Learning Causally Emergent Representations

Christos Kaplanis
Google DeepMind
kaplanis@google.com

Fernando E. Rosas*
Department of Informatics, University of Sussex
Centre for Psychedelic Research, Department of Brain Science, Imperial College London
Centre for Complexity Science, Imperial College London, London
Centre for Eudaimonia and Human Flourishing, University of Oxford
f.rosas@sussex.ac.uk

Pedro A.M. Mediano*
Department of Computing, Imperial College London
Department of Psychology, University of Cambridge
Division of Psychology and Language Sciences, University College London
p.mediano@imperial.ac.uk

Abstract

Cognitive processes usually take place at a macroscopic scale in systems characterised by emergent properties, which make the whole 'more than the sum of its parts.' While recent proposals have provided quantitative, information-theoretic metrics to detect emergence in time series data, it is often highly non-trivial to identify the relevant macroscopic variables *a priori*. In this paper we leverage recent advances in representation learning and differentiable information estimators to put forward a data-driven method to find emergent variables. The proposed method successfully detects emergent variables and recovers the ground-truth emergence values in a synthetic dataset. This proof-of-concept paves the ground for future analyses uncovering the emergent structure of cognitive representations in biological and artificial intelligence systems.

1 Introduction

Cognitive processes usually take place in systems made of multiple interacting parts, e.g. neurons composing the nervous system of an organism. Importantly, cognitive processes themselves don't seem to take place at a 'microscopic' level of individual units, but at 'macroscopic' levels involving assemblies of several coordinated units [7]. Hence, when trying to unveil the inner workings of a — natural or artificial — cognitive system, it is crucial to be able to identify relevant macroscopic variables that best characterise the corresponding cognitive processes.

The identification of macroscopic variables has traditionally been driven by intuition and expert knowledge. For example, the investigation of collective behaviour in statistical physics is based on macroscopic variables known as 'order parameters,' which are typically identified heuristically and then used to describe phase transitions and other phenomena of interest [18]. Unfortunately, identifying relevant macroscopic variables is often more an art than a science, being heavily dependent

*F.R. and P.M. are joint senior authors.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

on prior knowledge and expectations. Having automated procedures to identify relevant macroscopic variables of cognitive systems would open important avenues for investigating the inner workings of different cognitive architectures.

A promising approach to identify empirically useful macroscopic variables is provided by unsupervised representation learning [13, 4, 20]. For example, information maximisation has proven to be a powerful objective for learning representations within neural networks [5, 13]. In this paper we combine this approach with recent breakthroughs in our ability to formally characterise emergent phenomena [17, 11], which have proven to be not only theoretically sound but also empirically powerful [9, 15]. Building on this literature, in this paper we investigate the feasibility of leveraging recently proposed metrics of emergence to identify representations that display emergent properties. Our results show that causal emergence facilitates learning of more complex features of the data relative to pure mutual information maximisation.

2 Methods

2.1 Quantifying emergence

Consider a system composed of n parts, and let X_t^i denote the state of part i at time t. The information that the joint process carries from t to t' can be quantified by the mutual information $I(\boldsymbol{X}_t; \boldsymbol{X}_{t'})$, where $\boldsymbol{X}_t = (X_t^1, \dots, X_t^n)$. How can one characterise an emergent macroscopic variable of such system? Following Ref. [17], one can define *causally emergent* variables V_t as satisfying two key criteria:

(i) **Supervenience**: there exists a function (or *coarse-graining*) f such that $V_t = f(\boldsymbol{X}_t)$.

(ii) **Unique information**: V_t holds unique information about the future evolution of the system $\boldsymbol{X}_{t'}$ that cannot be not found in the individual $X_t^1, \dots, X_t^n$ by themselves.

Critically, the unique information of V_t about $\boldsymbol{X}_{t'}$ can be rigorously quantified using the framework of *Partial Information Decomposition* (PID, [21]), and its recent extension to time series data (ΦID, [10]). Emergence, therefore, is defined as the capability of a supervenient variable to provide predictive power that cannot be reduced to underlying microscale phenomena.

