Learning shapes neural geometry in the prefrontal cortex

Michał J. Wójcik[1,2@], Jake P. Stroud[3], Dante Wasmuht[2], Makoto Kusunoki[4], Mikiko Kadohisa[4], Mark J. Buckley[2], Nicholas E. Myers[2,5], Laurence T. Hunt[2,6], John Duncan[2,4] & Mark G. Stokes[2]

[1] Centre for Neural Circuits and Behaviour, Department of Physiology, Anatomy and Genetics, University of Oxford, Oxford, UK
[2] Department of Experimental Psychology, University of Oxford, Oxford, UK
[3] Department of Engineering, University of Cambridge, Cambridge, UK
[4] MRC Cognition and Brain Sciences Unit, University of Cambridge, Cambridge, UK
[5] School of Psychology, University of Nottingham, Nottingham, UK
[6] Department of Psychiatry, University of Oxford, Oxford, UK
@corresponding author: michal.wojcik@dpag.ox.ac.uk

Mark passed away on the 13th of January 2023. He was not only a valued colleague but also a friend and mentor to many of us. His brilliant mind and insightful contributions will be sorely missed.

***Abstract*. The relationship between the geometry of neural representations and the task being performed is a central question in neuroscience[1–6]. The primate prefrontal cortex (PFC) is a primary focus of inquiry in this regard, as under different conditions, PFC can encode information with geometries that either rely on past experience[7–13] or are experience agnostic[3,14–16]. One hypothesis is that PFC representations should evolve with learning[4,17,18], from a format that supports exploration of all possible task rules to a format that minimises the encoding of task-irrelevant features[4,17,18] and supports generalisation[7,8]. Here we test this idea by recording neural activity from PFC when learning a new rule ('XOR rule') from scratch. We show that PFC representations progress from being high dimensional, nonlinear and randomly mixed to low dimensional and rule selective, consistent with predictions from constrained optimised neural networks. We also find that this low-dimensional representation facilitates generalisation of the XOR rule to a new stimulus set. These results show that previously conflicting accounts of PFC representations can be reconciled by considering the adaptation of these representations across different stages of learning.**

Two seemingly discrepant accounts propose that PFC neural activity should track either low-[8–13,19] or high-dimensional[3,14–16] representations of the environment. Traditionally, it has been proposed that PFC cells are tuned adaptively to task-relevant information, leading to low-dimensional neural activity[13]. This results in the population displaying structured selectivity patterns, as commonly observed after training on a cognitive task (**Fig. 1a**, *low-dimensional*) [13]. A contrasting hypothesis suggests that the PFC may rely on high-dimensional, nonlinearly mixed representations of task features

to support complex cognition (**Fig. 1a**, *high-dimensional*)[3,14]. According to this notion, the PFC serves as a nonlinear kernel such that when a low-dimensional input is projected onto it, dimensionality expands, and a wide repertoire of responses can be generated[15,16].

Recently, it has been proposed that the PFC is capable of transitioning between high- and low-dimensional representations across learning, to accommodate the changing demands of the environment [4,17,18,20]. For example, early in learning, high-dimensional representations may allow flexible exploration of all possible input–output mappings ("contingencies") in order to discriminate which task rules are currently relevant [3,14,16]. This is because a high-dimensional representation allows for a high number of linearly separable task features (**Fig. 1a**). Conversely, once an animal has learnt that only one set of contingencies is relevant, a low-dimensional representation may be used to minimise energy expenditure[4,17–19]. Moreover, these low-dimensional representations may enable generalisation to novel contexts, because the same representational geometry can be reused across both familiar and new stimulus sets (**Supp. Fig. 1**)[7,8]. In other words, different stages of learning impose different demands on the neural population. Learning could thus shape neural dimensionality and progressively push neural activity towards different solutions along the trade-off between discriminability and generalisability, i.e., from a high-dimensional regime towards a low-dimensional regime[18,20].

Here, we tested this idea in two macaque monkeys which learnt an exclusive-or (XOR) rule – a problem that can be solved by a range of representations, from low- to high-dimensional (**Fig. 1a**)[19]. Importantly, we tracked how the dimensionality and geometry of PFC representations changed across multiple training sessions of an XOR rule that was entirely new to the animals at the start of recording (experiment 1) and during subsequent generalisation of this rule to a new stimulus set (experiment 2). We used a classical conditioning paradigm in which the nonlinear combination of the features of two objects presented in succession (XOR) predicted the outcome of the trial. Importantly, the animals were only required to fixate through both experiments. This classical conditioning design carries advantages relative to operant tasks in that changes in representation with learning are unlikely to be confounded with changes in behaviour[21–24]. Later, we also show that our results hold in a previously collected delayed match-to-sample task[25–27].

