

# DISENTANGLING WITH BIOLOGICAL CONSTRAINTS: A THEORY OF FUNCTIONAL CELL TYPES

**James C.R. Whittington\***  
Stanford & Oxford

**Will Dorrell**  
UCL

**Surya Ganguli**  
Stanford

**Timothy E.J. Behrens**  
UCL & Oxford

## ABSTRACT

Neurons in the brain are often finely tuned for specific task variables. Moreover, such disentangled representations are highly sought after in machine learning. Here we mathematically prove that simple biological constraints on neurons, namely nonnegativity and energy efficiency in both activity and weights, promote such sought after disentangled representations by enforcing neurons to become selective for single factors of task variation. We demonstrate these constraints lead to disentangling in a variety of tasks and architectures, including variational autoencoders. We also use this theory to explain why the brain partitions its cells into distinct cell types such as grid and object-vector cells, and also explain when the brain instead entangles representations in response to entangled task factors. Overall, this work provides a mathematical understanding of why, when, and how neurons represent factors in both brains and machines, and is a first step towards understanding of how task demands structure neural representations.

## 1 INTRODUCTION

Understanding why and how neurons behave is now foundational for both machine learning and neuroscience. Such understanding can lead to better, more interpretable artificial neural networks, as well as provide insights into how biological networks mediate cognition. A key to both these pursuits lies in understanding how neurons can best structure their firing patterns to solve tasks.

Neuroscientists have some understanding of how task demands affect both early single neuron responses (Olshausen & Field, 1996; Yamins et al., 2014; Ocko et al., 2018; McIntosh et al., 2016) and population level measures such as dimensionality (Gao et al., 2017; Stringer et al., 2019). However, there is little understanding of neural population structure in higher brain areas. As an example, we do not even understand why many different bespoke cellular responses exist for physical space, such as grid cells (Hafting et al., 2005), object-vector cells (Høydal et al., 2019), border vector cells (Solstad et al., 2008; Lever et al., 2009), band cells (Krupic et al., 2012), or many other cells (O’Keefe & Dostrovsky, 1971; Gauthier & Tank, 2018; Sarel et al., 2017; Deshmukh & Knierim, 2013). Each cell has a well defined, specific cellular response pattern to space, objects, *or* borders, as opposed to a mixed response to space, objects, *and* borders. Similarly, we don’t understand why neurons in inferior temporal cortex are aligned to axes of data generative factors (Chang & Tsao, 2017; Bao et al., 2020; Higgins et al., 2021), why visual cortical neurons are de-correlated (Ecker et al., 2010), why neurons in parietal cortex are selective only for specific tasks (Lee et al., 2022), why prefrontal neurons are apparently mixed-selective (Rigotti et al., 2013), and why grid cells sometimes warp towards rewarded locations (Boccaro et al., 2019) and sometimes don’t (Butler et al., 2019). In essence, why are some neural representations entangled and others not?

Machine learning has long endeavoured to build models that disentangle factors of variation (Hinton et al., 2011; Higgins et al., 2017a; Locatello et al., 2019). Such disentangled factors can facilitate compositional generalisation and reasoning (Higgins et al., 2018; 2017b; Whittington et al., 2021a), as well as lead to more interpretable outcomes in which individual neurons represent meaningful quantities. Unfortunately, building models that disentangle is challenging (Locatello et al., 2019), with disentangling often coming at the cost of accurately modelling the data itself.

---

\*Correspondence to: jcrwhittington@gmail.com

In this work we 1) prove simple biological constraints of **nonnegativity** and **minimising activity energy** lead to factorised representations in linear networks; 2) empirically show these constraints lead to disentangled representations in both linear and non-linear networks; 3) obtain competitive disentangling scores on a standard disentangling benchmark; 4) provide an understanding why neurons in the brain are characterised into specific cell types due to these same biological constraints; 5) empirically show these constraints lead to specific cell types; 6) suggest when and why neurons in the brain exhibit disentangling versus mixed-selectivity.

### 1.1 RELATED WORK

**Disentangling in machines.** Algorithms like PCA or ICA can extract linear factors of variation from data. However, in neural models, while one can learn the principle subspace (Pehlevan et al., 2015) or learn principle components sequentially (e.g. with Oja’s rule), learning single independent factors in individual neurons (i.e. disentangling) has only recently been achieved via specific weight regularization in linear autoencoders (Kunin et al., 2019). For disentangling in *non-linear* models, most modern methods use variational autoencoders (VAE; Kingma & Welling (2013)), with various choices that promote explicit factorisation of the learned latent space. For example  $\beta$ -VAEs (Higgins et al., 2017a) up-weight (by  $\beta$ ) the term in the VAE loss that encourages the posterior to be a factorised distribution. Other variations of disentangling VAEs (Burgess et al., 2018; Kim & Mnih, 2018; Ridgeway & Mozer, 2018; Kumar et al., 2018; Chen et al., 2018) similarly try to explicitly enforce a factorised aggregate posterior. Importantly, while it has been shown disentangling is not possible without inductive bias in the data or model (Locatello et al., 2019), numerous inductive biases have been shown to be sufficient for disentangling (i.e. conditioning the prior on a third variable Khemakhem et al. (2020); or local isometry and non-Gaussianity Horan et al. (2021)). Our work also relates to modern self-supervised learning (Bardes et al., 2022) which seeks decorrelated representations with non-zero variance. Again, most algorithms have an explicit term to force decorrelation. Here instead, we show that simple biological constraints of nonnegativity and energy efficiency lead to emergent disentangling without an explicit decorrelation term in the objective.

**Disentangling in brains.** There is no formal understanding of when and why neurons in the brain factorise across task parameters. Nevertheless, single latent dimensions from disentangling models do predict single neuron activity, implying biological neurons are disentangled (Higgins et al., 2021). One conjecture in neuroscience is that while representing task factors is important for generalisation, mixed-selectivity is important for efficient readout (Behrens et al., 2018; Rigotti et al., 2013; Bernardi et al., 2020). Here we show disentangled brain representations are preferred if the task is factorised into independent factors.

**Non-negativity.** Recent work on multitask learning in recurrent neural networks (RNNs) (Yang et al., 2019; Driscoll et al., 2022) demonstrated that neural populations, with a nonnegative activation function, partition themselves into task specific modules (Driscoll et al., 2022). Non-negativity is also important in obtaining hexagonal, not square, grid cells (Dordek et al., 2016; Whittington et al., 2021b; Sorscher et al., 2019; 2020). Also, nonnegative matrix factorisation empirically yields spatially localised factors for images (Lee & Seung, 2000). Our work demonstrates theoretically and empirically *why* nonnegativity leads to single neurons becoming selective for single factors.

## 2 LINEAR DISENTANGLING WITH BIOLOGICAL CONSTRAINTS

We first provide a theorem that suggests why the combined biological constraints of nonnegativity and energy efficiency lead to neural disentangling (proofs of all theorems are in App. A.2):

**Theorem 1.** Let  $e \in \mathbb{R}^k$  be a random vector whose  $k$  independent components denote  $k$  task factors. We assume each independent task factor  $e_i$  is drawn from a distribution that has mean 0, variance  $\sigma^2$ , and maximum and minimum values of  $\min(e_i) = -a$  and  $\max(e_i) = a$ . Also let  $z \in \mathbb{R}^n$  be a linear neural representation of the task factors given by

$$z = M e + b_z, \quad (1)$$

where  $M \in \mathbb{R}^{n \times k}$  are mixing weights and  $b_z \in \mathbb{R}^n$  is a bias. We further assume two constraints: (1) the neural representation is *nonnegative* with  $z_i \geq 0$  for all  $i = 1, \dots, n$ , and (2) the neural population variance is a nonzero constant,  $\sum_j \text{Var}(z_j) = C$ , so that the neural representation

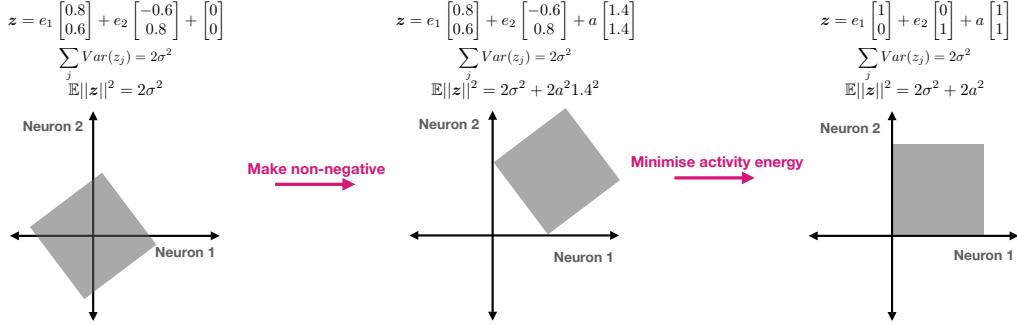


Figure 1: **Proof intuition.** Two uniformly distributed independent factors represented with two entangled neurons (left). The representation can be made nonnegative at the expense of activity energy (middle). Activity energy is minimised under a nonnegativity (and variance) constraint when the neurons are axis aligned to task factors (i.e. disentangled, right). Grey boxes denote uniform distributions over neural activity induced by uniform distributions over task factors. Note our proof does not require uniformity.

retains some information about the task variables. Under these two constraints we show that in the space of all possible neural representations (parameterised by  $M$  and  $b_z$ ), the representations that achieve minimal activity energy  $\mathbb{E}\|z\|^2$  also exhibit disentangling, by which we mean every neuron  $z_j$  is selective for at most one task parameter: i.e.  $|M_{jk}| |M_{jl}| = 0$  for  $k \neq l$ . (Proof in App. A.2.1).

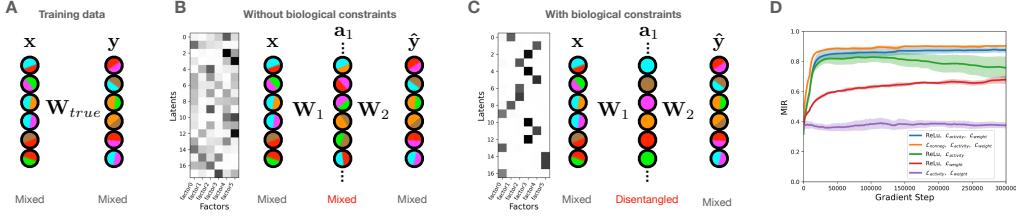
**Intuition.** The intuition underlying the proof of the theorem is shown in Fig.1, where the key idea can be seen with two neurons encoding two factors. In particular, the bias must make every  $z_i$  nonnegative for all values of  $e_1$  and  $e_2$ . But since  $e_1$  and  $e_2$  are independent, the minimum firing of neuron 1 for example obeys  $\min(z_1) = \min(0.8e_1 - 0.6e_2) = \min(0.8e_1) + \min(-0.6e_2) = -a(0.8 + 0.6)$ . Thus for neurons that mix factors, a larger bias term must be used to ensure nonnegativity, which leads to increased expected energy. Minimising this energy (subject to a constant variance) requires the smallest possible bias for each neuron, which occurs when each neuron is selective for a single task factor.

The above theorem, while simple, is restricted in two ways: (1) the independent task factors  $e_i$  are *directly* available to the network; (2) representational collapse (i.e. setting  $z = 0$ ) under energy minimisation is prevented solely by a variance constraint. We thus consider a more general setting where a neural circuit receives not the independent task factor vector  $e$ , but instead receives the mixed combination  $x = De$ . We further model the neural representation  $z$  as a linear generative model that can predict observed data  $x$  via  $x = Wz + b_x$ . Thus prediction, not variance constraint, now prevents collapsing neural representations (proof in Appendix A.2.2). Furthermore in Appendix A.2.3 we prove the following:

**Theorem 2.** Let  $x = De$  be observed entangled data, where the independent task factor vector  $e$  obeys the same distributional assumptions as in Theorem 1. Let a neural representation  $z$  exactly predict observed data via  $x = Wz + b_x$  with zero error. Then for all such data generation models (with parameters  $D$ ) and all such neural representations (with parameters  $W$  and  $b_x$ ), as long as: (1) the columns of  $D$  are (scaled) orthonormal; (2) the norm of the read-out weights  $\|W\|_F^2$  is finite; (3) the neural representation is nonnegative (i.e.  $z > 0$ ), then out of all such neural representations, the minimum energy representations are also disentangled ones. By this we mean that each neuron  $z_i$  will be selective for at most one hidden task factor  $e_j$ .

We note that  $D$  having (scaled) orthonormal columns may seem like a strong constraint, but it holds approximately for any random matrix  $D$  with many observations (dimensionality of  $x$ ) and few independent task factors (dimensionality of  $e$ ) (proof in Appendix A.2.4). Appendix A.2.5 discusses and provides intuition for when disentangling occurs as  $D$  takes more general forms.

Strikingly, the essential content of Theorem 2 is that any linear, nonnegative, optimally energetically efficient, generative neural representation that accurately predicts entangled observations that are linear mixtures of hidden task factors, will possess single neurons that are selective for individual task factors, despite *never* having *direct* access to them. In terms of applications, this theorem could apply in supervised or self-supervised machine learning settings in any neural network layer  $z$  that



**Figure 2: Shallow linear networks disentangle.** We train 1-hidden layer linear networks on linear data. **A)** Schematic showing both input and output are entangled linear mixtures of factors (colours). Neurons colours schematically denote which of the factors it codes for.  $\mathbf{W}_{true}$  is the true mapping between  $x$  and  $y$ . **B)** A model without biological constraints learns entangled internal representations. Mutual information matrix shown on left, schematic on right.  $\mathbf{W}_1$  and  $\mathbf{W}_2$  are the learnable weights projecting to and from the hidden layer. **C)** A model with our constraints learns disentangled representations. **D)** Several model variants, in which only those with all our constraints learn disentangled representations. Average and standard error shown for 5 random seeds.

is linearly read out from, or in neuroscience settings where  $z$  reflects a neural population from which one attempts to linearly decode task factors. While Theorem 2 holds in a simple setting, we will show through simulations that its essential content, namely that nonnegativity and energy efficiency together promote disentangling, holds in practice in much more complex multilayer neural networks.

### 3 DISENTANGLING IN MACHINES

We now present simulation results demonstrating that nonnegativity and energy efficiency (minimising either activity or weight energy) lead to single neuron selectivity for single task factors. We show this for supervised and unsupervised learning, both for linear and non-linear tasks and networks (simulation details in Appendix A.6).

**A measure for disentangled subspaces.** While our theory describes when single neurons become selective for single independent task factors, it does not limit the number of neurons selective for any given factor. For example in Theorem 2, four copies of the same neuron in  $z$ , each with half the activity, along with four copies of projecting weights each with half the values, predicts  $x$  just as well and has exactly the same energy in both  $z$  and  $\mathbf{W}$ <sup>1</sup>. More interestingly, an underlying task factor may not be one-dimensional, e.g. spatial location, in which case the subspace that codes for this factor have at least the same dimension. This phenomena cannot be captured by many metrics of disentangling (e.g. the popular mutual information gap; MIG; Chen et al. (2018)) since they score highly if each factor is represented in just *one* neuron. Thus we define a new metric (mutual information ratio; MIR) that instead scores highly if each neuron only cares about one factor (see Appendix A.1 for details).

