eLife

Reviewed Preprint
v2 • June 4, 2024
Revised by authors

Reviewed Preprint
v1 • December 12, 2023

Neuroscience

Signatures of Bayesian inference emerge from energy efficient synapses

James Malkin ✉, **Cian O'Donnell, Conor Houghton, Laurence Aitchison**

Faculty of Engineering, University of Bristol, Bristol, UK • 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.

eLife assessment

This **important** study provides deep insight into a ubiquitous, but poorly understood, phenomenon: synaptic noise (primarily due to failures). Through a combination of theoretical analysis, simulations, and comparison to existing experimental data, this paper makes a **compelling** case that synapses are noisy because reducing noise is expensive. It touches on probably the most significant feature of living organisms -- their ability to learn -- and will be of broad interest to the neuroscience community.

https://doi.org/10.7554/eLife.92595.2.sa2

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 $\boldsymbol{n}$ docked, readily releasable vesicles (see **Fig. 1a** for an illustration of a single synaptic connection with multi-vesicular release). An alternative interpretation of this model might consider n the number of uni-vesicular connections between two neurons. 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, μ, and variance, σ^2, of the PSP can be written (see **Fig. 1b**),

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

where $\boldsymbol{q}$ is considered a scaling variable. An assertion in our model is that variability in PSP strength is the result of variable numbers of vesicle release, not variability in $\boldsymbol{q}$; here, during any PSP, $\boldsymbol{q}$ is assumed constant across vesicles. While there is some suggestion that intra- and inter-site variability in $\boldsymbol{q}$ is a significant component of PSP variability (see Silver(2003)) we ultimately expect quantal variability to be small relative to the variability attributed to vesicular release. This is supported by the classic observation that PSP amplitude histograms have a multi-peak structure (Boyd and Martin, 1956; Holler et al., 2021); and by more direct measurement and modelling of vesicle release (Forti et al., 1997; Raghavachari and Lisman, 2004).

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, $\boldsymbol{p}$, 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 $\boldsymbol{\alpha\sigma^{-1/2}}$ (**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

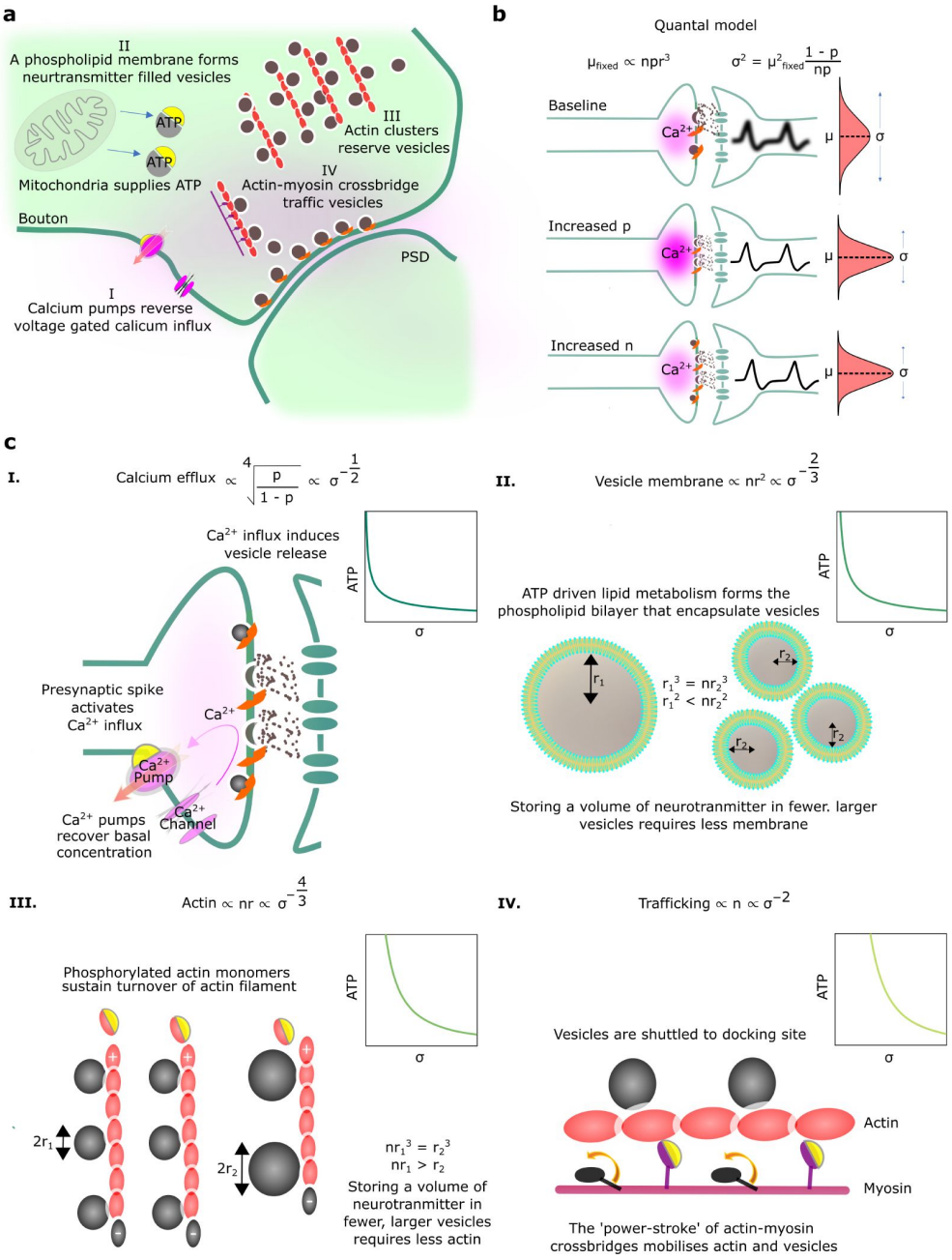


Figure 1.

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* ɑ 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. Note that *r* represents the vesicle radius. **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.

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 $\boldsymbol{\alpha\sigma}^{-2/3}$ (**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 $\boldsymbol{\alpha\sigma}^{-4/3}$ (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 microtubles 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 $\boldsymbol{\alpha\sigma}^{-2}$ (**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 (Atwell and Laughlin, 2001). These can be formalised as costs on the PSP mean, $\boldsymbol{\mu}$, 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,

$$\text{overall cost} = \text{performance cost} + \text{magnitude cost} + \text{reliability cost}. \tag{2}$$

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,

$$\text{magnitude cost} = \lambda \sum_i |\mu_i|, \tag{3}$$

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

Here, i indexes synapses, and recall that σ_i is the standard deviation of the ith 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,

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

where, recall, μ_i is the PSP mean and σ_i^2 is the PSP variance for the ith synapse. The output firing rate was given by,

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

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 μ_i and σ_i using Adam (Kingma and Ba, 2014) (see Methods for details on architecture and hyperparameters).

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 ρ motivated by the biophysical costs (the different columns).

In practice all the reliability costs and others we may have overlooked should together constitute an overall energetic reliability cost. However, it is difficult to estimate the specific contributions of different costs, that is the individual values of c. While Attwell and Laughlin (2001); Engl and Attwell (2015) estimate the ATP demands for various synaptic processes, it is difficult to relate these to the relative scale of each cost at a synapse level. Therefore, for simplicity, we kept each cost separate, training neural networks with just one choice of reliability cost; emphasising results shared across all costs. It is possible that one cost dominates all the others, but if that is not the case it will be necessary to use a more complicated reliability cost. However, since we have considered four costs with very different power-law behaviours, it is likely the behaviour will not be significantly different to what we have observed.

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 σ_i is optimised but kept the same across all synapses (grey lines). Second, we considered an heterogeneous noise setting, where σ_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.

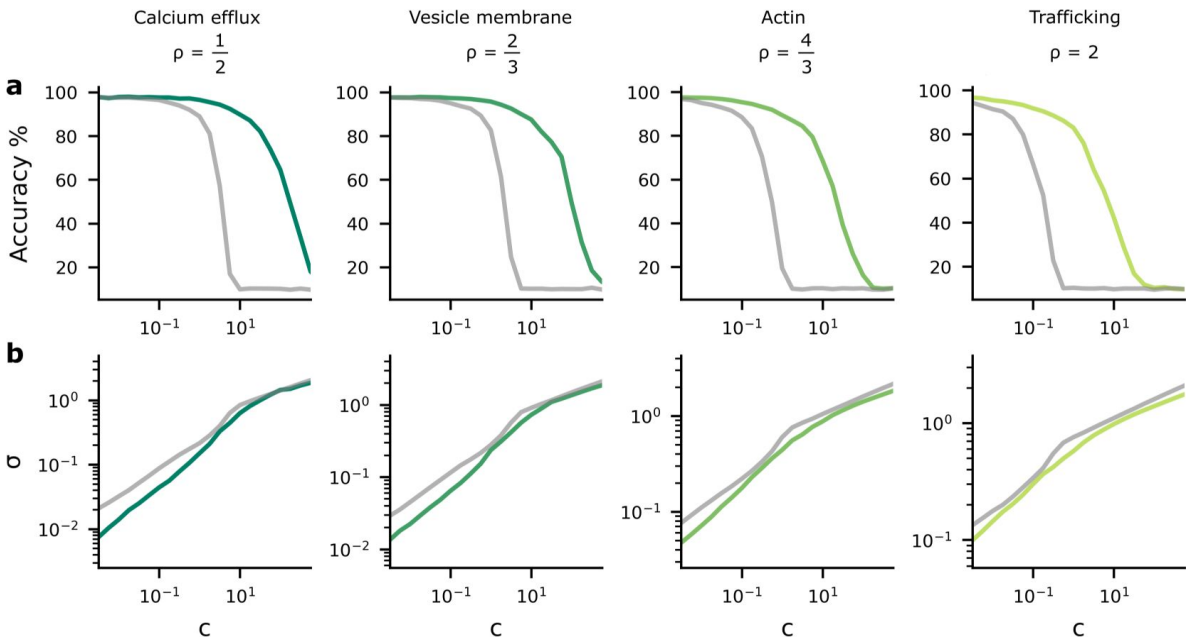


Figure 2.

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, *σ*) increases.

The findings in **Fig. 2** imply a tradeoff between accuracy and average noise level, σ, 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 ρ 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 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, H. This describes the shape of the objective as a function of the synaptic weights, the w_is: 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 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**). The Hessian values were estimated using the average-squared gradient, see **Appendix - Synapse importance and gradient magnitudes**. We found that as the importance of the synapse increased, the optimised noise

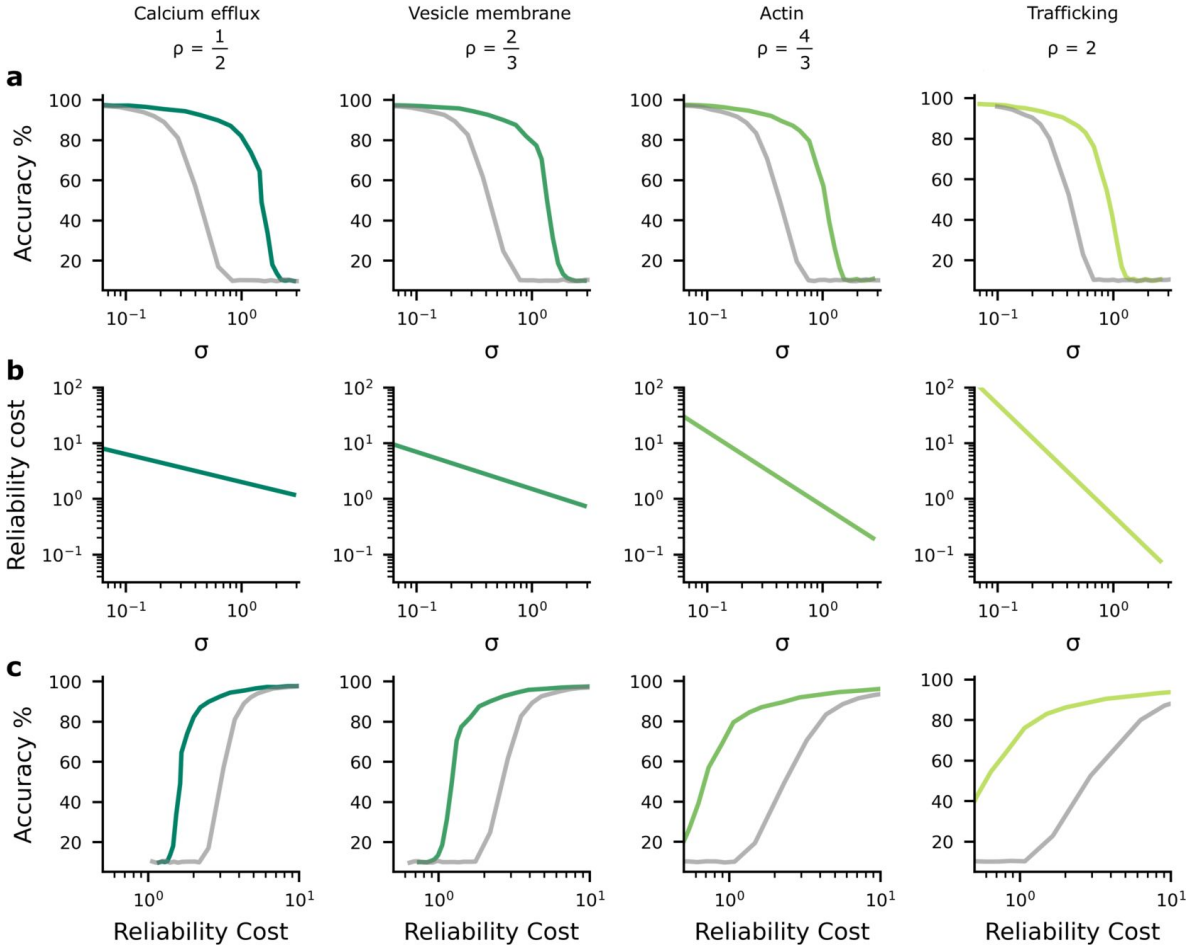


Figure 3.