We found that initially PFC selectivity was high-dimensional, nonlinear and randomly mixed. As learning progressed, PFC activity became low-dimensional, manifesting highly structured selectivity predominantly towards the nonlinear mixture of variables across the neural population. Next, we tested whether the learnt low-dimensional representation was reused in conditions where new unseen combinations of stimuli were presented. We found that during this generalisation of the XOR rule, the

representation of these new stimuli became aligned with the pre-existing low-dimensional representation of the old stimulus set, allowing the same neural code to be re-used.

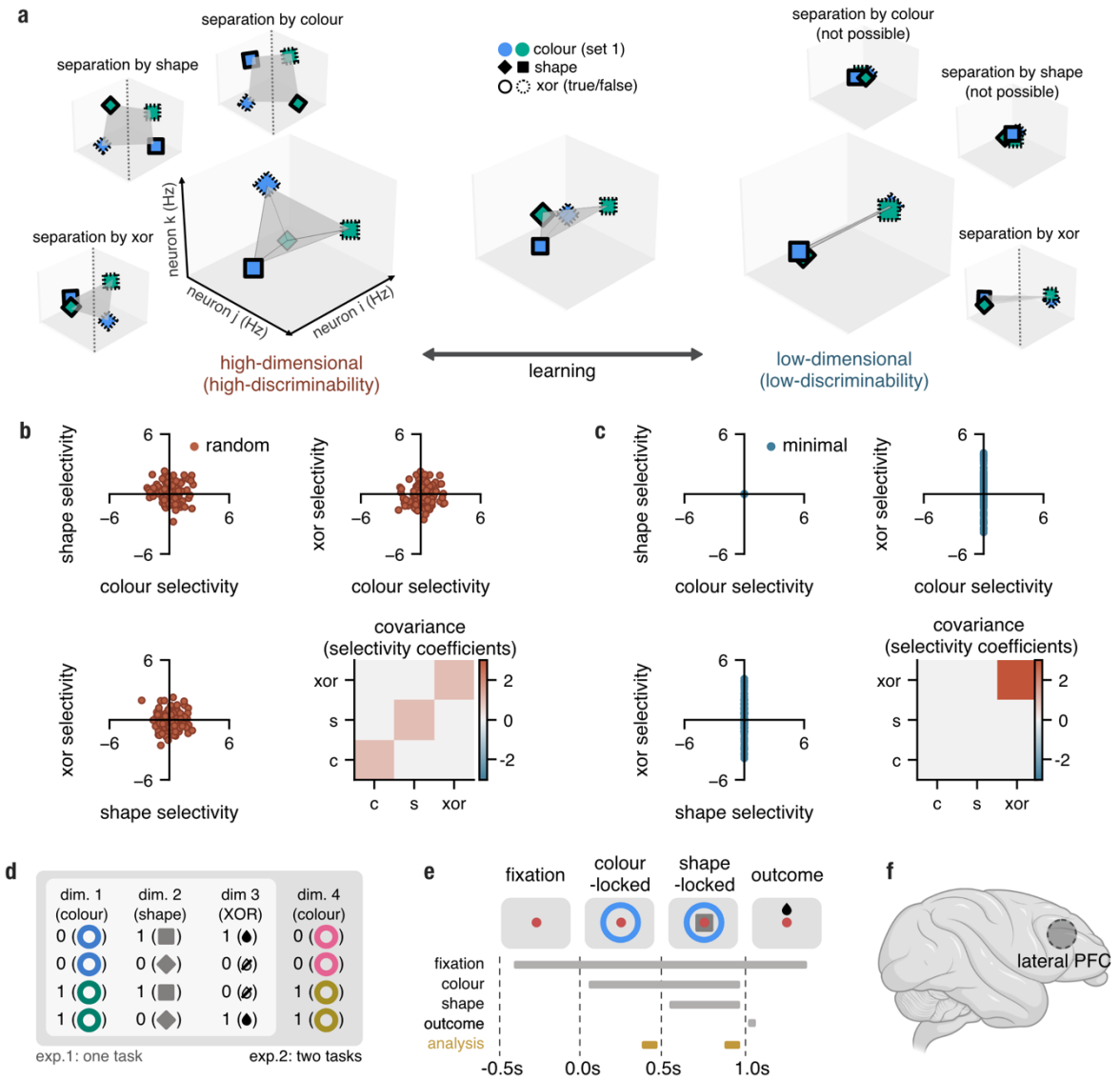


Figure 1. Potential effects of learning on neural geometry in the prefrontal cortex. Learning can reduce or expand neural dimensionality, changing how many linear decoding axes can be implemented on neural firing rates (discriminability). **a**, High-dimensional representations enable high discriminability. A high-dimensional regime allows the separation of all task features using three possible readout axes (left), whereas a low-dimensional representation only allows task-relevant features to be separated (right). **b**, Each neuron can be represented as a point in the 3-dimensional selectivity space spanned by colour, shape, and XOR (their interaction). In the random model, selectivity is distributed according to a spherical Gaussian distribution in this space (*Methods, generative models*); the covariance matrix is computed between the selectivity coefficients. **c**, Analogous to a but for the minimal model; neurons are selective only for the XOR (interaction between colour and shape), as this is the only feature that is necessary to solve the task. **d**, In Experiment 1, animals were incentivised to combine two passively viewed task features (colour and shape) in a non-linear fashion (XOR). For example, blue+square and green+diamond combinations were rewarded, whereas blue+diamond and green+square were not. In experiment 2 an additional set of colour–shape contingencies was introduced. Monkeys could generalise the learned rule from experiment 1 to the new set of colours in experiment 2. **e**, Timeline of task events in a single trial. **f**, Neural data was collected from the lateral

surface of the prefrontal cortex in two macaque monkeys (see *Methods* and **Supplementary Figure 8** for further details).

Generative models of nonlinear random and minimal selectivity

We first wanted to understand how the geometry of the neural representations could change over the course of learning. We thus explored the geometries produced by two generative models of neural selectivity with different discriminability-generalisability trade-offs[4,17]: (i) a high-dimensional geometry produced by non-linear random mixed selectivity[3,14–16,28] (high discriminability, low generalisability); and (ii) a low-dimensional geometry produced by structured, minimal