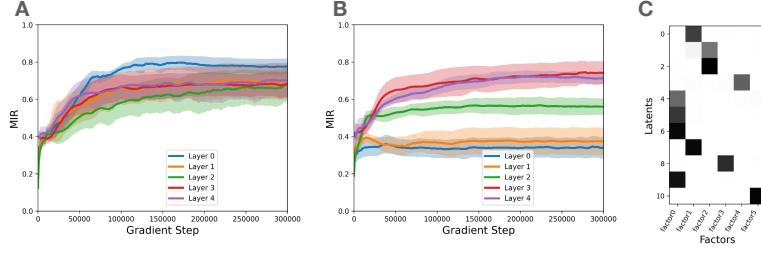
**Regularizers as constraints.** We impose nonnegativity via a ReLU activation function, or softly via explicit regularization  $\mathcal{L}_{\text{nonneg}} = \beta_{\text{nonneg}} \sum_i \max(-a_i, 0)$  where  $i$  indexes a neuron in the network, and  $\beta_{\text{nonneg}}$  determines the regularization strength. Similarly, we apply regularization to the activity energy and weight energy;  $\mathcal{L}_{\text{activity}} = \beta_{\text{activity}} \sum_l \|\mathbf{a}_l\|^2$  and  $\mathcal{L}_{\text{weight}} = \beta_{\text{weight}} \sum_l \|\mathbf{W}_l\|^2$ . The role of  $\mathcal{L}_{\text{weight}}$  is to promote activity (variance) in the network, otherwise activity could be reduced via  $\mathcal{L}_{\text{activity}}$ , and such reduced activity could be compensated for with arbitrarily large weights. The total loss we optimise is

$$\mathcal{L} = \underbrace{\mathcal{L}_{\text{nonneg}} + \mathcal{L}_{\text{activity}} + \mathcal{L}_{\text{weight}}}_{\text{Biological constraints}} + \underbrace{\mathcal{L}_{\text{prediction}}}_{\text{Functional constraints}} . \quad (2)$$

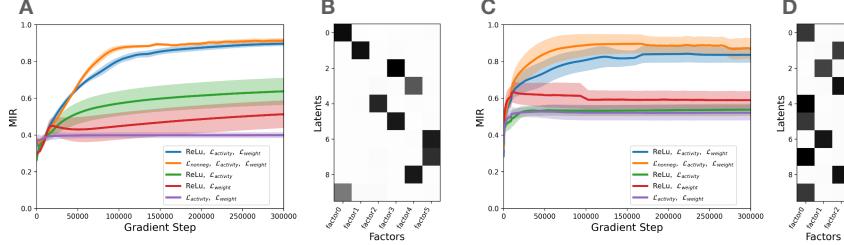
Here ‘functional constraints’, are any prediction losses the network has i.e. error in predicting target labels in supervised learning, or reconstruction error in autoencoders.

**Disentangling in supervised shallow neural networks.** First we consider a dataset  $\mathcal{D} = \{\mathbf{x}, \mathbf{y}\}$ , where  $\mathbf{x}$  is a orthogonal mixture of six i.i.d. random variables (hidden independent task factors; uniform distribution), and  $\mathbf{y}$  is a linear transform of  $\mathbf{x}$  (dimension 6; Fig. 2A). First we train shallow

<sup>1</sup>Learning dynamics may favour fewer neurons per factor as there are fewer weights to align.



**Figure 3: Deep non-linear networks disentangle.** We train 5-hidden layer nonlinear networks with our constraints on linear and non-linear data. **A)** For *linear* data, all layers in the network learn a disentangled representation. **B)** For *non-linear* data, only later layers learn a disentangled representation. **C)** Example mutual information matrix from the penultimate hidden layer.



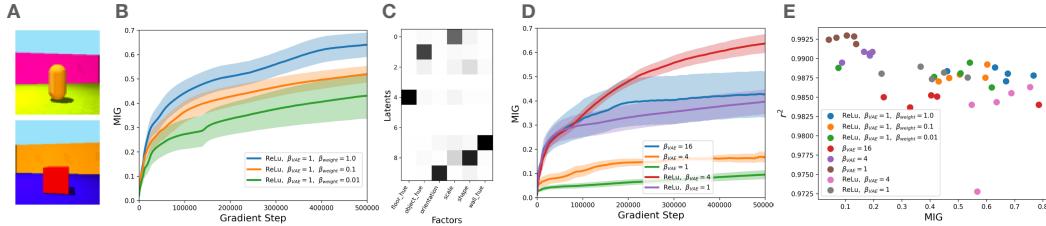
**Figure 4: Learning data generative factors with autoencoders.** **A)** Training linear autoencoders on linear data. Only models with our constraints learn disentangled representations. **B)** Example mutual information matrix from a high MIR model. **C)** Non-linear autoencoders trained on non-linear data. Only models with our constraints learn disentangled representations. **D)** Example mutual information matrix from a high MIR model. All learning curves show mean and standard error from 4 mean from 5 random seeds.

linear networks to read-out  $y$  from  $x$ . Networks without biological constraints exhibit mixed internal representations (Fig. 2B). However, with our constraints, networks learn distinct sub-networks for each task factor (Fig. 2C). Removing any one of our constraints leads to entangled representations (Fig. 2D). Lastly we note sparsity constraints do not induce disentangling (Appendix A.3). Thus the disentangling effect of ReLUs is not from sparsity, but instead from nonnegativity.

**Disentangling in supervised deep neural networks.** Training deep non-linear (ReLU) networks on this data also leads to distinct sub-networks, with all layers learning disentangled representations (Fig. 3A). However with non-linear data ( $x \leftarrow x^3$ ,  $y$  remaining the same), the early layers are mixed-selective, whereas the later layers are disentangled (Fig. 3B-C). Understanding why the final hidden layer disentangles is easy, since it linearly projects to the target and so our theory directly applies. By extrapolating our theory, we conjecture that our biological constraints encourage any layer to be as linearly related to task factors and as disentangled as possible. However, early layers cannot be linear in hidden task factors since they are required to perform non-linear computations on the non-linear data, and thus only once activity becomes linearly related to independent task factors in later layers does disentangling set in (as predicted by our linear theory).

**Disentangling in unsupervised neural networks.** We now consider unsupervised learning, i.e.  $\mathcal{D} = \{x\}$ , where  $x$  is a linear mixture of multiple independent task factors as in Theorem 2. Training 0-hidden layer autoencoders on this data, with our biological constraints, recovers the independent task factors in individual neural subspaces (Fig. 4A/B). Moreover, this only occurs when all constraints are present (Fig. 4A). Again, even though our theory applies to the linear setting, the same phenomena occur when training deep non-linear autoencoders on non-linear data; i.e. when  $x$  is a *non-linear* mixture of multiple i.i.d. random variables, i.e.  $\mathcal{D} = \{f(x)\}$  (Fig. 4C-D).

**Disentangling on a standard benchmark with VAEs.** We now consider a standard disentangling dataset (Fig. 5A; Kim & Mnih (2018)). To be consistent with, and to compare to, the disentangling literature we use a VAE and measure disentangling with the familiar mutual-information gap (MIG) metric (Chen et al., 2018). For nonnegativity we ask the mean of the posterior to be nonnegative (via



**Figure 5: Learning data generative factors with variational autoencoders.** **A)** We train on the Shapes3D dataset, with two example images shown. These images have 6 underlying factors. **B)** MIG scores are higher with higher weight regularization, and generally higher than any  $\beta$ -VAE (panel D). **C)** Mutual information matrix for a high scoring model. **D)**  $\beta$ -VAE MIG scores. Adding a ReLU improves MIG scores. **E)** MIG score against  $R^2$  shows models with our constraints lie in the Goldilocks region of high disentangling and high reconstruction. All learning curves show mean and standard error from 5 random seeds. Results from an additional dataset in Fig. 12

a ReLU)<sup>2</sup>, but we *do not* add a norm constraint as the VAE loss already one in its KL term between the Gaussian posterior and the Gaussian prior.

While state-of-the-art results are not our aim (instead we wish to elucidate that simple biological constraints lead to disentangling), our disentangling results (Fig. 5B) are competitive and often better than those in the literature (comparing to results of many models shown in Locatello et al. (2019)) even though those models explicitly ask for a factorised aggregate posterior. The particular baseline model we show here is  $\beta$ -VAE (Fig. 5D). We see that (1) our constraints lead to disentangling (Fig. 5B-C); (2) including a ReLU improves  $\beta$ -VAE disentangling (as predicted by nonnegativity arguments above, Fig. 5D); and (3) our constraints give results in the Goldilocks region of high disentangling and high reconstruction accuracy (Fig. 5E).

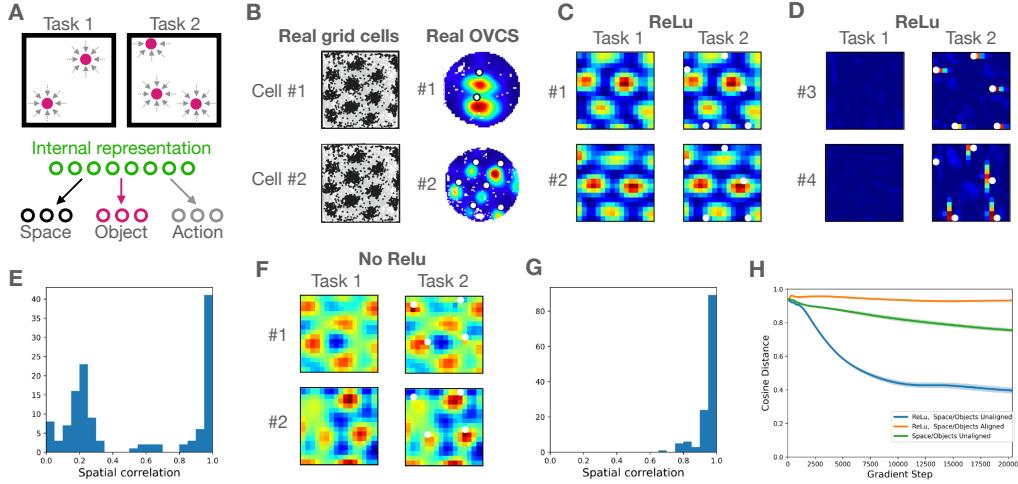
#### 4 DISENTANGLING IN BRAINS: A THEORY OF CELL TYPES

We next turn our attention to neuroscience, which is indeed the inspiration for our biological constraints. While we hope our general theory of neural representations will be useful for explaining representations across tasks and brain areas, for reasons stated below, we choose our first example from spatial processing in the hippocampal formation. We show our biological constraints lead to separate neural populations (modules) coding for separate task variables, but **only** when task variables correspond to independent factors of variation. Importantly, the modules consist of distinct functional cell types with similar firing properties, resembling grid (Hafting et al., 2005) and object-vector cells (Høydal et al., 2019) (GCs and OVCs).

We choose to focus on spatial representations for two reasons. Firstly, there is a significant puzzle about why neurons deep in the brain, synaptically far from the sensorimotor periphery, almost miraculously develop single cell representations for human-interpretable factors (e.g. GCs for location in space, and OVCs for relative location to objects). Such observations are not easily accounted for by standard neural network accounts that argue that representations are unlikely to be human-interpretable (Richards et al., 2019). Secondly, whilst these bespoke spatial representations are commonly observed to factorise into single cells, there are situations in which selectivity spans across multiple task variables (Boccaro et al., 2019; Hardcastle et al., 2017). For example, sometimes spatial firing patterns of GCs are warped by reward (Boccaro et al., 2019) and sometimes they are not (Butler et al., 2019). There is no theory for explaining why and when this happens.

**A factorised task for rodents.** We consider a task in which rodents must know where they are in space, but must also approach one of multiple objects. If objects appear in different places in different contexts, the task is factorised into two independent factors (Fig. 6A): ‘Where am I in allocentric spatial coordinates?’ and ‘Where am I in object-centric coordinates?’. By contrast, if objects always appear in the same locations, the task is not factorised (as spatial location can predict object location). Our theory says that solving this task requires two neural sub-spaces - one for

<sup>2</sup>Using a nonnegative posterior mean is odd when the prior is Gaussian, but it allows for easier comparison.



**Figure 6: Modules of distinct cell types form with nonnegativity and factorised tasks.** **A)** Top: Task schematic of an environment with objects that move location in different contexts. Bottom: We train a representation to predict 1) spatial location, 2) object location, and 3) correct action at every location. **B)** When rodents navigate such environments, GCs encode location in physical space, while OVCs encode relative location to objects. These plots are ratemaps; the average firing of a given cell at every location. **C-D)** Model ratemaps when the representation has a ReLU activation function. We see separate modules of GCs that do not change across tasks, and OVCs that move around to track objects. **E)** To quantify ‘module-ness’, we see how much each cell’s ratemaps changes between tasks (via a spatial correlation). Some cells don’t change (GCs), other cells do (OVCs). **F-G)** There are no clear modules without a ReLU activation. **H)** As another demonstration of disentangling, we compute the cosine distance between the population’s contribution to spatial versus object decoding.

allocentric location and one for location relative to objects - and that these sub-spaces should be represented in separate neural populations when the task is factorised, but not otherwise.

The standard neuroscience finding appears factorised with two distinct modules of non-overlapping cell populations: (1) GCs (Hafting et al., 2005) which represent allocentric space via hexagonal firing patterns (Fig. 6B left); and (2) OVCs (Høydal et al., 2019) which represent relative location to objects through firing fields at specific relative distances and orientations (Fig. 6B right).

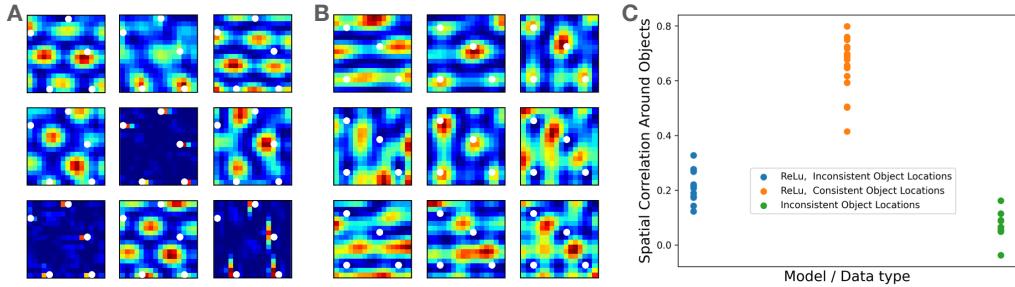
**Model with additional structural constraint.** Predicting allocentric spatial locations from egocentric self-motion cues is known as path integration (Burak & Fiete, 2009), and is believed to be a fundamental function of entorhinal cortex (where GCs and OVCs are found). GCs naturally emerge from training RNNs to path integrate under several additional biological constraints (Sorscher et al., 2019; 2020; Banino et al., 2018; Cueva & Wei, 2018). Hence to model this task (with locations *and* objects) we could train an RNN,  $\mathbf{z}$ , that predicts (1) what the spatial location,  $\mathbf{x}$ , will be and (2) whether we will encounter an object, after an action,  $\mathbf{a}$ , from the current location, and (3) what the expected action,  $\mathbf{a}$ , will be. However, here we adopt a far more general framework that does not limit future applications of our approach simply to sequential integration problems.