The performance-reliability cost tradeoff in ANN simulations

a. Accuracy decreases as the average PSP standard deviation, σ, 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 σ^2 leads to greater reliability costs, and this relationship is different for different biophysical mechanisms and hence values for ρ (columns). **c.** Higher accuracy therefore implies larger reliability cost.

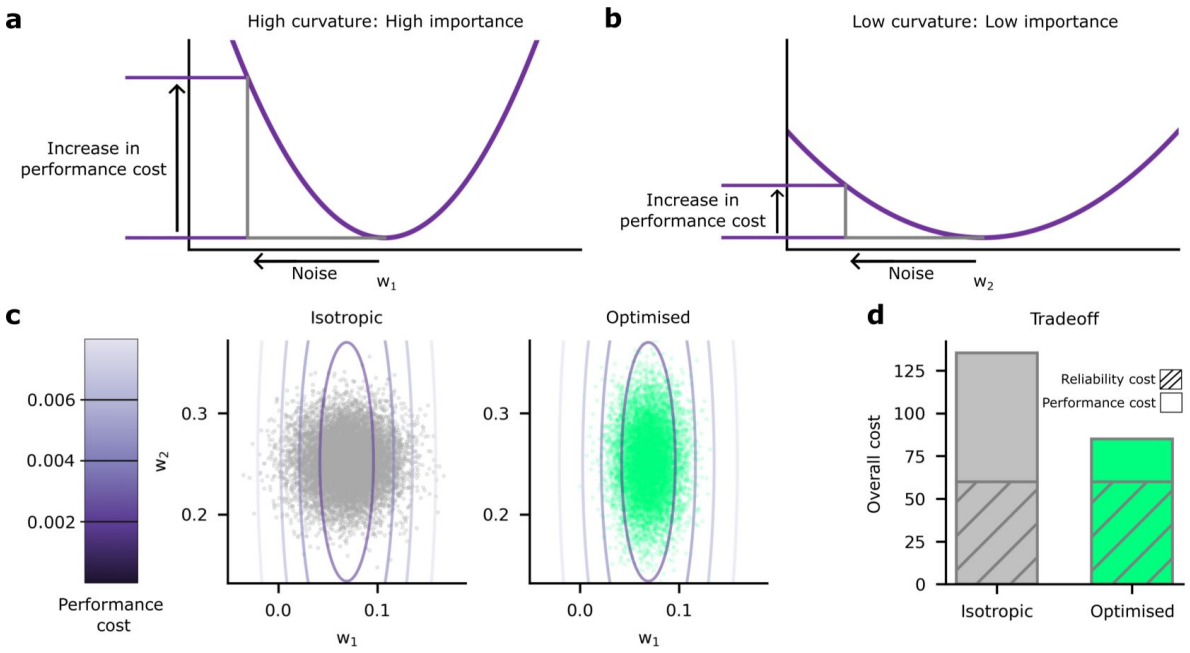


Figure 4.

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.

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, similar patterns have been observed in experimental data. In **Fig. 6a** we present the negative correlation between learning rate and synaptic reliability presented by Schug et al. (2021) from in-vitro measurements of V1 (layer 5) pyramidal synapses before and after STDP induced LTP conducted by Sjostrom et al. (2001). Furthermore, a relationship between input firing rate and synaptic variability was observed by Aitchison et al. (2021) using in-vivo functional recordings from V1 (layer 2/3) (Ko et al., 2013) (**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.

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. While the specific scheme for modulating the learning rate is ultimately an assumption, we believe modern deep learning offers strong guidance as to the optimal family of schemes for modulating the learning rate. In particular, modern, state-of-the-art, update rules for artificial neural networks almost always use an adaptive learning rate. These adaptive learning rates, $\boldsymbol{\eta_i}$, (including the most common such as Adam and variants) almost always use a normalising learning rate which decreases in response to high incoming gradients,

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

Specifically, the local learning rate for the $\boldsymbol{i}$th synapse, $\boldsymbol{\eta_i}$, is usually a base learning rate, $\boldsymbol{\eta}_{\text{base}}$, divided by the root-meansquare 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 (see **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; Jegminat et al., 2022).

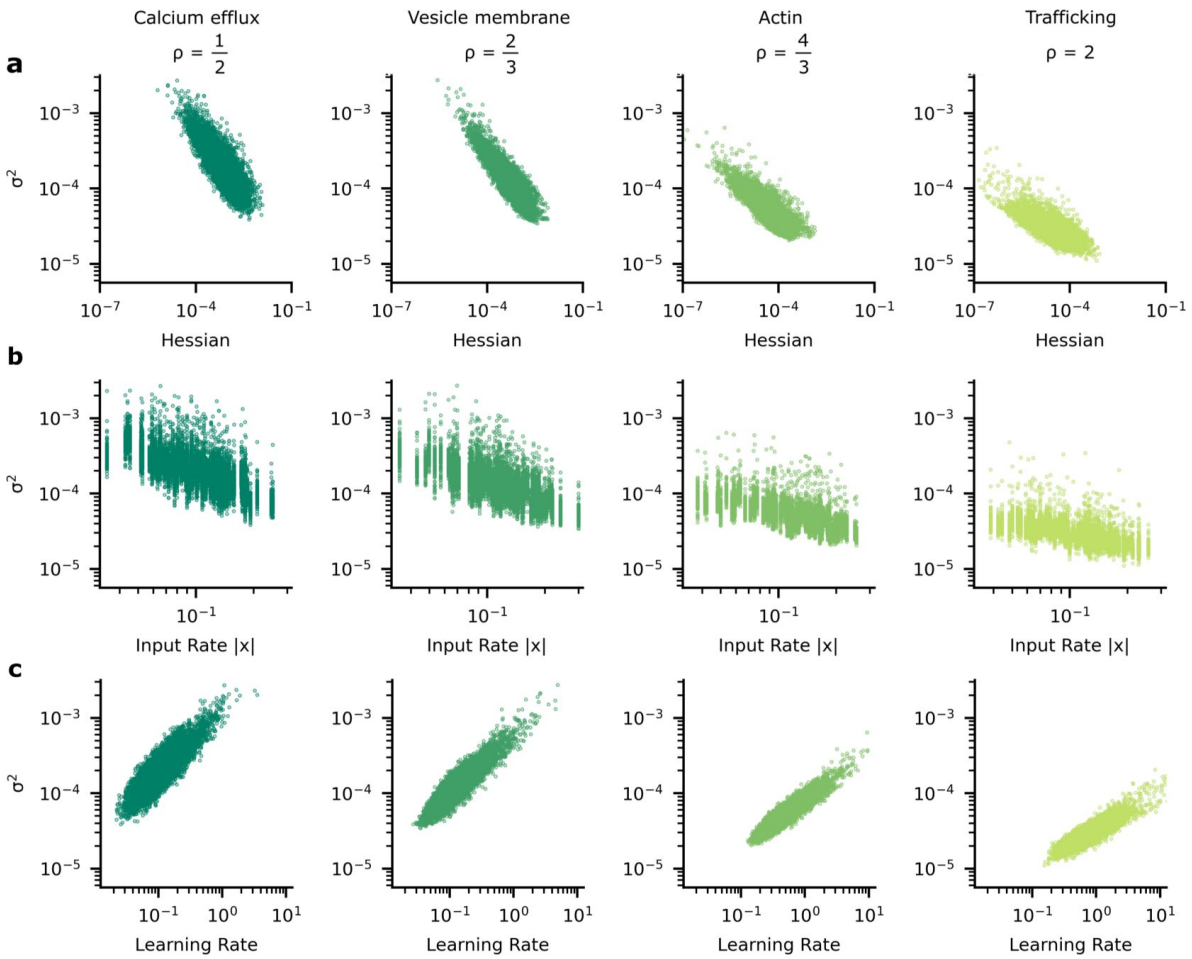


Figure 5.

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.

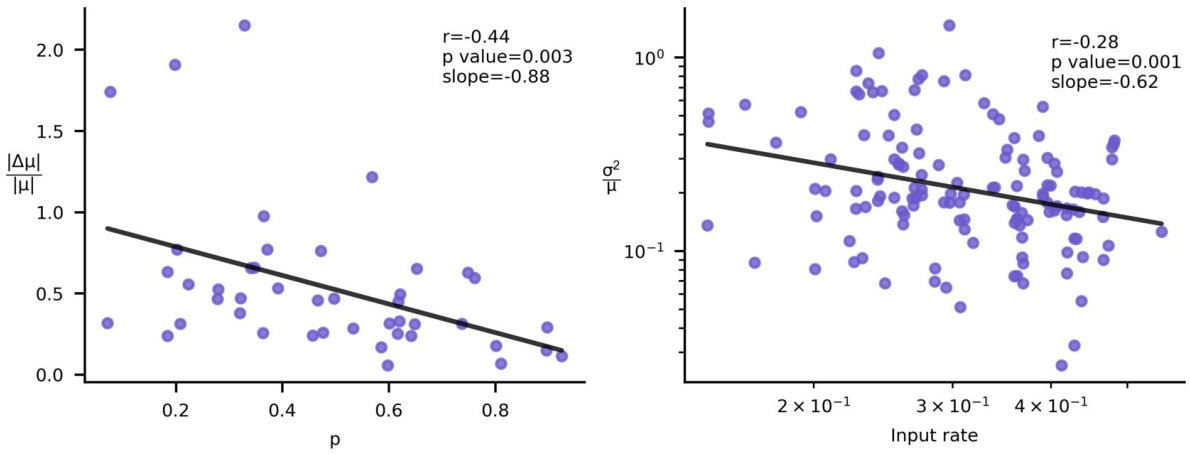


Figure 6.

Experimental signatures of Bayesian synapses

The Bayesian synapse hypothesis predicts relationships between synapse reliability, learning rate and input rate. **a**) Synapses with higher probability of release, *p* demonstrate smaller increases in synaptic mean following LTP induction. This pattern was originally observed by Schug et al. (2021)). **b)** As input firing rates are increased, normalised EPSP variability decreases with *slope* = *-0.62* (Aitchison etal., 2021)).

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**) match those arising from Bayesian synapses presented in **Fig. 6** (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 estimated and plotted the posterior variance against the optimised synaptic variability (**Fig. 7a**) (see Methods). We considered our four different biophysical mechanisms (values for $\boldsymbol{\rho}$; **Fig. 7a**, columns), and values for $\boldsymbol{c}$ (**Fig. 7a**, rows). In all cases, there was a clear correlation between the posterior and the optimised variability: synapses with larger posterior variance also had large optimised variance. To further assess this connection, we used the relation between the Hessian and posterior variance given by **Eq. 54c** and the analytic result given in **Appendix - Analytic predictions for σ_i** to plot the relationships between σ_i and the posterior variability, σ_{post} as a function of ρ (**Fig. 7b**;) and as a function of ρ (**Fig. 7c**). Again, these plots show a clear correlation between synapse variance 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 with slope equal to one. In contrast, these lines actually have a slope smaller than one, indicating that optimised variability is less heterogeneous than posterior variance (**Fig. 7bc**). Interestingly, the slope increases towards one as the associated ρ is decreased, this suggests that synapse variability best approximates the posterior when ρ is small.