selectivity (low discriminability, high generalisability)[10,12,29–33]. The former is a well-established geometry (e.g., inherent in reservoir computing models[15,16]) whereas the latter was inspired by recent theoretical work which suggests that complex (nonlinear) tasks may require highly structured population activity[34]. These models make distinct predictions about the distribution of selectivity to the task-relevant variables in a 3-dimensional selectivity space (colour, shape, and their nonlinear interaction, i.e., XOR). We refer to minimal or random selectivity describing the distribution of selectivity and respectively to high- and low-dimensional geometry referring to the representations that these selectivity distributions generate.

In selectivity space, each axis represents a units' response to one stimulus variable (e.g., high shape selectivity = higher firing rate for square than diamond, **Fig. 1b**). Hypothetically, one could imagine different distributions of variable encoding within this selectivity space. The properties of these distributions are determined by a covariance matrix: the diagonal elements describe the strength of coding of each variable (variance) whereas the off-diagonal entries determine the strength of the relationship between variables (covariance). In our generative models neural firing rates were simply constructed as a linear combination of task variables (colour, shape and XOR inputs represented using one-hot encodings; see *Methods, generative models* for details) with a specified covariance matrix of selectivities to the task variables. In line with previous studies, the high-dimensional model was constructed by allowing selectivity to linear (colour and shape) and nonlinear (XOR) features to be distributed randomly according to a spherical Gaussian distribution (**Fig. 1b**). In contrast, in the minimally structured XOR selectivity model, neurons were only selective for the nonlinear interaction (i.e., the XOR) term carried variance (**Fig. 1c**). One possible way of constructing such a model is considering biologically plausible limits on neural firing rates (minimising net firing rate). In line with this, we derived this model mathematically and demonstrated that it minimises total firing rate activity while maximising task performance (see **Supp. Materials, section 1**). Consistent with this, we found that feedforward networks trained using backpropagation to perform the task while also minimising a metabolic cost term converge to the minimal XOR selectivity

model (see **Supp. Fig. 1a-f**; *Methods, optimised feedforward networks*). Please note that alternative mechanisms, such as initialisation or presence of noise, could be also applied to learn a similar minimal selectivity model[10]. In line with prior accounts[17,35], XOR decoding in the low-dimensional model was more robust to noise (**Supp Fig. 2c**) and required fewer units to implement a stable readout (see **Supp. Fig. 3g**).