In particular, it was recently shown (Gao et al., 2021; Dorrell et al., 2022) that path integration constraints can be applied directly on the representation by adding a new constraint in the loss imposing

$$\mathbf{z}(\mathbf{x}) = f(\mathbf{W}_a \mathbf{z}(\mathbf{x} - \mathbf{a})). \quad (3)$$

Here  $f(\cdot)$  is an activation function and  $\mathbf{W}_a$  is a weight matrix that depends on the action  $\mathbf{a}$ . This surrogate constraint imposes potential path integration by ensuring that a motion  $\mathbf{a}$  in space  $\mathbf{x}$  imposes a lawful change in neural representation  $\mathbf{z}$ , thereby transforming the sequential path integration problem into the problem of directly estimating neural representations of space. Thus we minimise:

$$\mathcal{L} = \underbrace{\mathcal{L}_{\text{nonneg}} + \mathcal{L}_{\text{activity}} + \mathcal{L}_{\text{weight}} + \mathcal{L}_{\text{location}} + \mathcal{L}_{\text{actions}} + \mathcal{L}_{\text{objects}}}_{\text{Biological constraints}} + \underbrace{\mathcal{L}_{\text{path integration}}}_{\text{Structural constraints}} \quad (4)$$



**Figure 7: Entangled tasks lead to entangled representations and grid cell warping.** We show a representative selection of cells from a model with **A)** a factorised task and **B)** an entangled task. The phases of the firing fields in the entangled task are locked to object location - they have warped their firing fields. This is not the case for the factorised task (aside from object specific cells). **C)** To quantify phase locking for each of the task/model variant, we compute the average spatial correlation of patches around objects. Only the entangled task shows high correlation, i.e. the cell representations have warped around objects. Each point is a model trained from a random seed.

These are the same biological constraints as above, but now the functional constraints involve predicting location, object, and action, and an additional structural constraint imposes equation 3. Interestingly, the structural constraint leads to a pattern forming optimisation dynamics (see Appendix A.7 for mathematical details).

**Modules of distinct cell types when tasks are factorised and representations are nonnegative.** Just as our theory predicts, when training on tasks where objects and space are factorised (i.e. objects can be anywhere in space), under our biological constraints of nonnegativity and energy efficiency, distinct neural modules emerge, each selective for a single task factor (Fig. 6C-D). We see GC-like neurons that consistently represent space independent of object locations, and OVC-like neurons that recenter their representations around the moving objects or are inactive if no objects are present (further cells are shown in Fig. 13). To quantify whether the population really has two distinct cell types (two modules) we analyse the consistency of a cell's representation by taking its average spatial correlation between many different object configurations (Fig. 6E); some cells change a lot (i.e. OVCs) some cells don't change (i.e. GCs). Thus the representation has two disentangled modules. Without the nonnegativity constraint, the neural representation is entangled with mixed-selective neurons and no clear distinction between cells that change and those that don't (Fig. 6F-G).

**Grid cell warping and mixed-selectivity when tasks are entangled.** Experimental results show GCs sometimes warp their firing fields towards rewarded locations (Boccaro et al., 2019) and sometimes don't (Butler et al., 2019). Intriguingly, in the warping situation, the rodents exhibited stereotyped behaviour around consistent rewards. We now explain these neuroscience observations as a consequence of space becoming entangled with objects/rewards.

Modelling factorised versus entangled tasks (objects changing locations versus always staying in the same locations), produces very different GC behaviours. In the factorised case, grid fields are unrelated to objects (Fig. 7A), whereas in the entangled task grid fields warp to objects (Fig. 7B). We quantify this by measuring the average spatial correlation of patches around each object. Only when the task is entangled are fields consistently warped towards objects (Fig. 7C). Thus we have an explanation for GC warping; they warp when behaviour becomes stereotyped around consistently placed objects/rewards.

## 5 DISCUSSION

We have proven that simple biological constraints like nonnegativity and energy efficiency lead to disentangling, and empirically verified this in machine learning and neuroscience tasks, leading to a new understanding of functional cell types. We now consider some more neuroscience implications.

**Representing categories.** Our theory additionally says that representations of individual categories should be encoded in separate neural populations (disentangled; see Appendix A.4<sup>3</sup>). This provides a potential explanation for "grandmother cells" that selectively represent specific concepts or

categories (Quiroga et al., 2005), and have long puzzled proponents of distributed representations. It further potentially explains situations where animals have been trained on multiple tasks, and different neurons are found to engage in each task (Rainer et al., 1998; Roy et al., 2010; Asaad et al., 2000; Lee et al., 2022; Flesch et al., 2022).

**When to disentangle?** Our theory speaks to situations in which brains or networks must generalise to new combinations of learnt factors. In this situation, if the input-output (or input-latent-output) transformations are linear, biological constraints will cause complete disentangling of the networks. When the mappings are nonlinear, we show empirically that mixed selectivity exists, but gradually de-mixes as layers approach the output (in supervised), or latent (in unsupervised) network layers.

Optimising for low firing contrasts with previous ideas which instead optimise for linear read-out. This latter situation is akin to kernel regression, where mixed selectivity through random expansion increases the dimensionality of neural representations, allowing simple linear read-outs in the high dimensional space to perform arbitrary nonlinear operations on the original low dimensional space.

**Mixed-selective cells in the brain.** Mixed selectivity does clearly exist in the brain. For example, Kenyon cells in the Drosophila mushroom body increase the dimensionality of their inputs by an order of magnitude by close to random projections (Aso et al., 2014). This may allow linear read-out to behaviour via simple dopamine gating. Similarly rodent hippocampal cells encode conjunctions of spatial and sensory variables to allow rapid formation of new memories (Komorowski et al., 2009)]. More recently it has been suggested that PFC neurons have this same property, for the same reason (Rigotti et al., 2013). However, it is less clear that this is a general property of representations in associative cortex (including PFC), which can separate into different neuronal representations of different interpretable factors (Hirokawa et al., 2019; Bernardi et al., 2020) or tasks (Lee et al., 2022; Flesch et al., 2022).

One possibility is that in overtrained situations with only a relatively small number of categories or trial-types (where mixed selectivity has been observed), the task can effectively be solved by categorising the current trial into one of a few previous experiences. By contrast in tasks where combinatorial generalisation is required, the factored solution may be preferred.

**A program to understand how brain representations structure themselves.** This work is one piece of the puzzle. It tells us when neural circuits systems should represent different factors in different neurons. It does not tell us, however, how each factor itself should be represented. For example it does not tell us why GCs and OVCs look the ways they do. We believe that the same principles of nonnegativity, minimising neural activity, and representing structure, will be essential components obtaining this more general understanding. Indeed in a companion paper, we use the same constraints, along with formalising structure/path-integration using group and representations theory, to mathematically understand why grid cells look like grid cells (Dorrell et al., 2022). Similarly, our current understanding is limited to the optimal solution for factorised representations, but we anticipate similar ideas will be applicable to neural dynamics (Driscoll et al., 2022).

## 6 CONCLUSION

We introduced constraints inspired by biological neurons - nonnegativity and energy efficiency (w.r.t. either activity or weights) - and proved these constraints lead to linear factorised codes being disentangled. We empirically verified this in simulation, and showed the same constraints lead to disentangling with both non-linear data and nonlinear networks. We even achieve competitive disentangling scores on a baseline disentangling task, even though this was not our specific aim. We showed these biological constraints explain why neuroscientists observe bespoke cell types, e.g. GCs (Hafting et al., 2005), OVCs (Høydal et al., 2019), border vector cells (Solstad et al., 2008; Lever et al., 2009), since space, boundaries, and objects appear in a factorised form (i.e. occur in any independent combination), and so are optimally represented by different neural populations for each factor. These same principles explain why neurons in inferior temporal cortex are axis aligned to underlying factors of variation that generate the data they represent (Chang & Tsao, 2017; Bao et al., 2020; Higgins et al., 2021), why visual cortex neurons are decorrelated (Ecker et al., 2010), or why neurons in parietal cortex only selective for specific tasks (Lee et al., 2022). Lastly, we also

---

<sup>3</sup>We note this phenomena could also be accounted for with a sparsity constraint.

explained the confusing finding of grid fields warping towards rewards (Boccaro et al., 2019) as the space and rewards becoming entangled.

This work bridges the gap between single neuron and population responses, and offers an understanding of properties of neural representations in terms of task structure above and beyond just dimensionality. Additionally it demonstrates the utility of neurobiological considerations in designing machine learning algorithms. Overall, we hope this work demonstrates the promise of a unified research program that more deeply connects the neuroscience and machine learning communities to help in their combined quest to both understand and learn neural representations. Such a unified approach spanning brains and machines could help both sides, offering neuroscientists a deeper understanding of how cortical representations structure themselves, and offering machine learners novel ways to control and understand the representations their machines learn.

#### ACKNOWLEDGEMENTS

We thank Emile Mathieu and Andrew Saxe for helpful comments and advice on our manuscript. We thank the following funding sources: Sir Henry Wellcome Post-doctoral Fellowship (222817/Z/21/Z) to J.C.R.W.; the Gatsby Charitable Foundation to W.D.; the James S. McDonnell, Simons Foundations, NTT Research, and an NSF CAREER Award to S.G.; Wellcome Principal Research Fellowship (219525/Z/19/Z), Wellcome Collaborator award (214314/Z/18/Z), and Jean-François and Marie-Laure de Clermont-Tonnerre Foundation award (JSMF220020372) to T.E.J.B.; the Wellcome Centre for Integrative Neuroimaging and Wellcome Centre for Human Neuroimaging are each supported by core funding from the Wellcome Trust (203139/Z/16/Z, 203147/Z/16/Z).

## REFERENCES

- Wael F. Asaad, Gregor Rainer, and Earl K. Miller. Task-Specific Neural Activity in the Primate Prefrontal Cortex. *Journal of Neurophysiology*, 84(1):451–459, July 2000. doi: 10.1152/jn.2000.84.1.451. URL <https://journals.physiology.org/doi/full/10.1152/jn.2000.84.1.451>.
- Yoshinori Aso, Daisuke Hattori, Yang Yu, Rebecca M Johnston, Nirmala A Iyer, Teri-TB Ngo, Heather Dionne, LF Abbott, Richard Axel, Hiromu Tanimoto, and Gerald M Rubin. The neuronal architecture of the mushroom body provides a logic for associative learning. *eLife*, 3:e04577, December 2014. doi: 10.7554/eLife.04577. URL <https://doi.org/10.7554/eLife.04577>.
- Andrea Banino, Caswell Barry, Benigno Uria, Charles Blundell, Timothy Lillicrap, Piotr Mirowski, Alexander Pritzel, Martin J Chadwick, Thomas Degris, Joseph Modayil, Greg Wayne, Hubert Soyer, Fabio Viola, Brian Zhang, Ross Goroshin, Neil Rabinowitz, Razvan Pascanu, Charlie Beattie, Stig Petersen, Amir Sadik, Stephen Gaffney, Helen King, Koray Kavukcuoglu, Demis Hassabis, Raia Hadsell, and Dharshan Kumaran. Vector-based navigation using grid-like representations in artificial agents. *Nature*, 557(7705):429–433, May 2018. doi: 10.1038/s41586-018-0102-6. URL <http://www.nature.com/articles/s41586-018-0102-6>.
- Pinglei Bao, Liang She, Mason McGill, and Doris Y. Tsao. A map of object space in primate inferotemporal cortex. *Nature*, 583(7814):103–108, July 2020. doi: 10.1038/s41586-020-2350-5. URL <https://doi.org/10.1038/s41586-020-2350-5>.
- Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning. *arXiv preprint*, January 2022. URL <http://arxiv.org/abs/2105.04906>.
- Timothy E J Behrens, Timothy H Muller, James C. R. Whittington, Shirley Mark, Alon B Baram, Kimberly L Stachenfeld, and Zeb Kurth-nelson. What Is a Cognitive Map? Organizing Knowledge for Flexible Behavior. *Neuron*, 100(2):490–509, 2018. doi: 10.1016/j.neuron.2018.10.002. URL [https://www.cell.com/neuron/fulltext/S0896-6273\(18\)30856-0](https://www.cell.com/neuron/fulltext/S0896-6273(18)30856-0).
- Silvia Bernardi, Marcus K. Benna, Mattia Rigotti, Jérôme Munuera, Stefano Fusi, and C. Daniel Salzman. The Geometry of Abstraction in the Hippocampus and Prefrontal Cortex. *Cell*, 183(4):954–967.e21, November 2020. doi: 10.1016/j.cell.2020.09.031. URL <https://linkinghub.elsevier.com/retrieve/pii/S0092867420312289>.
- Charlotte N. Boccara, Michele Nardin, Federico Stella, Joseph O’Neill, and Jozsef Csicsvari. The entorhinal cognitive map is attracted to goals. *Science*, 363(6434):1443–1447, March 2019. doi: 10.1126/science.aav4837. URL <http://www.sciencemag.org/lookup/doi/10.1126/science.aav4837>.
- Yoram Burak and Ila R. Fiete. Accurate path integration in continuous attractor network models of grid cells. *PLoS Computational Biology*, 5(2):e1000291, February 2009. doi: 10.1371/journal.pcbi.1000291. URL <https://dx.plos.org/10.1371/journal.pcbi.1000291>.
- Christopher P. Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in \$\beta\$-VAE. *arXiv preprint*, April 2018. doi: 10.48550/arXiv.1804.03599. URL <http://arxiv.org/abs/1804.03599>.
- William N. Butler, Kiah Hardcastle, and Lisa M. Giocomo. Remembered reward locations restructure entorhinal spatial maps. *Science*, 363(6434):1447–1452, March 2019. doi: 10.1126/science.aav5297. URL <http://www.sciencemag.org/lookup/doi/10.1126/science.aav5297>.
- Le Chang and Doris Y. Tsao. The Code for Facial Identity in the Primate Brain. *Cell*, 169(6):1013–1028.e14, June 2017. doi: 10.1016/j.cell.2017.05.011. URL <http://dx.doi.org/10.1016/j.cell.2017.05.011>.

