxiv*, pages 2024–05, 2024.

[20] Lauren Bennett, William de Cothi, Laurenz Muessig, Fabio Ribeiro Rodrigues, Francesca Cacucci, Tom Wills, Yanjun Sun, Lisa M Giocomo, Colin Lever, Steven Poulter, et al. Unifying subicular function: A predictive map approach. *bioRxiv*, pages 2024–11, 2024.

[21] Guillaume Etter, Suzanne van der Veldt, Coralie-Anne Mosser, Michael E Hasselmo, and Sylvain Williams. A population code for idiothetic representations in the hippocampal-septal circuit. *bioRxiv*, pages 2023–11, 2023.

[22] G William Chapman, Andrew S Alexander, Frances S Chance, and Michael E Hasselmo. Self-supervised mapping and localization by predictive learning. In *2024 International Conference on Neuromorphic Systems (ICONS)*, pages 193–200. IEEE, 2024.

[23] Simen Storesund. Simulating transition cell learning under the transition scale-space model in a spiking neural network. Master's thesis, NTNU, 2024.

[24] G William Chapman, Andrew S Alexander, Frances S Chance, and Michael E Hasselmo. Predictive learning for self-supervised mapping and localization.

[25] Tom M George, William de Cothi, Kimberly L Stachenfeld, and Caswell Barry. Rapid learning of predictive maps with stdp and theta phase precession. *Elife*, 12:e80663, 2023.

[26] Tom M George. Theta sequences as eligibility traces: A biological solution to credit assignment. *arXiv preprint arXiv:2305.08124*, 2023.

[27] Tom M George, Kimberly L Stachenfeld, Caswell Barry, Claudia Clopath, and Tomoki Fukai. A generative model of the hippocampal formation trained with theta driven local learning rules. *Advances in Neural Information Processing Systems*, 36, 2024.

[28] Peter Dayan. Improving generalization for temporal difference learning: The successor representation. *Neural computation*, 5(4):613–624, 1993.

[29] André Barreto, Will Dabney, Rémi Munos, Jonathan J Hunt, Tom Schaul, Hado P van Hasselt, and David Silver. Successor features for transfer in reinforcement learning. *Advances in neural information processing systems*, 30, 2017.

[30] John O'Keefe and Jonathan Dostrovsky. The hippocampus as a spatial map: preliminary evidence from unit activity in the freely-moving rat. *Brain research*, 1971.

[31] Kimberly L Stachenfeld, Matthew M Botvinick, and Samuel J Gershman. The hippocampus as a predictive map. *Nature neuroscience*, 20(11):1643–1653, 2017.

[32] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[33] Ching Fang, Dmitriy Aronov, LF Abbott, and Emily L Mackevicius. Neural learning rules for generating flexible predictions and computing the successor representation. *elife*, 12:e80680, 2023.

[34] Jacopo Bono, Sara Zannone, Victor Pedrosa, and Claudia Clopath. Learning predictive cognitive maps with spiking neurons during behavior and replays. *Elife*, 12:e80671, 2023.

[35] Eleanor Spens and Neil Burgess. A generative model of memory construction and consolidation. *Nature Human Behaviour*, 8(3):526–543, 2024.

[36] Honi Sanders, César Rennó-Costa, Marco Idiart, and John Lisman. Grid cells and place cells: an integrated view of their navigational and memory function. *Trends in neurosciences*, 38(12):763–775, 2015.

[37] Bruce L McNaughton, Carol A Barnes, Jason L Gerrard, Katalin Gothard, Min W Jung, James J Knierim, H Kudrimoti, Y Qin, WE Skaggs, M Suster, et al. Deciphering the hippocampal polyglot: the hippocampus as a path integration system. *Journal of Experimental Biology*, 199(1):173–185, 1996.

[38] Yoram Burak and Ila R Fiete. Accurate path integration in continuous attractor network models of grid cells. *PLoS computational biology*, 5(2):e1000291, 2009.

[39] Christopher J Cueva and Xue-Xin Wei. Emergence of grid-like representations by training recurrent neural networks to perform spatial localization. *arXiv preprint arXiv:1803.07770*, 2018.

[40] Ben Sorscher, Gabriel Mel, Surya Ganguli, and Samuel Ocko. A unified theory for the origin of grid cells through the lens of pattern formation. *Advances in neural information processing systems*, 32, 2019.

[41] Peter Dayan, Geoffrey E Hinton, Radford M Neal, and Richard S Zemel. The helmholtz machine. *Neural computation*, 7(5):889–904, 1995.

[42] Geoffrey E Hinton, Peter Dayan, Brendan J Frey, and Radford M Neal. The" wake-sleep" algorithm for unsupervised neural networks. *Science*, 268(5214):1158–1161, 1995.

[43] Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 61(3):611–622, 1999.

[44] Byron M Yu, Krishna V Shenoy, and Maneesh Sahani. Expectation propagation for inference in non-linear dynamical models with poisson observations. In *2006 IEEE Nonlinear Statistical Signal Processing Workshop*. IEEE, 2006.