We established a metric to measure whether neural activity is better described by the random or minimal model. We first fitted a linear model regression to surrogate data generated by both our generative models, in which task variables (colour, shape, and colour x shape (XOR)) were used as predictors of each unit's firing rate (see *Methods, model section;* for similar analysis see[34,36]). Then we measured the average within-model distance between two covariance matrices drawn from the random model (**Supp. Fig. 2a**, red line; *Methods, eq. 4*) and the average between-model distance between the covariances drawn from the random and the minimal model (**Supp. Fig. 2b**, blue line, *Methods, eq. 5*). To test whether our measure captures learning dynamics, we constructed four artificial populations with varying proportions of random and minimal selectivity. As the proportion of the minimal model in this mixed population increased, it became more dissimilar to the average random model and more similar to the average minimal model (**Supp. Fig. 2a**, black line). A reflected version of these results held true when the minimal model was used as reference (**Supp. Fig. 2b**).

Subsequently, to gain insight into the task geometries that these models generated, we employed an established technique[3,14] and trained linear decoders to decode all three task variables in both models. As previously suggested[3,14], a randomly mixed selectivity model yielded a high-dimensional task representation, allowing for all variables, including the nonlinear XOR, to be decoded (**Supp. Fig. 2d**, red; cf. **Fig. 1a**, far left). In contrast, for the minimal model, only the XOR combination of shape and colour, and not shape or colour independently, could be decoded (**Supp. Fig. 2d**, blue; cf. **Fig. 1a**, far right).While both models can perform the task, we expected their representation of the XOR variable to differ fundamentally. On the one hand, the minimal model by design should represent the XOR in a format that generalises over all other task variables. On the other hand, this is not guaranteed in the random model. We verified this intuition using cross-generalised decoding, a method in which a linear decoder is trained to decode a given task variable (e.g., XOR = True vs. XOR = False) on a given set of task conditions (e.g., blue colour) and tested on a different set of task conditions[7] (e.g., green colour; see *Methods section, cross-generalised decoding*). We found that for the model, cross-generalised decoding was at chance-level for all task variables (**Supp. Fig. 2e**, red). This is because the random model, by design, exhibits no reliable structure in its representation of variables and therefore these dimensions are represented randomly in relation to each other. In contrast, the minimal model displayed maximal cross-generalised decoding for the XOR variable (**Supp. Fig. 2e**, blue), indicating that it can be decoded regardless of which set of task conditions the

decoder is trained and tested on. This suggests that the minimal model is able to represent the XOR in a highly cross-generalisable format (**Fig. 1a**, far right). Consequently, the minimal model also exhibits below-chance cross-generalised decoding for colour and shape (see **Supp. Fig. 12**). Next, we directly compared neural data at each stage of learning to the selectivity (random vs minimal) and neural geometry (low- vs high-dimensional) generated by these models.

A low-dimensional task relevant geometry emerges over learning

In experiment 1, the animals were trained to combine a colour stimulus (either blue or green) with a subsequently presented shape (either square or diamond) in a nonlinear fashion to predict the reward (the XOR between colour and shape) outcome of the trial (**Fig. 1d,e**). Shapes varied in width (narrow or broad), though this feature was task-irrelevant as it did not correlate with reward outcomes (**Supp. Fig. 5a**). Using a semi-chronic multielectrode system we sequentially recorded 376 neurons from the lateral PFC across both macaques (**Fig. 1f**; see *Methods, data acquisition and pre-processing*). Moving electrodes between sessions ensured that a new sample of neurons was obtained in each session. Importantly, to capture learning dynamics, we started recording from the first session in which the animals were exposed to the task. Experimental sessions were split into four learning stages for each animal separately. Data for each stage was combined across animals. Selectivity analyses were run in the time window (**Fig. 1e**, *analysis*) before the animals received feedback about the outcome of the trial (i.e., reward; for details see *Methods*, *data acquisition and pre-processing*).

We first assessed learning behaviourally by examining the animals' tendency to terminate non-rewarded trials before the potential reward onset (shape-locked period**,** **Fig. 1e**) by breaking fixation. This was quantified by calculating adaptive trial termination (ATT) - the ratio of terminated non-rewarded trials to terminated rewarded trials, which we tracked across different learning stages (see *Methods, adaptive trial termination* for details). Our findings indicate that, over the course of learning, the animals increasingly differentiated between rewarded and non-rewarded trials, suggesting an adaptive change in neural representations ($r = .56, p < 0.01$; **Fig. 2a**).