- Ricky T. Q. Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating Sources of Disentanglement in Variational Autoencoders. *Advances in Neural Information Processing Systems*, 31, 2018. URL <https://proceedings.neurips.cc/paper/2018/hash/1ee3dfcd8a0645a25a35977997223d22-Abstract.html>.
- Christopher J. Cueva and Xue-Xin Wei. Emergence of grid-like representations by training recurrent neural networks to perform spatial localization. *International Conference on Learning Representations*, 0:1–19, March 2018. URL <http://arxiv.org/abs/1803.07770>.
- Sachin S Deshmukh and James J Knierim. Influence of local objects on hippocampal representations: Landmark vectors and memory. *Hippocampus*, 23(4):253–67, April 2013. doi: 10.1002/hipo.22101. URL <http://www.ncbi.nlm.nih.gov/pubmed/23447419>.
- Yedidyah Dordek, Daniel Soudry, Ron Meir, and Dori Derdikman. Extracting grid cell characteristics from place cell inputs using non-negative principal component analysis. *eLife*, 5 (MARCH2016):1–36, 2016. doi: 10.7554/eLife.10094.
- W Dorrell, P Latham, TEJ Behrens, and JCR Whittington. Actionable Neural Representations: A Normative Account of Grid Cells. *In Prep*, 2022.
- Laura Driscoll, Krishna Shenoy, and David Sussillo. Flexible multitask computation in recurrent networks utilizes shared dynamical motifs. preprint, Neuroscience, August 2022. URL <http://biorxiv.org/lookup/doi/10.1101/2022.08.15.503870>.
- Alexander S. Ecker, Philipp Berens, Georgios A. Keliris, Matthias Bethge, Nikos K. Logothetis, and Andreas S. Tolias. Decorrelated Neuronal Firing in Cortical Microcircuits. *Science*, 327(5965): 584–587, January 2010. doi: 10.1126/science.1179867. URL <https://www.science.org/doi/abs/10.1126/science.1179867>.
- Yuguang Fang, K.A. Loparo, and Xiangbo Feng. Inequalities for the trace of matrix product. *IEEE Transactions on Automatic Control*, 39(12):2489–2490, December 1994. doi: 10.1109/9.362841. URL <https://ieeexplore.ieee.org/document/362841/>.
- Timo Flesch, Keno Juechems, Tsvetomira Dumbalska, Andrew Saxe, and Christopher Summerfield. Orthogonal representations for robust context-dependent task performance in brains and neural networks. *Neuron*, 110(7):1258–1270.e11, 2022. doi: 10.1016/j.neuron.2022.01.005. URL <https://doi.org/10.1016/j.neuron.2022.01.005>.
- Peiran Gao, Eric Trautmann, Byron M. Yu, Gopal Santhanam, Stephen I. Ryu, Krishna V. Shenoy, and Surya Ganguli. A theory of multineuronal dimensionality, dynamics and measurement. *bioRxiv preprint*, 0:214262, 2017. doi: 10.1101/214262. URL <https://www.biorxiv.org/content/early/2017/11/05/214262>.
- Ruiqi Gao, Jianwen Xie, Xue-Xin Wei, Song-Chun Zhu, and Ying Nian Wu. On Path Integration of Grid Cells: Group Representation and Isotropic Scaling. *Advances in Neural Information Processing Systems* 35, 0(NeurIPS):1–28, June 2021. URL <http://arxiv.org/abs/2006.10259>.
- Jeffrey L. Gauthier and David W. Tank. A Dedicated Population for Reward Coding in the Hippocampus. *Neuron*, 99(1):179–193.e7, 2018. doi: 10.1016/j.neuron.2018.06.008. URL <https://doi.org/10.1016/j.neuron.2018.06.008>.
- Torkel Hafting, Marianne Fyhn, Sturla Molden, May-britt Britt Moser, and Edvard I. Moser. Microstructure of a spatial map in the entorhinal cortex. *Nature*, 436(7052):801–806, 2005. doi: 10.1038/nature03721.
- Kiah Hardcastle, Niru Maheswaranathan, Surya Ganguli, and Lisa M. Giocomo. A Multiplexed, Heterogeneous, and Adaptive Code for Navigation in Medial Entorhinal Cortex. *Neuron*, 94 (2):375–387.e7, 2017. doi: 10.1016/j.neuron.2017.03.025. URL <http://dx.doi.org/10.1016/j.neuron.2017.03.025>.

- Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. *International Conference on Learning Representations*, 0, July 2017a. URL <https://openreview.net/forum?id=Sy2fzU9g1>.
- Irina Higgins, Arka Pal, Andrei A. Rusu, Loic Matthey, Christopher P Burgess, Alexander Pritzel, Matthew Botvinick, Charles Blundell, and Alexander Lerchner. DARLA: Improving Zero-Shot Transfer in Reinforcement Learning. *arXiv preprint*, 2017b. doi: 10.3109/17482620903223036. URL <http://arxiv.org/abs/1707.08475>.
- Irina Higgins, Nicolas Sonnerat, Loic Matthey, Arka Pal, Christopher P Burgess, Matko Bosnjak, Murray Shanahan, Matthew Botvinick, Demis Hassabis, and Alexander Lerchner. SCAN: Learning Hierarchical Compositional Visual Concepts. *arXiv preprint*, pp. 1–24, 2018. doi: 10.1186/s12884-017-1520-4. URL <http://arxiv.org/abs/1707.03389>.
- Irina Higgins, Le Chang, Victoria Langston, Demis Hassabis, Christopher Summerfield, Doris Tsao, and Matthew Botvinick. Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons. *Nature Communications*, 12(1):1–14, 2021. doi: 10.1038/s41467-021-26751-5.
- Geoffrey E. Hinton, Alex Krizhevsky, and Sida D. Wang. Transforming Auto-Encoders. *International Conference on Artificial Neural Networks*, 6791:44–51, 2011. doi: 10.1007/978-3-642-21735-7\_6. URL [http://link.springer.com/10.1007/978-3-642-21735-7\\_6](http://link.springer.com/10.1007/978-3-642-21735-7_6).
- Junya Hirokawa, Alexander Vaughan, Paul Masset, Torben Ott, and Adam Kepecs. Frontal cortex neuron types categorically encode single decision variables. *Nature*, 576(7787):446–451, December 2019. doi: 10.1038/s41586-019-1816-9. URL <https://www.nature.com/articles/s41586-019-1816-9>.
- Daniella Horan, Eitan Richardson, and Yair Weiss. When Is Unsupervised Disentanglement Possible? *Advances in Neural Information Processing Systems*, 34:5150–5161, 2021. URL <https://proceedings.neurips.cc/paper/2021/hash/29586cb449c90e249f1f09a0a4ee245a-Abstract.html>.
- Øyvind Arne Høydal, Emilie Ranheim Skytøen, Sebastian Ola Andersson, May-Britt Moser, and Edvard Ingjald Moser. Object-vector coding in the medial entorhinal cortex. *Nature*, 568(7752):400–404, April 2019. doi: 10.1038/s41586-019-1077-7. URL <http://www.nature.com/articles/s41586-019-1077-7>.
- Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational Autoencoders and Nonlinear ICA: A Unifying Framework. *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pp. 2207–2217, June 2020. URL <https://proceedings.mlr.press/v108/khemakhem20a.html>.
- Hyunjik Kim and Andriy Mnih. Disentangling by Factorising. *Proceedings of the 35th International Conference on Machine Learning*, pp. 2649–2658, July 2018. URL <https://proceedings.mlr.press/v80/kim18b.html>.
- Diederik P. Kingma and Jimmy Lei Ba. Adam: A Method for Stochastic Optimization. *arXiv preprint*, 0, 2014. doi: <http://doi.acm.org.ezproxy.lib.ucf.edu/10.1145/1830483.1830503>. URL <http://arxiv.org/abs/1412.6980>.
- Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. *arXiv preprint*, 0(MI):1–14, 2013. doi: 10.1051/0004-6361/201527329. URL <http://arxiv.org/abs/1312.6114>.
- Robert W. Komorowski, Joseph R. Manns, and Howard Eichenbaum. Robust Conjunctive Item-Place Coding by Hippocampal Neurons Parallels Learning What Happens Where. *Journal of Neuroscience*, 29(31):9918–9929, August 2009. doi: 10.1523/JNEUROSCI.1378-09.2009. URL <https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.1378-09.2009>.

Julia Julija Krupic, Neil Burgess, John O’Keefe, and John O’Keefe. Neural Representations of Location Composed of Spatially Periodic Bands. *Science*, 337(6096):853–857, August 2012. doi: 10.1126/science.1222403. URL <https://www.science.org/doi/10.1126/science.1222403>.

Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational Inference of Disentangled Latent Concepts from Unlabeled Observations. *International Conference on Learning Representations*, pp. 16, 2018.

Daniel Kunin, Jonathan M. Bloom, Aleksandrina Goeva, and Cotton Seed. Loss Landscapes of Regularized Linear Autoencoders. *arXiv preprint*, 2019. URL <http://arxiv.org/abs/1901.08168>.

Daniel D Lee and H Sebastian Seung. Algorithms for Non-negative Matrix Factorization. *Advances in Neural Information Processing Systems*, 0(1), 2000.

Julie J. Lee, Michael Krumin, Kenneth D. Harris, and Matteo Carandini. Task specificity in mouse parietal cortex. *Neuron*, August 2022. doi: 10.1016/j.neuron.2022.07.017. URL <https://www.sciencedirect.com/science/article/pii/S0896627322006626>.

Colin Lever, Stephen Burton, Ali Jeewajee, John O’Keefe, and Neil Burgess. Boundary vector cells in the subiculum of the hippocampal formation. *Journal of Neuroscience*, 29(31):9771–9777, 2009. doi: 10.1523/JNEUROSCI.1319-09.2009.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations, June 2019. URL <http://arxiv.org/abs/1811.12359>.

Lane McIntosh, Niru Maheswaranathan, Aran Nayebi, Surya Ganguli, and Stephen Baccus. Deep Learning Models of the Retinal Response to Natural Scenes. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL <https://proceedings.neurips.cc/paper/2016/hash/a1d33d0dfec820b41b54430b50e96b5c-Abstract.html>.

Samuel Ocko, Jack Lindsey, Surya Ganguli, and Stephane Deny. The emergence of multiple retinal cell types through efficient coding of natural movies. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. doi: 10.1101/458737. URL <https://papers.nips.cc/paper/2018/hash/d94fd74dcde1aa553be72c1006578b23-Abstract.html>.

John O’Keefe and J. Dostrovsky. The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Research*, 34(1):171–175, November 1971. doi: 10.1016/0006-8993(71)90358-1. URL <http://linkinghub.elsevier.com/retrieve/pii/0006899371903581>.

Bruno A. Olshausen and David J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images, 1996.

Cengiz Pehlevan, Tao Hu, and Dmitri B. Chklovskii. A Hebbian/Anti-Hebbian Neural Network for Linear Subspace Learning: A Derivation from Multidimensional Scaling of Streaming Data. *Neural Computation*, 27(7):1461–1495, July 2015. doi: 10.1162/NECO\_a\_00745. URL <http://arxiv.org/abs/1503.00669>.

R. Quian Quiroga, L. Reddy, G. Kreiman, C. Koch, and I. Fried. Invariant visual representation by single neurons in the human brain. *Nature*, 435(7045):1102–1107, June 2005. doi: 10.1038/nature03687. URL <https://www.nature.com/articles/nature03687>.

Gregor Rainer, Wael F. Asaad, and Earl K. Miller. Selective representation of relevant information by neurons in the primate prefrontal cortex. *Nature*, 393(6685):577–579, June 1998. doi: 10.1038/31235. URL <https://www.nature.com/articles/31235>.

Blake A. Richards, Timothy P. Lillicrap, Philippe Beaudoin, Yoshua Bengio, Rafal Bogacz, Amelia Christensen, Claudia Clopath, Rui Ponte Costa, Archy de Berker, Surya Ganguli, Colleen J. Gillon, Danijar Hafner, Adam Kepecs, Nikolaus Kriegeskorte, Peter Latham, Grace W. Lindsey, Kenneth D. Miller, Richard Naud, Christopher C. Pack, Panayiota Poirazi, Pieter Roelfsema, João Sacramento, Andrew Saxe, Benjamin Scellier, Anna C. Schapiro, Walter Senn, Greg Wayne, Daniel Yamins, Friedemann Zenke, Joel Zylberberg, Denis Therien, and Konrad P. Kording. A deep learning framework for neuroscience. *Nature Neuroscience*, 22(11):1761–1770, November 2019. doi: 10.1038/s41593-019-0520-2. URL <https://www.nature.com/articles/s41593-019-0520-2>.

Karl Ridgeway and Michael C Mozer. Learning Deep Disentangled Embeddings With the F-Statistic Loss. *Advances in Neural Information Processing Systems*, 31, 2018. URL <https://proceedings.neurips.cc/paper/2018/hash/2b24d495052a8ce66358eb576b8912c8-Abstract.html>.

Mattia Rigotti, Omri Barak, Melissa R Warden, Xiao-Jing Wang, Nathaniel D Daw, Earl K Miller, and Stefano Fusi. The importance of mixed selectivity in complex cognitive tasks. *Nature*, 497(7451):1–6, 2013. doi: 10.1038/nature12160. URL <http://dx.doi.org/10.1038/nature12160>.

Jefferson E. Roy, Maximilian Riesenhuber, Tomaso Poggio, and Earl K. Miller. Prefrontal Cortex Activity during Flexible Categorization. *Journal of Neuroscience*, 30(25):8519–8528, June 2010. doi: 10.1523/JNEUROSCI.4837-09.2010. URL <https://www.jneurosci.org/content/30/25/8519>.

Ayelet Sarel, Arseny Finkelstein, Liora Las, and Nachum Ulanovsky. Vectorial representation of spatial goals in the hippocampus of bats. *Science*, 355(6321):176–180, 2017. doi: 10.1126/science.aak9589.

Trygve Solstad, Charlotte N Boccara, Emilio Kropff, May-Britt Moser, and Edvard I Moser. Representation of Geometric Borders in the Entorhinal Cortex. *Science*, 322(5909):1865–1868, December 2008. doi: 10.1126/science.1166466. URL <https://www.sciencemag.org/lookup/doi/10.1126/science.1166466>.

Ben Sorscher, Gabriel C Mel, Surya Ganguli, and Samuel A Ocko. A unified theory for the origin of grid cells through the lens of pattern formation. *Advances in Neural Information Processing Systems* 32, 32(NeurIPS):10003–10013, 2019.

Ben Sorscher, Gabriel C Mel, Samuel A Ocko, Lisa Giocomo, and Surya Ganguli. A unified theory for the computational and mechanistic origins of grid cells. *bioRxiv preprint*, pp. 2020.12.29.424583, 2020. URL <https://doi.org/10.1101/2020.12.29.424583>.

Carsen Stringer, Marius Pachitariu, Nicholas Steinmetz, Matteo Carandini, and Kenneth D. Harris. High-dimensional geometry of population responses in visual cortex. *Nature*, 2019. doi: 10.1038/s41586-019-1346-5. URL <http://dx.doi.org/10.1038/s41586-019-1346-5>.

James C. R. Whittington, Rishabh Kabra, Loic Matthey, Christopher P. Burgess, and Alexander Lerchner. Constellation: Learning relational abstractions over objects for compositional imagination. *arXiv preprint*, 2021a. URL <http://arxiv.org/abs/2107.11153>.

James C. R. Whittington, Joseph Warren, and Tim E. J. Behrens. Relating transformers to models and neural representations of the hippocampal formation. *International Conference on Learning Representations*, September 2021b. URL <https://openreview.net/pdf?id=B8DV09B1YE0>.

Daniel L K Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 111(23):8619–8624, 2014. doi: 10.1073/pnas.1403112111.

Guangyu Robert Yang, Madhura R. Joglekar, H. Francis Song, William T. Newsome, and Xiao-Jing Wang. Task representations in neural networks trained to perform many cognitive tasks. *Nature Neuroscience*, 22(2):297–306, February 2019. doi: 10.1038/s41593-018-0310-2. URL <https://www.nature.com/articles/s41593-018-0310-2>.

## A APPENDIX

### A.1 DEFINITIONS

**Disentangling definition.** Firstly, before any proofs, we define disentangling as when single *model* neurons care about single ground truth factors. We do not mind if more than one model neuron cares about the same factor. We note that other definitions of disentangling are when *single* ground truth factors of variation are encoded in at most a *single* model neuron - we do not ask for this level of parsimony.

**Disentangling metric.** Our metric (mutual information ratio; MIR) measures the mutual information between neurons and factors,  $I_{n,f}$ , then calculates each neuron's preference (we exclude inactive neurons) for one factor over the rest via

$$r_n = \frac{\max_f(I_{n,f})}{\sum_f I_{n,f}} \quad (5)$$

We then divide by the number of (active) neurons,  $n_n$ , and normalise for the number of factors,  $n_f$ , i.e.

$$MIR = \frac{\frac{\sum_n r_n}{n_n} - \frac{1}{n_f}}{1 - \frac{1}{n_f}} \quad (6)$$

High MIR means single neurons show high preference for a single ground truth factor.

