Decoupled Predictive Coding Networks in Backpropagation

Mark A. Griffiths

December 24, 2025

Abstract

Backpropagation underpins modern deep learning but relies on point-estimate optimisation, providing no principled representation of epistemic uncertainty and limiting robustness and interpretability, while remaining vulnerable to distribution shift. Bayesian neural networks address some of these limitations but are difficult to scale, while predictive coding offers biologically grounded inference learning with explicit uncertainty at the cost of expensive iterative inference.

This PhD proposes a hybrid Bayesian learning framework that integrates predictive coding and backpropagation via a decoupled neural interface. A predictive coding network performs variational inference and generates oracle gradients corrected by a path–integral over inference dynamics, which are then amortised into a conventional backpropagation network. This yields an explicit variational free energy defined over backpropagation-trained models, enabling principled uncertainty estimation, causal structure learning, and free-energy–based pruning while preserving computational efficiency.

The project will develop the theoretical foundations of this framework, implement and evaluate it on benchmark datasets, and explore extensions to neuromorphic hardware and active inference.

Introduction and Background

Backpropagation (BP) is the dominant optimisation algorithm in contemporary machine learning, underpinning deep neural networks across science and industry [1, 2]. Its efficiency and compatibility with GPU hardware have enabled large-scale models, including modern large language models [3, 4, 5]. Variants of BP are now standard in applications ranging from Earth system science [6] to physics [7] and protein structure prediction [8].

Despite these successes, BP exhibits fundamental limitations. Its reliance on point-estimate optimisation provides no explicit representation of epistemic uncertainty, leading

to miscalibration, overconfidence, and degraded performance under distribution shift [9, 10]. These issues are well documented in out-of-distribution detection [11], continual learning [12, 13], and robustness to noise [10]. In safety-critical and scientific decision-making contexts, such failures can lead to systematically poor decisions when models encounter novel or ambiguous data [14, 15].

Beyond engineering concerns, BP is also widely regarded as biologically implausible. The algorithm requires precise back-transport of error derivatives through synaptic connections, for which no known neural mechanism exists [16, 17]. This disconnect between dominant machine-learning practice and biological learning motivates the investigation of alternative learning paradigms that are both uncertainty-aware and mechanistically grounded.

Bayesian learning addresses uncertainty by placing distributions over parameters [18, 19]. However, exact Bayesian neural networks (BNNs) are computationally intractable for deep models, while approximate methods often suffer from poor calibration, limited expressiveness, or unclear probabilistic semantics [20]. Predictive coding (PC), derived from the Free Energy Principle (FEP), offers a biologically inspired alternative in which learning proceeds via minimisation of variational free energy [34]. PC networks naturally encode uncertainty and admit causal interpretations, but rely on iterative inference dynamics that limit scalability.

This proposal argues that these trade-offs are not fundamental. Instead, it proposes that the inferential and uncertainty-aware advantages of predictive coding can be decoupled from its computational costs and transferred to BP networks via synthetic gradients. In doing so, it aims to preserve the scalability of BP while endowing models with explicit variational free energy, principled uncertainty estimation, and causal structure learning capabilities.

Research Scope and Questions

The central research question of this PhD is:

Can the inferential and uncertainty-modelling advantages of predictive coding be amortised into BP-trained networks in a principled and scalable manner?

This question is operationalised through the following objectives:

- To derive exact oracle gradients for post-inference variational free energy that account for inference-trajectory dependence.
- To design a decoupled learning architecture that transfers these gradients into standard BP training.
- To evaluate whether the resulting framework improves robustness, uncertainty calibration, and causal interpretability relative to BP and BNN baselines.

Bayesian Neural Networks

BNNs replace point estimates with distributions over parameters, enabling principled modelling of epistemic uncertainty [18, 19]. Exact inference relies on Markov Chain Monte Carlo methods, particularly Hamiltonian Monte Carlo and the No-U-Turn Sampler [21, 22, 23]. While theoretically attractive, these methods are computationally prohibitive for deep networks and further challenged by highly multimodal posterior landscapes [24].

Approximate Bayesian methods address scalability concerns. Monte Carlo dropout and deep ensembles estimate uncertainty through stochastic regularisation or repeated optimisation [25, 26]. Although empirically effective, these approaches do not correspond to principled posterior sampling and can yield miscalibrated uncertainty estimates [10, 27]. Variational Bayesian methods optimise tractable posterior approximations [28, 29], with extensions using normalising flows to improve expressiveness [30, 31]. Nonetheless, they remain constrained by the chosen variational family [20].