This strong, but not perfect, connection between the patterns of variability in Bayesian inference and energy-efficient networks motivated us to seek a formal connection between Bayesian and efficient synapses. As such, in the Appendix, we derive a theoretical connection between our overall performance cost and Bayesian inference (see **Appendix - Energy efficient noise and variational-Bayes for neural network weights**). Moreover, this connection is subsequently used to provide an explanation for why synapse variability aligns closer to posterior variance for small ρ (see **Eq. 51**). 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)

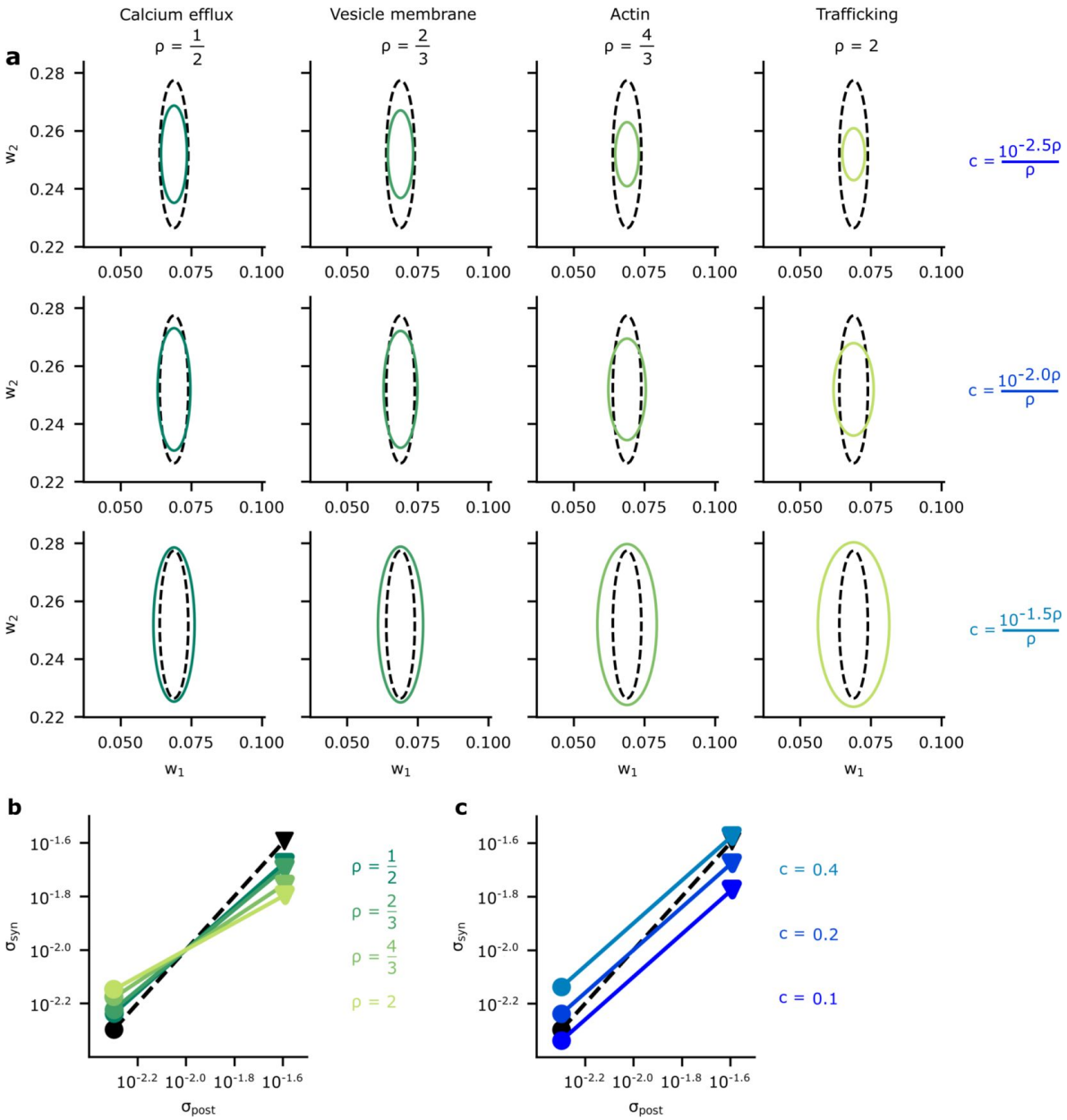


Figure 7.

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 ρ (columns) and c (rows) for an illustrative pair of synapses. Note that using fixed values of c for different ρ's dramatically changed the scale of the ellipses. Instead, we chose c as a function of ρ to ensure that the scale of the optimised noise variance was roughly equal across different ρ. This allowed us to highlight the key pattern: that smaller values for ρ give optimised variance closer to the true posterior variances, while higher values for ρ 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 ρ. 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.48** 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 ρ =1/2 across different settings for c)

(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 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, $\boldsymbol{c}$, which sets the magnitude of the reliability cost. In fact, the patterns of variability exhibited in **Fig. 5** are preserved as $\boldsymbol{c}$ is changed: this was demonstrated for values of $\boldsymbol{c}$ 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.

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 Hyvarinen, 2002; Knill and Pouget, 2004; Ma et al., 2006; Fiser et al., 2010; Berkes etal., 2011; Orban etal., 2016; Aitchison and Lengyel, 2016; Haefner etal., 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 ξ, ~ Normal(0, 1) (Kingma etal., 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 $\boldsymbol{\sigma}$s, they were re-parameterised using a softplus function with argument φ_i, with $\sigma_i = \text{softplus}(\varphi_i)$. The base learning rate in Eq.7 is $\eta_{\text{base}} = 5 \times 10^{-4}$. The μ_is were initialised homogeneously across the network from Uniform(−0.1, 0.1) and the σ_is 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 L_{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 $\boldsymbol{c}$, to accuracy and σ. A different neural network was trained for each c, after 50 training epochs the average σ across synapses was computed, and accuracy was evaluated on the test dataset. Plots in **Fig. 3** present mappings of this σ against accuracy and reliability cost. The reliability cost was computed using fixed $s = 1$ (see Appendix 48).

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.48) in these simulations.

To estimate the posterior used in **Fig. 7** we optimised a factorised Gaussian approximation to the posterior over weights using variational inference and Bayes by back-propagation (Blundell et al., 2015). We then took σ_{post} from two optimised weights. For the variance and slope comparisons between Bayesian and efficient synapses in **Fig. 7** we used the analytic results in **Appendix - Analytic predictions for σ_i**.

Source code used for simulations available at github.com/JamesMalkin/EfficientBayes

Acknowledgements

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).

Appendix 1 Reliability costs

The difficulty in determining reliability costs is that σ 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:

$$\sigma^2 = np(1-p)q^2. \tag{8}$$

However, these variables also determine the mean

$$\mu = npq \tag{9}$$

so a straighforward optimisation of σ under reliability costs will also change μ 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 σ is a function of μ, and the remaining variables. Eliminating q gives

$$\sigma = \mu\sqrt{\frac{1-p}{np}} \tag{10}$$

The idea is to consider energetic costs associated with p and n and relate these to σ while holding μ fixed. To simplify the biological motivation, we assume that during changes to the synapse aimed at manipulating the energetic cost, $\boldsymbol{q}$ will also change to compensate for any collateral changes in μ keeping μ constant. Hence, $\boldsymbol{q}$ a "compensatory variable". Moreover, there is biological evidence that q is the mechanism used in real synapses to fix μ (Turrigiano etal., 1998; Karunanithi etal., 2002). Fixing μ through a compensatory variable is termed homeostatic plasticity. Fixing μ through q is described as "quantal scaling". For reviews on homeostatic plasticity see Turrigiano and Nelson (2004); Davis and Müller (2015).

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 ρ. Of course, since we are considering costs for fixed μ the coefficient of variation $k = \sigma/\mu$ is proportional to σ; 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):

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

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

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

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 σ^{-2}. Thus

$$\text{Calcium pump cost} \propto \frac{1}{\sigma^{\frac{1}{2}}} \tag{13}$$

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+ 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 μ and p,

$$\text{Vesicle membrane cost} \propto nr^2 \tag{14}$$

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

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

Since this reliability cost depends on n, so p is regarded as constant, so σ^{-2} α n and hence $p = 2/3$.

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 filamants. 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

$$\text{Actin cost} \propto nr \tag{16}$$

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

Trafficking – ρ = 2

ATP fueled myosin motors drive trains of actin filament along with associated cargo such as vesicles and actin-myson 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:

$$\text{Trafficking cost} \propto np \tag{17}$$

so if n is regarded as the principal way this cost is changed, with p fixed then n α σ^{-2} and so $\mu = 2$. This is the view point we are taking to motivate examining reliability costs with $\mu = 2$. This is certainly useful in considering the range of model behaviours over a broad a range of μ 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 μ for varying ***np*** means $q \propto 1/np$ we have

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

which, is, again, of the form 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 μ and σ; 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

Here, we show that under a linear model, important synapses, as measured by the Hessian, have high input rates. Additionally, we show that energy efficient synapses implies that these important synapses have low optimized variability.

We consider a simplified linear model, with targets y, inputs X and weights w. We have,

$$\mathrm{P}(\boldsymbol{y}|\boldsymbol{w},\boldsymbol{X}) = \mathcal{N}(\boldsymbol{y};\boldsymbol{X}\boldsymbol{w},\epsilon^2\boldsymbol{I}). \tag{19}$$

We take the performance cost to be,

$$\text{performance cost} = -\log \mathrm{P}(\boldsymbol{y}|\boldsymbol{w},\boldsymbol{X}). \tag{20}$$

where N is the number of datapoints. We take the network weights to be drawn from a multivariate Gaussian, with diagonal covariance, $\sum$, i.e. $\Sigma_{ii} = \sigma_i^2$,

$$\mathrm{Q}(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w};\boldsymbol{\mu},\boldsymbol{\Sigma}). \tag{21}$$

All our derivations rely on looking at the quadratic form for the performance cost. In particular,

$$\text{performance cost} = \mathrm{E}_{\mathrm{Q}(\boldsymbol{w})}\left[\frac{1}{2\epsilon^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w})\right] \tag{22}$$

Expanding the brackets,

$$\text{performance cost} = \frac{1}{2\epsilon^2}(\boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{y}^T\boldsymbol{X}\,\mathrm{E}[w] + E[\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w}]), \tag{23}$$

and evaluating the expectations,

$$\text{performance cost} = \frac{1}{2\epsilon^2}[\boldsymbol{y}\boldsymbol{y}^T - 2\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\mu} + \mathrm{tr}(\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\mu}] \tag{24}$$

Using this form, we can identify the Hessian, and consider the tradeoff between the performance cost and reliability costs.

To measure synapse importance, we use the Hessian, i.e. the second derivative of the performance cost wrt the mean. From Eq. (24), we can identify this as,

$$H_{ii} = \frac{\partial^2 \text{performance cost}}{\partial \mu_i^2} = \frac{1}{\epsilon^2} \sum_t X_{ti}^2. \tag{25}$$

where t indexes time.

To understand the tradeoff between reliability and performance costs, note that the only term in the performance cost that depends on the variability is $\text{tr}(\boldsymbol{X}^T \boldsymbol{X} \boldsymbol{\Sigma})$, and this can be written,

$$\text{tr}(\boldsymbol{X}^T \boldsymbol{X} \boldsymbol{\Sigma}) = \sum_i \sum_t \sigma_i^2 X_{ti}^2. \tag{26}$$

Thus,

$$\frac{d}{d\sigma_i}[\text{performance cost} + \text{reliability cost}] = H_{ii}\sigma_i - s^\rho \sigma_i^{-\rho-1} = 0 \tag{27}$$

which means

$$\sigma_i^{\rho+2} = s^\rho / H_{ii} \tag{28}$$

or

$$\sigma_i = \sqrt[\rho+2]{s^\rho / H_{ii}} \tag{29}$$

Thus, more important synapses (as measured by the Hessian, H_{ii}) have lower variability if the synapse is energy efficient. Moreover, through Eq. (25), we expect synapses with higher input rates to be more important.