Quantifying unique information in high-dimensional systems can be highly non-trivial. Luckily, the ΦID formalism allows to derive simpler measures that provide sufficient criteria for emergence. In particular, it has been shown that the following is a sufficient condition for causal emergence [17]:

$$\Psi := I(V_t; V_{t+1}) - \sum_i I(X_t^i; V_{t+1}) > 0 \,. \tag{1}$$

Importantly, Ψ is comparatively easy to calculate, as it relies only on pairwise marginal distributions and on Shannon's mutual information. These key features allow the framework to be applicable on a wide range of scenarios, as illustrated by the applications reviewed in Ref. [11]. Note that here we take $t' = t + 1$, but in principle any $t' > t$ is valid.

2.2 Model architecture and information estimators

Our aim is to establish an automated procedure to identify emergent macroscopic variables V_t with respect to a microscopic substrate $\boldsymbol{X}_t$. For this, we investigate parametric coarse-grainings $V_t = f_\theta(\boldsymbol{X}_t)$ that can be optimised to maximise Ψ via a differentiable objective function.

A key ingredient to maximising Ψ is employing a suitable estimator of Shannon's mutual information. Although many popular estimators are not differentiable [6, 8], the literature offers a number of differentiable estimators [14, 19]. We use the Smoothed Mutual Information "Lower-bound" Estimator (SMILE) [19], which is one of a family of approaches that formulates mutual information estimation as a variational problem, and was specifically designed to address the issue of high variance in existing estimators such as the NWJ lower bound [12] and MINE [1]. The SMILE mutual information

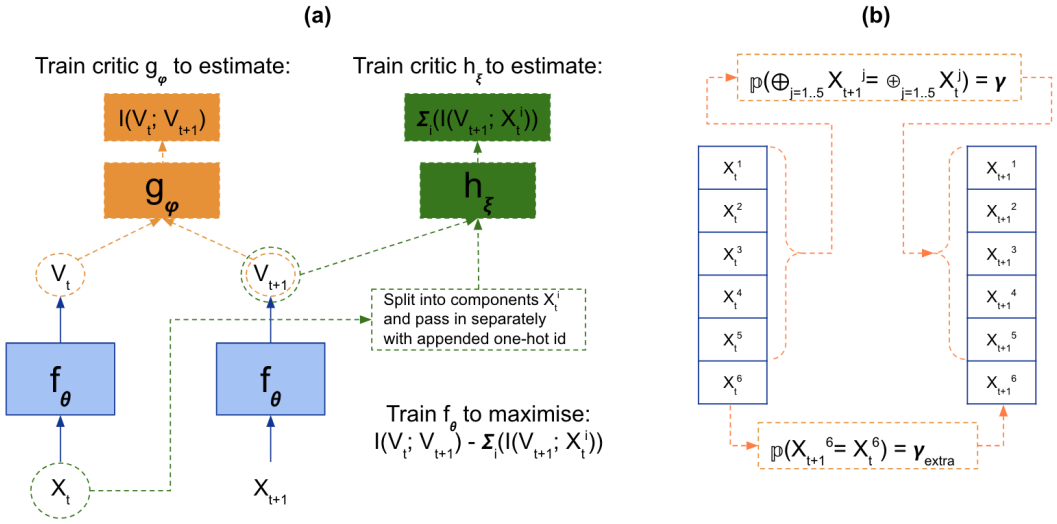

Figure 1: **Model architecture and data-generating process**. **a)** Architecture for learning causally emergent representations. **b)** The sequential bit-string data used for training, featuring auto-correlation of the parity of the first 5 bits and auto-correlation of the value of the 6$^{\text{th}}$ bit of the string.

estimator is given by

$$I(X;Y) = \mathbb{E}_{p(x,y)}\left[\log \frac{p(x,y)}{p(x)p(y)}\right]$$
$$\geq \mathbb{E}_{p(x,y)}\left[g_\varphi(x,y)\right] - \log \mathbb{E}_{p(x)p(y)}\left[\text{clip}\left(e^{g_\varphi(x,y)}, e^{-\tau}, e^{\tau}\right)\right] \triangleq I^{\text{S}}_\varphi(X;Y)\,, \quad (2)$$

where g_φ is a parameterised function that estimates the log density ratio $\log\left(p(x,y)/(p(x)p(y))\right)$, $\text{clip}(v,l,u) = \max\left(\min\left(v,u\right),l\right)$, and $\tau \geq 0$ is a hyperparameter. As $\tau \to \infty$, I^{S} converges to the MINE estimation [1], but a finite τ prevents the potentially exponential growth of the variance of the estimate with MI (which MINE suffers from [19]).

