

# Signatures of Bayesian inference emerge from energy efficient synapses

James Malkin<sup>1\*</sup>, Cian O'Donnell<sup>1,2</sup>, Conor Houghton<sup>1</sup>, Laurence Aitchison<sup>1</sup>

\*For correspondence:  
james.malkin@bristol.ac.uk (JM)

<sup>1</sup>Faculty of Engineering, University of Bristol, Bristol, UK; <sup>2</sup>Intelligent Systems Research Centre, School of Computing, Engineering, and Intelligent Systems, Ulster University, Derry/Londonderry, UK

---

## Abstract

Biological synaptic transmission is unreliable, and this unreliability likely degrades neural circuit performance. While there are biophysical mechanisms that can increase reliability, for instance by increasing vesicle release probability, these mechanisms cost energy. We examined four such mechanisms along with the associated scaling of the energetic costs. We then embedded these energetic costs for reliability in artificial neural networks (ANN) with trainable stochastic synapses, and trained these networks on standard image classification tasks. The resulting networks revealed a tradeoff between circuit performance and the energetic cost of synaptic reliability. Additionally, the optimised networks exhibited two testable predictions consistent with pre-existing experimental data. Specifically, synapses with lower variability tended to have 1) higher input firing rates and 2) lower learning rates. Surprisingly, these predictions also arise when synapse statistics are inferred through Bayesian inference. Indeed, we were able to find a formal, theoretical link between the performance-reliability cost tradeoff and Bayesian inference. This connection suggests two incompatible possibilities: evolution may have chanced upon a scheme for implementing Bayesian inference by optimising energy efficiency, or alternatively, energy efficient synapses may display signatures of Bayesian inference without actually using Bayes to reason about uncertainty.

---

## Introduction

The synapse is the major site of inter-cellular communication in the brain. The amplitude of synaptic post-synaptic potentials (PSPs) are usually highly variable or stochastic. This variability arises primarily presynaptically: the release of neurotransmitter from presynaptically-housed vesicles into the synaptic cleft has variable release probabilities and variable quantal sizes (*Lisman and Harris, 1993; Branco and Staras, 2009; Brock et al., 2020*). Unreliable synaptic transmission seems puzzling, especially in light of evidence for low-noise, almost failure-free transmission at some synapses (*Paulsen and Heggelund, 1994, 1996; Bellingham et al., 1998*). Moreover, the degree to which a synapse is unreliable does not just vary from one synapse type to another, there is also an heterogeneity of precision amongst synapses of the same type (*Murthy et al., 1997; Dobrunz and Stevens, 1997*). Given that there is capacity for more precise transmission, why is this capacity not used in more synapses?

Unreliable transmission degrades accuracy but *Laughlin et al. (1998)* showed that the synaptic connection from a photoreceptor to a retinal large monopolar cell could increase its precision by increasing the number of synapses, averaging the noise away, but this comes at the cost of extra energy per bit of information transmitted. Moreover, *Levy and Baxter (2002)* demonstrated that there is a value for the precision which optimises the energy cost of information transmission. In this paper, we explore this notion of a performance-energy tradeoff.

However, it is important to consider precision and energy cost in the context of neuronal computation; the brain does not

simply transfer information from neuron to neuron, it performs computation through the interaction between neurons. However, models outlining a synaptic energy-performance tradeoff, (*Laughlin et al., 1998; Levy and Baxter, 2002; Goldman, 2004; Harris et al., 2012, 2019; Karbowski, 2019*), predominantly consider information transmission between just two neurons and the corresponding information-theoretic view treats the synapse as an isolated conduit of information (*Shannon, 1948*). In contrast, in reality, a single synapse is just one unit of the computational machinery of the brain. As such, the performance of an individual synapse needs to be considered in the context of circuit performance. To perform computation in an energy-efficient way the circuit as a whole needs to allocate resources across different synapses to optimise the overall energy cost of computation (*Yu et al., 2016; Schug et al., 2021*).

Here, we consider the consequences of a tradeoff between network performance and energetic reliability costs that depend explicitly upon synapse precision. We estimate the energy costs associated with precision by considering the biological mechanisms underpinning synaptic transmission. By including these costs in a neural network designed to perform a classification task, we observe a heterogeneity in synaptic precision and find that this “allocation” of precision is related to signatures of synapse “importance”, which can be understood formally on the grounds of Bayesian inference.

## Results

We proposed energetic costs for reliable synaptic transmission and then measured their consequences in an artificial neural network.

### Biophysical costs

Here, we seek to understand the biophysical energetic costs of synaptic transmission, and how those costs relate to the reliability of transmission (Fig. 1a). We start by considering the underlying mechanisms of synaptic transmission. In particular, synaptic transmission begins with the arrival of a spike at the axon terminal. This triggers a large influx of calcium ions into the axon terminal. The increase in calcium concentration causes the release of neurotransmitter-filled vesicles docked at axonal release sites. The neurotransmitter diffuses across the synaptic cleft to the postsynaptic dendritic membrane. There, the neurotransmitter binds with ligand-gated ion channels causing a change in voltage, i.e. a postsynaptic potential. This process is often quantified using the *Katz and Miledi (1965)* quantal model of neurotransmitter release. Under this model, for each connection between two cells, there are  $n$  docked, readily releasable vesicles. When the presynaptic cell spikes, each docked vesicle releases with probability  $p$  and each released vesicle causes a postsynaptic potential of size  $q$ . Thus, the mean,  $\mu$ , and variance,  $\sigma^2$ , of the PSP can be written (see Fig. 1b),

$$\begin{aligned}\mu &= npq \\ \sigma^2 &= np(1-p)q^2.\end{aligned}\tag{1}$$

We considered four biophysical costs associated with improving the reliability of synaptic transmission, while keeping the mean fixed, and derived the associated scaling of the energetic cost with PSP variance.

**Calcium efflux.** Reliability is higher when the probability of vesicle release,  $p$ , is higher. As vesicle release is triggered by an increase in intracellular calcium, greater calcium concentration implies higher release probability. However, increased calcium concentration implies higher energetic costs. In particular, calcium that enters the synaptic bouton will subsequently need to be pumped out. We take the cost of pumping out calcium ions to be proportional to the calcium concentration, and take the relationship between release probability and calcium concentration to be governed by a Hill Equation, following *Sakaba and Neher (2001)*. The resulting relationship between energetic costs and reliability is cost  $\propto \sigma^{-1/2}$  (Fig. 1c (I); see Appendix - Reliability costs for further details).

**Vesicle membrane surface area.** There may also be energetic costs associated with producing and maintaining a large amount of vesicle membrane. *Purdon et al. (2002)* argues that phospholipid metabolism may take a considerable proportion of the brain’s energy budget. Additionally, costs associated with membrane surface area may arise because of leakage of hydrogen ions across vesicles (*Pulido and Ryan, 2021*). Importantly, a cost for vesicle surface area is implicitly a cost on reliability. In particular, we could obtain highly reliable synaptic release by releasing many small vesicles, such that stochasticity in individual vesicle release events averages out. However, the resulting many small vesicles have a far larger surface area than a single large vesicle, with the same mean PSP. Thus, a cost on surface area implies a relationship

between energetic costs and reliability; in particular cost  $\propto \sigma^{-2/3}$  (Fig. 1c (II); see Appendix - Reliability costs for further details).

**Actin.** Another cost for small but numerous vesicles arises from a demand for structural organisation of the vesicles pool by filaments such as actin (*Cingolani and Goda, 2008; Gentile et al., 2022*). Critically, there are physical limits to the number of vesicles that can be attached to an actin filament of a given length. In particular, if vesicles are smaller we can attach more vesicles to a given length of actin, but at the same time, the total vesicle volume (and hence the total quantity of neurotransmitter) will be smaller (Fig. 1c (III)). A fixed cost per unit length of actin thus implies a relationship between energetic costs and reliability of, cost  $\propto \sigma^{-4/3}$  (see Appendix - Reliability costs).

**Trafficking.** A final class of costs is proportional to the number of vesicles (*Laughlin et al., 1998*). One potential biophysical mechanism by which such a cost might emerge is from active transport of vesicles along actin filaments or microtubules to release sites (*Chenouard et al., 2020*). In particular, vesicles are transported by ATP-dependent myosin-V motors (*Bridgman, 1999*), so more vesicles require a greater energetic cost for trafficking. Any such cost proportional to the number of vesicles gives rise to a relationship between energetic cost and PSP variance of the form, cost  $\propto \sigma^2$  (Fig. 1c (IV); see Appendix - Reliability costs).

**Costs related to PSP mean/magnitude** While costs on precision are the central focus of this paper, it is certainly the case that other costs relating to the mean PSP magnitude constitute a major cost of synaptic transmission. For example, high amplitude PSPs require a large quantity of neurotransmitter, high probability of vesicle release, and a large number of post-synaptic receptors (*Attwell and Laughlin, 2001*). These can be formalised as costs on the PSP mean,  $\mu$ , and can additionally be related to L1 weight decay in a machine learning context (*Rosset and Zhu, 2006; Sacramento et al., 2015*).

### Reliability costs in artificial neural networks

Next, we sought to understand how these biophysical energetic costs of reliability might give rise to patterns of variability in a trained neural network. Specifically, we trained artificial neural networks (ANNs) using an objective that embodied a tradeoff between performance and reliability costs,

$$\text{overall cost} = \text{performance cost} + \text{magnitude cost} + \text{reliability cost}. \quad (2)$$

The “performance cost” term measures the network’s performance on the task, for instance in our classification tasks we used the usual cross-entropy cost. The “magnitude cost” term captures costs that depend on the PSP mean, while the “reliability cost” term captures costs that depend on the PSP precision. In particular,

$$\text{magnitude cost} = \lambda \sum_i |\mu_i|, \quad (3)$$

$$\text{reliability cost} = c \sum_i \sigma_i^{-\rho}. \quad (4)$$

Here,  $i$  indexes synapses, and recall that  $\sigma_i$  is the standard deviation of the  $i$ th synapse. The multiplier  $c$  in the reliability cost determines the strength of the reliability cost relative to the performance cost. Small values for  $c$  imply that the reliability cost term is less important, permitting precise transmission and higher performance. Large values for  $c$  give greater importance to the reliability cost encouraging energy efficiency by allowing higher levels of synaptic noise, causing detriment to performance (see Fig. 2).

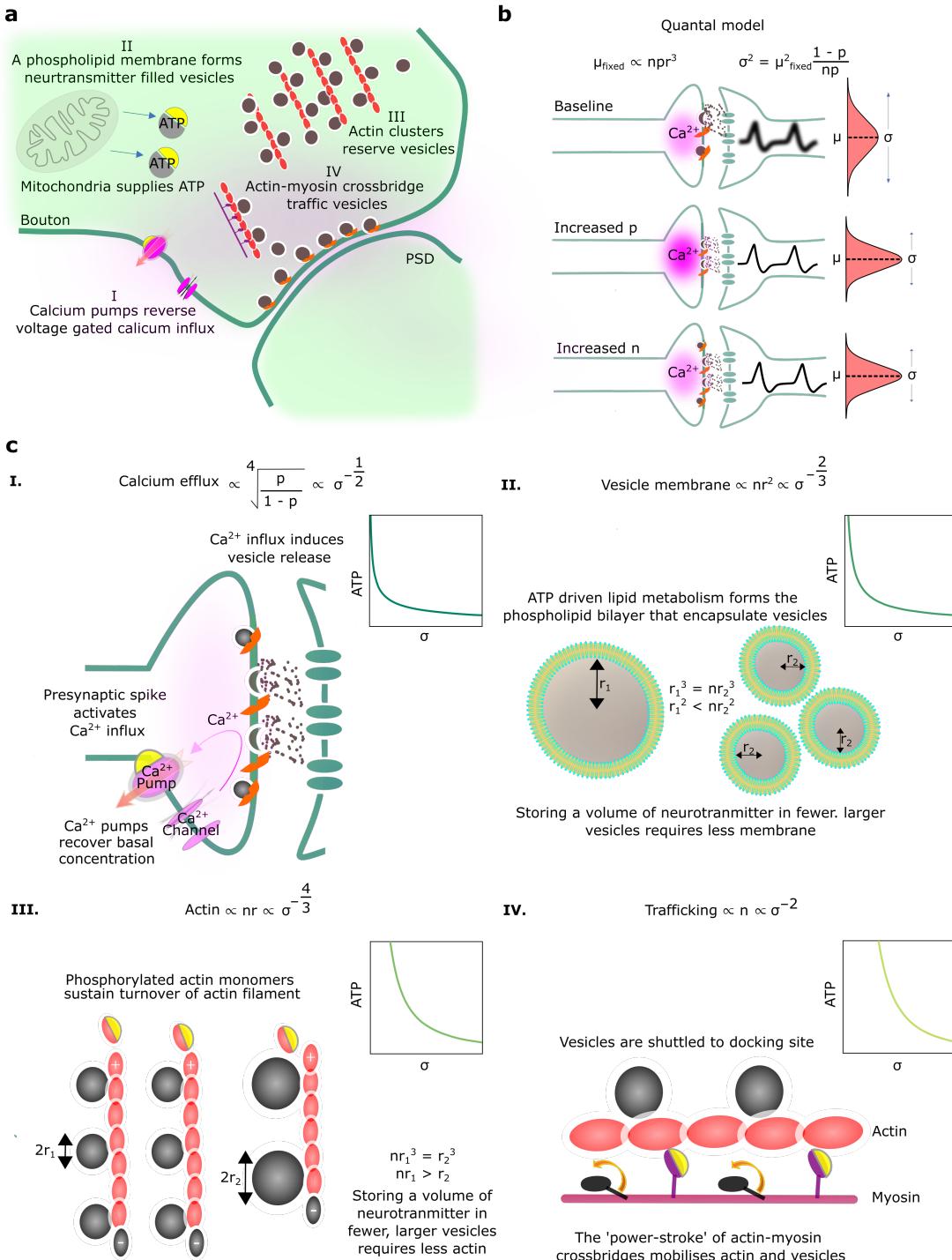
We trained fully-connected, rate-based neural network to classify MNIST digits. Stochastic synaptic PSPs were sampled from a Normal distribution,