## A.2 DETAILS OF DERIVATIONS

### A.2.1 PROOF OF CONSTRAINTS LEADING TO DISENTANGLING FOR A GIVEN POPULATION VARIANCE.

**Theorem.** Let  $\mathbf{e} \in \mathbb{R}^k$  be a random vector whose  $k$  independent components denote  $k$  task factors. We assume each independent task factor  $e_i$  is drawn from a distribution that has mean 0, variance  $\sigma^2$ , and maximum and minimum values of  $\min(e_i) = -a$  and  $\max(e_i) = a$ . Also let  $\mathbf{z} \in \mathbb{R}^n$  be a linear neural representation of the task factors given by

$$\mathbf{z} = \mathbf{M}\mathbf{e} + \mathbf{b}_z, \quad (7)$$

where  $\mathbf{M} \in \mathbb{R}^{n \times k}$  are mixing weights and  $\mathbf{b}_z \in \mathbb{R}^n$  is a bias. We further assume two constraints: (1) the neural representation is *nonnegative* with  $z_i \geq 0$  for all  $i = 1, \dots, n$ , and (2) the neural population variance is a nonzero constant,  $\sum_j Var(z_j) = C$ , so that the neural representation retains some information about the task variables. Under these two constraints we show that in the space of all possible neural representations (parameterised by  $\mathbf{M}$  and  $\mathbf{b}_z$ ), the representations that achieve minimal activity energy  $\mathbb{E}\|\mathbf{z}\|^2$  also exhibit disentangling, by which we mean every neuron  $z_j$  is selective for at most one task parameter: i.e.  $|M_{jk}|M_{jl}| = 0$  for  $k \neq l$ .

**Proof.** We aim to find a representation,  $\mathbf{z}$ , that minimises activity energy (the expected norm of  $\mathbf{z}$ ), is nonnegative, all for a fixed population variance,  $C$ . This is a constrained optimisation problem, and equates to understanding what  $\mathbf{M}$  and  $\mathbf{b}_z$  must look like in order to satisfy our constraints, and minimise  $\mathbb{E}\|\mathbf{z}\|^2$ .

$$\underset{\mathbf{M}, \mathbf{b}_z}{\text{minimise}} \mathbb{E}\|\mathbf{z}\|^2 \quad \text{s.t. } z_i \geq 0, \sum_j Var(z_j) = C \quad (8)$$

The total activity energy, i.e. the expected norm of  $\mathbf{z}$ , is

$$\begin{aligned} \mathbb{E}\|\mathbf{z}\|^2 &= \sum_j \mathbb{E}(z_j^2) = \sum_j Var(z_j) + (\mathbb{E}z_j)^2 \\ &= \sum_j Var((\mathbf{M}\mathbf{e}_t)_j) + \sum_j ((\mathbf{b}_z)_j)^2 \end{aligned} \quad (9)$$

We want to minimise this, under the constraint  $z_j \geq 0$ . To satisfy this constraint,  $(\mathbf{b}_z)_j$  must account for any negativity. This means that

$$(\mathbf{b}_z)_j \geq -\min(\mathbf{M}\mathbf{e})_j = \sum_k |M_{jk}|a \quad (10)$$

Where the last equality is because all  $e_k$  are i.i.d. and so the minimum of a sum of random variables is the sum of their minima. The modulus sign is since  $M_{jk}$  can be positive or negative, so we need to consider the maximum and minimum of  $e_k$ , and an assumption of ours was that the maximum and minimum had the same value  $a$ . Thus

$$(\mathbf{b}_z)_j = \sum_k |M_{jk}|a + p_j \quad (11)$$

Where  $p_j \geq 0$ . Intuitively,  $p_j = 0$  since anything else would increase activity energy. Formally,

$$\begin{aligned} \mathbb{E}\|\mathbf{z}\|^2 &= \sum_j Var((\mathbf{M}\mathbf{e}_t)_j) + \sum_j ((\mathbf{b}_z)_j)^2 \\ &= \sigma^2 \sum_{j,k} M_{jk}^2 + \sum_j (\sum_k |M_{jk}|a + p_j)^2 \\ &= \sigma^2 \sum_{j,k} M_{jk}^2 + \sum_j (\sum_k |M_{jk}|a)^2 + p_j^2 + 2p_j^2 \sum_k |M_{jk}|a \end{aligned} \quad (12)$$

Since both  $\sum_k |M_{jk}|a$  and  $p_j$  are greater than zero,  $p_j$  always increases activity energy, no matter what  $M_{jk}$  is. Thus we set it to zero, i.e.  $p_j = 0$ . Hence we can simplify the expected activity as

follows

$$\begin{aligned}
 \mathbb{E}\|\mathbf{z}\|^2 &= \sigma^2 \sum_{j,k} M_{jk}^2 + \sum_j (\sum_k |M_{jk}|a)^2 \\
 &= \sigma^2 \sum_{j,k} M_{jk}^2 + \sum_j (\sum_k |M_{jk}|a)(\sum_l |M_{jl}|a) \\
 &= \sigma^2 \sum_{j,k} M_{jk}^2 + a^2 \sum_{j,k} (M_{jk}^2 + \sum_{l \neq k} |M_{jk}||M_{jl}|) \\
 &= (1 + \frac{a^2}{\sigma^2}) \sum_j Var(z_j) + a^2 \sum_{j,k,l \neq k} |M_{jk}||M_{jl}| \\
 &= (1 + \frac{a^2}{\sigma^2})C + a^2 \sum_{j,k,l \neq k} |M_{jk}||M_{jl}|
 \end{aligned} \tag{13}$$

Now all the constraints have been incorporated, our only job is to minimise  $\mathbb{E}\|\mathbf{z}\|^2$ . This is done when  $\sum_{j,k,l \neq k} |M_{jk}||M_{jl}| = 0$ , and that only happens when  $|M_{jk}||M_{jl}| = 0$  for all  $j$  and  $l \neq k$ . In words, this means that neuron  $j$  in only receives information from one element of  $\mathbf{e}$ . This is disentangling.

#### A.2.2 PROOF THAT POPULATION VARIANCE IS BOUNDED BY READ-OUT WEIGHTS NORM

**Theorem.** Let  $\mathbf{x} = \mathbf{D}\mathbf{e}$  be observed entangled data, where  $\mathbf{x} \in \mathbb{R}^m$ ,  $\mathbf{D} \in \mathbb{R}^{m \times k}$ , and  $\mathbf{e} \in \mathbb{R}^k$  is a random vector whose  $k$  independent components denote  $k$  task factors. We assume each independent task factor  $e_i$  is drawn from a distribution that has mean 0, variance  $\sigma^2$ , and maximum and minimum values of  $\min(e_i) = -a$  and  $\max(e_i) = a$ . Let a neural representation  $\mathbf{z} \in \mathbb{R}^n$  exactly predict observed data via  $\mathbf{x} = \mathbf{W}\mathbf{z} + \mathbf{b}_x$  with zero error, i.e.  $\mathbf{x} = \mathbf{D}\mathbf{e} = \mathbf{W}\mathbf{z} + \mathbf{b}_x$ . Where  $\mathbf{W} \in \mathbb{R}^{m \times n}$   $m \geq n \geq k$ , are read-out weights and  $\mathbf{b}_x \in \mathbb{R}^m$  is an offset.

Then for all such data generation models (with parameters  $\mathbf{D}$ ) and all such neural representations (with parameters  $\mathbf{W}$  and  $\mathbf{b}_x$ ), as long as: (1) the smallest singular value of  $\mathbf{D}$  is non-zero,  $\sigma_{\min}(\mathbf{D}) > 0$ ; (2) the norm of the read-out weights  $\|\mathbf{W}\|_F^2$  is finite; then the population variance of  $\mathbf{z}$  is bounded from below by the norm of the read-out weights.

$$\sum_j Var(z_j) \geq n^2 \frac{\sigma_{\min}^2(\mathbf{D})}{\|\mathbf{W}\|_F^2} \tag{14}$$

**Proof.** Since  $\mathbf{x} = \mathbf{W}\mathbf{z} + \mathbf{b}_x$ , we have

$$\begin{aligned}
 \mathbf{z} &= \mathbf{W}^+ \mathbf{x} - \mathbf{W}^+ \mathbf{b}_x \\
 &= \mathbf{W}^+ \mathbf{D}\mathbf{e} - \mathbf{W}^+ \mathbf{b}_x
 \end{aligned} \tag{15}$$

Where  $\mathbf{W}^+$  is the Moore-Penrose pseudoinverse. Additionally since the read-out error is zero,  $\mathbf{W}$  must contain all the same information as  $\mathbf{D}$ , and so must be  $\mathbf{W} = \mathbf{D}\mathbf{F}^+$ , where  $\mathbf{F}^+ \in \mathbb{R}^{k \times n}$  is a matrix with rank  $k$  ( $n \geq k$ ). The pseudoinverse of  $\mathbf{W}$  is thus  $\mathbf{W}^+ = \mathbf{F}\mathbf{D}^+$ . The population variance becomes

$$\begin{aligned}
 \sum_j Var(z_j) &= Tr(Var(\mathbf{W}^+ \mathbf{D}\mathbf{e} - \mathbf{W}^+ \mathbf{b}_x)) \\
 &= Tr(Var(\mathbf{F}\mathbf{D}^+ \mathbf{D}\mathbf{e})) \\
 &= Tr(Var(\mathbf{F}\mathbf{e})) \\
 &= Tr(\mathbf{F}Var(\mathbf{e})\mathbf{F}^T) \\
 &= \sigma^2 Tr(\mathbf{F}\mathbf{F}^T) \\
 &= \sigma^2 \|\mathbf{F}\|_F^2
 \end{aligned} \tag{16}$$

Using the fact that the norm of a matrix's pseudoinverse is bounded by the norm of the matrix

$$\|\mathbf{F}^+\|_F^2 = Tr((\mathbf{F}^T \mathbf{F})^{-1}) \geq \frac{n^2}{\|\mathbf{F}\|_F^2} \tag{17}$$

Along with using the following trace identity (Fang et al., 1994)

$$\|\mathbf{F}^+\|_F^2 \sigma_{min}^2(\mathbf{D}) \leq \|\mathbf{W}\|_F^2 = \|\mathbf{D}\mathbf{F}^+\|_F^2 \leq \|\mathbf{F}^+\|_F^2 \sigma_{max}^2(\mathbf{D}) \quad (18)$$

Where  $\sigma_{min}(\mathbf{D})$  and  $\sigma_{max}(\mathbf{D})$  are the smallest and largest singular values of  $\mathbf{D}$ . Thus

$$\frac{1}{\|\mathbf{F}^+\|_F^2} \geq \frac{\sigma_{min}^2(\mathbf{D})}{\|\mathbf{W}\|_F^2} \quad (19)$$

Combining it all together

$$\begin{aligned} \sum_j Var(z_j) &= \sigma^2 \|\mathbf{F}\|_F^2 \\ &\geq \frac{n^2 \sigma^2}{\|\mathbf{F}^+\|_F^2} \\ &\geq \frac{n^2 \sigma^2 \sigma_{min}^2(\mathbf{D})}{\|\mathbf{W}\|_F^2} \end{aligned} \quad (20)$$

We note that the first inequality becomes an equality if and only if  $\mathbf{F}^+$  has (scaled) orthonormal *rows*, and the second inequality becomes an equality if and only if  $\mathbf{D}$  has (scaled) orthonormal *columns*.

#### A.2.3 PROOF OF DISENTANGLING WHEN DATA GENERATIVE MATRIX HAS (SCALED) ORTHONORMAL COLUMNS

**Theorem.** Let  $\mathbf{x} = \mathbf{D}\mathbf{e}$  be observed entangled data, where  $\mathbf{D} \in \mathbb{R}^{m \times k}$ , and  $\mathbf{e} \in \mathbb{R}^k$  is a random vector whose  $k$  independent components denote  $k$  task factors. We assume each independent task factor  $e_i$  is drawn from a distribution that has mean 0, variance  $\sigma^2$ , and maximum and minimum values of  $\min(e_i) = -a$  and  $\max(e_i) = a$ . Let a neural representation  $\mathbf{z} \in \mathbb{R}^n$  exactly predict observed data via  $\mathbf{x} = \mathbf{W}\mathbf{z} + \mathbf{b}_x$  with zero error, i.e.  $\mathbf{x} = \mathbf{D}\mathbf{e} = \mathbf{W}\mathbf{z} + \mathbf{b}_x$ . Where  $\mathbf{W} \in \mathbb{R}^{m \times n}$   $m \geq n \geq k$ , are read-out weights and  $\mathbf{b}_x \in \mathbb{R}^m$  is an offset.

Then for all such data generation models (with parameters  $\mathbf{D}$ ) and all such neural representations (with parameters  $\mathbf{W}$  and  $\mathbf{b}_x$ ), as long as: (1) the columns of  $\mathbf{D}$  are (scaled) orthonormal; (2) the norm of the read-out weights  $\|\mathbf{W}\|_F^2$  is finite; (3) the neural representation is nonnegative (i.e.  $\mathbf{z} > 0$ ), then out of all such neural representations, the minimum energy representations are also disentangled ones. By this we mean that each neuron  $z_i$  will be selective for at most one hidden task factor  $e_j$ .

**Proof.** We aim to find a representation,  $\mathbf{z}$ , that minimises activity energy (the expected norm of  $\mathbf{z}$ ), is nonnegative, and predicts data  $\mathbf{x}$  via  $\mathbf{x} = \mathbf{W}\mathbf{z} + \mathbf{b}_x$  with zero error ( $\mathbf{D}\mathbf{e} = \mathbf{W}\mathbf{z} + \mathbf{b}_x$ ). This is a constrained optimisation problem, and equates to understanding what  $\mathbf{W}$  and  $\mathbf{b}_x$  (and therefore  $\mathbf{z}$ ) must look like in order to satisfy our constraints, and minimise  $\mathbb{E}\|\mathbf{z}\|^2$ .

$$\underset{\mathbf{W}, \mathbf{b}_x}{\text{minimise}} \mathbb{E}\|\mathbf{z}\|^2 \quad \text{s.t. } z_i \geq 0, \mathbf{D}\mathbf{e} = \mathbf{W}\mathbf{z} + \mathbf{b}_x \forall \mathbf{e}, \|\mathbf{W}\|_F^2 = K \quad (21)$$

We note that this is the classic matrix factorisation setting. The difference here is that we ask the representation  $\mathbf{z}$  to be nonnegative.