Equipped with this estimator, we can now formulate our representation learning algorithm for causally emergent features. The architecture is schematically shown in Fig. 1a. Our model involves three learnable functions:

1. A representation network f_θ, that learns a supervenient variable $V_t = f_\theta(\boldsymbol{X}_t)$.
2. A critic for the macroscopic variable g_φ, that controls the estimation of $I(V_t; V_{t+1})$.
3. A critic for the microscopic variable h_ξ, that controls the estimation of $I(X_t^i; V_{t+1})$.

During training, φ and ξ are trained to estimate their respective mutual information quantities, while the representation parameters θ are trained to maximise the SMILE approximation to Ψ given by

$$\Psi^{\text{S}}(\theta,\varphi,\xi) := I^{\text{S}}_\varphi(f_\theta(\boldsymbol{X}_t); f_\theta(\boldsymbol{X}_{t+1})) - \sum_i I^{\text{S}}_\xi(X_t^i; f_\theta(\boldsymbol{X}_{t+1}))\,. \quad (3)$$

We refer to Ψ^{S} as the *emergence objective function*, and the first term in the RHS of Eq. (3) as the *predictive mutual information* [2] — since (by the data processing inequality [3]) it represents a lower bound on the joint mutual information between the past and future states of the whole system, $I(\boldsymbol{X}_t; \boldsymbol{X}_{t'})$. As a control condition, we also ran experiments with an objective function consisting of only the predictive information, removing the marginal mutual information terms.

2.3 Synthetic dataset

We evaluate our method for learning causally emergent representations by applying it to sequences of random bit-strings of length n with two constructed temporal correlations:

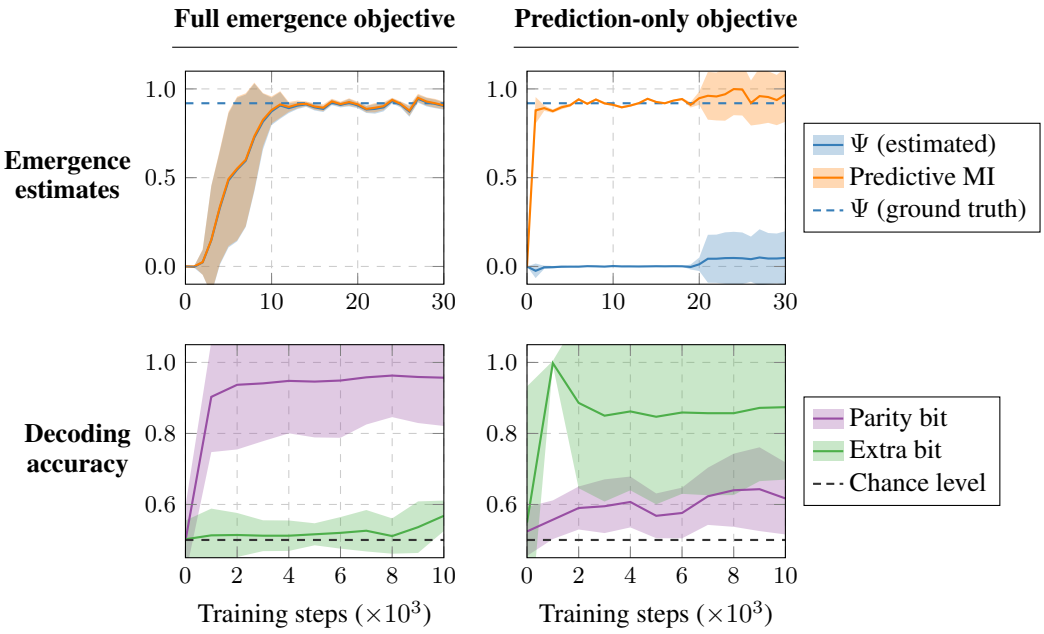

Figure 2: **The model recovers ground truth emergent features**. Using the emergence objective function (left column), the model finds the correct Ψ value and is able to recover the known emergent feature (parity bit). Using only predictive MI as the objective (right column), the model fails to discover any emergent features.