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, η_i, which decreases in response to high gradients at that synapse,

$$\eta_i = \frac{\eta_{\text{base}}}{\sqrt{\langle g_i^2 \rangle}}. \tag{30}$$

Specifically, the local learning rate for the ith synapse, η_i, is usually a base learning rate, η_{base}, divided by $\sqrt{\langle g_i^2 \rangle}$, the root-mean-square of the gradient for each datapoint or 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,

$$E_{P(y,\boldsymbol{x}|\boldsymbol{w})}\left[\frac{\partial^2 \text{performance cost}}{\partial \boldsymbol{\mu}^2}\right] = -E_{P(y,\boldsymbol{x}|\boldsymbol{w})}\left[\frac{\partial^2 \log P(y|\boldsymbol{w},\boldsymbol{x})}{\partial \boldsymbol{\mu}^2}\right] = E_{P(y,\boldsymbol{x}|\boldsymbol{w})}[\boldsymbol{g}\boldsymbol{g}^T], \tag{31}$$

where the expectation is evaluated over data generated by the model and g is defined,

$$\boldsymbol{g} = \frac{\partial}{\partial \boldsymbol{\mu}} \log P(y|\boldsymbol{w},\boldsymbol{x}) \tag{32}$$

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, $\boldsymbol{Q}(w)$, and fit the parameters of that approximate posterior using gradient descent on an objective, the ELBO (Blundell et al., 2015). In particular,

$$\log P(y|x) \geq \text{ELBO} = E_{Q(w)}[\log P(y|x,w) + \log P(w)] + H[Q(w)]. \tag{33}$$

where $\boldsymbol{H}[\boldsymbol{Q}(w)]$ 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 etal., 2015),

$$Q(w_i) = \mathcal{N}(w_i; \mu_i, \sigma_i^2). \tag{34}$$

Note that optimising the ELBO wrt the parameters of $\boldsymbol{Q}$ is difficult, because $\boldsymbol{Q}$ is the distribution over which the expectation is taken in Eq. (33). To circumvent this issue, we use the reparameterisation trick (Kingma etal., 2015), which involves writing the weights in terms of IID standard Gaussian variables, $\in$,

$$w_i = \mu_i + \sigma_i \epsilon_i. \tag{35}$$

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

$$\text{ELBO} = E_{\epsilon}\left[\log P(y|x, w{=}\mu{+}\sigma\epsilon) + \log P(w{=}\mu{+}\sigma\epsilon)\right] + H[Q(w)]. \tag{36}$$

Note that we use w, μ etc. without indicies 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),

$$\log P(y|x, w) = -\text{cross entropy}(y; f(x, w)) \tag{37}$$

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 $1/\lambda$,

$$\log P(w) = \sum_i \log \text{Laplace}(w_i; 0, \tfrac{1}{\lambda}) \tag{38}$$

$$= -N \log(2/\lambda) - \lambda \sum_i |w_i| \tag{39}$$

$$= \text{const} - \text{magnitude cost} \tag{40}$$

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. 33) with terms in our biological cost (Eq. 2). We thus have two terms left: the entropy in the ELBO (Eq. 33) 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. 33) 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. 33) as being an "entropic reliabity 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,

$$C_{\text{VI;i}} = -H[Q_i] = -\tfrac{1}{2} \ln 2\pi e \sigma_i^2. \tag{41}$$

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

$$C_{\text{VI;i}} = \frac{1}{\rho} \ln \left(\frac{s}{\sigma_i}\right)^{\rho} - \tfrac{1}{2} \ln 2\pi e s^2, \tag{42}$$

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

$$\log\left(\frac{s}{\sigma}\right)^{\rho} \leq \left(\frac{s}{\sigma_i}\right)^{\rho} - 1 \tag{43}$$

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

$$C_{\mathrm{VI;i}} \leq \frac{1}{\rho}\left(\left(\frac{s}{\sigma_i}\right)^{\rho} - 1\right) - \tfrac{1}{2}\log 2\pi e s^2 = C_{\mathrm{BI;i}} = \text{reliability cost}_i + \text{const}_i. \tag{44}$$

Here, $C_{\mathrm{BI;i}}$ is the biological reliability cost for a synapse, which formally bounds the entropic reliability cost from VI,

$$\text{const}_i = -\tfrac{1}{\rho} - \tfrac{1}{2}\log 2\pi e s^2 \qquad \text{reliability cost}_i = \frac{1}{\rho}\left(\frac{s}{\sigma_i}\right)^{\rho} = c\sigma_i^{-\rho} \tag{45}$$

We can also sum these quantities across synapses,

$$\text{const} = \sum_i \text{const}_i \qquad \text{reliability cost} = \sum_i \text{reliability cost}_i. \tag{46}$$

$$C_{\mathrm{VI}} = \sum_i C_{\mathrm{VI;i}} \qquad C_{\mathrm{BI}} = \sum_i C_{\mathrm{BI;i}}. \tag{47}$$

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

$$c = s^{\rho}/\rho. \tag{48}$$

Note that while both reliability cost_i and $C_{\mathrm{BI;i}}$ represent the biological reliability costs, they are slightly different in that $C_{\mathrm{BI;i}}$ includes an additive constant. Importantly, this additive constant is independent of σ_i and μ_i, so 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,

$$\begin{aligned} &\log P(y|x) \geq \mathrm{ELBO} = E_{Q(w)}[\log P(y|x,w) + \log P(w)] - C_{\mathrm{VI}} \\ &\log P(y|x) \geq \mathrm{ELBO} \geq E_{Q(w)}[\log P(y|x,w) + \log P(w)] - C_{\mathrm{BI}} = -\text{overall cost} - \text{const} \end{aligned} \tag{49}$$

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 ρ.

Interpreting *c*, *ρ* and *s*

As discussed in the main text, c and ρ 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 ρ on the tightness of the bound of the biological reliability cost on the variational reliability cost. In particular, we consider settings of s and ρ 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 $\rho \to 0$.

$$\lim_{x\to 0} \frac{1}{x}(z^x - 1) = \log z \tag{50}$$

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

$$\lim_{\rho\to 0} \frac{1}{\rho}\left(\left(\frac{s}{\sigma_i}\right)^{\rho} - 1\right) = \log \frac{s}{\sigma_i} \tag{51}$$

Thus,

$$\lim_{\rho\to 0} C_{\mathrm{BI;i}} = \log \frac{s}{\sigma_i} - \frac{1}{2}\log 2\pi e s^2 = -\frac{1}{2}\log 2\pi e \sigma_i^2 = C_{\mathrm{VI;i}}. \tag{52}$$

This explains the apparent improvement in predictive performance (**Appendix 4-Fig. 1**) and in matching the posteriors (**Fig. 7**) with lower values of ρ.

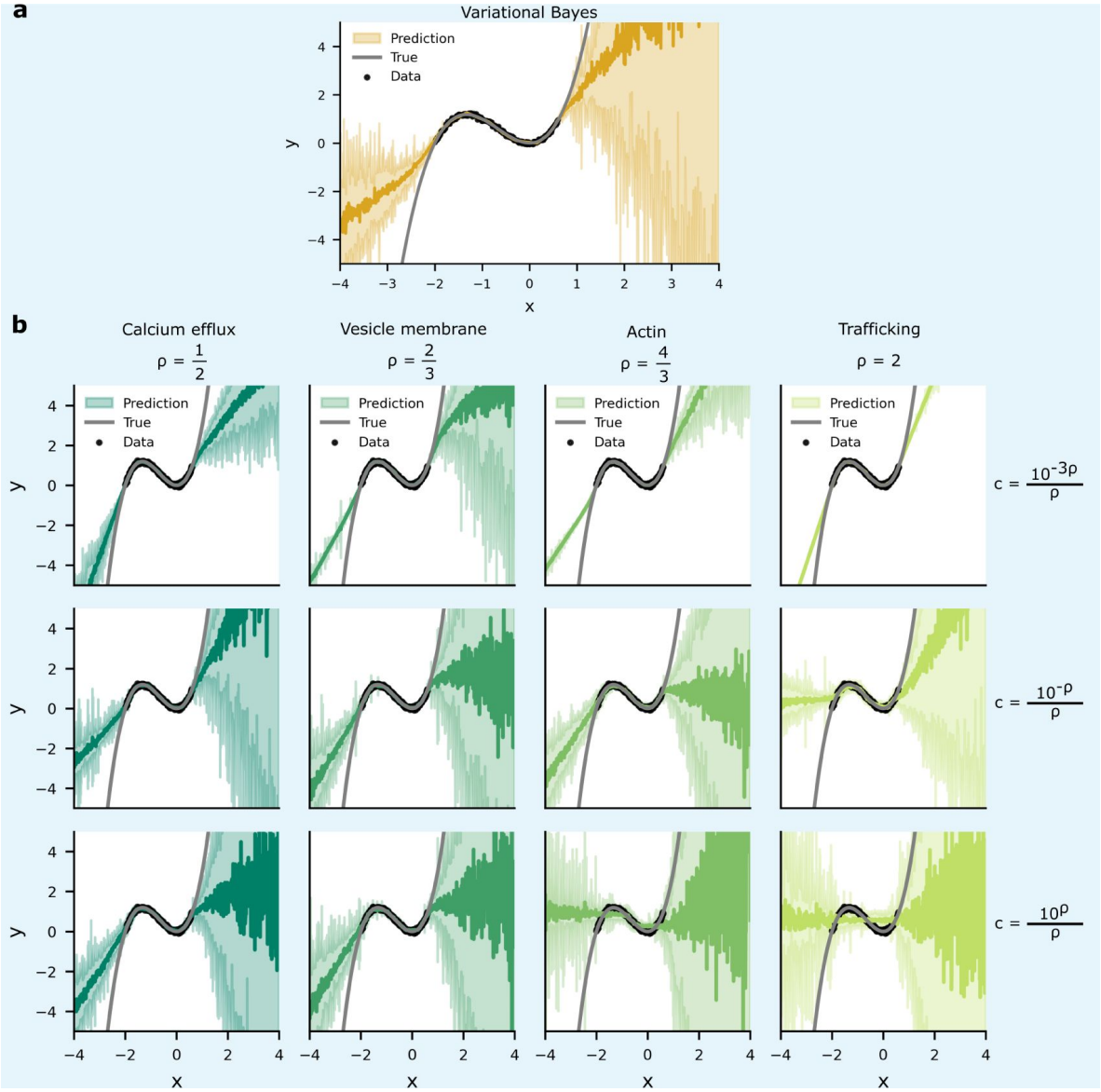
Second, the biological reliability cost becomes equal to the ideal entropic reliability cost when $s = \sigma_i$,

$$C_{\mathrm{BI;i}}(s = \sigma_i) = -\frac{1}{2}\log 2\pi e s^2 = -\frac{1}{2}\log 2\pi e \sigma_i^2 = C_{\mathrm{VI;i}}. \tag{53}$$

as the first term in **Eq. 44** 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 ρ and c (**Eq. 48**), and ρ 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 σ_i are not clear.