$$w_i \sim \text{Normal}(\mu_i, \sigma_i). \quad (5)$$

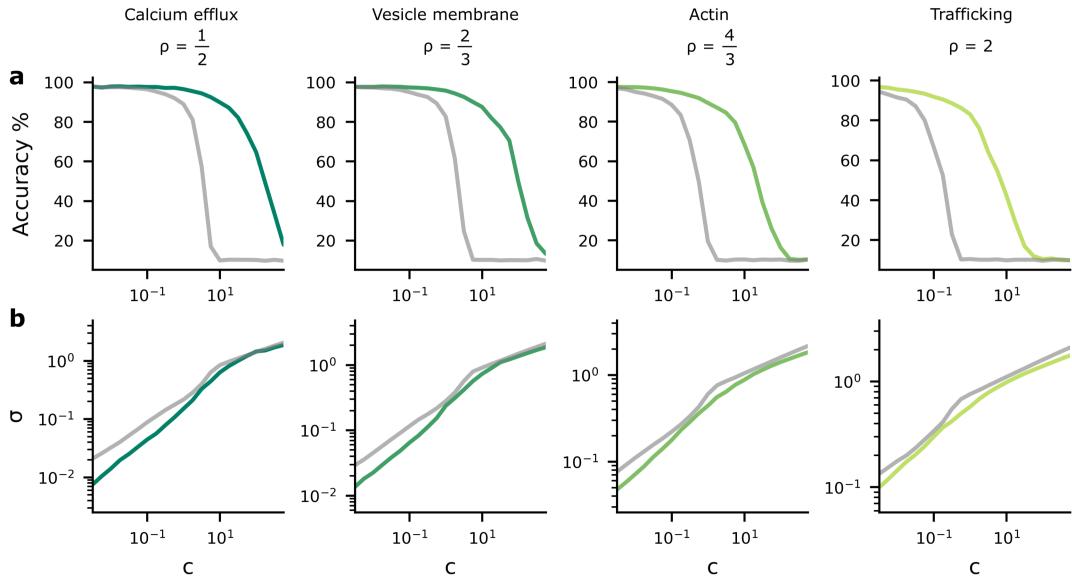
where, recall,  $\mu_i$  is the PSP mean and  $\sigma_i^2$  is the PSP variance for the  $i$ th synapse. The output firing rate was given by,

$$\text{firing rate} = f \left( \sum_i w_i x_i - w_0 \right). \quad (6)$$

Here,  $\sum_i w_i x_i - w_0$  can be understood as the somatic membrane potential, and  $f$  represents the relationship between somatic membrane potential and firing rate; we used ReLU (*Fukushima, 1975*). We optimised network parameters  $\mu_i$  and  $\sigma_i$  using Adam (*Kingma and Ba, 2014*) (see Methods for details on architecture and hyperparameters).



**Figure 1. Physiological reliability costs - a.** Physiological processes that influence PSP precision. **b.** A binomial model of vesicle release. For fixed PSP mean, increasing  $p$  or  $n$  decreases PSP variance. We have substituted  $q \propto r^3$  to reflect that vesicle volume scales quantal size (Karunanithi et al., 2002). **c.** Four different biophysical costs of reliable synaptic transmission. **I**) Calcium pumps reverse the calcium influx that triggers vesicle release. A high probability of vesicle release requires a large influx of calcium, and extruding this calcium is costly. **II**) An equivalent volume of neurotransmitter can be stored in few large vesicles or shared between many smaller vesicles. Sharing a fixed volume of neurotransmitter among many small vesicles reduces PSP variability but increases vesicle surface area, creating greater demand for phospholipid metabolism and hence greater energetic costs. **III**) Actin filament support the structure of vesicle clusters at the terminal. Many and large vesicles require more actin and higher rates of ATP dependent actin turnover. **IV**) There are biophysical costs that scale with the number of vesicles (Laughlin et al., 1998; Attwell and Laughlin, 2001) e.g. vesicle trafficking driven by myosin-V active transport along actin filaments.



**Figure 2. Accuracy and PSP variance as we change the tradeoff between reliability and performance costs.** We changed the tradeoff by modifying  $c$ , in Eq. 4, which multiplies the reliability cost. **a.** As the reliability cost multiplier,  $c$ , increases, the accuracy decreases considerably. The green lines show the heterogeneous noise setting where the noise level is optimised on a per-synapse basis, while the grey lines show the homogeneous noise setting, where the noise is optimised, but forced to be the same for all synapses. **b.** When the reliability cost multiplier,  $c$  increases, the synaptic noise level (specifically, the average standard deviation,  $\sigma$ ) increases.

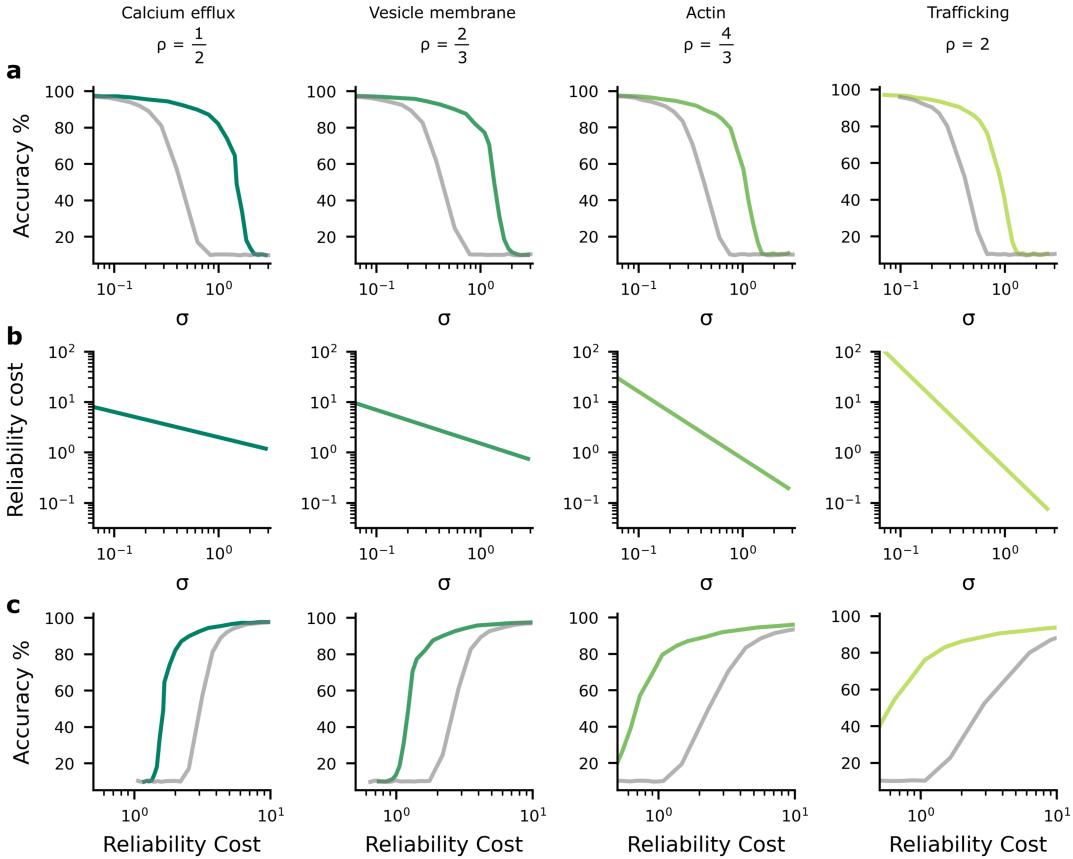
### The tradeoff between accuracy and reliability costs in trained networks

Next we sought to understand how the tradeoff between accuracy and reliability cost manifests in trained networks. Perhaps the critical parameter in the objective, (Eq. 2 and Eq. 4) was  $c$ , which controlled the importance of the reliability cost relative to the performance cost. We trained networks with a variety of different values of  $c$ , and with four values for  $\rho$  motivated by the biophysical costs (the different columns). As expected, we found that as  $c$  increased, performance fell (Fig. 2a) and the average synaptic standard deviation increased (Fig. 2b). Importantly, we considered two different settings. First, we considered an homogeneous noise setting, where  $\sigma_i$  is optimised but kept the same across all synapses (grey lines). Second, we considered an heterogeneous noise setting, where  $\sigma_i$  is allowed to vary across synapses, and is optimised on a per-synapse basis. We found that heterogeneous noise (i.e. allowing the noise to vary on a per-synapse basis) improved accuracy considerably for a fixed value of  $c$ , but only reduced the average noise slightly.

The findings in Fig. 2 imply a tradeoff between accuracy and average noise level,  $\sigma$ , as we change  $c$ . If we explicitly plot the accuracy against the noise level using the data from Fig. 2, we see that as the synaptic noise level increases, the accuracy decreases (Fig. 3a). Further, the synaptic noise level is associated with a reliability cost (Fig. 3b), and this relationship changes in the different columns as they use different values of  $\rho$  associated with different biological mechanisms that might give rise to the dominant biophysical reliability cost. Thus, there is also a relationship between accuracy and reliability costs (Fig. 3c), with accuracy increasing as we allow the system to invest more energy in becoming more reliable, which implies a higher reliability cost. Again, we plotted both the homogeneous (grey lines) and heterogeneous noise cases (green lines). We found that heterogeneous noise allowed for considerably improved accuracy at a given average noise standard deviation or a given reliability cost.

### Energy-efficient patterns of synapse variability

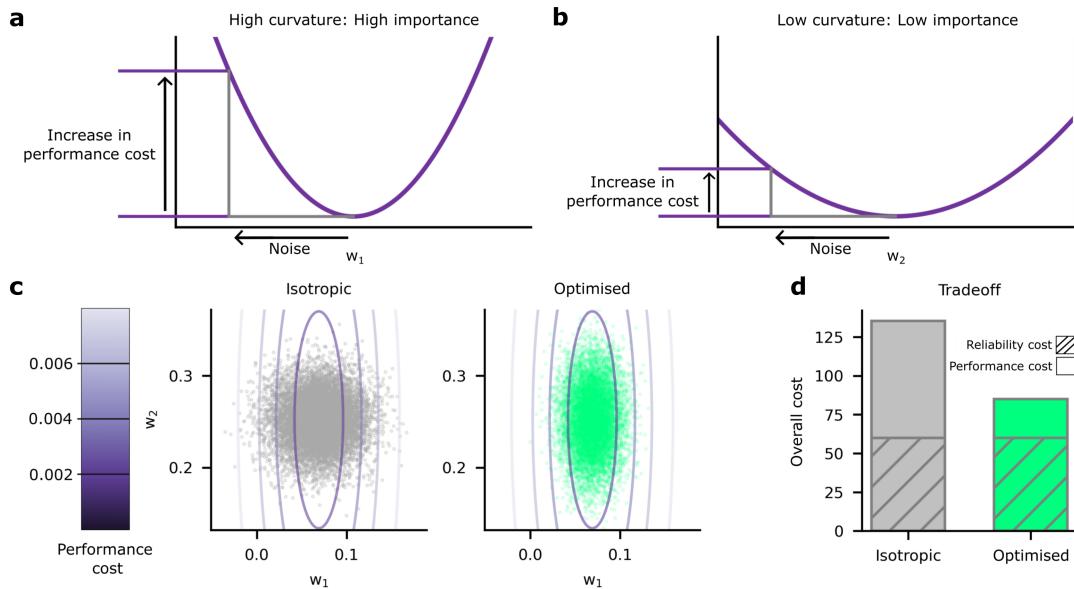
We found that the heterogeneous noise setting, where we individually optimise synaptic noise on a per-synapse basis, performed considerably better than the homogeneous noise setting (Fig. 3). This raised an important question: how does the network achieve such large improvements by optimising the noise levels on a per-synapse basis? We hypothesised that the system invests a lot of energy in improving the reliability for “important” synapses, i.e. synapses whose weights have a large impact on predictions and accuracy (Fig. 4a). Conversely, the system allows unimportant synapses to have high variability, which reduces reliability costs (Fig. 4b). To get further intuition, we compared both  $w_1$  and  $w_2$  on the same



**Figure 3. The performance-reliability cost tradeoff in ANN simulations - a.** Accuracy decreases as the average PSP standard deviation,  $\sigma$ , increases. The grey lines are for the homogeneous noise setting where the PSP variance is optimised but isotropic (i.e. the same across all synapses), while the green lines is for the heterogeneous noise setting, where the PSP variances are optimised individually on a per-synapse basis. **b.** Increasing reliability by reducing  $\sigma^2$  leads to greater reliability costs, and this relationship is different for different biophysical mechanisms and hence values for  $\rho$  (columns). **c.** Higher accuracy therefore implies larger reliability cost.

plot (Fig. 4c). Specifically, we put the important synapse,  $w_1$  from Fig. 4a, on the horizontal axis, and the unimportant synapse,  $w_2$  from Fig. 4b, on the vertical axis. In Fig. 4c, the relative importance of the synapse is now depicted by how the cost increases as we move away from the optimal value of the weight. Specifically, the cost increases rapidly as we move away from the optimal value of  $w_1$ , but increases much more slowly as we move away from the optimal value of  $w_2$ . Now, consider deviations in the synaptic weight driven by homogeneous synaptic variability (Fig. 4c left, grey points). Many of these points have poor performance (i.e. a high performance cost), due to relatively high noise on the important synapse (i.e.  $w_1$ ). Next, consider deviations in the synaptic weight driven by heterogeneous, optimised variability (Fig. 4c left, green points). Critically, optimising synaptic noise reduces variability for the important synapse, and that reduces the average performance cost by eliminating large deviations on the important synapse. Thus, for the same overall reliability cost, heterogeneous, optimised variability can achieve much lower performance costs, and hence much lower overall costs than homogeneous variability (Fig. 4d).