[45] Byron M Yu, John P Cunningham, Krishna V Shenoy, and Maneesh Sahani. Neural decoding of movements: From linear to nonlinear trajectory models. In *Neural Information Processing: 14th International Conference, ICONIP 2007, Kitakyushu, Japan, November 13-16, 2007, Revised Selected Papers, Part I 14*. Springer, 2008.

[46] Byron M Yu, John P Cunningham, Gopal Santhanam, Stephen Ryu, Krishna V Shenoy, and Maneesh Sahani. Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. *Advances in Neural Information Processing Systems*, 2008.

[47] Jakob H Macke, Lars Buesing, John P Cunningham, Byron M Yu, Krishna V Shenoy, and Maneesh Sahani. Empirical models of spiking in neural populations. *Advances in Neural Information Processing Systems*, 2011.

[48] Andrew Zammit Mangion, Ke Yuan, Visakan Kadirkamanathan, Mahesan Niranjan, and Guido Sanguinetti. Online variational inference for state-space models with point-process observations. *Neural Computation*, 2011.

[49] Mijung Park, Gergo Bohner, and Jakob H Macke. Unlocking neural population non-stationarities using hierarchical dynamics models. *Advances in Neural Information Processing Systems*, 2015.

[50] Yuanjun Gao, Evan W Archer, Liam Paninski, and John P Cunningham. Linear dynamical neural population models through nonlinear embeddings. *Advances in Neural Information Processing Systems*, 2016.

[51] Chethan Pandarinath, Daniel J O'Shea, Jasmine Collins, Rafal Jozefowicz, Sergey D Stavisky, Jonathan C Kao, Eric M Trautmann, Matthew T Kaufman, Stephen I Ryu, Leigh R Hochberg, et al. Inferring single-trial neural population dynamics using sequential auto-encoders. *Nature methods*, 2018.

[52] Daniel Hernandez, Antonio Khalil Moretti, Ziqiang Wei, Shreya Saxena, John Cunningham, and Liam Paninski. Nonlinear evolution via spatially-dependent linear dynamics for electrophysiology and calcium data. *arXiv preprint arXiv:1811.02459*, 2018.

[53] Lea Duncker, Gergo Bohner, Julien Boussard, and Maneesh Sahani. Learning interpretable continuous-time models of latent stochastic dynamical systems. In *International conference on machine learning*. PMLR, 2019.

[54] Zhe Dong, Bryan Seybold, Kevin Murphy, and Hung Bui. Collapsed amortized variational inference for switching nonlinear dynamical systems. In *International Conference on Machine Learning*, pages 2638–2647. PMLR, 2020.

[55] Ding Zhou and Xue-Xin Wei. Learning identifiable and interpretable latent models of high-dimensional neural activity using pi-vae. *Advances in Neural Information Processing Systems*, 2020.

[56] Rabia Gondur, Usama Bin Sikandar, Evan Schaffer, Mikio Christian Aoi, and Stephen L Keeley. Multi-modal gaussian process variational autoencoders for neural and behavioral data. *arXiv preprint arXiv:2310.03111*, 2023.

[57] Amber Hu, David Zoltowski, Aditya Nair, David Anderson, Lea Duncker, and Scott Linderman. Modeling latent neural dynamics with gaussian process switching linear dynamical systems. *arXiv preprint arXiv:2408.03330*, 2024.

[58] 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, 2005.

[59] Apostolos P Georgopoulos, Andrew B Schwartz, and Ronald E Kettner. Neuronal population coding of movement direction. *Science*, 233(4771):1416–1419, 1986.

[60] David H Hubel and Torsten N Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1):106, 1962.

[61] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22, 1977.

[62] Steffen Schneider, Jin Hwa Lee, and Mackenzie Weygandt Mathis. Learnable latent embeddings for joint behavioural and neural analysis. *Nature*, 617(7960):360–368, 2023.

[63] Neil Lawrence. Gaussian process latent variable models for visualisation of high dimensional data. *Advances in neural information processing systems*, 16, 2003.

[64] Jack Wang, Aaron Hertzmann, and David J Fleet. Gaussian process dynamical models. *Advances in neural information processing systems*, 18, 2005.

[65] 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*, 2022.

[66] Raeed H Chowdhury, Joshua I Glaser, and Lee E Miller. Area 2 of primary somatosensory cortex encodes kinematics of the whole arm. *eLife*, page e48198, 2020.

[67] Eloy Parra-Barrero and Sen Cheng. Learning to predict future locations with internally generated theta sequences. *PLOS Computational Biology*, 19(5):e1011101, 2023.

[68] David J Foster and Matthew A Wilson. Hippocampal theta sequences. *Hippocampus*, 17(11):1093–1099, 2007.

[69] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[70] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[71] Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pages 35151–35174. PMLR, 2023.

[72] Mohamady El-Gaby, Adam Loyd Harris, James CR Whittington, William Dorrell, Arya Bhomick, Mark E Walton, Thomas Akam, and Timothy EJ Behrens. A cellular basis for mapping behavioural structure. *Nature*, pages 1–10, 2024.

[73] James CR Whittington, William Dorrell, Timothy EJ Behrens, Surya Ganguli, and Mohamady El-Gaby. A tale of two algorithms: Structured slots explain prefrontal sequence memory and are unified with hippocampal cognitive maps. *Neuron*, 2024.