Appendix 5 Analytic predictions for σ_i

At various points in the main text, we note a connection between the Hessian, synapse importance and optimal variability. We start with **Eq. (29)**, which relates the optimal, energy efficient noise variance, σ_i^2 to the Hessian, H_{ii} which gives **Eq. (54a)**. Then, we combine this with the form for



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 $f(x)$.

the Hessian (**Eq. 25**), which gives **Eq. (54b)**. Finally, we note Hessian describes the log-likelihood (**Eq. 25** and **Eq. 20**). Thus, assuming the prior variance is large, we have $H_{ii} = \sigma_{\text{post;i}}$, which gives **Eq. (54c)**,

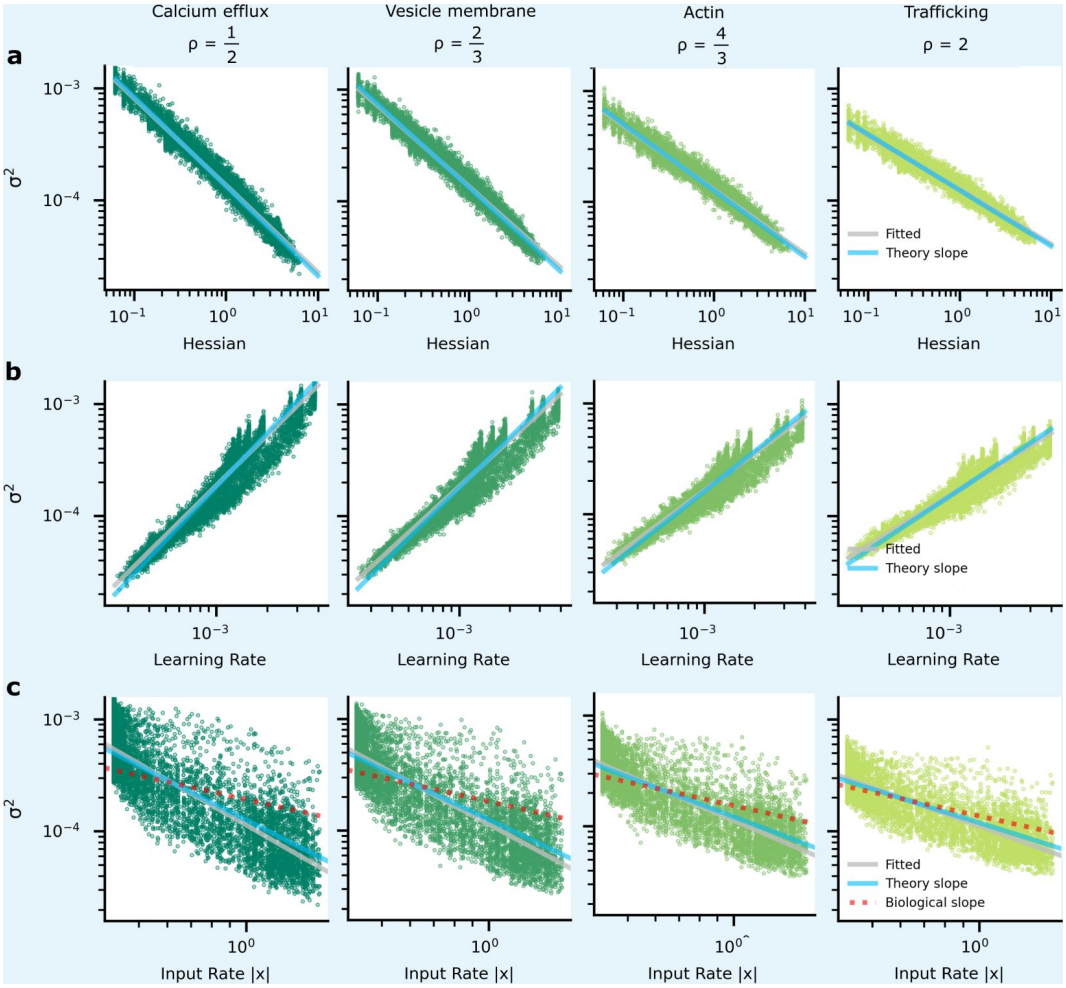
$$H_{ii} \propto \langle g_i^2 \rangle \propto \frac{1}{\eta_i^2} \qquad \log H_{ii} = -2 \log \eta_i + \text{const} \qquad \log \sigma_i^2 = \frac{4}{\rho+2} \log \eta_i + \text{const} \tag{54a}$$

$$H_{ii} \propto \langle x_i^2 \rangle \qquad \log H_{ii} = \log \langle x_i^2 \rangle \qquad \log \sigma_i^2 = -\frac{2}{\rho+2} \log \langle x_i^2 \rangle + \text{const} \tag{54b}$$

$$H_{ii} \propto \frac{1}{\sigma^2_{\text{post};i}} \qquad \log H_{ii} = -\log \sigma^2_{\text{post};i} + \text{const} \qquad \log \sigma_i^2 = \frac{2}{\rho+2} \log \sigma^2_{\text{post};i} + \text{const}\,. \tag{54c}$$

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 ρ. 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.

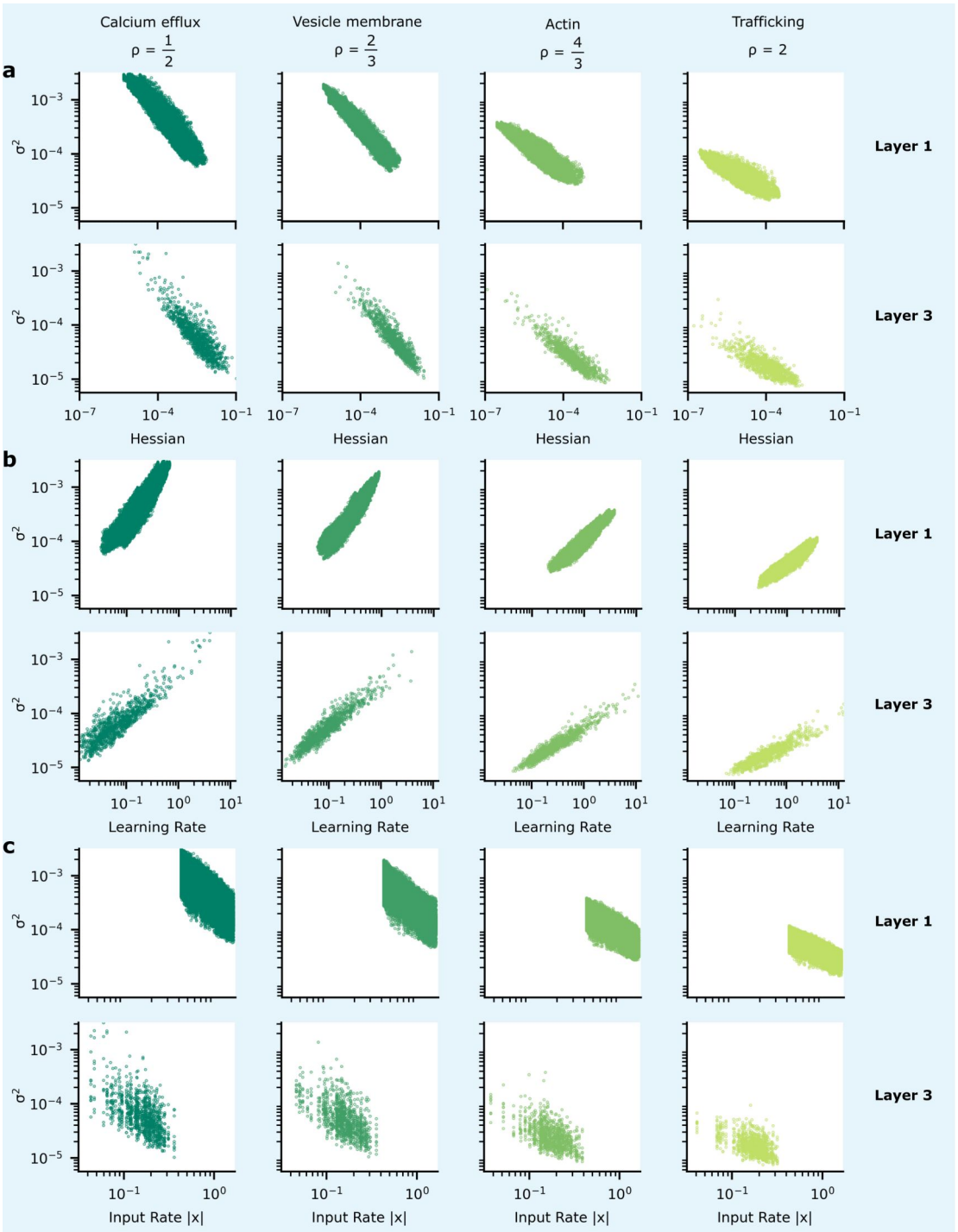
Appendix 6 Supplementary figures



Appendix 5—figure 1

Comparing analytic predictions for synapse variability with simulations and experimental data in a zero-hidden-layer network for MNIST classification.

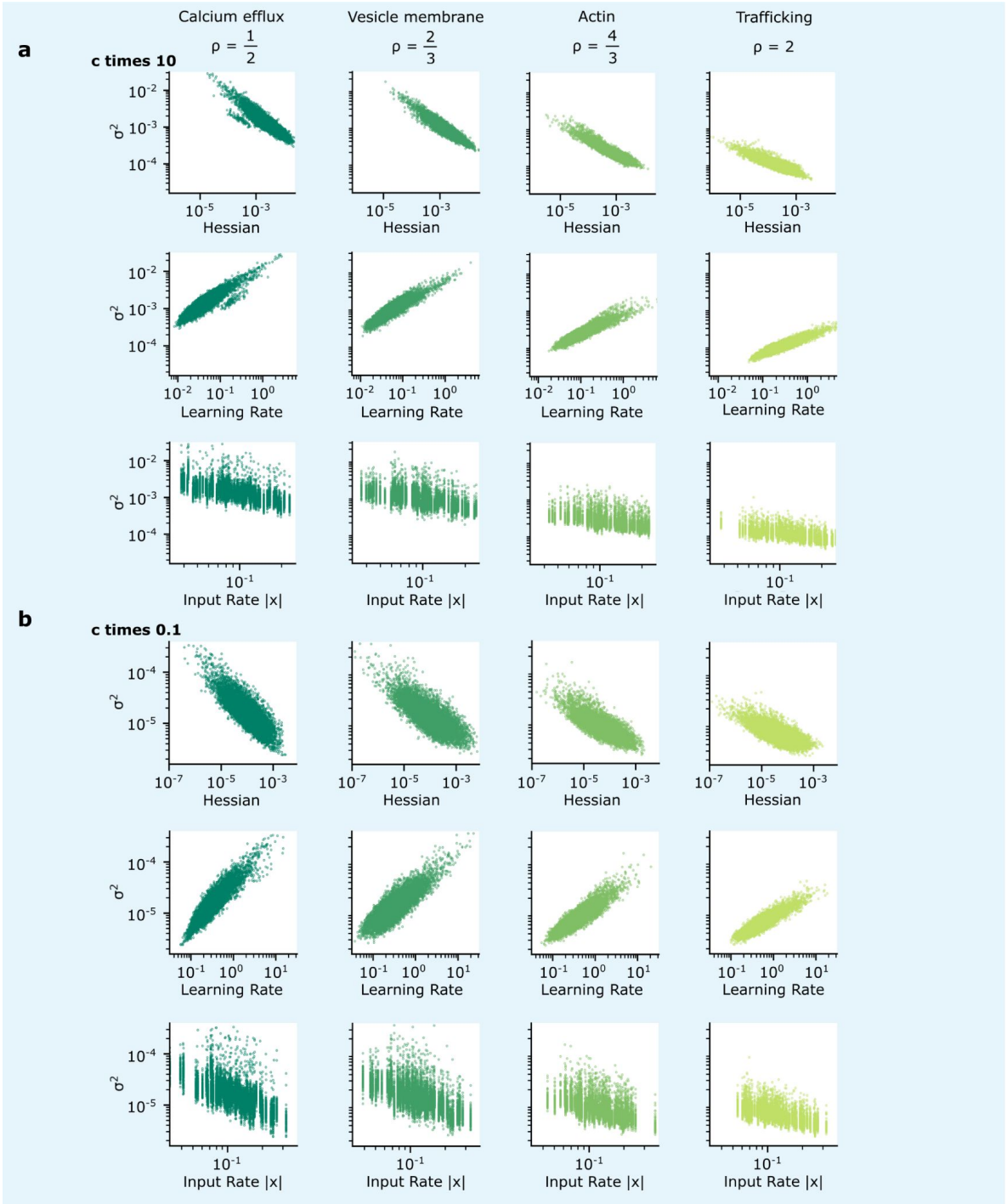
The green dots show simulated synapses, and the grey line is fitted to these simulated points. The blue line is from our analytic predictions, while the red dashed line is taken from experimental data (**Fig. 6b**).



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.



Appendix 6—figure 2

Patterns of synapse variability are robust to changes in the reliability.

We show that the patterns of variability of synapses connecting the two hidden layers presented in **Fig. 5** are preserved over a wide range of *c*. **a,b** When the reliability cost multiplier, *c*, is either increased (**a**) or decreased (**b**) by a factor of ten, overall synapse variability increases or decreases accordingly, but the qualitative correlations seen in **Fig. 5** are preserved.