To investigate experimental predictions arising from optimised, heterogeneous variability, we needed a way to formally assess the “importance” of synapses. We used the “curvature” of the performance cost: namely the degree to which small deviations in the weights from their optimal values will degrade performance. If the curvature is large (Fig. 4a), then small deviations in the weights, e.g. those caused by noise, can drastically reduce performance. In contrast, if the curvature is smaller (Fig. 4b), then small deviations in the weights cause a much smaller reduction in performance. As a formal measure of the curvature of the objective, we used the Hessian matrix,  $\mathbf{H}$ . This describes the shape of the objective as a function of the synaptic weights, the  $w_i$ s: specifically, it is the matrix of second derivatives of the objective, with respect to the weights, and measures the local curvature of objective. We were interested in the diagonal elements,  $H_{ii}$ ; the second



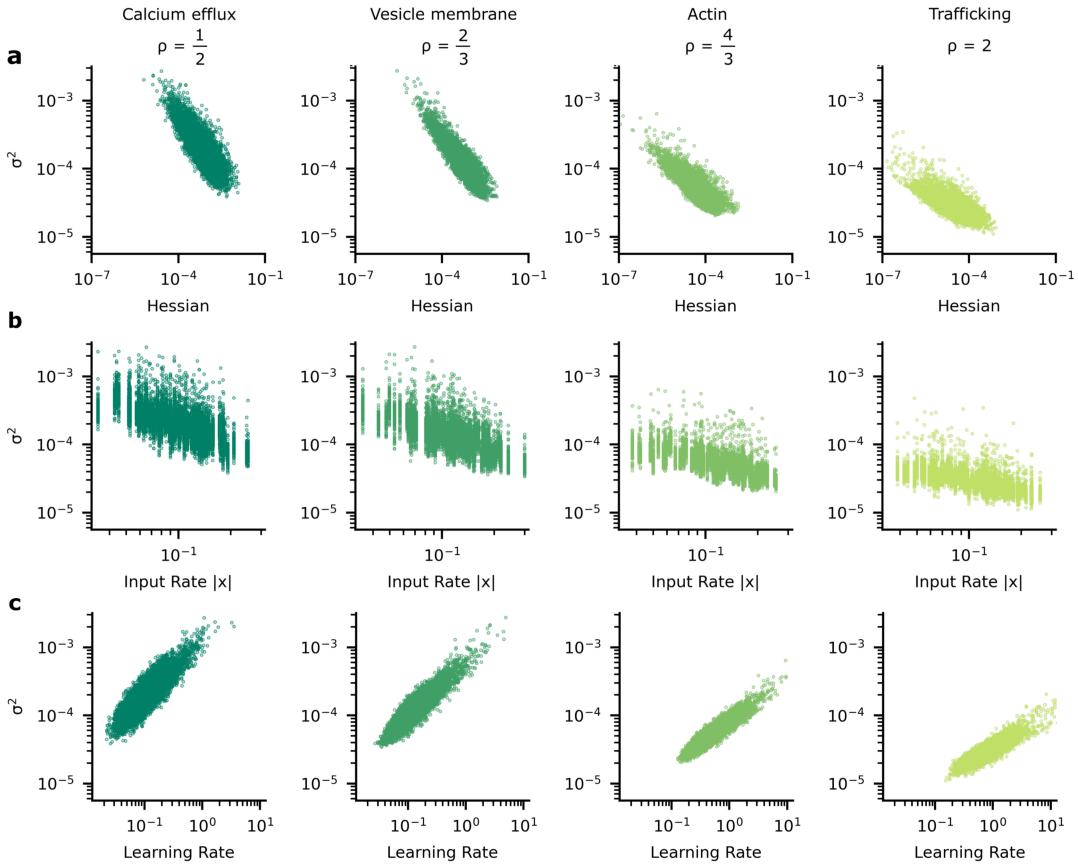
**Figure 4. Schematic depiction of the impact of synaptic noise on synapses with different importance.** **a.** First, we considered an important synapse for which small deviations in the weight,  $w_1$ , e.g. driven by noise, imply a large increase in the performance cost. This can be understood as a high curvature of the performance cost as a function of  $w_1$ . **b.** Next we considered an unimportant synapse, for which deviations in the weights cause far less increase in performance cost. **c.** A comparison of the impacts of homogeneous and optimised heterogeneous variability for synapses  $w_1$  and  $w_2$  from a and b. The performance cost is depicted using the purple contours, and realisations of the PSPs driven by synaptic variability are depicted in the grey/green points. The grey points (left) depict homogeneous noise while the green points (right) depict optimised, heterogeneous noise. **d.** The noise distributions in panel c are chosen to keep the same reliability cost (diagonally hatched area); but the homogeneous noise setting has far a higher performance cost, primarily driven by larger noise in the important synapse,  $w_1$ .

derivatives of the objective with respect to  $w_i$ .

We began by looking at how the optimised synaptic noise varied with synapse importance, as measured by the curvature or, more formally, the Hessian (Fig. 5a). We found that as the importance of the synapse increased, the optimised noise level decreased. These patterns of synapse variability make sense because noise is more detrimental at important synapses and so it is worth investing energy to reduce the noise in those synapses.

However, this relationship (Fig. 5a) between the importance of a synapse and the synaptic variability is not experimentally testable, as we are not able to directly measure synapse importance. That said, we are able to obtain two testable predictions. First, the input rate in our simulations was negatively correlated with optimised synaptic variability (Fig. 5b). Second, the optimised synaptic variability was larger for synapses with larger learning rates (Fig. 5c). Critically, both of these patterns have been observed in experimental data. The relationship between input firing rate and synaptic variability was first observed by *Aitchison et al. (2021)* using data from *Ko et al. (2013)* (Fig. 6a). The relationship between learning rate and synaptic variability was first observed by *Schug et al. (2021)*, using data from *Sjöström et al. (2003)* as processed by *Costa et al. (2017)* (Fig. 6b).

To understand why these patterns of variability emerge in our simulations and in data, we need to understand the connection between synapse importance, synaptic inputs (Fig. 5b, Fig. 6b) and synaptic learning (Fig. 5c, Fig. 6a). Perhaps the easiest connection is between the synapse importance and the input firing rate. If the input cell never fires, then the synaptic weight cannot affect the network output, and the synapse has zero importance (and also zero Hessian (see Appendix - High input rates and high precision at important synapses)). This would suggest a tendency for synapses with higher input firing rates to be more important, and hence to have lower variability. This pattern is indeed borne out in our simulations (Fig. 5b; also see Supplementary - Appendix 6–Fig. 1), though of course there is a considerable amount of noise: there are a few important synapses with low input rates, and vice-versa.



**Figure 5. The heterogeneous patterns of synapse variability in ANNs optimised by the tradeoff** - We present data patterns on logarithmic axis between signatures of synapse importance and variability for 10,000 (100 neuron units, each with 100 synapses) synapses that connect two hidden layers in our ANN. **a.** Synapses whose corresponding diagonal entry in the Hessian is large have smaller variance. **b.** Synapses with high variance have faster learning rates **c.** As input firing rate increases, synapse variance decreases.

Next, we consider the connection between learning rate and synapse importance. To understand this connection, we need to choose a specific scheme for modulating the learning rate as a function of the inputs. Modern, state-of-the-art, update rules for artificial neural networks often use an adaptive learning rate. These adaptive learning rates,  $\eta_i$ , (including the most common such as Adam and variants) almost always use a normalising learning rate which decreases in response to high incoming gradients,

$$\eta_i = \frac{\eta_{\text{base}}}{\sqrt{\langle g_i^2 \rangle}}. \quad (7)$$

Specifically, the local learning rate for the  $i$ th synapse,  $\eta_i$ , is usually a base learning rate,  $\eta_{\text{base}}$ , divided by the root-mean-square gradient at this synapse  $\sqrt{\langle g_i^2 \rangle}$ . Critically, the root-mean-square gradient turns out to be strongly related to synapse importance. Intuitively, important synapses with greater impact on network predictions will have larger gradients Appendix - Synapse importance and gradient magnitudes.

In-vivo performance requires selective formation, stabilisation and elimination of long term plasticity (LTP) (Yang *et al.*, 2009), raising the questions as to which biological mechanisms are able to provide this selectivity. Reducing updates at historically important synapses is one potential approach to determining which synapses should have their strengths adjusted and which should be stabilised. Adjusting learning rates based on synapse importance enables fast, stable learning (LeCun *et al.*, 2002; Kingma and Ba, 2014; Khan *et al.*, 2018; Aitchison, 2020; Martens, 2020).

For our purposes, the crucial point is that when training using an adaptive learning rate such as Eq. 7, important synapses have higher root-mean-squared gradients, and hence lower learning rates. Here we use a specific set of update rules which

uses this adaptive learning rate (i.e. Adam (*Kingma and Ba, 2014; Yang and Li, 2022*)). Thus, we can use learning rate as a proxy for importance, allowing us to obtain the predictions tested in Fig. 5b which match Fig. 5a/c.

### The connection to Bayesian inference

Surprisingly, our experimental predictions obtained for optimised, heterogeneous synaptic variability (Fig. 5,6) match those arising from Bayesian synapses (i.e. synapses that use Bayes to infer their weights (*Aitchison et al., 2021*)). Our first prediction was that lower variability implies a lower learning rate. The same prediction also arises if we consider Bayesian synapses. In particular, if variability and hence uncertainty is low, then a Bayesian synapse is very certain that it is close to the optimal value. In that case, new information should have less impact on the synaptic weight, and the learning rate should be lower. Our second prediction was that higher presynaptic firing rates imply less variability. Again, this arises in Bayesian synapses: Bayesian synapses should become more certain and less variable if the presynaptic cell fires more frequently. Every time the presynaptic cell fires, the synapse gets a feedback signal which gives a small amount of information about the right value for that synaptic weight. So the more times the presynaptic cell fires, the more information the synapse receives, and the more certain it becomes.

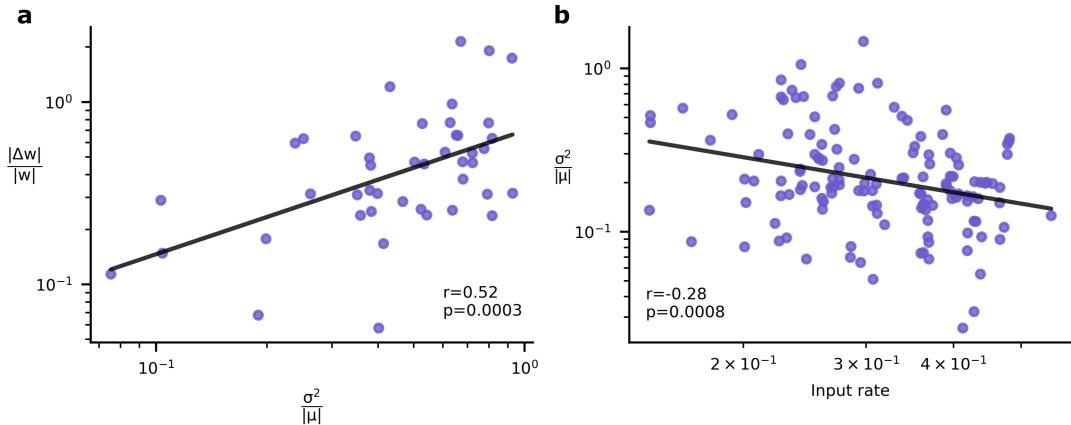
This match between observations for our energy-efficient synapses and previous work on Bayesian synapses led us to investigate potential connections between energy efficiency and Bayesian inference. Intuitively, there turns out to be a strong connection between synapse importance and uncertainty. Specifically, if a synapse is very important, then the performance cost changes dramatically when there are errors in that synaptic weight. That synapse therefore receives large gradients, and hence strong information about the correct value, rapidly reducing uncertainty.

To assess the connection between Bayesian posteriors and energy efficient variability in more depth, we plotted the posterior variance against the optimised synaptic variability (Fig. 7a). We considered our four different biophysical mechanisms (values for  $\rho$ ; Fig. 7a, columns), and values for  $c$  (Fig. 7a, rows). In all cases, there was a clear correlation between the posterior and the optimised variability: directions with larger posterior variance also had large optimised variability. To more formally assess this connection, we derived the relationships between the optimised noise,  $\sigma_i$  and the posterior variable,  $\sigma_{\text{post}}$  as a function of  $\rho$  (Fig. 7b;) and as a function of  $c$  (Fig. 7c). Again, these plots show a clear correlation between optimised variability and posterior variance; though the relationship is far from perfect. For a perfect relationship, we would expect the lines in Fig. 7bc to all lie along the diagonal. In contrast, these lines actually have a slope smaller than one, indicating that optimised variability is less heterogeneous than posterior variance (Fig. 7bc).

Finally, in the Appendix, we derive a formal connection between our overall performance cost and Bayesian inference. Specifically, variational inference, a well-known procedure for performing (approximate) Bayesian inference in NNs (*Hinton and van Camp, 1993; Graves, 2011; Blundell et al., 2015*). Variational inference optimises the “evidence lower bound objective” (ELBO) (*Barber and Bishop, 1998; Jordan et al., 1999; Blei et al., 2017*), which surprisingly turns out to resemble our performance cost. Specifically, the ELBO includes a term which encourages the entropy of the approximating posterior distribution (which could be interpreted as our noise distribution) to be larger. This resembles a reliability cost, as our reliability costs also encourage the noise distribution to be larger. Critically, the biological power-law reliability cost has a different form from the ideal, entropic reliability cost. However, we are able to derive a formal relationship: the biological power-law reliability costs bounds the ideal entropic reliability cost. Remarkably, this implies that our overall cost (Eq. 2) bounds the ELBO, so reducing our cost (Eq. 2) tightens the ELBO bound and gives an improved guarantee on the quality of Bayesian inference.