Predictive Coding and Inference Learning

Predictive coding provides a Bayesian framework in which learning and inference are unified through minimisation of variational free energy [32, 33]. PC networks introduce explicit error units encoding mismatches between predictions and observations, yielding local update rules derived from free-energy gradients. Under Gaussian assumptions, free energy reduces to a precision-weighted sum of squared prediction errors [33].

During supervised learning, PC networks perform iterative inference until equilibrium, after which parameters are updated [39]. At equilibrium, predictive coding recovers backpropagation gradients on arbitrary computation graphs [41]. This process, termed Inference Learning (IL), is associated with robustness to small datasets, reduced parameter interference, and improved stability under distribution shift [34, 38]. However, the computational cost of iterative inference limits the direct applicability of PC to large-scale machine learning [40].

Decoupled Predictive Coding Networks

Architecture

This project proposes a hybrid learning framework that combines predictive coding (PC) and backpropagation (BP) through a decoupled neural interface [42]. A predictive coding network performs iterative inference over latent states and computes gradients derived from variational free energy (FE), while a conventional BP network uses these gradients, adaptively blended with its own task-loss gradients, for optimisation. Parameters are

shared between the two systems, but inference and optimisation are functionally separated.

A key methodological feature is the inclusion of a path–integral correction that accounts for how parameters influence the inference trajectory, rather than only the terminal inferred state. This correction can be interpreted through stochastic optimal control formulations, where gradients are recovered by integrating sensitivity terms along state trajectories [43], and through the Free Energy Principle (FEP), where learning corresponds to minimising an accumulated free-energy action over time [34]. Importantly, this formulation allows gradients to remain well-defined under truncated or pre-equilibrium inference, offering potential computational flexibility.

The framework induces an explicit variational free energy over a BP-trained network via inference dynamics. This, in principle, enables structure learning through free-energy–based pruning, where structural relevance is evaluated in terms of a parameter's influence on inference dynamics rather than posterior variance alone [44, 34]. Preliminary experiments suggest increased robustness to certain forms of structured noise on MNIST and CIFAR-10 [47, 48], although these findings remain exploratory and dataset-dependent.

Free-Energy–Based Structure Learning and Causal Pruning

An advantage of the proposed approach is that it yields an explicit variational free energy defined over its effective weight space via latent-state inference. Since free energy trades off data fit against model complexity, it provides a principled objective against which structural modifications may be evaluated [34, 35]. In this context, free-energy–based pruning corresponds to removing connections whose ablation decreases free energy, indicating improved model evidence and reduced explanatory redundancy. Structure learning can therefore be explored both as a form of causal analysis and as a potential mechanism for adaptive model simplification [36, 37].

This formulation is intended to complement, rather than replace, structure learning in BNNs. BNNs typically infer structure indirectly from posterior uncertainty—through sparsity-inducing priors, posterior variance thresholds, or marginal likelihood comparisons [19, 18, 29]. By contrast, the proposed framework evaluates structural relevance through the effect of parameters on inference trajectories and accumulated free energy. In this setting, connections may be retained or pruned based on their contribution to stabilising inference dynamics and reducing prediction error propagation.

As a result, structure learning can be integrated into the learning process itself rather than applied post hoc, potentially supporting continual or online adaptation under changing data regimes. Structural relevance admits a limited causal interpretation, in the sense that parameters are assessed by their *interventional* effect on inference dynamics rather than by correlational statistics derived from static posterior uncertainty, consistent with

formal interventionist accounts of causality [51], invariance-based causal discovery in machine learning [52], and dynamical formulations of causal influence in active inference [53, 54]. The inclusion of the path–integral correction helps ensure that these assessments remain meaningful even when inference is truncated [46].

Neuromorphic Hardware and Translational Opportunities

The reliance of inference learning (IL) on local error processing and explicit latent-state dynamics suggests potential compatibility with neuromorphic and hybrid-neuromorphic hardware [49, 50]. Although iterative inference introduces additional computational cost, this may be partially mitigated through parallelisation on architectures where error units and local update rules can be implemented directly in hardware [49, 50]. In such settings, predictive coding may offer alternative trade-offs between computational efficiency and accuracy relative to standard backpropagation, particularly in regimes where update locking limits parallelism [40].

Beyond hardware considerations, the framework provides a basis for exploratory work in causal and uncertainty-aware modelling. One motivating application concerns the analysis of neural pathways, such as the influence of arousal on neuromodulatory control of retinal ganglion cell sensitivity. Within a static BP-trained model, such processes could potentially be examined through uncertainty-aware pruning and inference-driven structure learning, where connections are evaluated by their effect on latent-state inference trajectories rather than predictive accuracy alone. This allows interventional hypotheses to be explored alongside correlational structure, by assessing how structural modifications alter inference dynamics and accumulated free energy [34, 45].