References

Aitchison L (2020) **Bayesian filtering unifies adaptive and non-adaptive neural network optimization methods** *Advances in Neural Information Processing Systems* **33**:18173–18182

Aitchison L, Jegminat J, Menendez JA, Pfister JP, Pouget A, Latham PE (2021) **Synaptic plasticity as Bayesian inference** *Nature Neuroscience* **24**:565–571

Aitchison L, Lengyel M (2016) **The Hamiltonian brain: Efficient probabilistic inference with excitatory-inhibitory neural circuit dynamics** *PLoS computational biology* **12**

Attwell D, Laughlin SB (2001) **An energy budget for signaling in the grey matter of the brain** *Journal of Cerebral Blood Flow & Metabolism* **21**:1133–1145

Barber D, Bishop CM (1998) **Ensemble learning in Bayesian neural networks** *Nato ASI Series F Computer and Systems Sciences* **168**:215–238

Bellingham MC, Lim R, Walmsley B (1998) **Developmental changes in EPSC quantal size and quantal content at a central glutamatergic synapse in rat** *The Journal of Physiology* **511**:861–869

Berkes P, Orban G, Lengyel M, Fiser J (2011) **Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment** *Science* **331**:83–87

Blei DM, Kucukelbir A, McAuliffe JD (2017) **Variational inference: A review for statisticians** *Journal of the American statistical Association* **112**:859–877

Blundell C, Cornebise J, Kavukcuoglu K, Wierstra D (2015) **Weight uncertainty in neural network** *International conference on machine learning* :1613–1622

Bondy AG, Haefner RM, Cumming BG (2018) **Feedback determines the structure of correlated variability in primary visual cortex** *Nature neuroscience* **21**:598–606

Boyd I, Martin A (1956) **The end-plate potential in mammalian muscle** *The Journal of physiology* **132**

Branco T, Staras K (2009) **The probability of neurotransmitter release: variability and feedback control at single synapses** *Nature Reviews Neuroscience* **10**:373–383

Bridgman P (1999) **Myosin Va movements in normal and dilute-lethal axons provide support for a dual filament motor complex** *The Journal of Cell Biology* **146**:1045–1060

Brock JA, Thomazeau A, Watanabe A, Li SSY, Sjöström PJ (2020) **A practical guide to using CV analysis for determining the locus of synaptic plasticity** *Frontiers in Synaptic Neuroscience* **12**

Chenouard N, Xuan F, Tsien RW (2020) **Synaptic vesicle traffic is supported by transient actin filaments and regulated by PKA and NO** *Nature communications* **11**

Cingolani LA, Goda Y (2008) **Actin in action: the interplay between the actin cytoskeleton and synaptic efficacy** *Nature Reviews Neuroscience* **9**:344–356

Davis GW, Müller M (2015) **Homeostatic control of presynaptic neurotransmitter release** *Annual review of physiology* **77**:251–270

Dobrunz LE, Stevens CF (1997) **Heterogeneity of release probability, facilitation, and depletion at central synapses** *Neuron* **18**:995–1008

Echeveste R, Aitchison L, Hennequin G, Lengyel M (2020) **Cortical-like dynamics in recurrent circuits optimized for sampling-based probabilistic inference** *Nature neuroscience* **23**:1138–1149

Engl E, Attwell D (2015) **Non-signalling energy use in the brain** *The Journal of physiology* **593**:3417–3429

Festa D, Aschner A, Davila A, Kohn A, Coen-Cagli R (2021) **Neuronal variability reflects probabilistic inference tuned to natural image statistics** *Nature communications* **12**

Fiser J, Berkes P, Orban G, Lengyel M (2010) **Statistically optimal perception and learning: from behavior to neural representations** *Trends in cognitive sciences* **14**:119–130

Fisher RA (1922) **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* **222**:309–368

Forti L, Bossi M, Bergamaschi A, Villa A, Malgaroli A (1997) **Loose-patch recordings of single quanta at individual hippocampal synapses** *Nature* **388**:874–878

Fukushima K (1975) **Cognitron: A self-organizing multilayered neural network** *Biological cybernetics* **20**:121–136

Gentile JE, Carrizales MG, Koleske AJ (2022) **Control of synapse structure and function by actin and its regulators** *Cells* **11**

Goldman MS (2004) **Enhancement of information transmission efficiency by synaptic failures** *Neural Computation* **16**:1137–1162

Gramlich MW, Klyachko VA (2017) **Actin/Myosin-V-and activity-dependent inter-synaptic vesicle exchange in central neurons** *Cell Reports* **18**:2096–2104

Graves A (2011) **Practical variational inference for neural networks** *Advances in Neural Information Processing Systems*

Haefner RM, Berkes P, Fiser J (2016) **Perceptual decision-making as probabilistic inference by neural sampling** *Neuron* **90**:649–660

Harris JJ, Jolivet R, Attwell D (2012) **Synaptic energy use and supply** *Neuron* **75**:762–777

Harris JJ, Engl E, Attwell D, Jolivet RB (2019) **Energy-efficient information transfer at thalamocortical synapses** *PLoS computational biology* **15**

Heidelberger R, Heinemann C, Neher E, Matthews G (1994) **Calcium dependence of the rate of exocytosis in a synaptic terminal** *Nature* **371**:513–515

Hinton G, van Camp D (1993) **Keeping neural networks simple by minimising the description length of weights** *Proceedings of COLT-93* :5–13

Holler S, Köstinger G, Martin KA, Schuhknecht GF, Stratford KJ (2021) **Structure and function of a neocortical synapse** *Nature* **591**:111–116

Hoyer P, Hyvärinen A (2002) **Interpreting neural response variability as Monte Carlo sampling of the posterior** *Advances in neural information processing systems*

Jegminat J, Surace SC, Pfister JP (2022) **Learning as filtering: Implications for spike-based plasticity** *PLoS computational biology* **18**

Jordan MI, Ghahramani Z, Jaakkola TS, Saul LK (1999) **An introduction to variational methods for graphical models** *Machine learning* **37**:183–233

Karbowski J (2019) **Metabolic constraints on synaptic learning andmemory** *Journal ofNeurophysiology* **122**:1473–1490

Karunanithi S, Marin L, Wong K, Atwood HL (2002) **Quantal size and variation determined by vesicle size in normal and mutant Drosophila glutamatergic synapses** *Journal of Neuroscience* **22**:10267–10276

Katz B, Miledi R (1965) **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* **161**:483–495

Khan M, Nielsen D, Tangkaratt V, Lin W, Gal Y, Srivastava A (2018) **Fast and scalable bayesian deep learning by weight-perturbation in adam** *International conference on machine learning* :2611–2620

Kingma DP, Ba J (2014) **Adam: A method for stochastic optimization**

Kingma DP, Salimans T, Welling M (2015) **Variational dropout and the local reparameterization trick** *Advances in Neural Information Processing Systems*

Knill DC, Pouget A (2004) **The Bayesian brain: the role of uncertainty in neural coding and computation** *TRENDS in Neurosciences* **27**:712–719

Ko H, Cossell L, Baragli C, Antolik J, Clopath C, Hofer SB, Mrsic-Flogel TD (2013) **The emergence of functional microcircuits in visual cortex** *Nature* **496**:96–100

Lange RD, Chattoraj A, Beck JM, Yates JL, Haefner RM (2021) **A confirmation bias in perceptual decision-making due to hierarchical approximate inference** *PLoS Computational Biology* **17**

Lange RD, Haefner RM (2017) **Characterizing and interpreting the influence of internal variables on sensory activity** *Current opinion in neurobiology* **46**:84–89

Lange RD, Haefner RM (2022) **Task-induced neural covariability as a signature of approximate Bayesian learning and inference** *PLoS computational biology* **18**

Laughlin SB, de Ruyter van Steveninck RR, Anderson JC (1998) **The metabolic cost of neural information** *Nature Neuroscience* **1**:36–41

LeCun Y, Bottou L, Orr GB, Müller KR (2002) **Efficient backprop** *Neural networks: Tricks of the trade* :9–50

Levy WB, Baxter RA (2002) **Energy-efficient neuronal computation via quantal synaptic failures** *Journal of Neuroscience* **22**:4746–4755

Lillicrap TP, Santoro A, Marris L, Akerman CJ, Hinton G (2020) **Backpropagation and the brain** *Nature Reviews Neuroscience* **21**:335–346

Lisman JE, Harris KM (1993) **Quantal analysis and synaptic anatomy—integrating two views of hippocampal plasticity** *Trends in Neurosciences* **16**:141–147

Ma WJ, Beck JM, Latham PE, Pouget A (2006) **Bayesian inference with probabilistic population codes** *Nature neuroscience* **9**:1432–1438

MacKay DJ (1992) **The evidence framework applied to classification networks** *Neural computation* **4**:720–736

MacKay DJ (1992) **A practical Bayesian framework for backpropagation networks** *Neural Computation* **4**:448–472

Martens J (2020) **New insights and perspectives on the natural gradient method** *The Journal of Machine Learning Research* **21**:5776–5851

Murphy KP (2012) **Machine learning: a probabilistic perspective**

Murthy VN, Sejnowski TJ, Stevens CF (1997) **Heterogeneous release properties of visualized individual hippocampal synapses** *Neuron* **18**:599–612

Orban G, Berkes P, Fiser J, Lengyel M (2016) **Neural variability and sampling-based probabilistic representations in the visual cortex** *Neuron* **92**:530–543

Paulsen O, Heggelund P (1996) **Quantal properties of spontaneous EPSCs in neurones of the guinea-pig dorsal lateral geniculate nucleus** *The Journal of Physiology* **496**:759–772

Paulsen O, Heggelund P (1994) **The quantal size at retinogeniculate synapses determined from spontaneous and evoked EPSCs in guinea-pig thalamic slices** *The Journal of Physiology* **480**:505–511

Pulido C, Ryan TA (2021) **Synaptic vesicle pools are a major hidden resting metabolic burden of nerve terminals** *Science Advances* **7**

Purdon A, Rosenberger T, Shetty H, Rapoport S (2002) **Energy consumption by phospholipid metabolism in mammalian brain** *Neurochemical Research* **27**:1641–1647

Raghavachari S, Lisman JE (2004) **Properties of quantal transmission at CA1 synapses** *Journal of neurophysiology* **92**:2456–2467

Richards BA, Lillicrap TP (2019) **Dendritic solutions to the credit assignment problem** *Current opinion in neurobiology* **54**:28–36

Rosset S, Zhu J (2006) **Sparse, Flexible and Efficient Modeling using L 1 Regularization** *Feature Extraction: Foundations and Applications* :375–394

Sacramento J, Ponte Costa R, Bengio Y, Senn W (2018) **Dendritic cortical microcircuits approximate the backpropagation algorithm** *Advances in neural information processing systems*

Sacramento J, Wichert A, van Rossum MC (2015) **Energy efficient sparse connectivity from imbalanced synaptic plasticity rules** *PLoS Computational Biology* **11**

Sakaba T, Neher E (2001) **Quantitative relationship between transmitter release and calcium current at the calyx of held synapse** *Journal of Neuroscience* **21**:462–476

Schug S, Benzing F, Steger A (2021) **Presynaptic stochasticity improves energy efficiency and helps alleviate the stability-plasticity dilemma** *eLife* **10**

Shannon CE (1948) **A mathematical theory of communication** *The Bell system technical journal* **27**:379–423

Shivkumar S, Lange R, Chattoraj A, Haefner R (2018) **A probabilistic population code based on neural samples** *Advances in neural information processing systems*

Silver RA (2003) **Estimation of nonuniform quantal parameters with multiple-probability fluctuation analysis: theory, application and limitations** *Journal of neuroscience methods* **130**:127–141

Sjostrom PJ, Turrigiano GG, Nelson SB (2001) **Rate, timing, and cooperativity jointly determine cortical synaptic plasticity** *Neuron* **32**:1149–1164

Turrigiano GG, Leslie KR, Desai NS, Rutherford LC, Nelson SB (1998) **Activity-dependent scaling of quantal amplitude in neocortical neurons** *Nature* **391**:892–896

Turrigiano GG, Nelson SB (2004) **Homeostatic plasticity in the developing nervous system** *Nature reviews neuroscience* **5**:97–107

Whittington JC, Bogacz R (2017) **An approximation of the error backpropagation algorithm in a predictive coding network with local hebbian synaptic plasticity** *Neural computation* **29**:1229–1262

Yang G, Pan F, Gan WB (2009) **Stably maintained dendritic spines are associated with lifelong memories** *Nature* **462**:920–924

Yang Y, Li P (2022) **Synaptic Dynamics Realize First-order Adaptive Learning and Weight Symmetry**

Yu L, Zhang C, Liu L, Yu Y (2016) **Energy-efficient population coding constrains network size of a neuronal array system** *Scientific reports* **6**

Editors

Reviewer #1 (Public Review):

Summary:

Given the cost of producing action potentials and transmitting them along axons, it has always seemed a bit strange that there are synaptic failures: when a spike arrives at a synapse, about half the time nothing happens. This paper proposes a perfectly reasonable explanation: reducing failures (or, more generally, reducing noise) is costly. Four possible mechanisms are proposed, each associated with a different cost, with costs of the form 1/sigma_i^rho where sigma_i is the failure-induced variability at synapse i and rho is an exponent. The four different mechanisms produce four different values of rho.