Firstly, we note that since the columns of  $\mathbf{D}$  are (scaled) orthonormal, then the singular values are all equal:  $\sigma_{min}(\mathbf{D}) = \sigma_{max}(\mathbf{D}) = \sigma(\mathbf{D})$ . This result is useful as now the inequality in equation 19 becomes an equality

$$\frac{1}{\|\mathbf{F}^+\|_F^2} = \frac{\sigma^2(\mathbf{D})}{\|\mathbf{W}\|_F^2} \quad (22)$$

Now we prove disentangling. Since the read-out error is zero,  $\mathbf{W}$  must contain all the same information as  $\mathbf{D}$  and so must be  $\mathbf{W} = \mathbf{D}\mathbf{F}^+$ . Using this, and repeating a similar process to the first

proof (Appendix A.2.1), we get

$$\begin{aligned}
 \mathbb{E}\|\mathbf{z}\|^2 &= (1 + \frac{a^2}{\sigma^2}) \sum_j Var(z_j) + a^2 \sum_{j,k,l \neq k} |(\mathbf{W}^\dagger \mathbf{D})_{jk}| |(\mathbf{W}^\dagger \mathbf{D})_{jl}| \\
 &= (\sigma^2 + a^2) \|\mathbf{F}\|_F^2 + a^2 \sum_{j,k,l \neq k} |F_{jk}| |F_{jl}| \\
 &\geq (\sigma^2 + a^2) \frac{n^2}{\|\mathbf{F}\|_F^2} + a^2 \sum_{j,k,l \neq k} |F_{jk}| |F_{jl}| \\
 &= (\sigma^2 + a^2) \frac{n^2 \sigma^2(\mathbf{D})}{\|\mathbf{W}\|_F^2} + a^2 \sum_{j,k,l \neq k} |F_{jk}| |F_{jl}|
 \end{aligned} \tag{23}$$

Where the inequality is due to equation 17. This inequality becomes an equality if and only if  $\mathbf{F}^\dagger$  has (scaled) orthonormal rows (i.e.  $\mathbf{F}$  has (inverse scaled) orthonormal columns). This means that for a fixed  $\|\mathbf{W}\|_F^2$ , the first term in equation 23 is minimised when  $\mathbf{F}^\dagger$  has (scaled) orthonormal rows, or equally when  $\mathbf{F}$  has (inverse scaled) orthonormal columns. This is interesting to us, as we can additionally make the second term in equation 23 go to zero when  $\mathbf{F}$  is a *particular* type of matrix with (inverse scaled) orthonormal columns. Thus we will have fully optimised equation 23, and this will be our solution. First, rewriting  $\mathbf{F}$  as a matrix with (inverse scaled) orthonormal columns

$$\mathbf{F} = \frac{1}{\alpha} \mathbf{O} \quad \text{and} \quad \mathbf{F}^\dagger = \alpha \mathbf{O}^T \tag{24}$$

Where  $\mathbf{O}$  is a matrix with orthonormal columns. The particular  $\mathbf{F}$  that minimises the second term in equation 23, is when  $\mathbf{O}$  only has, at most, a single non-zero element per row as this sets  $|F_{jk}| |F_{jl}| = 0$ . This is disentangling and is easily seen by

$$\begin{aligned}
 \mathbf{z} &= \mathbf{W}^\dagger \mathbf{x} - \mathbf{W}^\dagger \mathbf{b}_x \\
 &= \mathbf{W}^\dagger \mathbf{D} \mathbf{e} - \mathbf{W}^\dagger \mathbf{b}_x \\
 &= \alpha \mathbf{O} \mathbf{D}^\dagger \mathbf{D} \mathbf{e} - \alpha \mathbf{O} \mathbf{D}^\dagger \mathbf{b}_x \\
 &= \alpha \mathbf{O} \mathbf{e} - \alpha \mathbf{O} \mathbf{D}^\dagger \mathbf{b}_x
 \end{aligned} \tag{25}$$

Since  $\mathbf{O}$  only has a single non-zero element per row, then each  $\mathbf{z}_i$  must only contain a single element (random variable) from  $\mathbf{e}$ . We note that nonnegativity is easily achieved by  $\mathbf{b}_x$  learning to take a value that satisfies nonnegativity, i.e.  $\alpha \mathbf{O}^T \mathbf{D}^\dagger \mathbf{b}_x = \min \alpha \mathbf{O}^T \mathbf{e}$ . Since  $\mathbf{O}$  and  $\mathbf{D}$  are full ( $k$ ) rank, this is always possible.

We also offer an alternate proof in Appendix A.2.5 when analysing ‘Simplification 1’.

#### A.2.4 PROOF THAT A TALL RANDOM MATRIX WITH FINITE WIDTH HAS (SCALED) APPROXIMATELY ORTHONORMAL COLUMNS

**Theorem.** Let  $\mathbf{D} \in \mathbb{R}^{m \times k}$  be a random matrix with elements  $D_{ij}$  that are i.d.d. with expectation 0 and variance  $\kappa^2/m^2$ ). Then as  $m \rightarrow \infty$ , with finite  $k$ ,  $\mathbf{D}^T \mathbf{D} \rightarrow \mathbf{I}$ .

**Proof.**

$$\begin{aligned}
 \lim_{m \rightarrow \infty} (\mathbf{D}^T \mathbf{D})_{ik} &= \lim_{m \rightarrow \infty} \sum_j D_{ij}^T D_{jk} \\
 &= m^2 \mathbb{E} D_{ji} D_{jk} \\
 &= \begin{cases} \kappa^2, & i = k \\ 0, & i \neq k \end{cases}
 \end{aligned} \tag{26}$$

Thus  $\mathbf{D}$  has orthonormal columns, scaled by  $\kappa$ , and so its singular values are all identical;  $\sigma_{\min}(\mathbf{D}) = \sigma_{\max}(\mathbf{D}) = \kappa^2$ .

#### A.2.5 WHEN THE DATA GENERATIVE MATRIX DOES NOT HAVE (SCALED) ORTHONORMAL COLUMNS

While we do not prove the general case, we offer some intuition here that suggest disentangled representations are favoured in many situations. Consider the general setting in which  $\mathbf{D} \in \mathbb{R}^{m \times k}$

has the following singular value decomposition (SVD):

$$\mathbf{D} = \mathbf{U}\Sigma\mathbf{V} \quad (27)$$

Where  $\Sigma \in \mathbb{R}^{m \times k}$  is a rectangular diagonal matrix with positive entries, and  $\mathbf{U} \in \mathbb{R}^{m \times m}$  and  $\mathbf{V} \in \mathbb{R}^{k \times k}$  are orthogonal matrices. As in the above proofs, since  $\mathbf{x}$  is predicted with zero error from  $\mathbf{z}$ , via

$$\mathbf{x} = \mathbf{D}\mathbf{e} = \mathbf{W}\mathbf{z} + \mathbf{b}_x \quad (28)$$

Thus  $\mathbf{W}$  must be of the form  $\mathbf{W} = \mathbf{D}\mathbf{F}^+$ , where  $\mathbf{F}^+ \in \mathbb{R}^{k \times n}$  is a rank  $k$  matrix, since  $\mathbf{z} \in \mathbb{R}^n$ ,  $\mathbf{x} \in \mathbb{R}^m$ , and  $\mathbf{e} \in \mathbb{R}^k$ ,  $m \geq n \geq k$ . We define the SVD of  $\mathbf{F}$  as

$$\mathbf{F} = \mathbf{O}\Lambda\mathbf{R}^T, \quad \mathbf{F}^+ = \mathbf{R}\Lambda^{-1}\mathbf{O}^T \quad (29)$$

Where  $\Lambda \in \mathbb{R}^{k \times n}$  is a rectangular diagonal matrix with positive entries,  $\lambda_i$ ,  $\Lambda^{-1} \in \mathbb{R}^{n \times k}$  is a rectangular diagonal matrix with positive entries,  $\frac{1}{\lambda_i}$ , and where  $\mathbf{O} \in \mathbb{R}^{n \times n}$  and  $\mathbf{R} \in \mathbb{R}^{k \times k}$  are orthogonal matrices. From equation 16 the population variance is

$$\begin{aligned} \sum_j Var(z_j) &= \sigma^2 \|\mathbf{F}\|_F^2 \\ &= \sigma^2 \sum_i \lambda_i^2 \end{aligned} \quad (30)$$

Where  $\sigma$  is the variance of the random variables  $e_i$ . The norm of the weights  $\|\mathbf{W}\|_F^2$  is

$$\begin{aligned} \|\mathbf{W}\|_F^2 &= \|\mathbf{D}\mathbf{F}^+\|_F^2 \\ &= \|\mathbf{U}\Sigma\mathbf{V}^T\mathbf{R}\Lambda^{-1}\mathbf{O}^T\|_F^2 \\ &= \|\Sigma\mathbf{V}^T\mathbf{R}\Lambda^{-1}\|_F^2 \\ &= \sum_{ij} \frac{\sigma_i^2(D)}{\lambda_j^2} \sum_{kl} V_{ki} V_{li} R_{kj} R_{lj} \end{aligned} \quad (31)$$

While we have been previously interested in finding the minimal activity energy with fixed  $\|\mathbf{W}\|_F^2$ , we instead allow  $\|\mathbf{W}\|_F^2$  to change and instead minimise  $\|\mathbf{W}\|_F^2$ :  $\mathcal{L} = \mathbb{E}\|\mathbf{z}\|^2 + \beta_w \|\mathbf{W}\|_F^2$ , where  $\beta_w$  is the strength of regularization on the weights. Thus using equation 23, we have

$$\mathcal{L} = (\sigma^2 + a^2) \sum_i \lambda_i^2 + a^2 \sum_{j,k,l \neq k} |F_{jk}| |F_{jl}| + \beta_w \sum_{ij} \frac{\sigma_i^2(D)}{\lambda_j^2} \sum_{kl} V_{ki} V_{li} R_{kj} R_{lj} \quad (32)$$

This is the overall thing we want to optimise. However, for now we do not consider the middle term - the interaction induced by the nonnegativity constraint - and instead consider the following optimisation problem

$$\underset{\mathbf{V}, \mathbf{R}, \Lambda}{\text{minimise}} \mathcal{L}' = (\sigma^2 + a^2) \sum_i \lambda_i^2 + \beta_w \sum_{ij} \frac{\sigma_i^2(D)}{\lambda_j^2} \sum_{kl} V_{ki} V_{li} R_{kj} R_{lj} \quad (33)$$

Under the constraints that  $\mathbf{V}$  and  $\mathbf{R}$  remain orthogonal and  $\Lambda$  remains rectangular diagonal with positive entries. This is difficult in general, but we can simplify to gain intuition and insights.

**Simplification 1.** The first simplification is when  $\Sigma$  is a scaled identity matrix. This is the same simplification we used in Proof 3 (Appendix A.2.3), i.e. the singular values of  $\mathbf{D}$  are all equal,  $\sigma_i^2(D) = \sigma^2(D)$ .

**Simplification 2.** The second simplification is that  $\mathbf{V}$  is the identity matrix. This corresponds to data being generated via  $\mathbf{D} = \mathbf{U}\Sigma$ , i.e. the random variables are scaled and then orthogonally projected.

Analysing **simplification 1** first, to build up intuition (and offer an alternative proof of Theorem 3),  $\mathcal{L}'$  becomes

$$\begin{aligned}
 \mathcal{L}' &= (\sigma^2 + a^2) \sum_i \lambda_i^2 + \beta_w \sum_{ij} \frac{\sigma_i^2(D)}{\lambda_j^2} \sum_{kl} V_{ki} V_{li} R_{kj} R_{lj} \\
 &= (\sigma^2 + a^2) \sum_i \lambda_i^2 + \beta_w \sigma^2(D) \sum_j \frac{1}{\lambda_j^2} \sum_{kl} R_{kj} R_{lj} \sum_i V_{ki} V_{li} \\
 &= (\sigma^2 + a^2) \sum_i \lambda_i^2 + \beta_w \sigma^2(D) \sum_j \frac{1}{\lambda_j^2} \sum_k R_{kj}^2 \\
 &= (\sigma^2 + a^2) \sum_i \lambda_i^2 + \beta_w \sigma^2(D) \sum_j \frac{1}{\lambda_j^2}
 \end{aligned} \tag{34}$$

Where we exploited the fact that  $\mathbf{R}$  and  $\mathbf{V}$  are orthogonal matrices. Optimising this is easy to do, we can simply take derivatives to get

$$\lambda_i^4 = \frac{\beta_w \sigma^2(D)}{\sigma^2 + a^2} \tag{35}$$

This result is independent of  $\mathbf{R}$  and  $\mathbf{O}$ , and so we are free to choose whatever orthogonal matrices we like. This is good news for us as the real game we are in is minimising  $\mathcal{L}$  (equation 32) which contains an additional term involving  $\mathbf{R}$  and  $\mathbf{O}$ :  $a^2 \sum_{j,k,l \neq k} |F_{jk}| |F_{jl}| = a^2 \sum_{j,k,l \neq k} |(\mathbf{O}\Lambda\mathbf{R}^T)_{jk}| |(\mathbf{O}\Lambda\mathbf{R}^T)_{jl}|$ . Thus if we can set  $|(\mathbf{O}\Lambda\mathbf{R}^T)_{jk}| |(\mathbf{O}\Lambda\mathbf{R}^T)_{jl}| = 0$ , then we will have fully minimised  $\mathcal{L}$ . This is easy enough to do (keeping  $\mathbf{R}$  and  $\mathbf{O}$  orthogonal) and is achieved when  $\mathbf{R}$  is a permutation matrix, and  $\mathbf{O}$  has at most one non-zero element per row (or equally  $\mathbf{O}$  has at most one non-zero element per column). This corresponds to disentangled representations.

Returning to **simplification 2**, where  $\mathbf{V}$  is the identify matrix, now  $\mathcal{L}'$  becomes:

$$\begin{aligned}
 \mathcal{L}' &= (\sigma^2 + a^2) \sum_i \lambda_i^2 + \beta_w \sum_{ij} \frac{\sigma_i^2(D)}{\lambda_j^2} \sum_{kl} V_{ki} V_{li} R_{kj} R_{lj} \\
 &= (\sigma^2 + a^2) \sum_i \lambda_i^2 + \beta_w \sum_{ij} \frac{\sigma_i^2(D)}{\lambda_j^2} R_{ij}^2
 \end{aligned} \tag{36}$$

Thus we have the following constrained optimisation problem

$$\underset{\mathbf{R}, \Lambda}{\text{minimise}} \sigma^2 \sum_i \lambda_i^2 + \beta_w \sum_{ij} \frac{\sigma_i^2(D) R_{ij}^2}{\lambda_j^2} \quad \text{s.t. } \lambda_i > 0, \mathbf{R}^T \mathbf{R} = \mathbf{I} \tag{37}$$

Here we let intuition take over. Taking inspiration from the simplification 1, our ansatz is that  $\mathbf{R}$  is a permutation matrix and  $\lambda_i^4 = \frac{\beta_w \sigma_i^2(D)}{\sigma^2 + a^2}$  (where  $i$  and  $j$  are related by the permutation). To justify that  $\mathbf{R}$  should be a permutation matrix in this case, we note that as  $\lambda_i^2$  gets smaller to keep  $\|\mathbf{W}\|_F^2 = \sum_{ij} \sigma_i^2(D) V_{ij} \frac{1}{\lambda_j^2}$  as small as possible, the  $R_{ij}$  must ensure the low valued  $\lambda_j^2$  are matched with the low valued  $\sigma_i^2(D)$ . Intuitively this is done when  $\mathbf{V}$  is a permutation matrix. Once  $\mathbf{R}$  is chosen as a permutation matrix, showing that  $\lambda_j^4 = \frac{\beta_w \sigma_j^2(D)}{\sigma^2 + a^2}$  is simple - just take derivatives and set to zero. While this is not a full proof, if one derives the KKT conditions, this solution is at least a consistent solution, i.e. it is a minima but may not be the global minima. We note that for specific (small) values of  $\beta_w$  it will be the global minima.

As before, our actual aim is to minimise  $\mathcal{L}$ . Since we are free to choose  $\mathbf{O}$ , and we choose it to make the cross terms in equation 32 go to zero, which is when  $\mathbf{O}$  has at most one non-zero element per column. This corresponds to disentangled representations.