A further area of interest relates to spike-timing-dependent plasticity (STDP), in which synaptic change depends on the temporal relationship between pre- and postsynaptic activity. Although the proposed framework is static at the level of task training, predictive coding–derived gradients depend on the temporal evolution of latent-state inference rather than solely on terminal prediction error. This algorithmic temporal sensitivity provides a formal analogue of STDP, in which parameter updates depend on the relative ordering of signals during inference, without implying a literal model of biological spike timing [33, 39]. More generally, the causal sensitivity afforded by inference-based learning may be relevant across a range of translational neuroscience settings, where understanding how uncertainty and structure shape inference dynamics is of central interest.

Extensions to Control Theory and Active Inference

Inference learning is equivalent, in the static limit, to variational Bayesian filtering under fixed equilibria [33]. Building on this equivalence, the framework could be extended to

dynamic generative models by introducing generalised coordinates of motion, allowing inference to track slowly varying equilibria over time [34].

In this setting, PC with explicit control variables admits an interpretation in terms of quadratic optimal control, where latent states and control inputs are inferred jointly prior to gradient computation. Under linear–Gaussian assumptions, this reduces to Kalman filtering, which may be used to amortise the production of synthetic gradients by propagating uncertainty-aware state estimates along inference trajectories [43]. These extensions motivate longer-term investigation of links between the proposed framework, control theory, and active inference, rather than forming a core deliverable of the PhD [36, 38].

More generally, in full active inference on hierarchical networks with nonlinear control dynamics and continuous-time inference, Kalman–Bucy filtering provides a principled approximation scheme for amortising gradient signals during BP training. In this setting, filtering dynamics serve to integrate state inference, control, and learning within a unified variational objective, while preserving compatibility with gradient-based optimisation [36, 46, 38].

Conclusions

This proposal outlines a scalable Bayesian learning framework that combines predictive coding and backpropagation to address fundamental limitations of point-estimate optimisation. By amortising inference-learning gradients through a decoupled PC network and incorporating a path–integral correction, the proposed approach endows BP-trained models with explicit variational free energy, principled uncertainty estimation, and causal structure learning.

The framework is complementary to existing Bayesian neural network approaches, computationally feasible, and well aligned with current priorities in robust, interpretable, and biologically inspired machine learning. Extensions to neuromorphic hardware and active inference provide clear and well-scoped directions for doctoral research.

PhD Timeline

Year 1: Theoretical development, baseline implementation, and benchmarking. **Year 2:** Scaling, robustness analysis, and structure-learning experiments. **Year 3:** Control and active inference extensions; thesis writing.

References

[1] Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning representations by back-propagating errors. *Nature* **323**, 533–536 (1986).

[2] Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning*. MIT Press (2016).

[3] Vaswani, A. *et al.* Attention is all you need. In *Advances in Neural Information Processing Systems* (2017).

[4] Brown, T. B. *et al.* Language models are few-shot learners. In *Advances in Neural Information Processing Systems* (2020).

[5] OpenAI. ChatGPT: Optimizing language models for dialogue. *OpenAI Technical Report* (2022).

[6] Reichstein, M. *et al.* Deep learning and process understanding for data-driven Earth system science. *Nature* **566**, 195–204 (2019).

[7] Carleo, G. *et al.* Machine learning and the physical sciences. *Reviews of Modern Physics* **91**, 045002 (2019).

[8] Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

[9] Guo, C., Pleiss, G., Sun, Y. & Weinberger, K. Q. On calibration of modern neural networks. In *Proceedings of ICML* (2017).

[10] Ovadia, Y. *et al.* Can you trust your model's uncertainty? In *Advances in Neural Information Processing Systems* (2019).

[11] Hendrycks, D. & Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples. *arXiv preprint arXiv:1610.02136* (2016).

[12] McCloskey, M. & Cohen, N. J. Catastrophic interference in connectionist networks. *Psychology of Learning and Motivation* **24** (1989).

[13] Kirkpatrick, J. *et al.* Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences* **114**, 3521–3526 (2017).

[14] Kendall, A. & Gal, Y. What uncertainties do we need in Bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems* (2017).

[15] Amodei, D. *et al.* Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565* (2016).

[16] Crick, F. The recent excitement about neural networks. *Nature* **337**, 129–132 (1989).

[17] Lillicrap, T. P. *et al.* Backpropagation and the brain. *Nature Reviews Neuroscience* **21**, 335–346 (2020).

[18] MacKay, D. J. C. A practical Bayesian framework for backpropagation networks. *Neural Computation* **4**, 448–472 (1992).

[19] Neal, R. M. *Bayesian Learning for Neural Networks*. Springer (1996).

[20] Wilson, A. G. & Izmailov, P. Bayesian deep learning and a probabilistic perspective of generalization. In *Advances in Neural Information Processing Systems* (2020).