What is interesting about the study is that the model makes experimental predictions about the relationship between learning rate, variability and presynaptic firing rate. Those predictions are consistent with experimental data, making it a strong candidate model. The fact that the predictions come from reasonable biological mechanisms make it a very strong candidate model and suggest several experiments to test it further.

Interestingly, the predictions made by this model are nearly indistinguishable from the predictions made by a normative model (Synaptic plasticity as Bayesian inference. Aitchison it al., Nature Neurosci. 24:565-571 (2021). As pointed out by the authors, working out whether the brain is using Bayesian inference to tune learning rules, or it just looks like it's Bayesian inference but the root cause is cost minimization, will be an interesting avenue for future research.

Finally, the authors relate their cost of reliability to the cost used in variational Bayesian inference. Intriguingly, the biophysical cost provides an upper bound on the variational cost. This is intellectually satisfying, as it answers a "why" question: why would evolution evolve to produce the kind of costs seen in the brain?

Strengths:

This paper provides a strong mix of theoretical analysis, simulations and comparison to experiments. And the extended appendices, which are very easy to read, provide additional mathematical insight.

Weaknesses:

None.

https://doi.org/10.7554/eLife.92595.2.sa1

Reviewer #2 (Public Review):

Summary

This manuscript argues about the similarity between two frameworks describing synaptic plasticity. In the Bayesian inference perspective, due to the noise and the limited available pre- and postsynaptic information, synapses can only have an estimate of what should be their weight. The belief about those weights is described by their mean and variance. In the energy efficient perspective, synaptic parameters (individual means and variances) are adapted such that the neural network achieves some task while penalizing large mean weights as well as small weight variances. Interestingly, the authors show both numerically and analytically the strong link between those two frameworks. In particular, both frameworks predict that (a) synaptic variances should decrease when the input firing rate increases and (b) that the learning rate should increase when the weight variances increase. Both predictions have some experimental support.

Strengths

(1) Overall, the paper is very well written and the arguments are clearly presented.

(2) The tight link between the Bayesian inference perspective and the energy efficiency perspective is elegant and well supported, both with numerical simulations as well as with analytical arguments.

(3) I also particularly appreciate the derivation of the reliability cost terms as a function of the different biophysical mechanisms (calcium efflux, vesicle membrane, actin and trafficking). Independently of the proposed mapping between the Bayesian inference perspective and the energy efficiency perspective, those reliability costs (expressed as power-law relationships) will be important for further studies on synaptic energetics.

Weaknesses

(1) As recognised by the authors, the correspondence between the entropy term in the variational inference description and the reliability cost in the energetic description is strong, but not perfect. Indeed, the entropy term scales as -log(sigma) while reliability cost scales as sigma^(-rho).

(2) Even though this is not the main point of the paper, I appreciate the effort made by the authors to look for experimental data that could in principle validate the Bayesian/energetic frameworks. A stronger validation will be an interesting avenue for future research.

https://doi.org/10.7554/eLife.92595.2.sa0

Author response:

The following is the authors' response to the original reviews.

> *Weaknesses*
>
> *(1) The authors face a technical challenge (which they acknowledge): they use two numbers (mean and variance) to characterize synaptic variability, whereas in the brain there are three numbers (number of vesicles, release probability, and quantal size). Turning biological constraints into constraints on the variance, as is done in the paper, seems somewhat arbitrary. This by no means invalidates the results, but it means that future experimental tests of their model will be somewhat nuanced.*

Agreed. There are two points to make here.

First, the mean and variance are far more experimentally accessible than n, p and q. The EPSP mean and variance is measured directly in paired-patch experiments, whereas getting n, p and q either requires far more extensive experimentation, or making strong assumptions. For instance, the data from Ko et al. (2013) gives the EPSP mean and variance, but not (directly) n, p and q. Thus, in some ways, predictions about means and variances are easier to test than predictions about n, p and q.

That said, we agree that in the absence of an extensive empirical accounting of the energetic costs at the synapse, there is inevitably some arbitrariness as we derive our energetic costs. That was why we considered four potential functional forms for the connection between the variance and energetic cost, which covered a wide range of sensible forms for this energetic cost. Our results were robust to this wide range functional forms, indicating that the patterns we describe are not specifically due to the particular functional form, but arise in many settings where there is an energetic cost for reliable synaptic transmission.

> *(2) The prediction that the learning rate should increase with variability relies on an optimization scheme in which the learning rate is scaled by the inverse of the magnitude of the gradients (Eq. 7). This seems like an extra assumption; the energy efficiency framework by itself does not predict that the learning rate should increase with variability. Further work will be needed to disentangle the assumption about the optimization scheme from the energy efficiency framework.*

Agreed. The assumption that learning rates scale with synapse importance is separate. However, it is highly plausible as almost all modern state-of-the-art deep learning training runs use such an optimization scheme, as in practice it learns far faster than other older schemes. We have added a sentence to the main text (line 221), indicating that this is ultimately an assumption.

> *Major*
>
> *(1) The correspondence between the entropy term in the variational inference description and the reliability cost in the energetic description is a bit loose. Indeed, the entropy term scales as −log(σ) while reliability cost scales as σ−ρ. While the authors do make the point that σ−ρ upper bounds −log(σ) (up to some constant), those two cost terms are different. This raises two important questions:*
>
> *a. Is this difference important, i.e. are there scenarios for which the two frameworks would have different predictions due to their different cost functions?*
>
> *b. Alternatively, is there a way to make the two frameworks identical (e.g. by choosing a proposal distribution Q(w) different from a Gaussian distribution (and tuneable by a free parameter that could be related to ρ) and therefore giving rise to an entropy term consistent with the reliability cost of the energy efficiency framework)?*

To answer b first, there is no natural way to make the two frameworks identical (unless we assume the reliability cost is proportional to log_σsyn_, and we don't think there's a biophysical mechanism that would give rise to such a cost). Now, to answer a, in Fig. 7 we extensively assessed the differences between the energy efficient *σsyn* and the Bayesian *σpost*. In Fig.7bc, we find that *σsyn* and *σpost* are positively correlated in all models. This positive correlation indicates that the qualitative predictions made by the two frameworks (Bayesian inference and energy efficiency) are likely to be very similar. Importantly though, there are systematic differences highlighted by Fig. 7ab. Specifically, the energy efficient *σsyn* tends to vary less than the Bayesian *σpost*. This appears in Fig. 7b which shows the relationship between *σsyn* (on the y-axis) and *σpost* (on the x-axis). Specifically, this plot has a slope that is smaller than one for all our models of the biophysical cost. Further, the pattern also appears in the covariance ellipses in Fig. 7a, in that the Bayesian covariance ellipses tend to be long and thin, while the energy efficient covariance ellipsis are rounder. Critically though both covariance ellipses show the same pattern in that there is more noise along less important directions (as measured by the Hessian).

We have added a sentence (line 273) noting that the search for a theoretical link is motivated by our observations in Fig. 7 of a strong, but not perfect link between the pattern of variability predicted by Bayesian and energy-efficient synapses.

> *(2) Even though I appreciate the effort of the authors to look for experimental evidence, I still find that the experimental support (displayed in Fig. 6) is moderate for three reasons.*
>
> *a. First, the experimental and simulation results are not displayed in a consistent way. Indeed, Fig 6a displays the relative weight change* |Dw|/w *as a function of the normalised variability* σ2/|μ| *in experiments whereas the simulation results in Fig 5c*

> *display the variance* σ2 as a function of the learning rate. Also, Fig 6b displays the normalised variability σ2/|μ| *as a function of the input rate whereas Fig 5b displays the variance _σ_2 as a function of the input rate. As a consequence the comparison between experimental and simulation results is difficult.*
>
> *b. Secondly, the actual power-law exponents in the experiments (see Fig 6a resp. 6b) should be compared to the power-law exponents obtained in simulation (see Fig 5c resp. Fig 5b). The difficulty relies here on the fact that the power-law exponents obtained in the simulations directly depend on the (free) parameter* ρ*. So far the authors precisely avoided committing to a specific* ρ*, but rather argued that different biophysical mechanisms lead to different reliability exponents* ρ*. Therefore, since there are many possible exponents* ρ *(and consequently many possible power-law exponents in simulation results in Fig 5), it is likely that one of them will match the experimental data. For the argument to be stronger, one would need to argue which synaptic mechanism is dominating and therefore come up with a single prediction that can be falsified experimentally (see also point 4 below).*
>
> *c, Finally, the experimental data presented in Fig6 are still "clouds of points". A coefficient of* r = *0*_.52 (in Fig 6a) is moderate evidence while the coefficient of r = −0._*26 (in Fig 6b) is weak evidence.*

The key thing to remember is that our paper is not about whether synapses are "really" Bayesian or energy efficient (or both/neither). Instead, the key point of our paper, as expressed in the title, is to show that the experimental predictions of Bayesian synapses are very similar to the predictions from energy efficient synapses. And therefore energy efficient synapses are very difficult to distinguish experimentally from Bayesian synapses. In that context, the two plots in Fig. 6 are not really intended to present evidence in favour of the energy efficiency / Bayesian synapses. In fact, Fig. 6 isn't meant to constitute a contribution of the paper at all, instead, Fig. 6 serves merely as illustrations of the kinds of experimental result that have (Aitchison et al. 2021) or might (Schug et al. 2021) be used to support Bayesian synapses. As such, Fig. 6 serves merely as a jumping-off point for discussing how very similar results might equally arise out of Bayesian and energy-efficiency viewpoints.

We have modified our description of Fig. 6 to further re-emphasise that the panels in Fig. 6 is not our contribution, but is taken directly from Schug et al. 2021 and Aitchison et al. 2021 (we have also modified Fig 6 to be precisely what was plotted in Schug et al. 2021, again to re-emphasise this point). Further, we have modified the presentation to emphasise that these plots serve merely as jumping off points to discuss the kinds of predictions that we might consider for Bayesian and energy efficient synapses.

This is important, because we would argue that the "strength of support" should be assessed for our key claim, made in the title, that "Signatures of Bayesian inference emerge from energy efficient synapses".

a) To emphasise that these are previously published results, we have chosen axes to matchthose used in the original work (Aitchison et al. 2021) and (Schug et al. 2021).

b) We agree that a close match between power-law exponents would constitute strong evidencefor energy-efficiency / Bayesian inference, and might even allow us to distinguish them. We did consider such a comparison, but found it was difficult for two reasons. First, while the confidence intervals on the slopes exclude zero, they are pretty broad. Secondly, while the slopes in a one-layer network are consistent and match theory (Appendix 5) the slopes in deeper networks are far more inconsistent. This is likely to be due to a number of factors such as details of the optimization algorithm and initialization. Critically, if details of the optimization algorithm matter in simulation, they may also matter in the brain. Therefore, it is not clear to us that a comparison of the actual slopes is can be relied upon.

To reiterate, the point of our article is not to make judgements about the strength ofevidence in previously published work, but to argue that Bayesian and energy efficient synapses are difficult to distinguish experimentally as they produce similar predictions. That said, it is very difficult to make blanket statements about the strength of evidence for an effect based merely on a correlation coefficient. It is perfectly possible to have moderate correlation coefficients along with very strong evidence of an effect (and e.g. very strong p-values), e.g. if there is a lot of data. Likewise, it is possible to have a very large correlation coefficient along with weak evidence of an effect (e.g. if we only have three or four datapoints, which happen to lie in a straight line). A small correlation coefficient is much more closely related to the effect-size. Specifically, the effect-size, relative to the “noise", which usually arises from unmeasured factors of variation. Here, we know there are many, many unmeasured factors of variation, so even in the case that synapses are really Bayesian / energy-efficient, the best we can hope for is low correlation coefficients

> *As mentioned in the public review, a weakness in the paper is the derivation of the constraints on σi given the biophysical costs, for two reasons.*
>
> *a.First, it seemed a bit arbitrary whether you hold n fixed or p fixed.*
>
> *b.Second, at central synapses, n is usually small – possibly even usually 1: REF(Synaptic vesicles transiently dock to refill release sites, Nature Neuroscience 23:1329-1338, 2020); REF(The ubiquitous nature of multivesicular release Trends Neurosci. 38:428-438, 2015). Fixing n would radically change your cost function. Possibly you can get around this because when two neurons are connected there are multiple contacts (and so, effectively, reasonably large n). It seems like this is worth discussing.*

a) Ultimately, we believe that the “real” biological cost function is very complex, and most likely cannot be written down in a simple functional form. Further, we certainly do not have the experimental evidence now, and are unlikely to have experimental evidence for a considerable period into the future to pin down this cost function precisely. In that context, we are forced to resort to two strategies. First, using simplifying assumptions to derive a functional form for the cost (such as holding n or p fixed). Second, considering a wide range of functional forms for the cost, and ensuring our argument works for all of them.

b) We appreciate the suggestion that the number of connections could be used as a surrogate where synapses have only a single release site. As you suggest we can propose an alternative model for this case where n represents the number of connections between neurons. We have added this alternative interpretation to our introduction of the quantal model under title “Biophysical costs". For a fixed PSP mean we could either have many connections with small vesicles or less connections with larger vesicles. Similarly for the actin cost we would certainly require more actin if the number of connections were increased.