## Discussion

Comparing the brain’s computational roles with associated energetic costs provides a useful means for deducing properties of efficient neurophysiology. Here, we applied this approach to PSP variability. We began by looking at the biophysical mechanisms of synaptic transmission, and how the energy costs for transmission might vary with synaptic reliability. We modified a standard ANN to incorporate unreliable synapses and trained this on a classification task using an objective that combined classification accuracy and an energetic cost on reliability. This led to a performance-reliability cost tradeoff and heterogeneous patterns of synapse variability that correlated with input rate and learning rate. We noted that these patterns of variability have been previously observed in data (see Fig. 6). Remarkably, these are also the patterns of variability predicted by Bayesian synapses (*Aitchison et al., 2021*) (i.e. when distributions over synaptic weights correspond with the



**Figure 6. Experimental results are consistent with predictions arising from optimising a tradeoff between performance and reliability costs** - The patterns of synapse variability in biological synapses compared with signatures of synapse importance. **a)** As the normalised EPSP variability increases, the absolute percentage change in the weight during a plasticity protocol also increases with slope = 1.47 (data from (Sjöström *et al.*, 2003) as processed by (Costa *et al.*, 2017); this pattern has previously been observed using this data by (Schug *et al.*, 2021)). **b)** As input firing rate is increased, normalised EPSP variability decreases with slope = −0.71 (data from (Ko *et al.*, 2013), and this pattern was previously observed from this data by (Aitchison *et al.*, 2021)).

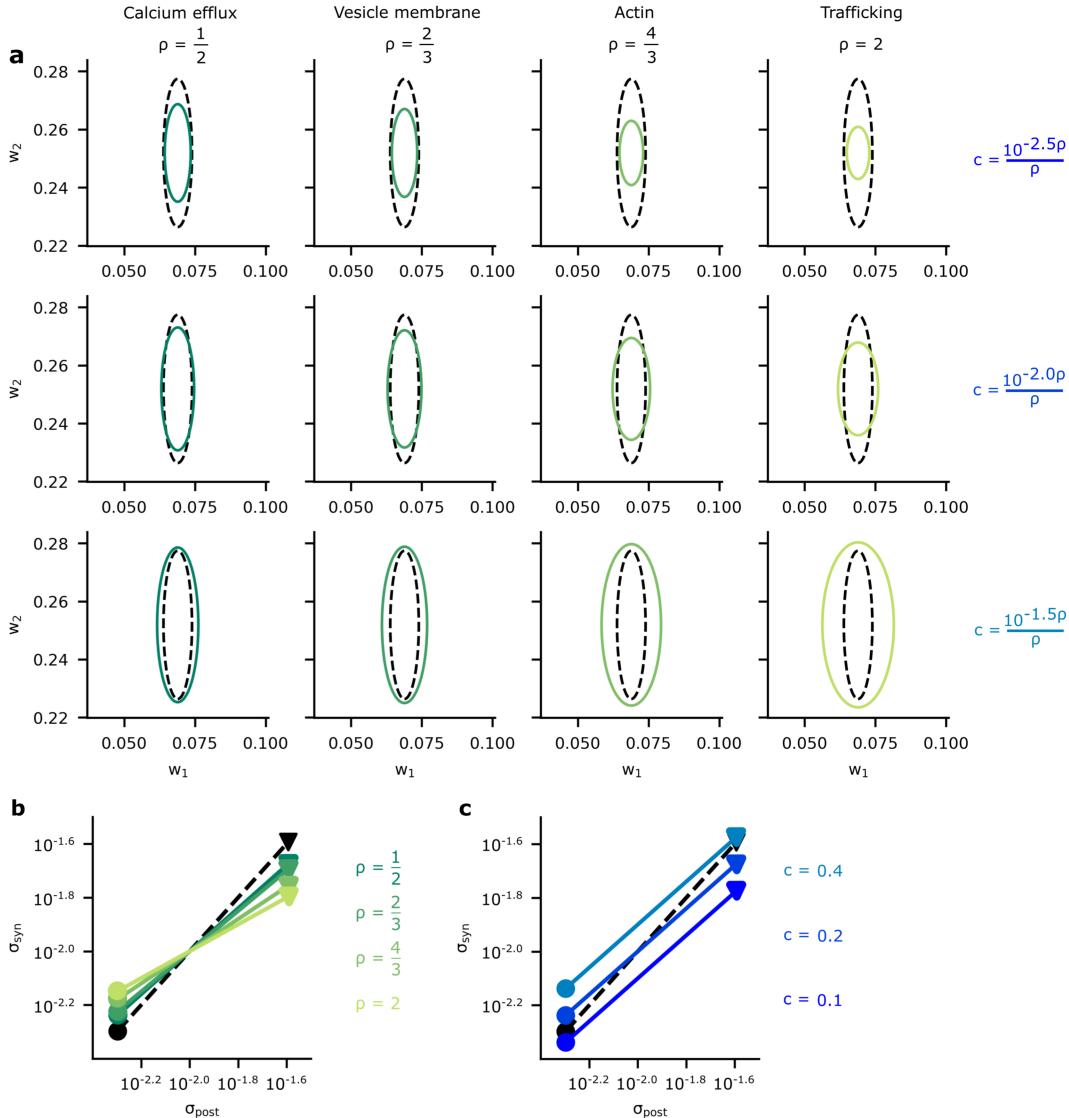
Bayesian posterior). Finally, we showed empirical and formal connections between the synaptic variability implied by Bayesian synapses and our performance-reliability cost tradeoff.

The reliability cost in terms of the synaptic variability (Eq. 4) is a critical component of the numerical experiments we present here. While the precise form of the cost is inevitably uncertain, we attempted to mitigate the uncertainty by considering a wide range of functional forms for the reliability cost. In particular, we considered four biophysical mechanisms, corresponding to four power-law exponents, ( $\rho = \frac{1}{2}, \frac{2}{3}, \frac{4}{3}, 2$ ). Moreover, these different power-law costs already cover a reasonably wide-range of potential penalties and we would expect the results to hold for many other forms of reliability cost as the intuition behind the results ultimately relies merely on there being *some* penalty for increasing reliability.

The biophysical cost also includes a multiplicative factor,  $c$ , which sets the magnitude of the reliability cost. In fact, the patterns of variability exhibited in Fig. 5 are preserved as  $c$  is changed: this was demonstrated for values of  $c$  which are ten times larger and ten times smaller, Supplementary - Appendix 6–Fig. 2. This multiplicative factor should be understood as being determined by the properties of the physics and chemistry underpinning synaptic dynamics, for example it could represent the quantity of ATP required by the metabolic costs of synaptic transmission (although this factor could vary e.g. in different cell types).

Our artificial neural networks used backpropagation to optimise the mean and variance of synaptic weights. While there are a number of schemes by which biological circuits might implement backpropagation (Whittington and Bogacz, 2017; Sacramento *et al.*, 2018; Richards and Lillicrap, 2019), it is not yet clear whether backpropagation is implemented by the brain (see Lillicrap *et al.* (2020) for a review on the plausibility of propagation in the brain). Regardless, backpropagation is merely the route we used in our ANN setting to reach an energy efficient configuration. The patterns we have observed are characteristic of an energy-efficient network and therefore should not depend on the learning rule that the brain uses to achieve energy-efficiency.

Our results in ANNs used MNIST classification as an example of a task; this may appear somewhat artificial, but all brain areas ultimately do have a task: to maximise fitness (or reward as a proxy for fitness). Moreover, our results all ultimately arise from trading off biophysical reliability costs against the fact that if a synapse is important to performing a task, then variability in that synapse substantially impairs performance. Of course performance, in different brain areas, might mean reward, fitness or some other measures. In contrast, if a synapse is unimportant, variability in that synapse impairs performance less. In all tasks there will be some synapses that are more, and some synapses that are less important, and our task, while relatively straightforward, captures this important property.



**Figure 7. A comparison of optimised synaptic variability and posterior variance.** - **a.** Posterior variance (grey-dashed ellipses) plotted alongside optimised synaptic variability (green ellipses) for different values of  $\rho$  (columns) and  $c$  (rows). Note that using fixed values of  $c$  for different  $\rho$ 's dramatically changed the scale of the ellipses. Instead, we chose  $c$  as a function of  $\rho$  to ensure that the scale of the optimised noise variance was roughly equal across different  $\rho$ . This allowed us to highlight the key pattern: that smaller values for  $\rho$  give optimised variance closer to the true posterior variances, while higher values for  $\rho$  tended to make the optimised synaptic variability more isotropic. **b.** To understand this pattern more formally, we plotted the synaptic variability as a function of the posterior variance for different values of  $\rho$ . Note that we set  $c$  to  $c = \frac{10^{-2.0\rho}}{\rho}$  to avoid large additive offsets (see Connecting the entropy and the biological reliability cost–Eq.51 for details). **c.** The synaptic variability as a function of the posterior variance for different values of  $c$ : [0.112, 0.2, 0.356] (3 DP). As  $c$  increases (lighter blues) we penalise reliability more, and hence the optimised synaptic noise variability increases. (Here we fixed  $\rho = 1/2$  across different settings for  $c$ )

Our results have important implications for the understanding of Bayesian inference in synapses. In particular, we show that energy efficiency considerations give rise to two phenomena that are consistent with predictions outlined in previous work on Bayesian synapses (*Aitchison et al., 2021*). First, that normalised variability decreases for synapses with higher presynaptic firing rates. Second, that synaptic plasticity is higher for synapses with higher variability.

Specifically, these findings suggest that synapses connect their uncertainty in the value of the optimal synaptic weight (see *Aitchison et al., 2021*, for details) to variability. This is in essence a synaptic variant of the “sampling hypothesis”.

Under the sampling hypothesis, neural activity is believed to represent a potential state of the world, and variability is believed to represent uncertainty (Hoyer and Hyvärinen, 2002; Knill and Pouget, 2004; Ma et al., 2006; Fiser et al., 2010; Berkes et al., 2011; Orbán et al., 2016; Aitchison and Lengyel, 2016; Haefner et al., 2016; Lange and Haefner, 2017; Shivkumar et al., 2018; Bondy et al., 2018; Echeveste et al., 2020; Festa et al., 2021; Lange et al., 2021; Lange and Haefner, 2022). This variability in neural activity, representing uncertainty in the state of the world, can then be read out by downstream circuits to inform behaviour. Here, we showed that a connection between synaptic uncertainty and variability can emerge simply as a consequence of maximising energy efficiency. This suggest that Bayesian synapses may emerge without any necessity for specific synaptic biophysical implementations of Bayesian inference.

Importantly though, while the brain might use synaptic noise for Bayesian computation, these results are also consistent with an alternative interpretation: that the brain is not Bayesian, it just looks Bayesian because it is energy efficient. To distinguish between these two interpretations, we ultimately need to know whether downstream brain areas exploit or ignore information about uncertainty that arises from synaptic variability.

## Materials and methods

The ANN simulations were run in PyTorch with feedforward, fully-connected neural networks with two hidden layers of width 100. The input dimension of 784 corresponded to the number of pixels in the greyscale MNIST images of handwritten digits, while the output dimension of ten corresponded to the number of classes. We used the reparameterisation trick to backpropagate with respect to the mean and variance of the weights, in particular, we set  $w_i = \mu_i + \sigma_i \xi$  where  $\xi \sim \text{Normal}(0, 1)$  (Kingma et al., 2015). MNIST classification was learned through optimisation of Gaussian parameters with respect to a cross-entropy loss in addition to reliability costs using minibatch gradient descent under Adam optimisation with a minibatch size of 20. To prevent negative values for the  $\sigma$ s, they were re-parameterised using a softplus function with argument  $\phi_i$ , with  $\sigma_i = \text{softplus}(\phi_i)$ . The base learning rate in Eq. 7 is  $\eta_{\text{base}} = 5 \times 10^{-4}$ . The  $\mu_i$ s were initialised homogeneously across the network from Uniform( $-0.1, 0.1$ ) and the  $\sigma_i$ s were initialised homogeneously across the network at  $10^{-4}$ . Hyperparameters were chosen via grid search on the validation dataset to enable smooth learning, high performance and rapid convergence. In the objective  $\mathcal{L}_{\text{BI}}$  used to train our simulations, we also add an L1 regularisation term over synaptic weights,  $\lambda |\mu|_1$ , where  $\lambda = 10^{-4}$ .

Plots in Fig. 2 present mappings from hyperparameter  $c$ , to accuracy and  $\sigma$ . A different neural network was trained for each  $c$ , after 50 training epochs the average  $\sigma$  across synapses was computed, and accuracy was evaluated on the test dataset. Plots in Fig. 3 present mappings of this  $\sigma$  against accuracy and reliability cost. The reliability cost was computed using fixed  $s = 1$  (see Appendix 51). To compute the Hessian in Fig. 5 and elsewhere, we used the empirical Fisher information approximation (Fisher, 1922),  $H \approx g^2$ . This was evaluated by taking the average  $g^2$  at  $w^* = \mu$  over ten epochs after full training for 50 epochs. The average learning rate  $\gamma |g|^{-1}$  and the average input rate  $|x|$  were also evaluated over ten epochs following training. The data presented illustrate these variables with regard to the weights of the second hidden layer. We set hyperparameter  $s = 0.001$  (see Eq. 51) in these simulations.

The experimental data in Fig. 6a were originally from data collected from plasticity experiments conducted by Sjöström et al. (2003). Costa et al. (2017) use this data to derive quantal parameters of a binomial model. Using the data processed by Costa et al. (2017) we calculated the normalised PSP variance before and after the experimental protocol, following the approach taken by Schug et al. (2021).

For geometric comparisons between the distribution over synapses and the Bayesian posterior presented in Fig. 7 we used the analytic results in Appendix - Analytic predictions for  $\sigma_i$ .

Source code used for simulations available at [github.com/JamesMalkin/EfficientBayes](https://github.com/JamesMalkin/EfficientBayes)

## Acknowledgments

We are grateful to Dr Stewart whose philanthropy supported GPU compute used in this project. JM was funded by the Engineering and Physical Sciences Research Council (2482786). COD was funded by the Leverhulme Trust (RPG-2019-229) and Biotechnology and Biological Sciences Research Council (BB/W001845/1). CH is supported by the Leverhulme Trust (RF-2021-533).