1. The *parity* of the first $n-1$ bits is auto-correlated across time, such that
$$\mathbb{P}\{\oplus_{i=1}^{n-1} X_{t+1}^i = \oplus_{i=1}^{n-1} X_t^i\} = \gamma_{\text{parity}} > \frac{1}{2}\,,$$
where $\oplus$ represents modulo-2 addition.
2. The last (or *extra*) bit in the bit-string X^n is auto-correlated across time, such that
$$\mathbb{P}\{X_{t+1}^n = X_t^n\} = \gamma_{\text{extra}} > \frac{1}{2}\,.$$

Since parity is a synergistic function of the bits of a bit-string (i.e. it cannot be predicted from each of the input bits individually [16]), and since the parity predicts some information about the future evolution of the system, $V_t = \oplus_{i=1}^{n-1} X_t^i$ is an emergent feature of the system.

Despite its simplicity, this dataset has two key advantages: there is a known emergent feature (the parity), and one can calculate the mutual information and the emergence capacity analytically.[2] These properties will allow us to verify that the model has successfully extracted the expected emergent properties and that mutual information is being accurately estimated. A schematic of the data-generating process is shown in Fig. 1b.

3 Results

Results show that our proposed architecture can accurately estimate the ground-truth value of Ψ in the synthetic dataset, confirming it is able to learn causally emergent representations (Fig. 2). To interpret the contents of the learned representation, we trained decoders with standard supervised learning to predict both the parity of the first $n-1$ bits (parity bit) and the last auto-correlated bit (extra bit). We found that the parity bit could be decoded with high accuracy but the extra bit could not, confirming that the learned representation indeed corresponded to an emergent feature.

As expected, when the marginal MI terms are removed from the objective function (Fig. 2, right column), the model is no longer able to obtain the correct Ψ value — and, interestingly, only the

[2] Specifically, $\Psi = 1 - H_2(\gamma_{\text{parity}})$ and $I(\boldsymbol{X}_t; \boldsymbol{X}_{t+1}) = 2 - H_2(\gamma_{\text{parity}}) - H_2(\gamma_{\text{extra}})$, where $H_2(p)$ represents the entropy of a Bernoulli distribution with parameter p. Here we set $\gamma_{\text{parity}} = \gamma_{\text{extra}} = 0.99$.

extra bit (but not the parity bit) is encoded in the representation. We hypothesise that, in the absence of the regularisation induced by the marginal MI, the system's inductive biases lead it towards learning "low-order" (i.e. non-emergent) representations. Note that, despite having a constraint removed, the model without marginal MI loss is unable to extract the full predictive information of the system (which equals approximately $1.84\,\mathrm{bit}$), showing that using the full emergence loss could incentivise the system to learn features that provide information about the system's dynamics that would otherwise be ignored. We obtain qualitatively similar results with a noisier version of the same data generating process (Supp. Fig. 3).

Moreover, we have also observed (Supp. Fig. 4) runs where f_θ learns a new, unexpected emergent feature that encodes the combination of the parity and extra bit, showing the capability of the model to discover emergent features that were not originally designed.

4 Conclusion

In this paper, we proposed a machine learning method for discovering emergent variables in time series data that leverages a recent information-theoretic characterisation of emergence [17], as well as advances in mutual information estimation from data with neural networks [19]. Our results provide a proof-of-concept for the method's viability by applying it to time series of auto-correlated bit strings, showing that it can be used to successfully learn the parity of a subset of bits — a feature that is known to have emergent character over the bits. Interestingly, a pure information maximisation objective struggled to learn this feature, suggesting that our method facilitates the identification of complex features of the data. In future work, it would be interesting to see if our method is competitive with recent representation learning methods on standard benchmarks [13, 4, 20].

References

[1] M. I. Belghazi, A. Baratin, S. Rajeswar, S. Ozair, Y. Bengio, A. Courville, and R. D. Hjelm. MINE: Mutual information neural estimation. *arXiv:1801.04062*, 2018.

[2] W. Bialek, I. Nemenman, and N. Tishby. Predictability, complexity, and learning. *Neural computation*, 13(11):2409–2463, 2001.

[3] T. M. Cover and J. Thomas. *Elements of Information Theory*. John Wiley & Sons, 1999.

[4] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, et al. Bootstrap your own latent – A new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020.

[5] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*, 2018.

[6] A. Kraskov, H. Stögbauer, and P. Grassberger. Estimating mutual information. *Physical Review E*, 69(6):066138, 2004.