> *Minor*
>
> *(1) A few additional references could further strengthen some claims of the paper:*
>
> *Davis, Graeme W., and Martin Muller. “Homeostatic Control of Presynaptic Neurotransmitter Release." Annual Review of Physiology 77, no. 1 (February 10, 2015): 251-70. https://doi.org/10.1146/annurev-physiol-021014-071740. This paper provides elegant experimental support for the claim (in line 538 now 583) that μ is kept constant and q acts as a compensatory variable.*
>
> *Jegminat, Jannes, Simone Carlo Surace, and Jean-Pascal Pfister. “Learning as Filtering: Implications for Spike-Based Plasticity." Edited by Blake A Richards. PLOS Computational*

> *Biology 18, no. 2 (February 23, 2022): e1009721. https://doi.org/10.1371/journal.pcbi.1009721.*
>
> *This paper also showed that a lower uncertainty implies a lower learning rate (see e.g. in line 232), but in the context of spiking neurons.*

Figure 1 of the the first suggested paper indeed shows that quantal size is a candidate for homeostatic scaling (fixing μ). This review also references lots of further evidence of quantal scaling and evidence for both presynaptic and postsynaptic scaling of q leaving space for speculation on whether vesicle radius or postsynaptic receptor number is the source of a compensatory q. On line 583 we have added a few lines pointing to the suggested review paper.

The second reference demonstrates Bayesian plasticity in the context of STDP, proposing learning rates tuned to the covariance in spike timing. We have added this as extra support for assuming an optimisation scheme that tunes learning rates to synapse importance and synapse variability (line 232).

> *In the numerical simulations, the reliability cost is implemented with a single power-law expression (reliability cost* $= c\sigma^{-\rho}$*). However, in principle, all the reliability costs will play in conjunction, i.e. reliability cost* $= \sum_i c_i \sigma^{-\rho_i}$*. While I do recognise that it may be difficult to estimate the biophysical values of the various* ci*, it might be still relevant to comment on this.*

Agreed. Limitations in the literature meant that we could only form a cursory review of the relative scale of each cost using estimates by Atwell, (2001), Engl, (2015). On line 135 we have added a paragraph explaining the rationale for considering each cost independently.

> *(3) In Eq. 8: _σ_2 doesn't depend on variability in* q*, which would add another term; barring algebra mistakes, it's* $\sigma^2 = n^2p^2var[q] + np(1-p) < q^2 >= np(1+(n-1)p)var[q] + np(1-p) < q >^2$*. It seems worth mentioning why you didn't include it. Can you argue that it's a small effect?*

Agreed. Ultimately, we dropped this term because we expected it to be small relative to variability in vesicle release, and because it would be difficult to quantify In practice, the variability is believed to be contributed mostly by variability in vesicle release. The primary evidence for this is histograms of EPSP amplitudes which show classic multi-peak structure, corresponding to one, two three etc. EPSPs. Examples of these plots include:

- "The end-plate potential in mammalian muscle", Boyd and Martin (1956); Fig. 8.

- "Structure and function of a neocortical synapse", Holler-Rickauer et al. (2019); Extended Figure 5.

> *(3) On pg. 7 now pg. 8, when the Hessian is introduced, why not say what it is? Or at least the diagonal elements, for which you just sum up the squared activity. That will make it much less mysterious. Or are we relying too much on the linear model given in App 2? If so, you should tell us how the Hessian was calculated in general. Probably in an appendix.*

With the intention of maintaining the interest of a wide audience we made the decision to avoid a mathematical definition of the Hessian, opting instead for a written definition i.e. line 192 - "*Hii*; the second derivatives of the objective with respect to *wi*." and later on a schematic (Fig. 4) for how the second derivative can be understood as a measure of curvature and synapse importance. Nonetheless, this review point has made us aware that the estimated Hessian values plotted in Fig. 5a have been insufficiently explained so we have added a reference on line 197 to the appendix section where we show how we estimated the diagonal values of the Hessian.

> *(4) Fig. 5: assuming we understand things correctly, Hessian ∝ |x|2. Why also plot _σ_2 versus |x|? Or are we getting the Hessian wrong?*

The Hessian is proportional to $\sum_t |x_t|^2$. If you assume that time steps are small and neurons spike, then $x_t \in \{0, 1\}$, and $|x_t|^2 = |x_t|$. it is difficult to say what timestep is relevant in practice.

> *(5) To get Fig. 6a, did you start with Fig. Appendix 1-figure 4 from Schug et al, and then use $\sigma^2/\mu = (1-p)q$, drop the* q*, and put 1 −* p *on the x-axis? Either way, you should provide details about where this came from. It could be in Methods.*

We have modified Fig. 6 to use the same axes as in the original papers.

> *(6) Lines 190-3: "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 Sjostrom et al. (2003) as processed by Costa et al. (2017) (Fig. 6b)." We believer 6a and 6b should be interchanged in that sentence.*

Thank you. We have switched the text appropriately.

> *(7) What is posterior variance? This seems kind of important.*

This refers to the "posterior variance" obtained using a Bayesian interpretation of the problem of obtaining good synaptic weights (Aitchison et al. 2021). In our particular setting, we estimate posterior variances by setting up the problem as variational inference: see Appendix 4 and 5, which is now referred to in line 390.

> *(8) Lines 244-5: "we derived the relationships between the optimized noise,* σi *and the posterior variable,* σpost *as a function of* ρ *(Fig. 7b;) and as a function of c (Fig. 7c)." You should tell the reader where you derived this. Which is Eq. 68c now 54c. Except you didn't actually derive it; you just wrote it down. And since we don't know what posterior variance is, we couldn't figure it out.*

If *H* is the Hessian of the log-likelihood, and if the prior is negligable relative to the the likelihood, then we get Eq. 69c. We have added a note on this point to the text.

> *(9) We believe Fig. 7a shows an example pair of synapses. Is this typical? And what about Figs. 7b and c. Also an example pair? Or averages? It would be helpful to make all this clear to the reader.*

Fig. 7a shows an illustrative pair of synapses, chosen to best display the relative patterns of variability under energy efficient and Bayesian synapses. We have noted this point in the legend for Fig. 7. Fig. 7bc show analytic relationships between energy efficient and Bayesian synapses, so each line shows a whole continuum of synapses(we have deleted the misleading points at the ends of the lines in Fig. 7bc).

> *(10) The y-axis of Fig 6a refers to the synaptic weight as w while the x-axis refers to the mean synaptic weight as mu. Shouldn't it be harmonised? It would be particularly nice if both were divided by* μ*, because then the link to Fig. 5c would be more clear.*

We have changed the y-axis label of Fig. 6a from w to μ. Regarding the normalised variance, we did try this but our Gaussian posteriors allowed the mean to become small in our simulations, giving a very high normalised variance. To remedy this we would likely need to assume a log- posterior, but this was out of scope for the present work.

> *(11) Line 250 (now line 281): "Finally, in the Appendix". Please tell us which Appendix. Also, why not point out here that the bound is tightest at small* ρ*?*

We have added the reference to the the section of the appendix with the derivation of the biological cost as a bound on the ELBO. We have also referenced the equation that gives the limit of the biological cost as ρ tends to zero.

> *(12) When symbols appear that previously appeared more than about two paragraphs ago, please tell us where they came from. For instance, we spent a lot of time hunting for* ηi*. And below we'll complain about undefined symbols. Which might mean we just missed them; if you told us where they were, that problem would be eliminated.*

We have added extra references for the symbols in the text following Eq. 69.

> *(13) Line 564, typo (we think): should be σ−2.*

Good spot. This has been fixed.

> *(14) A bit out of order, but we don't think you ever say explicitly that* r *is the radius of a vesicle. You do indicate it in Fig. 1, but you should say it in the main text as well.*

We have added a note on this to the legend in Fig. 1.

> *(15) Eq. 14: presumably there's a cost only if the vesicle is outside the synapse? Probably worth saying, since it's not clear from the mechanism.*

Looking at Pulido and Ryan (2021) carefully, it is clear that they are referring to a cost for vesicles inside the presynaptic side of the synapse. (Importantly, vesciles don't really exist outside the synapse; during the release process, the vesicle membrane becomes part of the cell membrane, and the contents of the vesicle is ejected into the synaptic cleft).

> *(16) App. 2: why solve for mu, and why compute the trace of the Hessian? Not that it hurts, but things are sort of complicated, and the fewer side points the better.*

Agreed, we have removed the solution for μ, and the trace, and generally rewritten Appendix 2 to clarify definitions, the Hessian etc.

> *(17) Eq. 35: we believe you need a minus sign on one side of the equation. And we don't believe you defined p(d|w). Also, are you assuming g = partial log p(d|w)/partial w? This should be stated, along with its implications. And presumably, it's not really true; people just postulate that* p(d|w) *∝ exp(−log_loss_)?*

We have replaced p(d|w) with p(y, x|w), and we replaced "overall cost" with log P(y|w, x). Yes, we are also postulating that p(y|w, x) ∝ exp(−log loss), though in our case that does make sense as it corresonds to a squared loss.

As regards the minus sign, in the orignal manuscript, we had the second derivative of the cost. There is no minus sign for the cost, as the Hessian of the cost at the mode is positive semi-definite. However, once we write the expression in terms of a log-likelihood, we do need a minus sign (as the Hessian of the log-likelihood at a mode is negative semi-definite).

> *(18) Eq. 47 now Eq. 44: first mention of* CBi;i*?*

We have added a note describing CB around these equations.

> *(19) The "where" doesn't make sense for Eqs. 49 and 50; those are new definitions.*

We have modified the introduction of these equations to avoid the problematic "where".

> *(20) Eq. 57 and 58 are really one equation. More importantly: where does Eq. 58 come from? Is this the* H *that was defined previously? Either way, you should make that clear.*

We have removed the problematic additional equation line number, and added a reference to where H comes from.

> *(21) In Eq. 59 now Eq. 60 aren't you taking the trace of a scalar? Seems like you could skip this.*

We have deleted this derivation, as it repeats material from the new Appendix 2.

> *(22) Eq. 66 is exactly the same as Eq. 32. Which is a bit disconcerting. Are they different derivations of the same quantity? You should comment on this.*

We have deleted lots of the stuff in Appendix 5 as, we agree, it repeats material from Appendix 2 (which has been rewritten and considerably clarified).

> *(23) Eq. 68 now 54, left column: please derive. we got:*
>
> *gai = gradient for weight i on trial* $\mathrm{a} = (v_a - w_a x_{ai}) x_{ai} / \epsilon^2$
>
> *where the second equality came from Eq. 20. Thus*
>
> $< g_{ai}^2 > = < (v_a - w_a x_{ai})^2 x_{ai}^2 > / \epsilon^4 \approx < (v_a - w_a x_{ai})^2 >< x_{ai}^2 > / \epsilon^4 = < x_{ai}^2 > / \epsilon^2 \propto H_i i$
>
> *Is that correct? If so, it's a lot to expect of the reader. Either way, a derivation would be helpful.*

We agree it was unnecessary and overly complex, so we have deleted it.

> *(24) App 5–Figure 2: presumably the data for panel b came from Fig. 6a, with the learning rate set to Δw/w? And the data for panel c from Fig. 6b? This (or the correct statement, if this is wrong) should be mentioned.*

Yes, the data for panel c came from Fig. 6b. We have deleted the data in panel b, as there are some subtleties in interpretation of the learning rates in these settings.

> *(25) line 952 now 946: typo, “and the from".*

Corrected to “and from".

https://doi.org/10.7554/eLife.92595.2.sa3