## References

- Aitchison L. Bayesian filtering unifies adaptive and non-adaptive neural network optimization methods. *Advances in Neural Information Processing Systems*. 2020; 33:18173–18182.
- Aitchison L, Jegminat J, Menendez JA, Pfister JP, Pouget A, Latham PE. Synaptic plasticity as Bayesian inference. *Nature Neuroscience*. 2021; 24(4):565–571.
- Aitchison L, Lengyel M. The Hamiltonian brain: Efficient probabilistic inference with excitatory-inhibitory neural circuit dynamics. *PLoS computational biology*. 2016; 12(12):e1005186.
- Attwell D, Laughlin SB. An energy budget for signaling in the grey matter of the brain. *Journal of Cerebral Blood Flow & Metabolism*. 2001; 21(10):1133–1145.
- Barber D, Bishop CM. Ensemble learning in Bayesian neural networks. *Nato ASI Series F Computer and Systems Sciences*. 1998; 168:215–238.
- Bellingham MC, Lim R, Walmsley B. Developmental changes in EPSC quantal size and quantal content at a central glutamatergic synapse in rat. *The Journal of Physiology*. 1998; 511(3):861–869.
- Berkes P, Orbán G, Lengyel M, Fiser J. Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. *Science*. 2011; 331(6013):83–87.
- Blei DM, Kucukelbir A, McAuliffe JD. Variational inference: A review for statisticians. *Journal of the American statistical Association*. 2017; 112(518):859–877.
- Blundell C, Cornebise J, Kavukcuoglu K, Wierstra D. Weight uncertainty in neural network. In: *International conference on machine learning* PMLR; 2015. p. 1613–1622.
- Bondy AG, Haefner RM, Cumming BG. Feedback determines the structure of correlated variability in primary visual cortex. *Nature neuroscience*. 2018; 21(4):598–606.
- Branco T, Staras K. The probability of neurotransmitter release: variability and feedback control at single synapses. *Nature Reviews Neuroscience*. 2009; 10(5):373–383.
- Bridgman P. Myosin Va movements in normal and dilute-lethal axons provide support for a dual filament motor complex. *The Journal of Cell Biology*. 1999; 146(5):1045–1060.
- Brock JA, Thomazeau A, Watanabe A, Li SSY, Sjöström PJ. A practical guide to using CV analysis for determining the locus of synaptic plasticity. *Frontiers in Synaptic Neuroscience*. 2020; 12:11.
- Chenouard N, Xuan F, Tsien RW. Synaptic vesicle traffic is supported by transient actin filaments and regulated by PKA and NO. *Nature communications*. 2020; 11(1):5318.
- Cingolani LA, Goda Y. Actin in action: the interplay between the actin cytoskeleton and synaptic efficacy. *Nature Reviews Neuroscience*. 2008; 9(5):344–356.
- Costa RP, Padamsey Z, D'Amour JA, Emptage NJ, Froemke RC, Vogels TP. Synaptic transmission optimization predicts expression loci of long-term plasticity. *Neuron*. 2017; 96(1):177–189.
- Dobrunz LE, Stevens CF. Heterogeneity of release probability, facilitation, and depletion at central synapses. *Neuron*. 1997; 18(6):995–1008.
- Echeveste R, Aitchison L, Hennequin G, Lengyel M. Cortical-like dynamics in recurrent circuits optimized for sampling-based probabilistic inference. *Nature neuroscience*. 2020; 23(9):1138–1149.
- Festa D, Aschner A, Davila A, Kohn A, Coen-Cagli R. Neuronal variability reflects probabilistic inference tuned to natural image statistics. *Nature communications*. 2021; 12(1):3635.
- Fiser J, Berkes P, Orbán G, Lengyel M. Statistically optimal perception and learning: from behavior to neural representations. *Trends in cognitive sciences*. 2010; 14(3):119–130.
- Fisher RA. On the mathematical foundations of theoretical statistics. *Philosophical transactions of the Royal Society of London Series A, containing papers of a mathematical or physical character*. 1922; 222(594-604):309–368.
- Fukushima K. Cognitron: A self-organizing multilayered neural network. *Biological cybernetics*. 1975; 20(3-4):121–136.
- Gentile JE, Carrizales MG, Koleske AJ. Control of synapse structure and function by actin and its regulators. *Cells*. 2022; 11(4):603.

- Goldman MS.** Enhancement of information transmission efficiency by synaptic failures. *Neural Computation*. 2004; 16(6):1137–1162.
- Gramlich MW, Klyachko VA.** Actin/Myosin-V-and activity-dependent inter-synaptic vesicle exchange in central neurons. *Cell Reports*. 2017; 18(9):2096–2104.
- Graves A.** Practical variational inference for neural networks. *Advances in Neural Information Processing Systems*. 2011; 24.
- Haefner RM, Berkes P, Fiser J.** Perceptual decision-making as probabilistic inference by neural sampling. *Neuron*. 2016; 90(3):649–660.
- Harris JJ, Jolivet R, Attwell D.** Synaptic energy use and supply. *Neuron*. 2012; 75(5):762–777.
- Harris JJ, Engl E, Attwell D, Jolivet RB.** Energy-efficient information transfer at thalamocortical synapses. *PLoS computational biology*. 2019; 15(8):e1007226.
- Heidelberger R, Heinemann C, Neher E, Matthews G.** Calcium dependence of the rate of exocytosis in a synaptic terminal. *Nature*. 1994; 371(6497):513–515.
- Hinton G, van Camp D.** Keeping neural networks simple by minimising the description length of weights. 1993. In: *Proceedings of COLT-93*; 1993. p. 5–13.
- Hoyer P, Hyvärinen A.** Interpreting neural response variability as Monte Carlo sampling of the posterior. *Advances in neural information processing systems*. 2002; 15.
- Jordan MI, Ghahramani Z, Jaakkola TS, Saul LK.** An introduction to variational methods for graphical models. *Machine learning*. 1999; 37:183–233.
- Karbowksi J.** Metabolic constraints on synaptic learning and memory. *Journal of Neurophysiology*. 2019; 122(4):1473–1490.
- Karunanithi S, Marin L, Wong K, Atwood HL.** Quantal size and variation determined by vesicle size in normal and mutant Drosophila glutamatergic synapses. *Journal of Neuroscience*. 2002; 22(23):10267–10276.
- Katz B, Miledi R.** The measurement of synaptic delay, and the time course of acetylcholine release at the neuromuscular junction. *Proceedings of the Royal Society of London Series B Biological Sciences*. 1965; 161(985):483–495.
- Khan M, Nielsen D, Tangkaratt V, Lin W, Gal Y, Srivastava A.** Fast and scalable bayesian deep learning by weight-perturbation in adam. In: *International conference on machine learning* PMLR; 2018. p. 2611–2620.
- Kingma DP, Ba J.** Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980. 2014; .
- Kingma DP, Salimans T, Welling M.** Variational dropout and the local reparameterization trick. *Advances in Neural Information Processing Systems*. 2015; 28.
- Knill DC, Pouget A.** The Bayesian brain: the role of uncertainty in neural coding and computation. *TRENDS in Neurosciences*. 2004; 27(12):712–719.
- Ko H, Cossell L, Baragli C, Antolik J, Clopath C, Hofer SB, Mrsic-Flogel TD.** The emergence of functional microcircuits in visual cortex. *Nature*. 2013; 496(7443):96–100.
- Lange RD, Chattoraj A, Beck JM, Yates JL, Haefner RM.** A confirmation bias in perceptual decision-making due to hierarchical approximate inference. *PLoS Computational Biology*. 2021; 17(11):e1009517.
- Lange RD, Haefner RM.** Characterizing and interpreting the influence of internal variables on sensory activity. *Current opinion in neurobiology*. 2017; 46:84–89.
- Lange RD, Haefner RM.** Task-induced neural covariability as a signature of approximate Bayesian learning and inference. *PLoS computational biology*. 2022; 18(3):e1009557.
- Laughlin SB, de Ruyter van Steveninck RR, Anderson JC.** The metabolic cost of neural information. *Nature Neuroscience*. 1998; 1(1):36–41.
- LeCun Y, Bottou L, Orr GB, Müller KR.** Efficient backprop. In: *Neural networks: Tricks of the trade* Springer; 2002.p. 9–50.
- Levy WB, Baxter RA.** Energy-efficient neuronal computation via quantal synaptic failures. *Journal of Neuroscience*. 2002; 22(11):4746–4755.
- Lillicrap TP, Santoro A, Marris L, Akerman CJ, Hinton G.** Backpropagation and the brain. *Nature Reviews Neuroscience*. 2020; 21(6):335–346.

- Lisman JE, Harris KM. Quantal analysis and synaptic anatomy—integrating two views of hippocampal plasticity. *Trends in Neurosciences*. 1993; 16(4):141–147.
- Ma WJ, Beck JM, Latham PE, Pouget A. Bayesian inference with probabilistic population codes. *Nature neuroscience*. 2006; 9(11):1432–1438.
- MacKay DJ. The evidence framework applied to classification networks. *Neural computation*. 1992; 4(5):720–736.
- MacKay DJ. A practical Bayesian framework for backpropagation networks. *Neural Computation*. 1992; 4(3):448–472.
- Martens J. New insights and perspectives on the natural gradient method. *The Journal of Machine Learning Research*. 2020; 21(1):5776–5851.
- Murphy KP. Machine learning: a probabilistic perspective. MIT press; 2012.
- Murthy VN, Sejnowski TJ, Stevens CF. Heterogeneous release properties of visualized individual hippocampal synapses. *Neuron*. 1997; 18(4):599–612.
- Orbán G, Berkes P, Fiser J, Lengyel M. Neural variability and sampling-based probabilistic representations in the visual cortex. *Neuron*. 2016; 92(2):530–543.
- Paulsen O, Heggelund P. Quantal properties of spontaneous EPSCs in neurones of the guinea-pig dorsal lateral geniculate nucleus. *The Journal of Physiology*. 1996; 496(3):759–772.
- Paulsen O, Heggelund P. The quantal size at retinogeniculate synapses determined from spontaneous and evoked EPSCs in guinea-pig thalamic slices. *The Journal of Physiology*. 1994; 480(3):505–511.
- Pulido C, Ryan TA. Synaptic vesicle pools are a major hidden resting metabolic burden of nerve terminals. *Science Advances*. 2021; 7(49):eabi9027.
- Purdon A, Rosenberger T, Shetty H, Rapoport S. Energy consumption by phospholipid metabolism in mammalian brain. *Neurochemical Research*. 2002; 27(12):1641–1647.
- Richards BA, Lillicrap TP. Dendritic solutions to the credit assignment problem. *Current opinion in neurobiology*. 2019; 54:28–36.
- Rosset S, Zhu J. Sparse, Flexible and Efficient Modeling using L 1 Regularization. *Feature Extraction: Foundations and Applications*. 2006; p. 375–394.
- Sacramento J, Ponte Costa R, Bengio Y, Senn W. Dendritic cortical microcircuits approximate the backpropagation algorithm. *Advances in neural information processing systems*. 2018; 31.
- Sacramento J, Wichert A, van Rossum MC. Energy efficient sparse connectivity from imbalanced synaptic plasticity rules. *PLoS Computational Biology*. 2015; 11(6):e1004265.
- Sakaba T, Neher E. Quantitative relationship between transmitter release and calcium current at the calyx of held synapse. *Journal of Neuroscience*. 2001; 21(2):462–476.
- Schug S, Benzing F, Steger A. Presynaptic stochasticity improves energy efficiency and helps alleviate the stability-plasticity dilemma. *eLife*. 2021; 10:e69884.
- Shannon CE. A mathematical theory of communication. *The Bell system technical journal*. 1948; 27(3):379–423.
- Shivkumar S, Lange R, Chatteraj A, Haefner R. A probabilistic population code based on neural samples. *Advances in neural information processing systems*. 2018; 31.
- Sjöström PJ, Turrigiano GG, Nelson SB. Neocortical LTD via coincident activation of presynaptic NMDA and cannabinoid receptors. *Neuron*. 2003; 39(4):641–654.
- Turrigiano GG, Leslie KR, Desai NS, Rutherford LC, Nelson SB. Activity-dependent scaling of quantal amplitude in neocortical neurons. *Nature*. 1998; 391(6670):892–896.
- Whittington JC, Bogacz R. An approximation of the error backpropagation algorithm in a predictive coding network with local hebbian synaptic plasticity. *Neural computation*. 2017; 29(5):1229–1262.
- Yang G, Pan F, Gan WB. Stably maintained dendritic spines are associated with lifelong memories. *Nature*. 2009; 462(7275):920–924.
- Yang Y, Li P. Synaptic Dynamics Realize First-order Adaptive Learning and Weight Symmetry. *arXiv preprint arXiv:221209440*. 2022; .
- Yu L, Zhang C, Liu L, Yu Y. Energy-efficient population coding constrains network size of a neuronal array system. *Scientific reports*. 2016; 6(1):19369.

## Appendix 1

### Reliability costs

The difficulty in determining reliability costs is that  $\sigma$  depends on three variables:  $n$ , the number of vesicles,  $p$  the probability of release and  $q$ , the quantal size, which measures the amount of neurotransmitter in each vesicle:

$$\sigma^2 = np(1-p)q^2. \quad (8)$$

However, these variables also determine the mean

$$\mu = npq \quad (9)$$

so a straightforward optimisation of  $\sigma$  under reliability costs will also change  $\mu$  and one pitfall is to accidentally consider only those changes in reliability that derive from changes in mean. A solution to this problem is to eliminate one of the variables so that  $\sigma$  is a function of  $\mu$ , and the remaining variables. Eliminating  $q$  gives

$$\sigma = \mu \sqrt{\frac{1-p}{np}} \quad (10)$$

The idea is to consider energetic costs associated with  $p$  and  $n$  and relate these to  $\sigma$  while holding  $\mu$  fixed. To simplify the biological motivation, we assume that during changes to synapse aimed at manipulating the energetic cost,  $q$  will also change to compensate for any collateral changes in  $\mu$  making  $\mu$  constant and  $q$  a “compensatory variable”. Moreover, it is possible, biologically, that  $q$  is the mechanism used in real synapses to fix  $\mu$  (*Turrigiano et al., 1998; Karunamithi et al., 2002*) and so using this is justifiable.