**No simplifications.** Reminding ourselves of the full objective

$$\mathcal{L} = (\sigma^2 + a^2) \sum_i \lambda_i^2 + a^2 \sum_{j,k,l \neq k} |F_{jk}| |F_{jl}| + \beta_w \sum_{ij} \frac{\sigma_i^2(D)}{\lambda_j^2} \sum_{kl} V_{ki} V_{li} R_{kj} R_{lj} \tag{38}$$

In this case  $\mathbf{V}$  is an arbitrary orthogonal matrix, and  $\sigma_i^2(D)$  are the (not necessarily equal) singular values of  $D$ . One potential ansatz is to assume  $\mathbf{R}$  is a permutation matrix once again, which then reduces the problem to simplification 2. Again this offers a minima, but not necessarily the global minima.

In sum, we can see that there is always a pressure to disentangle due to the middle term in the loss. However it will have to trade-off against the weight regularization term (last term). Thus we posit that for small values of  $\beta_w$  disentangling is preferred, even for arbitrary  $D$ .

### A.3 SPARSITY DOES NOT PROMOTE DISENTANGLING

Readers may wonder if imposing a sparsity constraint would encourage disentangling, as ReLUs promote sparsity as well as enforcing nonnegativity. This is not the case, and can be understood intuitively, and also with numerical simulation. For intuition, a sparsity constraint creates a diamond shaped iso-contour in neural space (Fig. 8) left, and so encourages the firing rate distribution to fall inside that diamond, which in the case our our random variables, means maximal entangling. We confirm this intuition in simulation (data and setup the same as Fig. 2) with a variety of sparsity regularization strengths (Fig. 8 right). Thus the reason ReLUs promote nonnegativity is not because they induce sparsity (which they do), but because they enforce nonnegativity.

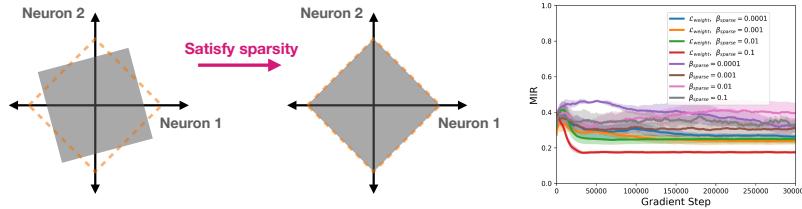


Figure 8: **Sparsity intuition and simulations.** Sparsity does not encourage disentangling. Here the best situation is maximal entangling. Simulations verify that sparsity constraints do not induce disentangling.

#### A.4 REPRESENTING CATEGORICAL DISTRIBUTIONS

We note that this is a fundamentally different problem to representing factors, since only one category is active at any one time; categories are anti-correlated. This contrasts to factors as all factors can be active independently. So even though the proofs are similar, they are about different situations.

**Theorem.** Given a nonnegative representation,  $z \in \mathbb{R}^n$ , that linearly reads-out a population,  $x \in \mathbb{R}^m$  which is itself a linear projection of a categorical variable (one-hot variable)

$$x = Wz \quad \text{and} \quad x = Cc \quad (39)$$

Where  $c \in \mathbb{R}^c$  is a one-hot representation (with each one-hot vector having equal probability) with each element,  $c_i$ , representing category identity (1 when present, 0 otherwise, only one category active at once),  $W \in \mathbb{R}^{m \times n}$  are read-out weights with fixed norm,  $\|W\|_F^2$ , and  $C \in \mathbb{R}^{m \times c}$  are the data generative weights which is a rank  $c$  matrix with all singular values equal and positive;  $\sigma_{\min}(C) = \sigma_{\max}(C) = \sigma(C) > 0$ . Then in order to minimise the expected activity energy,  $\mathbb{E}\|z\|^2$ ,  $z$  must be structured so that each neuron only represents at most a single category.

**Intuition.** In order to linearly read-out a categorical variable from a neural population,  $z$ , then  $z$  must take be a linear combination of a one-hot vector ( $c$ ; each element in the one-hot vector corresponding to each category).

$$z = Mc \quad (40)$$

For this representation to be nonnegative, all entries of  $M$  must be nonnegative. This means that the population response,  $z_i$ , for each category,  $i$ , must be a vector in the positive orthant (Fig. 9). Assuming we keep the expected activity energy constant, i.e.

$$\mathbb{E}\|z\|^2 = \frac{1}{c} \sum_i \|z_{ci}\|^2 = \frac{1}{c} \|M\|_F^2 = \text{constant} \quad (41)$$

Then the easiest representation to read-out from is the one-hot representation, since in all other situations the vectors are closer together which require larger read-out weights to disambiguate. Alternatively, if there is noise on  $z$  then the categories will be more often confused if the  $z_c$  vectors are closer to each other. A one-hot representation is the best, and it is a disentangled representation for each category.

**Proof.** We aim to show when minimising  $\mathbb{E}\|z\|^2$ , when  $z$  is nonnegative ( $z_i \geq 0$ ), and for a finite norm of the read-out weights,  $\|W\|_F^2$ , then it is optimal for each element,  $z_i$  in  $z$  to represent at most a single category. This is a constrained optimisation problem

$$\underset{W}{\text{minimise}} \mathbb{E}\|z\|^2 \quad \text{s.t. } z_i \geq 0, Cc = Wz \quad \forall c \quad (42)$$

Since we read-out with zero error,  $Cc = Wz$ , then  $W$  must contain all the information of  $C$ , i.e.  $W = Cf^+$ , where  $f^+ \in \mathbb{R}^{c \times n}$  is a rank  $c$  matrix. Thus  $z$  must be

$$\begin{aligned} z &= W^+x \\ &= Ff^+Cc \\ &= Fc \end{aligned} \quad (43)$$

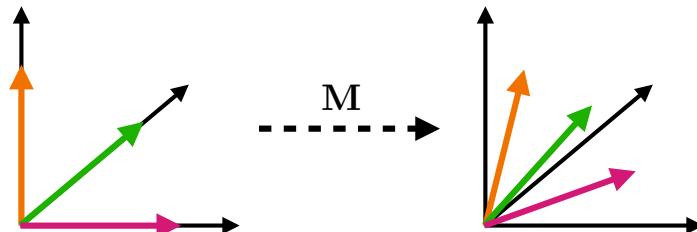


Figure 9: Schematic for representing categories. Each axis is neural firing rate. Each colour is a population activity for a single category, i.e.  $z_c$ .

For this to be nonnegative, since  $\mathbf{c}$  is one-hot, then  $F_{ij} \geq 0$ . Additionally, the expected activity energy becomes

$$\begin{aligned}\mathbb{E}\|\mathbf{z}\|^2 &= \mathbb{E}\|\mathbf{Fc}\|^2 \\ &= \frac{1}{c}\|\mathbf{F}\|_F^2 \\ &\geq \frac{1}{c} \frac{n^2}{\|\mathbf{F}^+\|_F^2} \\ &= \frac{1}{c} \frac{n^2 \sigma^2(\mathbf{C})}{\|\mathbf{W}\|_F^2}\end{aligned}\tag{44}$$

Where the final equality is from equation 22, and where the inequality is from equation 17. This inequality becomes an equality if and only if  $\mathbf{F}^+$  has (scaled) orthonormal rows (or  $\mathbf{F}$  has (scaled) orthonormal columns). Thus to minimise  $\mathbb{E}\|\mathbf{z}\|^2$  we want  $\mathbf{F}^+$  to have (scaled) orthonormal rows or equally  $\mathbf{F}$  to have (scaled) orthonormal columns. To satisfy nonnegativity we wanted  $F_{ij} \geq 0$ . There is only one type of matrix with (scaled) orthonormal columns that has nonnegative entries, and that is when  $\mathbf{F}$  has rows that contain at most one non-zero element, and so each neuron only ‘attends’ to one category; categories are disentangled.

### A.5 SIMULATION RESULTS

We train an autoencoder on data from 6 categories. The data from each category is noiseless, so we essentially just have 6 data samples that we train on. Each category vector is sampled from a uniform distribution. We train an autoencoder with a deep encoder (hidden layers: [500,300,100]) and a shallow decoder (0-hidden layers), and with latent dimension of 10.

Along with variants of our constraints, we also train a network with a sparsity inducing loss:  $\mathcal{L}_{\text{sparsity}} = \beta_{\text{sparsity}} \sum_i |a_i|$ , where  $a_i$  is the activity of a neuron in the latent layer. All other hyper-parameters are as described in the autoencoder section of Table 1, and we set  $\beta_{\text{sparsity}}$  to be the same as  $\beta_{\text{nonneg}}$ .

We see that only our constraints, and the sparsity inducing loss, achieves disentangling of categories (Fig. 10).

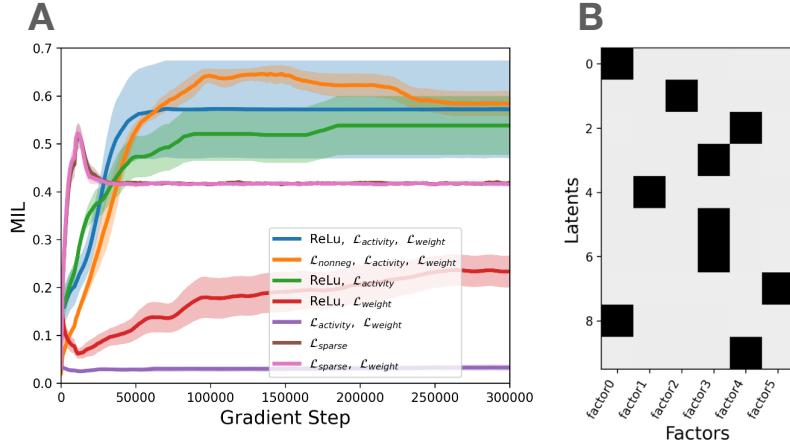


Figure 10: **Learning data generative factors with autoencoders.** **A)** Training linear autoencoders on linear data. Only models with our constraints, or with sparsity, learn disentangled representations. **B)** Example mutual information matrix from a high MIR model. All learning curves show mean and standard error from 4 mean from 5 random seeds.

## A.6 NETWORK ARCHITECTURES AND OPTIMISATION DETAILS

All code will be released on publication. All hyper-parameters are shown in Table 1. We describe any additional details here. We use the Adam optimiser (Kingma & Ba, 2014). Supervised nets use square error loss. Unsupervised nets use sigmoid cross entropy loss.

**Supervised networks.**  $\mathcal{L}_{\text{nonneg}}$ ,  $\mathcal{L}_{\text{activity}}$  and  $\mathcal{L}_{\text{weight}}$  are applied to all layers. The shallow Network is a 1-layer MLP with hidden dimensions of [188]. The deep networks are 5-layer MLP with hidden dimensions of [188, 186, 184, 182, 180]. We use a squared error loss for  $\mathcal{L}_{\text{prediction}}$ .

**Autoencoders.** We apply  $\mathcal{L}_{\text{nonneg}}$  and  $\mathcal{L}_{\text{activity}}$  to the latent layer only. We apply  $\mathcal{L}_{\text{weight}}$  to all layers. The shallow autoencoder has no hidden layers in the encoder or decoder. The deep autoencoder has an encoder with hidden dimensions of [500, 300, 100] and a decoder with hidden dimensions of [100, 300, 500]. The output has a sigmoid and we use a sigmoid cross entropy loss for  $\mathcal{L}_{\text{prediction}}$ .

**Variational Autoencoder.** We use an architecture described in Higgins et al. (2017a) (Table 2). We use the standard  $\beta$ -VAEs loss with the following additions. We apply  $\mathcal{L}_{\text{nonneg}}$  to the latent layer only. We apply  $\mathcal{L}_{\text{weight}}$  to the dense layers in the decoder only. We *do not* apply  $\mathcal{L}_{\text{activity}}$  as an effective norm constrain term already exists in the  $\beta$ -VAEs loss.

	<b>Supervised Net</b>		<b>Autoencoder</b>		<b>VAE</b>
	<i>Shallow</i>	<i>Deep</i>	<i>Shallow</i>	<i>Deep</i>	<i>Deep</i>
input dimension	6	6	50	50	(64, 64, 3)
output dimension	6	6	n/a	n/a	n/a
# parameters	2450	138574	1570	414870	766295
# (hidden) layers	1	5	enc: 0 dec: 0	enc: 3 dec 3	enc: 6 dec 6
latent dimension	n/a		10		10
learning rate	3e-3		1e-4		1e-4
batch size	128		64		64
# gradient updates	300000		300000		500000
$\beta_{\text{activity}}$	1e-3		5e-3		n/a
$\beta_{\text{weight}}$	1e-4		1e-3		0.01 → 1.0
$\beta_{\text{nonneg}}$	2.0		0.5		100.0
$\beta_{\text{VAE}}$	n/a		n/a		1.0

Table 1: The values of various hyper-parameters. enc/dec: encoder/decoder. n/a: not applicable.

<b>Encoder</b>	<b>Decoder</b>
Input: $64 \times 64 \times$ number of channels	Input: 10
$4 \times 4$ conv, 32 ReLU, stride 2	FC, 256 ReLU
$4 \times 4$ conv, 32 ReLU, stride 2	FC, $4 \times 4 \times 64$ ReLU
$4 \times 4$ conv, 64 ReLU, stride 2	$4 \times 4$ upconv, 64 ReLU, stride 2
$4 \times 4$ conv, 64 ReLU, stride 2	$4 \times 4$ upconv, 32 ReLU, stride 2
FC 256	$4 \times 4$ upconv, 32 ReLU, stride 2
F2 $2 \times 10$	$4 \times 4$ upconv, number of channels, stride 2

Table 2: Encoder and decoder architecture for the variational autoencoder experiments.

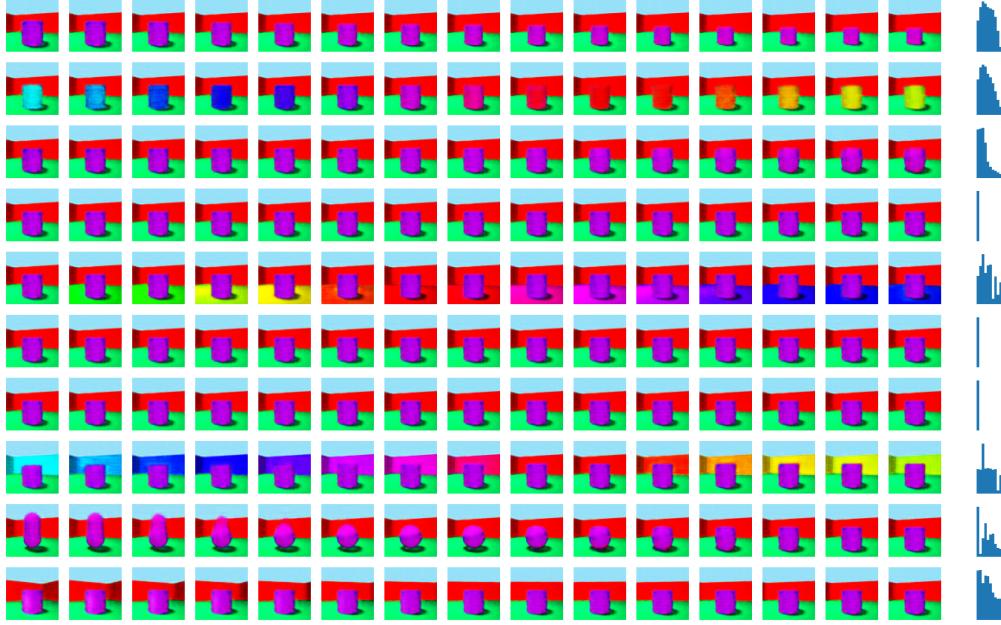


Figure 11: **Latent space traversals.** An image is encoded and then the value of a single latent dimensions is changed, and the resulting image is generated. Each row shows this image for when a single latent dimension is traversed. The histogram shows the marginal distribution of that latent variable. From the images, we can see that each latent dimension predominantly cares about a single ground truth factor.