[21] Neal, R. M. MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo*. Chapman & Hall / CRC (2011).

[22] Betancourt, M. A conceptual introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434* (2017).

[23] Hoffman, M. D. & Gelman, A. The No-U-Turn sampler. *Journal of Machine Learning Research* **15**, 1593–1623 (2014).

[24] Welling, M. & Teh, Y. W. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of ICML* (2011).

[25] Gal, Y. & Ghahramani, Z. Dropout as a Bayesian approximation. In *Proceedings of ICML* (2016).

[26] Lakshminarayanan, B., Pritzel, A. & Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems* (2017).

[27] Foong, A. Y. K. *et al.* Expressiveness of approximate inference in Bayesian neural networks. In *Advances in Neural Information Processing Systems* (2019).

[28] Graves, A. Practical variational inference for neural networks. In *Advances in Neural Information Processing Systems* (2011).

[29] Blundell, C. *et al.* Weight uncertainty in neural networks. In *Proceedings of ICML* (2015).

[30] Louizos, C. & Welling, M. Multiplicative normalizing flows. In *Proceedings of ICML* (2017).

[31] Rezende, D. J. & Mohamed, S. Variational inference with normalizing flows. In *Proceedings of ICML* (2015).

[32] Rao, R. P. N. & Ballard, D. H. Predictive coding in the visual cortex. *Nature Neuroscience* **2**, 79–87 (1999).

[33] Friston, K. J. A theory of cortical responses. *Philosophical Transactions of the Royal Society B* **360**, 815–836 (2005).

[34] Friston, K. J. The free-energy principle. *Nature Reviews Neuroscience* **11**, 127–138 (2010).

[35] Friston, K., Kilner, J. & Harrison, L. A free energy principle for the brain. *Journal of Physiology–Paris* **100**, 70–87 (2006).

[36] Friston, K. J. *et al.* Active inference: A process theory. *Neural Computation* **29**, 1–49 (2017).

[37] Tschantz, A., Seth, A. K. & Buckley, C. L. Learning action-oriented models through active inference. *PLOS Computational Biology* **16**, e1007805 (2020).

[38] Parr, T., Pezzulo, G. & Friston, K. J. *Active Inference*. MIT Press (2022).

[39] Whittington, J. C. R. & Bogacz, R. An approximation of backpropagation in predictive coding networks. *Neural Computation* **29**, 1–38 (2017).

[40] Millidge, B., Seth, A. K. & Buckley, C. L. Predictive coding: A theoretical and experimental review. *arXiv preprint arXiv:2006.10129* (2020).

[41] Millidge, B., Tschantz, A. & Buckley, C. L. Predictive coding approximates backpropagation. *Neural Computation* **34**, 1329–1368 (2022).

[42] Jaderberg, M. *et al.* Decoupled neural interfaces using synthetic gradients. In *Proceedings of ICML* (2017).

[43] Theodorou, E., Buchli, J. & Schaal, S. A generalized path integral control approach. *Journal of Machine Learning Research* **11**, 3137–3181 (2010).

[44] Friston, K. J., Litvak, V., Oswal, A., Razi, A., Stephan, K. E., van Wijk, B. C. M., Ziegler, G. & Zeidman, P. Bayesian model reduction and empirical Bayes for group (DCM) studies. *NeuroImage* **128**, 413–431 (2016).

[45] Frässle, S., Lomakina, E. I., Razi, A., Friston, K. J., Buhmann, J. M. & Stephan, K. E. Regression DCM for fMRI. *NeuroImage* **155**, 406–421 (2018).

[46] Millidge, B., Tschantz, A., Seth, A. K. & Buckley, C. L. On the relationship between predictive coding and backpropagation. *Biological Cybernetics* **116**, 131–145 (2022).

[47] LeCun, Y. *et al.* Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **86**, 2278–2324 (1998).

[48] Krizhevsky, A. Learning multiple layers of features from tiny images. Technical Report, University of Toronto (2009).

[49] Davies, M. *et al.* Loihi: A neuromorphic manycore processor with on-chip learning. *IEEE Micro* **38**, 82–99 (2018).

[50] Furber, S. B. *et al.* The SpiNNaker project. *Proceedings of the IEEE* **102**(5), 652–665 (2014).

[51] Pearl, J. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edn (2009).

[52] Peters, J., Janzing, D. & Schölkopf, B. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press (2017).

[53] Friston, K. J. Life as we know it. *Journal of the Royal Society Interface* **10**, 20130475 (2013).

[54] Parr, T. & Friston, K. J. Generalised free energy and active inference. *Biological Cybernetics* **113**, 495–513 (2019).