In what follows four different energy costs are considered, the first depends on  $p$ , the next two on  $n$ , the situation for the final one is less clear, but in each case we derive a reliability cost in the form  $\sigma^{-\rho}$  for some value of  $\rho$ . Of course, since we are considering costs for fixed  $\mu$  the coefficient of variation  $k = \sigma/\mu$  is proportional to  $\sigma$ ; it may be helpful to think of these calculations as finding the relationship between the cost and  $k$ .

#### **Calcium efflux – $\rho = \frac{1}{2}$**

Calcium influx into the synapse is an essential part of the mechanism for vesicle release. Presynaptic calcium pumps act to restore calcium concentrations in the synapse; this pumping is a significant portion of synaptic transmission costs (*Attwell and Laughlin, 2001*). By rearranging the Hill equation defined by *Sakaba and Neher 2001*, it can be shown that vesicle release has an interaction coefficient of four, this means the odds of release per vesicle are related to intracellular calcium amplitude via the fourth power (*Heidelberger et al., 1994; Sakaba and Neher, 2001*):

$$\frac{p}{1-p} \propto [Ca]^4. \quad (11)$$

To recover basal synaptic calcium concentration, the calcium influx is reversed by ATP-driven calcium pumps, where  $[Ca] \propto ATP$ :

$$\text{Calcium pump cost} \propto \sqrt[4]{\frac{p}{1-p}}. \quad (12)$$

Since this physiological cost does not depend on  $n$  we assume it is fixed and so the odds of release  $p/(1-p)$  is proportional to  $\sigma^2$ . Thus

$$\text{Calcium pump cost} \propto \frac{1}{\sigma^{\frac{1}{2}}} \quad (13)$$

or  $\rho = 1/2$ .

### Vesicle membrane – $\rho = \frac{2}{3}$

There is a cost associated with the total area of vesicle membrane. Evidence in *Pulido and Ryan 2021* suggest stored vesicles emit charged H<sup>+</sup> ions, with the number emitted proportional to the number of v-glut, glutamate transporters, on the surface of vesicles. v-ATPase pumps reverse this process maintaining the pH of the cell. It is suggested that this cost is 44% of the resting synaptic energy consumption. In addition, metabolism of the phospholipid bilayer that form the membrane of neurotransmitter filled vesicles has been identified as a major energetic cost (*Purdon et al., 2002*). Provided the total volume is the same, release of the same amount of neurotransmitter into the synaptic cleft can involve many smaller vesicles or fewer larger ones. However, while having many small vesicles will be more reliable, it requires a greater surface area of costly membrane. With fixed  $\mu$  and  $p$ ,

$$\text{Vesicle membrane cost} \propto nr^2 \quad (14)$$

Since  $r^2 \propto q^{2/3}$  and using  $\mu = npq$  this give

$$\text{Vesicle membrane cost} \propto \left( \frac{n\mu^2}{p^2} \right)^{1/3} \quad (15)$$

Since this reliability cost depends on  $n$ , so  $p$  is regarded as constant, so  $\sigma^2 \propto n$  and hence  $\rho = 2/3$ .

### Actin – $\rho = \frac{4}{3}$

Actin polymers are an energy costly structural filament that support the structural organisation of vesicle reserve pools (*Cingolani and Goda, 2008*), with the vesicles strung out along the filaments. We assume each vesicle to require a length of actin roughly proportional its diameter, this means that the total length of actin is proportional to  $nr$ . Hence

$$\text{Actin cost} \propto nr \quad (16)$$

The calculation then proceeds much as for the membrane cost, but with  $r$  instead of  $r^2$  giving  $\rho = 4/3$ .

### Trafficking – $\rho = 2$

ATP fueled myosin motors drive trains of actin filament along with associated cargo such as vesicles and actin-myosin trafficking moves vesicles from vesicle reserve pools to release sites sustaining the readily releasable pool (RRP) following vesicle release (*Bridgman, 1999; Gramlich and Klyachko, 2017*). This gives a cost for vesicle recruitment proportional to  $np$ , the number of vesicles released:

$$\text{Trafficking cost} \propto np \quad (17)$$

so if  $n$  is regarded as the principal way this cost is changed, with  $p$  fixed then  $n \propto \sigma^{-2}$  and so  $\rho = 2$ . This is the view point we are taking to motivate examining reliability costs with  $\rho = 2$ . This is certainly useful in considering the range of model behaviours over a broad a range of  $\rho$  values.

Nonetheless, it is sensible to ask whether the likely biological mechanism behind a varying trafficking cost is one which changes  $np$  itself. In this case, since a constant  $\mu$  for varying  $nr$  means  $q \propto 1/np$  we have

$$\text{Trafficking cost} \propto \frac{1-p}{\sigma^2} \quad (18)$$

which, is, again, of the form  $\text{cost} \propto \sigma^{-\rho}$  provided  $p$  is small. For larger  $p$ , however, it depends on exactly how  $p$  changes as  $np$  changes.

Generally, throughout these calculation we have supposed that  $q$  is a compensatory variable and that either  $p$  or  $n$  changes in the process that changes the cost at a synapse. The benefit of this is that we are able to model costs directly in terms of reliability; in the future, though, it might be interesting to consider models which use  $n$ ,  $p$  and  $q$  instead of  $\mu$  and  $\sigma$ ; this would certainly be convenient for the sort of comparisons we are making here and interesting, although two formulations seem equivalent, this does not mean that the learning dynamics will be the same.

## Appendix 2

### High input rates and high precision at important synapses

Using a generalised linear model, we demonstrate that synapse variability optimised under the performance-reliability cost tradeoff depends on the input rate and the Hessian. To do this, consider a set of predictor-response pairs  $(y_a, \mathbf{x}_a)$  where  $a$  labels the data points, and each data point  $\mathbf{x}_a$  is a vector over features we label with index  $i$ . Logistic regression is an attempt to approximate  $y_a$  as a function of  $\mathbf{x}_a$ :

$$f(y_a) \approx \mathbf{x}_a^T \mathbf{w} \quad (19)$$

where  $\mathbf{w}$  are the regression coefficients which correspond with synaptic weights,  $f$  is the link function, in the example of logistic regression this would be the logit function. One approach to regression is to minimise the normalised square error in this approximation:

$$\text{performance cost} = E_{Q(\mathbf{w})} [-\log P(y|\mathbf{w}, \mathbf{x})] = E_{Q(\mathbf{w})} \left[ \frac{1}{2\epsilon^2} \sum_a (f(y_a) - \mathbf{x}_a^T \mathbf{w})^2 \right] + \text{const} \quad (20)$$

Network weights are drawn independently from a normal distribution,  $Q(\mathbf{w})$ ,

$$w_i \sim \text{Normal}(\mu_i, \sigma_i) \quad (21)$$

To solve for optimal  $\sigma$  in addition to  $\mu$  we optimise  $\mathcal{L}_{\text{BI}} = \text{performance cost} + \text{reliability cost}$ .

$$\mathcal{L}_{\text{BI}}(\boldsymbol{\mu}, \boldsymbol{\sigma}) = E_{Q(\mathbf{w})} \left[ \frac{1}{2\epsilon^2} \sum_a (v_a - \mathbf{x}_a^T \mathbf{w})^2 \right] + \frac{1}{\rho} s^\rho \sum_i \sigma_i^{-\rho} \quad (22)$$

where  $v_a = f(y_a)$ . Also note that  $\epsilon^2$  is the expected square error, the variance of  $P(y|\mathbf{w}, \mathbf{x})$ . For brevity we have disregarded the magnitude cost since it is a cost on the synapse mean while our focus in this section is on how synapse variance relates to input rate and synapse importance.

Expanding the brackets and using the commutativity of the dot-product,  $\mathbf{w}^T \mathbf{x}_a \mathbf{x}_a^T \mathbf{w} = \mathbf{x}_a^T \mathbf{w} \mathbf{w}^T \mathbf{x}_a$ , this gives

$$\mathcal{L}_{\text{BI}}(\boldsymbol{\mu}, \boldsymbol{\sigma}) = \frac{1}{2\epsilon^2} \sum_a [-2v_a \mathbf{x}_a^T E[\mathbf{w}] + \mathbf{x}_a^T E[\mathbf{w} \mathbf{w}^T] \mathbf{x}_a] + \frac{1}{\rho} s^\rho \sum_i \sigma_i^{-\rho} + \text{const} \quad (23)$$

Now  $E[\mathbf{w} \mathbf{w}^T] = \boldsymbol{\mu} \boldsymbol{\mu}^T + \boldsymbol{\Sigma}$  where  $\boldsymbol{\Sigma}$  is the multivariate covariance matrix for the  $\sigma_i$ ; since these are independent  $\Sigma_{ii} = \sigma_i$  and the other entries are zero. With this

$$\mathcal{L}_{\text{BI}} = \frac{1}{2\epsilon^2} \sum_a (-2v_a \mathbf{x}_a^T \boldsymbol{\mu} + \mathbf{x}_a^T (\boldsymbol{\mu} \boldsymbol{\mu}^T + \boldsymbol{\Sigma}) \mathbf{x}_a) + \frac{1}{\rho} s^\rho \sum_i \sigma_i^{-\rho} + \text{const} \quad (24)$$

To obtain the Hessian we first take the first derivative with respect to  $\boldsymbol{\mu}$ ,

$$\nabla_{\boldsymbol{\mu}} \mathcal{L}_{\text{BI}} = \frac{1}{2\epsilon^2} \sum_a [-2v_a \mathbf{x}_a^T + \mathbf{x}_a^T (\nabla_{\boldsymbol{\mu}} \boldsymbol{\mu} \boldsymbol{\mu}^T) \mathbf{x}_a] = \frac{1}{\epsilon^2} \sum_a [-v_a \mathbf{x}_a^T + \boldsymbol{\mu}^T \mathbf{x}_a \mathbf{x}_a^T] \quad (25)$$

We can set the derivative to zero and solve to get the least squares formula:

$$\boldsymbol{\mu} = \frac{1}{\epsilon^2} \mathbf{A}^{-1} \mathbf{c} \quad (26)$$

where for convenience we write  $\mathbf{A} = \sum_a \mathbf{x}_a \mathbf{x}_a^T$  and  $\mathbf{c} = \sum_a v_a \mathbf{x}_a$ . Note that  $\mathbf{A}$  is the auto-correlation matrix of the neural network input evaluated over data-points.

The second derivative gives the Hessian for synaptic weights:

$$H_{ij} = \frac{d^2}{d\mu_i d\mu_j} \mathcal{L}_{\text{BI}} = \frac{1}{\epsilon^2} A_{ij} \quad (27)$$

We can use this to calculate the trace of the Hessian,

$$\text{tr } \mathbf{H} = \frac{1}{\epsilon^2} \text{tr } \mathbf{A} = \frac{1}{\epsilon^2} \sum_i \sum_a x_{ai}^2 \quad (28)$$

To find  $\sigma_i$  that optimises the tradeoff we take the derivative with respect to  $\sigma_i$ ; writing in index form we have

$$\mathcal{L}_{\text{BI}} = \frac{1}{2\epsilon^2} \left[ \sum_{ij} A_{ij}(\mu_i \mu_j + \Sigma_{ij}) \right] + \frac{1}{\rho} s^\rho \sum_i \sigma_i^{-\rho} + \text{const} \quad (29)$$

so

$$\frac{d}{d\sigma_i} \mathcal{L}_{\text{BI}} = \frac{1}{2\epsilon^2} 2A_{ii}\sigma_i - s^\rho \sigma_i^{-\rho-1} = 0 \quad (30)$$

which means

$$\sigma_i^{\rho+2} = \frac{\epsilon^2 s^\rho}{A_{ii}} \quad (31)$$

or

$$\sigma_i = \sqrt[\rho+2]{\epsilon^2 s^\rho / A_{ii}} = \sqrt[\rho+2]{s^\rho / H_{ii}} = \sqrt[\rho+2]{\epsilon^2 s^\rho / \sum_a x_{ai}^2} \quad (32)$$

Hence, for generalised linear models, the  $\sigma_i$  that solves or optimises the tradeoff is related to the size of synapse input rate,  $\sum_a x_{ai}^2$  through a power law relationship depending on  $\rho$ , with larger input giving smaller variance. This relationship is particularly clear for  $\rho = 2$  since this simplifies the expression for  $\sigma_i^2$ :

$$\sigma_i^2 = s\epsilon H_{ii}^{-\frac{1}{2}} = s\epsilon \left( \sum_a x_{ai}^2 \right)^{-\frac{1}{2}}. \quad (33)$$

## Appendix 3

### Synapse importance and gradient magnitudes

In our ANN simulations we train synaptic weights using the most established adaptive optimisation scheme, Adam (*Kingma and Ba, 2014*); which has recently been realised using biologically plausible mechanisms (*Yang and Li, 2022*). Adam uses a synapse-specific learning rate,  $\eta_i$ , which decreases in response to high gradients at that synapse,

$$\eta_i = \frac{\eta_{\text{base}}}{\sqrt{\langle g_i^2 \rangle}}. \quad (34)$$

Specifically, the local learning rate for the  $i$ th synapse,  $\eta_i$ , is usually a base learning rate,  $\eta_{\text{base}}$ , divided by  $\sqrt{\langle g_i^2 \rangle}$ , the root-mean-square of the gradient for each datapoint / minibatch.

The key intuition is that if the gradients for each datapoint/minibatch are large, that means that this synapse is important, as it has a big impact on the predictions for every datapoint. In fact, these mean-square gradients can be related to our formal measure of synapse importance, the Hessian. Specifically, for data generated from the model, the Hessian (or Fisher information *Fisher, 1922*) is equivalent to the mean squared gradient, where the gradient is taken over each datapoint separately,

$$E_{P(\mathbf{d}|\mathbf{w})} \left[ \frac{\partial^2 \log P(\mathbf{d}|\mathbf{w})}{\partial^2 \mathbf{w}} \right] = E_{P(\mathbf{d}|\mathbf{w})}[\mathbf{g}\mathbf{g}^T] \quad (35)$$

Of course, in practice, the data is not drawn from the model, so the relationship between the squared gradients and the Hessian computed for real data is only approximate. But it is close enough to induce a relationship between synapse importance (measured as the diagonal of the Hessian) and learning rates (which are inversely proportional to the root-mean-square gradients) in our simulations.