We next applied our linear decoding analyses to the neural recordings to establish whether the emergence of this behavioural strategy was accompanied by changes in the geometry of neural representations. At the beginning of training (learning stage 1, grey lines), the animals exhibited a high-dimensional geometry that allowed for colour, shape and XOR to be decoded, reminiscent of the high-dimensional model (**Fig. 2b-d)**. This was especially prominent in the period prior to the reward delivery (shaded area). Over the course of learning (stage 1 vs stage 4) we observed a reduction in colour decoding ($cluster\ 1: 0.112 - 0.182s, p = 0.05;\ \ cluster\ 2: 0.826 - 0.927s, p =$

0.01; $cluster\ 3: 1.007 - 1.500s, p < 0.001$) and shape decoding ($cluster\ 1: 0.564 - 0.917s, p < 0.001$; $cluster\ 2: 1.047 - 1.349s, p < 0.001$) but not XOR decoding (**Fig. 2b-d,** grey vs black lines). The shape stimuli also had a feature that was irrelevant for the prediction of the outcome: width (**Supp. Fig. 5a**). Similarly to colour and shape, the decoding of this feature also decreased over learning (**Supp. Fig. 5b**; $cluster\ 1: 0.585 - 1.369s, p < 0.001$). The reduction of input features (colour, shape) and stable output feature (XOR) decoding was predicted by a transition from a high-dimensional to a low-dimensional model **(Supp. Fig. 2d**, red vs blue).