[7] M. Levin. The computational boundary of a "self": Developmental bioelectricity drives multicellularity and scale-free cognition. *Frontiers in Psychology*, 10:2688, 2019.

[8] J. T. Lizier. JIDT: An information-theoretic toolkit for studying the dynamics of complex systems. *Frontiers in Robotics and AI*, 1:11, 2014.

[9] A. I. Luppi, P. A. Mediano, F. E. Rosas, J. Allanson, J. D. Pickard, G. B. Williams, M. M. Craig, P. Finoia, A. R. Peattie, P. Coppola, et al. Reduced emergent character of neural dynamics in patients with a disrupted connectome. *Neuroimage*, 269:119926, 2023.

[10] P. A. Mediano, F. E. Rosas, A. I. Luppi, R. L. Carhart-Harris, D. Bor, A. K. Seth, and A. B. Barrett. Towards an extended taxonomy of information dynamics via Integrated Information Decomposition. *arXiv:2109.13186*, 2021.

[11] P. A. Mediano, F. E. Rosas, A. I. Luppi, H. J. Jensen, A. K. Seth, A. B. Barrett, R. L. Carhart-Harris, and D. Bor. Greater than the parts: A review of the information decomposition approach to causal emergence. *Philosophical Transactions of the Royal Society A*, 380(2227), 2022.

[12] X. Nguyen, M. J. Wainwright, and M. I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

[13] A. v. d. Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv:1807.03748*, 2018.

[14] B. Poole, S. Ozair, A. van den Oord, A. A. Alemi, and G. Tucker. On variational lower bounds of mutual information. In *NeurIPS Workshop on Bayesian Deep Learning*, 2018.

[15] A. M. Proca, F. E. Rosas, A. I. Luppi, D. Bor, M. Crosby, and P. A. Mediano. Synergistic information supports modality integration and flexible learning in neural networks solving multiple tasks. *arXiv:2210.02996*, 2022.

[16] F. Rosas, P. Mediano, M. Gastpar, and H. Jensen. Quantifying high-order interdependencies via multivariate extensions of the mutual information. *Physical Review E*, 100(3), 2019.

[17] F. E. Rosas, P. A. Mediano, H. J. Jensen, A. K. Seth, A. B. Barrett, R. L. Carhart-Harris, and D. Bor. Reconciling emergences: An information-theoretic approach to identify causal emergence in multivariate data. *PLoS Computational Biology*, 16(12):e1008289, 2020.

[18] R. Solé. *Phase Transitions*, volume 3. Princeton University Press, 2011.

[19] J. Song and S. Ermon. Understanding the limitations of variational mutual information estimators. In *International Conference on Learning Representations*, 2019.

[20] N. Tomasev, I. Bica, B. McWilliams, L. Buesing, R. Pascanu, C. Blundell, and J. Mitrovic. Pushing the limits of self-supervised ResNets: Can we outperform supervised learning without labels on ImageNet? *arXiv:2201.05119*, 2022.

[21] P. L. Williams and R. D. Beer. Nonnegative decomposition of multivariate information. *arXiv:1004.2515*, 2010.

A Hyperparameters

Table 1: Hyperparameters for causal emergence representation learning.

Hyperparameter	Value
Number of bits in X_t	5 + 1
Number of training steps	30000
Parity autocorrelation γ_{parity}	0.99
Extra bit autocorrelation γ_{extra}	0.99
Batch size	100
Representation network layer sizes	[64, 16, 1]
Critic networks layer sizes	[128, 64, 8]
Optimiser	Adam
Representation network learning rate	1e-5
Critic networks learning rate	1e-4

Table 2: Hyperparameters for supervised learning of parity or extra bit using frozen representation after every 1000 steps of causal emergence training.