## Appendix 4

### Energy efficient noise and variational-Bayes for neural network weights

**Introduction to Variational Bayes for neural network weights** One approach to performing Bayesian inference for the weights of a neural network is to use variational Bayes. In variational Bayes, we introduce a parametric approximate posterior,  $Q(w)$ , and fit the parameters of that approximate posterior using gradient descent on an objective, the ELBO (*Blundell et al., 2015*). In particular,

$$\log P(y|x) \geq \text{ELBO} = E_{Q(w)}[\log P(y|x, w) + \log P(w)] + H[Q(w)]. \quad (36)$$

where  $H[Q(w)]$  is the entropy of the approximate posterior. Maximising the ELBO is particularly useful for selecting models as it forms a lower-bound on the marginal-likelihood, or evidence,  $\log P(y|x)$  (*MacKay, 1992a; Barber and Bishop, 1998*). When using variational Bayes for neural network weights, we usually use Gaussian approximate posteriors (*Blundell et al., 2015*),

$$Q(w_i) = \mathcal{N}(w_i; \mu_i, \sigma_i^2). \quad (37)$$

Note that optimising the ELBO wrt the parameters of  $Q$  is difficult, because  $Q$  is the distribution over which the expectation is taken in Eq. (36). To circumvent this issue, we use the reparameterisation trick (*Kingma et al., 2015*), which involves writing the weights in terms of IID standard Gaussian variables,  $\epsilon$ ,

$$w_i = \mu_i + \sigma_i \epsilon_i. \quad (38)$$

Thus, we can write the ELBO as an expectation over  $\epsilon$ , which has a fixed IID standard Gaussian distribution,

$$\text{ELBO} = E_\epsilon [\log P(y|x, w=\mu+\sigma\epsilon) + \log P(w=\mu+\sigma\epsilon)] + H[Q(w)]. \quad (39)$$

Note that we use  $w$ ,  $\mu$  etc. without indices in this expression to indicate all the weights / mean weights.

**Identifying the log-likelihood and log-prior.** Following the usual practice in deep learning, we assume the likelihood is formed by a categorical distribution, with probabilities obtained by applying the softmax function to the output of the neural network. The log-likelihood for a categorical distribution with softmax probabilities is the negative cross-entropy (*Murphy, 2012*),

$$\log P(y|x, w) = -\text{cross entropy}(y; f(x, w)) \quad (40)$$

where  $f(x, w)$  is the output of the network with weights,  $w$  and inputs,  $x$ . We can additionally identify a Laplace prior with the magnitude cost. Specifically, if we take the prior to be Laplace, with scale  $1/\lambda$ ,

$$\log P(w) = \sum_i \log \text{Laplace}(w_i; 0, \frac{1}{\lambda}) \quad (41)$$

$$= -N \log(2/\lambda) - \lambda \sum_i |w_i| \quad (42)$$

$$= \text{const} - \text{magnitude cost} \quad (43)$$

where we identify the magnitude cost using Eq. (3).

**Connecting the entropy and the biological reliability cost.** Now, we have identified the log-likelihood and log-prior terms in the ELBO (Eq. 36) with terms in our biological cost (Eq. 2). We thus have two terms left: the entropy in the ELBO (Eq. 36) and the reliability cost in the biological cost (Eq. 2). Critically, the entropy term also acts as a reliability cost, in that it encourages more variability in the weights (as the entropy is positive in the ELBO (Eq. 36) and we are trying to maximise the ELBO, the entropy term gives a bonus for more variability). Intuitively, we can think of the entropy term in the ELBO (Eq. 36) as being an ‘‘entropic reliability cost’’, as compared to the ‘‘biological reliability cost’’ in (Eq. 2).

This intuitive connection suggests that we might be able to find a more formal link. Specifically, we can define an entropic reliability cost as simply the negative entropy,

$$C_{VI;i} = -H[Q_i] = -\frac{1}{2} \ln 2\pi e \sigma_i^2. \quad (44)$$

Our goal is to write the biological reliability cost,  $C_{BI;i}$ , as a bound on  $C_{VI;i}$ . By rearranging and introducing  $s$  and  $\rho$ ,

$$C_{VI;i} = \frac{1}{\rho} \ln \left( \frac{s}{\sigma_i} \right)^\rho - \frac{1}{2} \ln 2\pi e s^2, \quad (45)$$

and noting that  $\log a \leq a - 1$ ,

$$\log \left( \frac{s}{\sigma_i} \right)^\rho \leq \left( \frac{s}{\sigma_i} \right)^\rho - 1 \quad (46)$$

we can demonstrate that any reliability cost expressed as a generic power-law forms an upper bound on the entropic cost,

$$C_{VI;i} \leq \frac{1}{\rho} \left( \left( \frac{s}{\sigma_i} \right)^\rho - 1 \right) - \frac{1}{2} \ln 2\pi e s^2 = C_{BI;i} = \text{reliability cost}_i + \text{const}_i. \quad (47)$$

where,

$$\text{const}_i = -\frac{1}{\rho} - \frac{1}{2} \ln 2\pi e s^2 \quad \text{reliability cost}_i = \frac{1}{\rho} \left( \frac{s}{\sigma_i} \right)^\rho = c \sigma_i^{-\rho} \quad (48)$$

$$\text{const} = \sum_i \text{const}_i \quad \text{reliability cost} = \sum_i \text{reliability cost}_i. \quad (49)$$

$$C_{VI} = \sum_i C_{VI;i} \quad C_{BI} = \sum_i C_{BI;i}. \quad (50)$$

The parameter  $c$  sets the importance of the reliability cost within the performance-reliability cost tradeoff (see Fig. 2)

$$c = s^\rho / \rho. \quad (51)$$

Note that while both  $\text{reliability cost}_i$  and  $C_{BI;i}$  represent the biological reliability costs, they are slightly different in that  $C_{BI;i}$  includes an additive constant. Importantly, this additive constant is independent of  $\sigma_i$  and  $\mu_i$ , it does not affect learning and can be ignored.

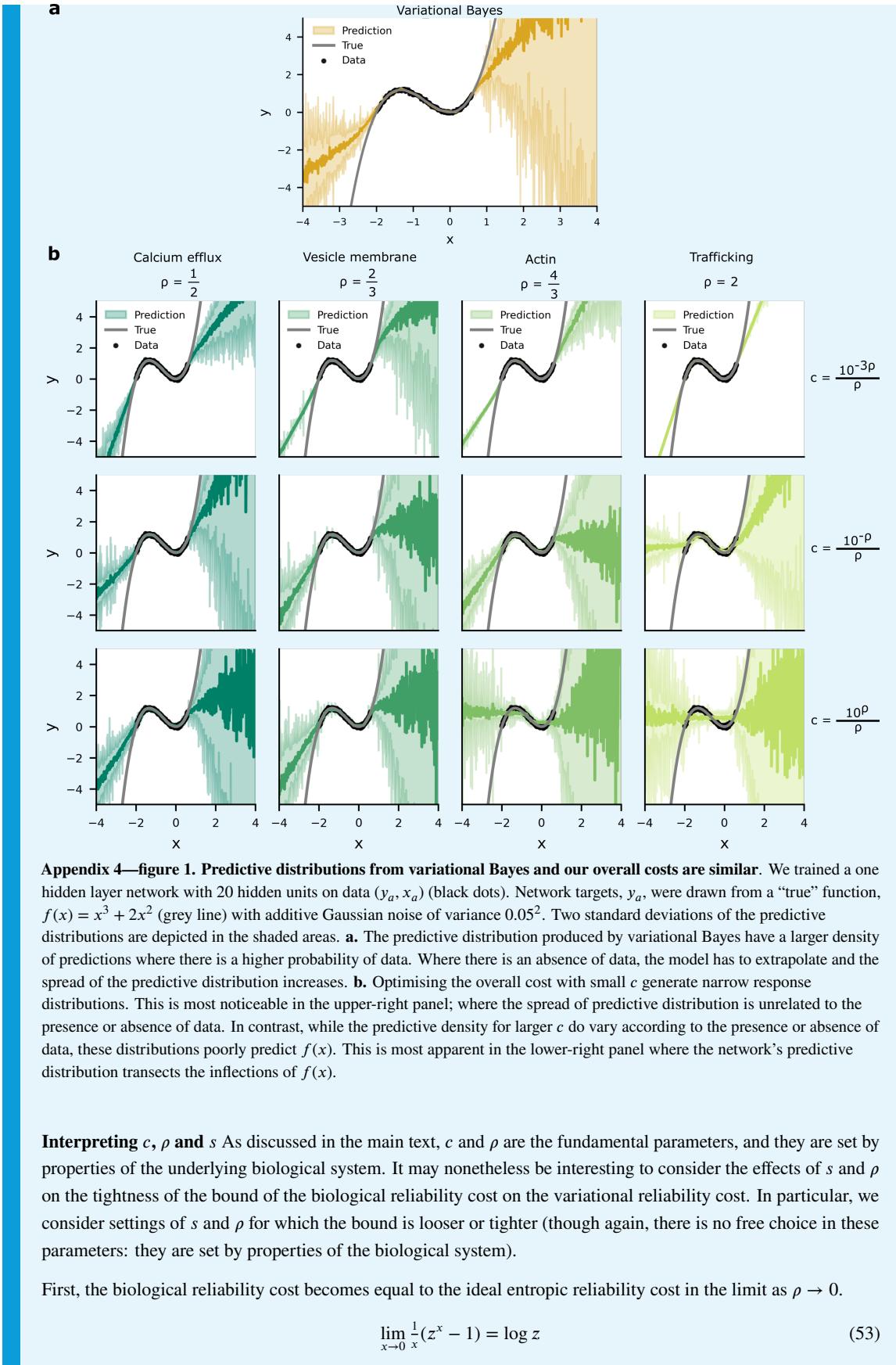
Given that the biological reliability cost forms a bound on the ideal entropic reliability cost we can consider using the biological reliability cost in place of the entropic reliability cost,

$$\log P(y|x) \geq \text{ELBO} = E_{Q(w)}[\log P(y|x, w) + \log P(w)] - C_{VI} \quad (52)$$

$$\log P(y|x) \geq \text{ELBO} \geq E_{Q(w)}[\log P(y|x, w) + \log P(w)] - C_{BI} = -\text{overall cost} - \text{const}$$

we find that our overall biological cost (Eq. 2) forms a bound on the ELBO, which itself forms a bound on the evidence,  $\log P(y|x)$ . Thus, pushing down the overall biological cost (Eq. 2) pushes up a bound on the model evidence.

**Predictive probabilities arising from biological reliability costs** Given the connections between our overall cost and the ELBO, we expect that optimising our overall cost will give a similar result to variational Bayes. To check this connection, we plotted the distribution of predictions induced by noisy weights arising from variational Bayes (Appendix 4–Fig. 1a) and our overall costs (Appendix 4–Fig. 1b). Variational Bayes maximises the ELBO, therefore its predictive distribution is optimised to reflect the data distribution from which data is drawn (*MacKay, 1992b*). We found comparable patterns for the predictive distributions learned through variational Bayes and our overall costs, albeit with some breakdown in predictive performance with higher values for  $\rho$ .



**Appendix 4—figure 1. Predictive distributions from variational Bayes and our overall costs are similar.** We trained a one hidden layer network with 20 hidden units on data  $(y_a, x_a)$  (black dots). Network targets,  $y_a$ , were drawn from a “true” function,  $f(x) = x^3 + 2x^2$  (grey line) with additive Gaussian noise of variance  $0.05^2$ . Two standard deviations of the predictive distributions are depicted in the shaded areas. **a.** The predictive distribution produced by variational Bayes have a larger density of predictions where there is a higher probability of data. Where there is an absence of data, the model has to extrapolate and the spread of the predictive distribution increases. **b.** Optimising the overall cost with small  $c$  generate narrow response distributions. This is most noticeable in the upper-right panel; where the spread of predictive distribution is unrelated to the presence or absence of data. In contrast, while the predictive density for larger  $c$  do vary according to the presence or absence of data, these distributions poorly predict  $f(x)$ . This is most apparent in the lower-right panel where the network’s predictive distribution transects the inflections of  $f(x)$ .

**Interpreting  $c$ ,  $\rho$  and  $s$**  As discussed in the main text,  $c$  and  $\rho$  are the fundamental parameters, and they are set by properties of the underlying biological system. It may nonetheless be interesting to consider the effects of  $s$  and  $\rho$  on the tightness of the bound of the biological reliability cost on the variational reliability cost. In particular, we consider settings of  $s$  and  $\rho$  for which the bound is looser or tighter (though again, there is no free choice in these parameters: they are set by properties of the biological system).

First, the biological reliability cost becomes equal to the ideal entropic reliability cost in the limit as  $\rho \rightarrow 0$ .

$$\lim_{x \rightarrow 0} \frac{1}{x} (z^x - 1) = \log z \quad (53)$$

Thus, taking  $\rho = x$ , and  $z = s/\sigma_i$ ,

$$\lim_{\rho \rightarrow 0} \frac{1}{\rho} \left( \left( \frac{s}{\sigma_i} \right)^\rho - 1 \right) = \log \frac{s}{\sigma_i} \quad (54)$$

Thus,

$$\lim_{\rho \rightarrow 0} C_{\text{BI};i} = \log \frac{s}{\sigma_i} - \frac{1}{2} \log 2\pi e s^2 = -\frac{1}{2} \log 2\pi e \sigma_i^2 = C_{\text{VI};i}. \quad (55)$$

This explains the apparent improvement in predictive performance (Appendix 4–Fig. 1) and in matching the posteriors (Fig. 7) with lower values of  $\rho$ .