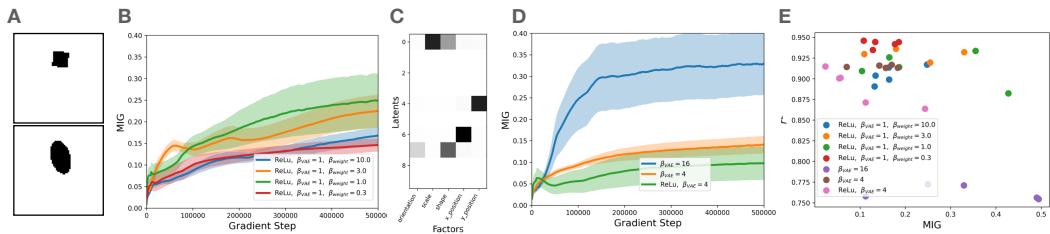


Figure 12: **Same as Fig. 5, but for the dSprites dataset.** **A)** We train on the dSprites dataset, with two example images shown. These images have 5 underlying factors. **B)** MIG scores are higher with higher weight regularization, and generally higher than any  $\beta$ -VAE (panel D). **C)** Mutual information matrix for a high scoring model. We note it has dropped a latent dimension. **D)**  $\beta$ -VAE MIG scores. We note the high disentangling  $\beta$ -VAE models have bad reconstruction as they drop latent dimensions. **E)** MIG score against  $R^2$  shows models with our constraints lie in the Goldilocks region of high disentangling and high reconstruction. All learning curves show mean and standard error from 5 random seeds.

### A.7 PATTERN FORMING DETAILS

We use a discrete 16x16 world, (so 256 locations;  $n_l = 256$ ) and optimise an independent representation,  $\mathbf{z}(\mathbf{x}) \in \mathbb{R}^{n_c}$ , at each location. We now detail each component of the following loss

$$\mathcal{L} = \underbrace{\mathcal{L}_{\text{nonneg}} + \mathcal{L}_{\text{activity}} + \mathcal{L}_{\text{weight}}}_{\text{Biological constraints}} + \underbrace{\mathcal{L}_{\text{location}} + \mathcal{L}_{\text{actions}} + \mathcal{L}_{\text{objects}}}_{\text{Functional constraints}} + \underbrace{\mathcal{L}_{\text{path integration}}}_{\text{Structural constraints}} \quad (45)$$

**Biological losses.** These are exactly the same as in the main paper, but we must also average over all locations,  $\mathbf{x}$

$$\mathcal{L}_{\text{nonneg}} = \frac{\beta_{\text{nonneg}}}{n_l} \sum_x \sum_i \max(-z_i(\mathbf{x}), 0) \quad (46)$$

$$\mathcal{L}_{\text{weight}} = \beta_{\text{weight}} \sum_t \|\mathbf{W}_t\|^2 \quad (47)$$

$$\mathcal{L}_{\text{activity}} = \frac{\beta_{\text{activity}}}{n_l} \sum_x \|\mathbf{z}(\mathbf{x})\|^2 \quad (48)$$

where  $z_i(\mathbf{x})$  is a neuron in representation  $\mathbf{z}(\mathbf{x})$ ,  $t$  indexes the task (i.e. object, action, location prediction), and the  $\beta$  values determines the regularization strength.

**Location loss.** The representation predicts location (a one-hot encoding describing each of the 256 locations) via a linear transformation, which is then fed into a softmax cross-entropy loss. In particular, the logits for each location,  $\mathbf{x}$ , are  $\mathbf{W}_l \mathbf{z}(\mathbf{x})$ , where  $\mathbf{W}_l \in \mathbb{R}^{n_l \times n_c}$ . If we denote each row of  $\mathbf{W}_l$  as  $\mathbf{l}_x$ , noting that this row ‘corresponds’ to location  $\mathbf{x}$  in the one-hot encoding, then the loss is as follows

$$\mathcal{L}_{\text{location}} = -\frac{\beta_{\text{location}}}{n_l} \sum_{\mathbf{x}} \ln \frac{e^{\mathbf{l}_x \cdot \mathbf{z}(\mathbf{x})}}{\sum_{\mathbf{x}'} e^{\mathbf{l}_{\mathbf{x}'} \cdot \mathbf{z}(\mathbf{x})}} \quad (49)$$

**Object loss.** An object is either present or not present at each location, so we use a sigmoid cross-entropy loss. In particular, the logits for each location  $\mathbf{x}$  is  $\mathbf{W}_o \mathbf{z}(\mathbf{x})$ , where  $\mathbf{W}_o \in \mathbb{R}^{1 \times n_c}$ . If  $\mathbb{1}_{\text{object at } \mathbf{x}}$  returns a 1 if an object is present at location,  $\mathbf{x}$ , and 0 otherwise, then the loss is as follows

$$\mathcal{L}_{\text{object}} = -\frac{\beta_{\text{object}}}{n_l} \sum_{\mathbf{x}} \mathbb{1}_{\text{object at } \mathbf{x}} \ln \sigma(\mathbf{W}_o \mathbf{z}(\mathbf{x})) + \mathbb{1}_{\text{object not at } \mathbf{x}} \ln(1 - \sigma(\mathbf{W}_o \mathbf{z}(\mathbf{x}))) \quad (50)$$

**Action loss.** More than one action can be correct at a location, so we use a sigmoid cross-entropy loss for each of the 4 actions. In particular, the logits at location,  $\mathbf{x}$ , are  $\mathbf{W}_t \cdot \mathbf{z}(\mathbf{x})$ , where  $\mathbf{W}_t \in \mathbb{R}^{n_a \times n_c}$ , where  $n_a$  is the number of actions (4 in our case - North, South, East, West). We denote each row of  $\mathbf{W}_t$  as  $\mathbf{t}_a$ , noting this row ‘corresponds’ to action  $a$  in the action encoding. If  $\mathbb{1}_{a=\mathbf{a}(\mathbf{x})}$  returns a 1 if action,  $a$ , is a correct action at location  $\mathbf{x}$ , and 0 otherwise, then the loss is as follows

$$\mathcal{L}_{\text{action}} = -\frac{\beta_{\text{action}}}{n_l} \sum_{\mathbf{x}} \sum_a \mathbb{1}_{a=\mathbf{a}(\mathbf{x})} \ln \sigma(\mathbf{t}_a \cdot \mathbf{z}(\mathbf{x})) + (1 - \mathbb{1}_{a=\mathbf{a}(\mathbf{x})}) \ln(1 - \sigma(\mathbf{t}_a \cdot \mathbf{z}(\mathbf{x}))) \quad (51)$$

**Structural loss.** We use a squared error loss for the structural constraint, which asks for neighbouring representations to be related to each other by an action matrix  $\mathbf{W}_a$  for each action,  $a$ . This is just like a path integration loss. This loss is done for every location,  $\mathbf{x}$ , and each of the 4 actions,  $a$ .

$$\mathcal{L}_{\text{path integration}} = \frac{\beta_{\text{path integration}}}{n_l} \sum_{\mathbf{x}} \sum_a \|\mathbf{z}(\mathbf{x}) - f(\mathbf{W}_a \mathbf{z}(\mathbf{x} - \mathbf{d}_a))\|^2 \quad (52)$$

Where  $\mathbf{W}_a \in \mathbb{R}^{n_c \times n_c}$  is a weight matrix that depends on action,  $a$ , (i.e. there are 4 trainable weights matrices - one for each action).  $\mathbf{d}_a$  means the displacement in the underlying space (the space of  $\mathbf{x}$ ), that the action  $a$  corresponds to.

**Pattern forming dynamics.** The overall loss can be optimised with respect to the weights. However, it can also be optimised directly with respect to  $\mathbf{z}$ . This is particularly interesting for us, as it allows our representation to be dynamic and change rapidly for a single task, and not just slowly via learning

over many tasks. This is a necessity for us as we need to represent objects which may move between tasks. To optimise both  $\mathbf{z}$  (task particularities) and weights (task generalities), we do so in two stages. First, we optimise with respect to  $\mathbf{z}$  to ‘infer’ a representation for the current task. Second, we optimise with respect to the weights to learn parameters that are general across tasks.

When optimising with respect to  $\mathbf{z}$  we only optimise two terms in the loss:  $\mathcal{L}_{\text{objects}}$  and  $\mathcal{L}_{\text{path integration}}$ . We optimise the first term so the system has the ability to know where the objects are. We optimise the second term so that information can be propagated around (effectively via path integration).

The dynamics of the  $\mathcal{L}_{\text{objects}}$  are:

$$\frac{d\mathcal{L}_{\text{objects}}}{d\mathbf{z}(\mathbf{x})} = -\mathbf{W}_{\text{objects}}^T (\mathbb{I}_{\text{object at } \mathbf{x}} - \sigma(\mathbf{W}_o \mathbf{z}(\mathbf{x}))) \quad (53)$$

This says if you get the object prediction wrong, then update  $\mathbf{z}$  to better predict the object. We restrict this update to only take place where the object is, so it is just an object signal. This update is equivalent to a rodent observing that it is at an object.

The dynamics of the  $\mathcal{L}_{\text{path integration}}$  are:

$$\begin{aligned} \frac{d\mathcal{L}_{\text{path integration}}}{d\mathbf{z}(\mathbf{x})} &= \sum_a -(\mathbf{z}(\mathbf{x}) - f(\mathbf{W}_a \mathbf{z}(\mathbf{x} - \mathbf{d}_a))) \\ &\quad + \mathbf{W}_a^T (\mathbf{z}(\mathbf{x} + \mathbf{d}_a) - f(\mathbf{W}_a \mathbf{z}(\mathbf{x}))) \odot f'(\mathbf{W}_a \mathbf{z}(\mathbf{x})) \end{aligned} \quad (54)$$

The two terms in the above equation can be easily understood. The first says that the representation at each location,  $\mathbf{z}(\mathbf{x})$ , should be updated according to what its neighbours think it should (this is the same update rule as path integration!). The second term says the representation at each location,  $\mathbf{z}(\mathbf{x})$ , should be updated if it did not predict its neighbours correctly. This equation tells representations to update based on their neighbours. This is just like a **cellular automata**, but instead of a discrete value being updated on the basis of its neighbours, it is a whole population vector whose elements can be continuous. Indeed, just like cellular automata, it is also possible to initialise a single ‘cell’ (location) of the cellular automata, and have that representations propagate throughout the space. In this case, it’s just like path integration, but spreading through all space at once. We note, however, that in our simulations we initialise representations at all locations (for each task).

We note that while we simulated this on a discrete grid, the same principles apply to continuous cases. In this case the sums over location/actions need to be replaced with integrals.

This is a very general approach for understanding representations. The structural loss does not have to relate to the rules of path integration. It can be anything. It could be the rules of a graph. It could be rules of topology. It could have one set of rules at some locations and another set of rules at other locations. The rules don’t have to be neighbouring representations telling each other what to be, it could also be long range rules too. If there are structure or rules in the world or behaviour, our constraints say that representations should understand that structure or rules. In mathematics this is known as a homeomorphism. In sum, understanding representations via constraints is very general.

**Hyper-parameters.**  $\beta_{\text{nonneg}}, \beta_{\text{weight}}, \beta_{\text{activity}}, \beta_{\text{location}}, \beta_{\text{object}}, \beta_{\text{action}}, \beta_{\text{path integration}}$  are chosen as  $1e-2, 1e-7, 1e-3, 1e-2, 1e-1, 1e-2, 40$ . The learning rate is  $12e-4$ . In the Euler updates for pattern formation, the step size for path integration is 0.1, and the step size for objects is 1. We additionally clip the norm of the update in pattern formation for stability.

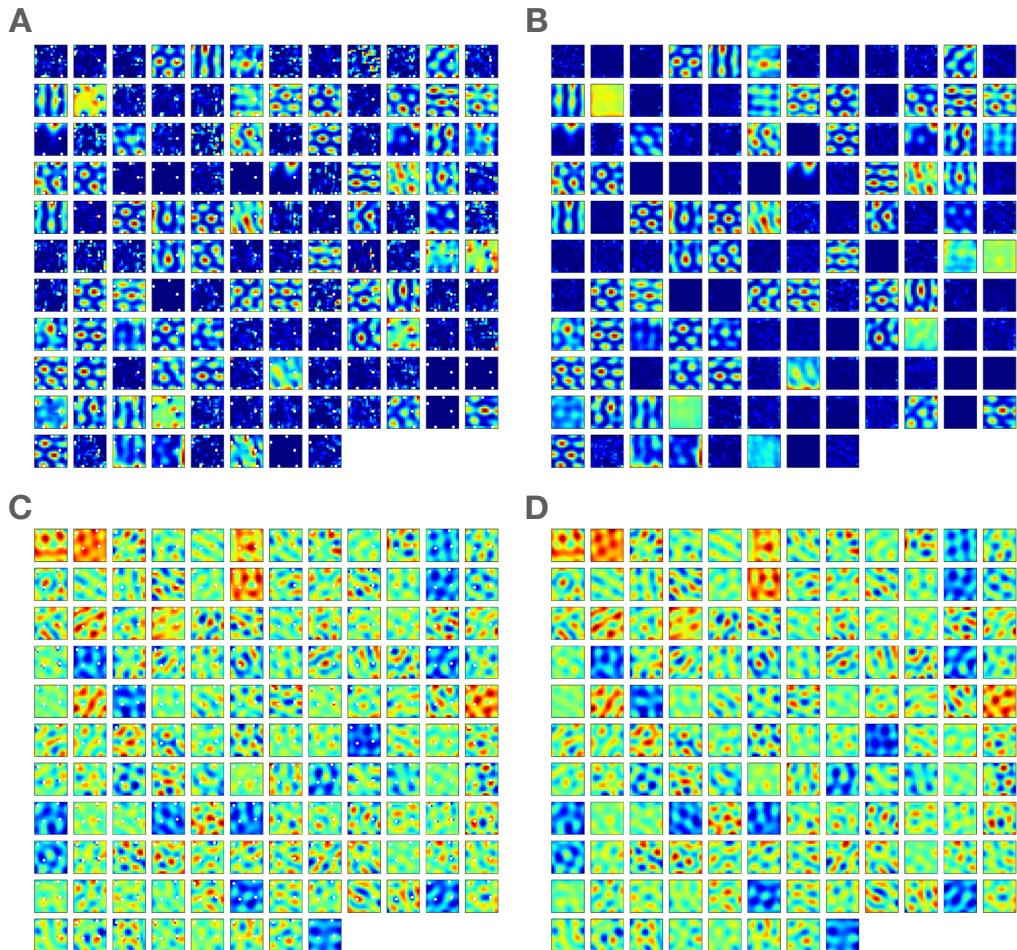


Figure 13: All cells. A/B) Ratemaps from a ReLU model. A) A task with no objects. B) Task with objects. C/D) Ratemaps from a model with linear activation function. C) A task with no objects. D) Task with objects.