Hyperparameter	Value
Network layer sizes	[128, 32, 1]
Learning rate	1e-4
L2 regularisation coefficient	1e-4
Batch size	100
Number of training steps	10000

B Supplementary Figures

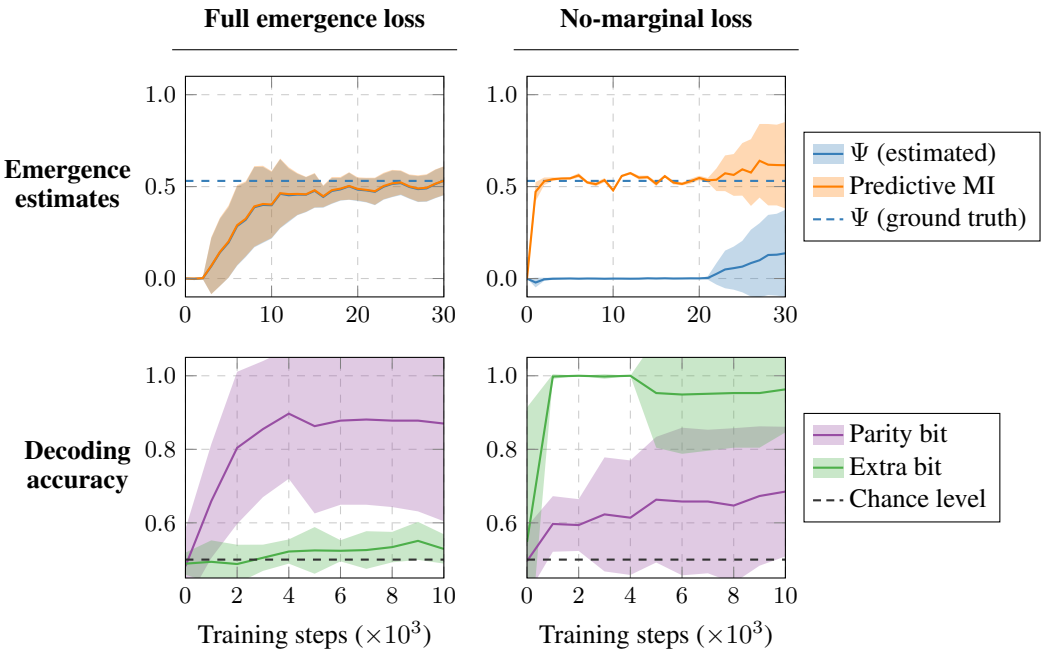

Figure 3: **Replication of main results with noisier data**. Same as in Fig. 2, but with $\gamma_{\text{parity}} = \gamma_{\text{extra}} = 0.9$.

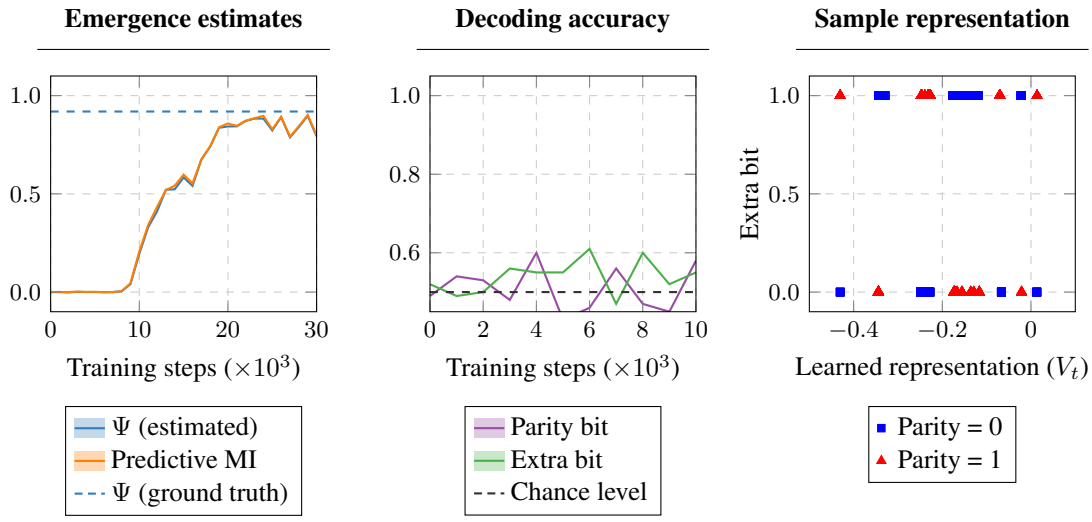

Figure 4: **Sample run where the system discovers an unexpected emergent feature**. This phenomenon, observed in approximately 5% of runs, the system estimates the correct Ψ value (left), but neither the parity or the extra bit can be decoded from the learnt representation (middle). Visual inspection of the representation reveals that the system learnt a synergistic combination of the parity and extra bit, which is itself also emergent — despite not being explicitly designed into the synthetic dataset.