Second, the biological reliability cost becomes equal to the ideal entropic reliability cost when  $s = \sigma_i$ ,

$$C_{\text{BI};i}(s = \sigma_i) = -\frac{1}{2} \log 2\pi e s^2 = -\frac{1}{2} \log 2\pi e \sigma_i^2 = C_{\text{VI};i}. \quad (56)$$

as the first term in Eq. 47 cancels. However,  $s$  cannot be set individually across synapses, but is instead roughly constant, with a value set by underlying biological constraints. In particular,  $s$  can be written as a function of  $\rho$  and  $c$  (Eq. 51), and  $\rho$  and  $c$  are quantities that are roughly constant across synapses, with their values set by biological constraints. Thus, biological implications of a tightening bound as  $s$  tends to  $\sigma_i$  are not clear.

## Appendix 5

### Analytic predictions for $\sigma_i$

At various points in the main text, we note a connection between the Hessian, synapse importance and optimal variability. To develop a formal understanding of these connections, we perform a second-order Taylor expansion of the performance cost and magnitude cost, considering only the quadratic term, as that is the only term that depends on the variances (the other terms are incorporated in const),

$$\text{performance cost} + \text{magnitude cost} \quad (57)$$

$$= E_{Q(\mathbf{w})} \left[ -\frac{1}{2} (\mathbf{w} - \boldsymbol{\mu})^T \mathbf{H} (\mathbf{w} - \boldsymbol{\mu}) \right] + \text{const} \quad (58)$$

Writing this as a Trace,

$$= E_{Q(\mathbf{w})} \left[ -\frac{1}{2} \text{Tr} ((\mathbf{w} - \boldsymbol{\mu})^T \mathbf{H} (\mathbf{w} - \boldsymbol{\mu})) \right] + \text{const} \quad (59)$$

Cyclically permuting the Trace,

$$= E_{Q(\mathbf{w})} \left[ -\frac{1}{2} \text{Tr} (\mathbf{H} (\mathbf{w} - \boldsymbol{\mu}) (\mathbf{w} - \boldsymbol{\mu})^T) \right] + \text{const} \quad (60)$$

Noting that the weights are the only stochastic quantities,

$$= -\frac{1}{2} \text{Tr} (\mathbf{H} E_{Q(\mathbf{w})} [(\mathbf{w} - \boldsymbol{\mu}) (\mathbf{w} - \boldsymbol{\mu})^T]) + \text{const} \quad (61)$$

And noting that the biological synaptic noise (e.g. from vesicle release) is independent,

$$= -\frac{1}{2} \sum_i H_{ii} \sigma_i^2 + \text{const}. \quad (62)$$

Now, combining the reliability cost, with the performance and magnitude cost gives the overall cost,

$$\text{overall cost} = \sum_i \left[ -\frac{1}{2} H_{ii} \sigma_i^2 - c \sigma_i^{-\rho} \right] + \text{const}. \quad (63)$$

Note that under this quadratic approximation, the dependence of the overall cost on  $\sigma_i$  has become decoupled across synapses. We can therefore compute the optimal synaptic noise level,  $\sigma_i$  as a function only of the Hessian for that synapse,  $H_{ii}$ . In particular, we solve for where the gradient of the cost is zero,

$$0 = \frac{d}{d\sigma_i} \text{overall cost} = -H_{ii} \sigma_i + c \rho \sigma_i^{-\rho-1} \quad (64)$$

Solving for  $\sigma_i$  this give

$$\sigma_i^{\rho+2} = \frac{c\rho}{H_{ii}} \quad (65)$$

or

$$\sigma_i = \sqrt[\rho+2]{\frac{c\rho}{H_{ii}}}. \quad (66)$$

Thus,

$$\log \sigma_i^2 = -\frac{2}{\rho+2} \log H_{ii} + \text{const}. \quad (67)$$

which is borne out in Appendix 5–Fig. 1a. Now that we know that optimal synapse variance scales with the Hessian, we can combine this with other quantities that scale with the Hessian; such as the learning rate, input

rate or the Bayesian posterior variance. Allowing us to derive how these quantities scale with optimised synapse variance.

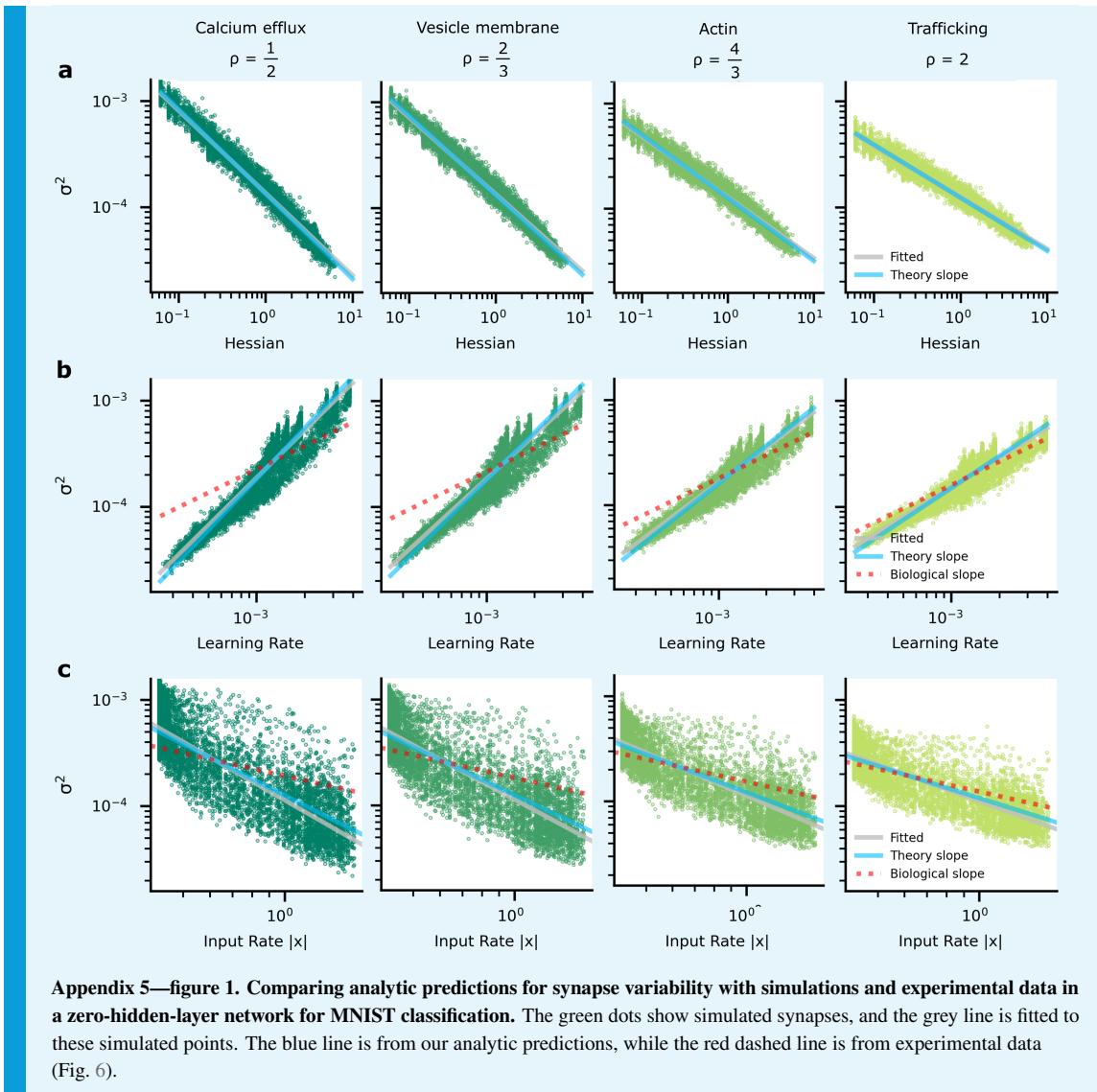
$$H_{ii} \propto \langle g_i^2 \rangle \propto \frac{1}{\eta_i^2} \quad \log H_{ii} = -2 \log \eta_i + \text{const} \quad \log \sigma_i^2 = \frac{4}{\rho+2} \log \eta_i + \text{const} \quad (68a)$$

$$H_{ii} \propto \langle x_i^2 \rangle \quad \log H_{ii} = \log \langle x_i^2 \rangle \quad \log \sigma_i^2 = -\frac{2}{\rho+2} \log \langle x_i^2 \rangle + \text{const} \quad (68b)$$

$$H_{ii} \propto \frac{1}{\sigma_{\text{post};i}^2} \quad \log H_{ii} = -\log \sigma_{\text{post};i}^2 + \text{const} \quad \log \sigma_i^2 = \frac{2}{\rho+2} \log \sigma_{\text{post};i}^2 + \text{const}. \quad (68c)$$

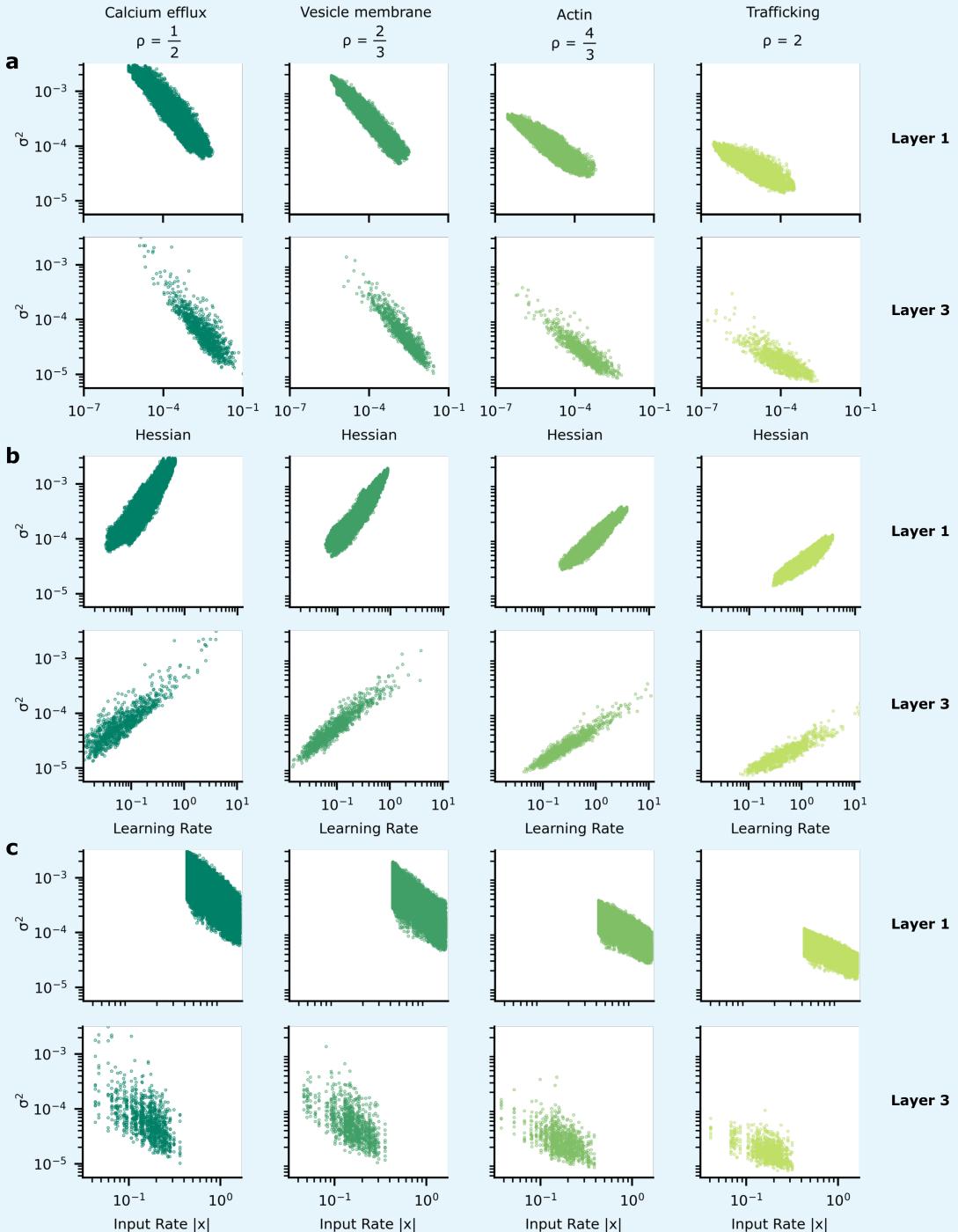
The first row uses the analytic relationship between the mean squared gradient and the Hessian (Eq. 35) to obtain a relationship between the optimal synaptic noise level and the learning rate providing the slope in Appendix 5–Fig. 1b. The second row expresses the analytic relationship between the Hessian and the mean square of the input rate (see Appendix - High input rates and high precision at important synapses), and gives the analytic relationship in Appendix 5–Fig. 1c). The third row expresses the analytic relationship between the Hessian and the Bayesian posterior variance, used in Fig. 7.

To test these predictions, we performed a simpler simulation classifying MNIST in a network with no hidden layers. We found that the analytic results closely matched the simulations, and that the biological slopes tend to match the  $\rho = 2$  better than the other values for  $\rho$ . However, while the direction of the slope was consistent in deeper networks, the exact value of the slope was not consistent (Fig. 5 and Supplementary - Appendix 6–Fig. 1), so it is unclear whether we can draw any strong conclusions here.



## Appendix 6

### Supplementary figures



**Appendix 6—figure 1. Patterns of synapse variability for the remaining layers of the neural network used to provide our results.** In Fig. 5 we showed the heterogeneous patterns of synapse variability for the synapses connecting the two hidden layers of our ANN. Here we exhibit the equivalent plots for the other synapses, those between the input and the first hidden layer (layer 1) and from the final hidden layer to the output layer (layer 3). As in Fig. 5 we show the relationship between synaptic variance and **a.** the Hessian; **b.** learning rate; and **c.** input rate. The patterns do not appear substantially different from layer to layer.