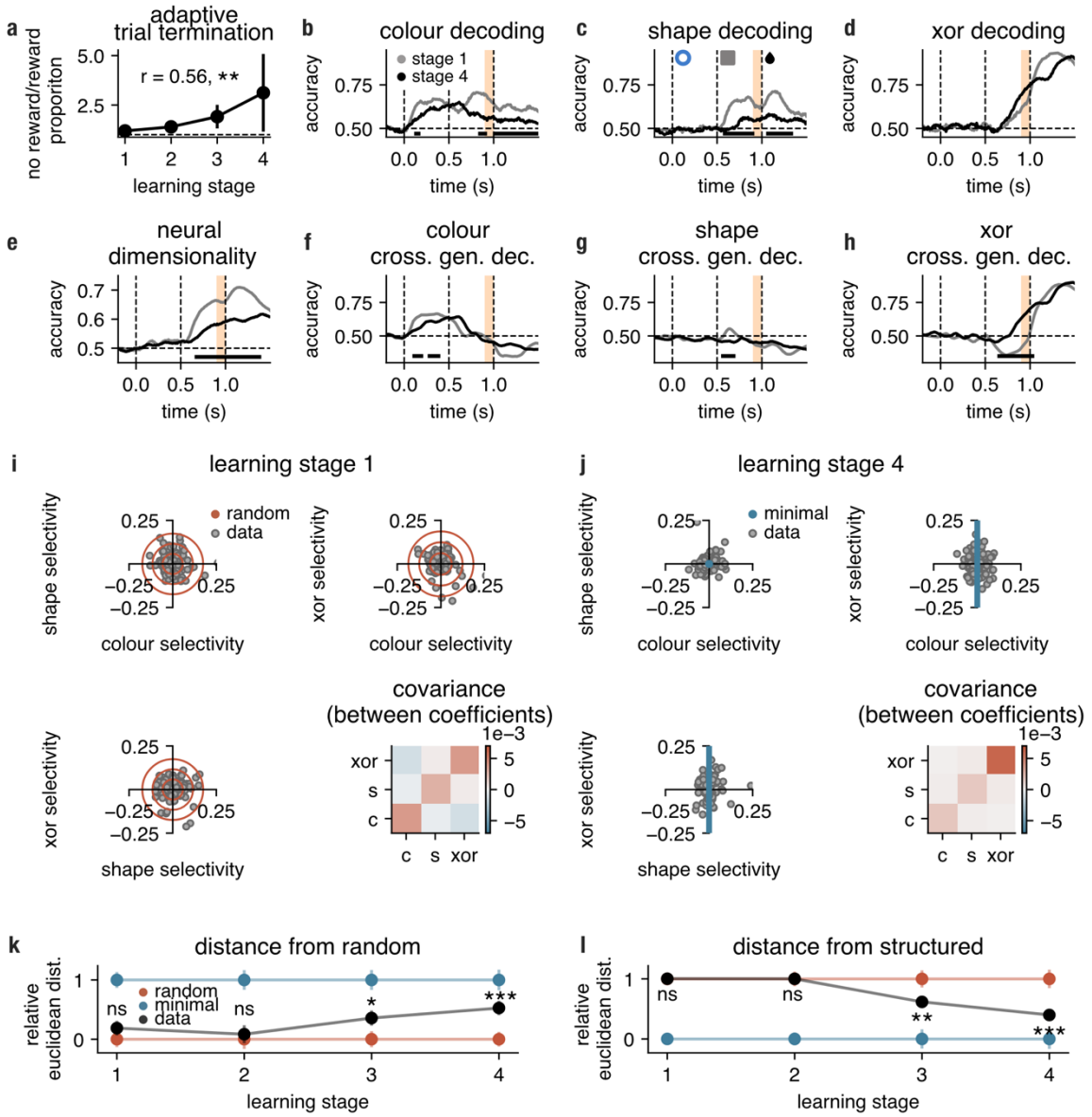


Figure 2. Analysis of neural representations in macaque PFC over learning. a, The tendency of animals to terminate trials in the shape-locked trial period plotted as a ratio of termination numbers in not rewarded and rewarded trials (illustrated as a function of learning). **b-d**, Time resolved linear decoding in stage 1(grey) and stage 4(black) of colour, shape, and XOR, respectively; horizontal bars indicate significant difference in decoding between stage 1 and stage 4 computed using a cluster-based permutation test; the pale orange area indicates the time window for which subsequent analyses were run. Vertical three dashed lines show the onset of the colour, shape and the outcome, respectively. **e**,

Time resolved neural dimensionality as measured by shattering dimensionality (all possible task dichotomies excluding colour, shape, width and XOR). **f-h**, Time resolved cross-generalised linear decoding of task variables for learning stages 1 (grey) and 4 (black). **i**,**j**, Neural selectivities in learning stages 1 and 4, respectively. Each point represents the selectivity of one neuron for each of the task variables (colour, shape and XOR). We show all 3 possible pairs of the 3 axes. In panel **i**, the 3 red rings show spherical Gaussian distributions corresponding to 1, 2, and 3 standard deviations of the data. In panel **j**, the 3 blue rings show ellipses from the minimal model that correspond to 1, 2, and 3 standard deviations of the data (note that they all lie on top of one another along the XOR selectivity axis). The right panels show the covariance matrix of the selectivities computed from the data for stage 1 of learning (**i**) and stage 4 of learning (**j**). **k**, Relative Euclidean distance between the covariance matrix of selectivity coefficients from the data and the covariance matrix expected from random selectivity (with matched total variance) as a function of learning (*Methods*, *measuring similarity between selectivity distributions*). Red and blue, respectively, error bars show mean ($\pm 1\ s.d.$ over 1000 randomly drawn models) of the relative Euclidean distance between the covariance matrix of the random, respectively minimal, model (with matched total variance to the data) and the covariance matrix expected from random selectivity; black error bars show standard deviation of relative distance between the observed covariance and random covariance ($\pm 1\ s.d.$ over 1000 random models). **l**, Same as panel d but we show the relative distance from the covariance matrix expected from minimal selectivity (with matched total variance). All p-values were calculated from permutation tests (***, $p < 0.01$; **, $p < 0.01$; *, $p < 0.05$; †, $p < 0.1$; $n.s.$, not significant).

We next explicitly tested whether the dimensionality of neural representations changed over learning (as measured with shattering dimensionality; see ref. [14] and *Methods, decoding* details). A neural representation described by three binary input dimensions (colour, shape and width) results in 35 dichotomies (division into two sets of four stimuli) that can be theoretically decoded. We found that the mean decoding accuracy of all dimensions (excluding colour, shape, width and XOR) decreased significantly over learning ($cluster\ 1: 0.655 - 1.399s, p < 0.001$). Additionally, a principal component analysis computed on condition averages revealed that the proportion of variance explained by the first principal component increased as a function of learning ($M_{stage\ 1} = 0.466\ vs\ M_{stage\ 4} = 0.581, p < 0.05$, one-sided; **Supp. Fig. 3f**; for details see *Methods, Principal component analysis*). Furthermore, as colour, shape, and width form a cube in the input space (when one-hot encoding is used) these dichotomies can be split into linear (a hyperplane can be used to divide all vertices into two sets) and nonlinear dimensions (a nonlinear function needs to be used to differentiate two sets of 4 vertices). We found that the decoding of linear dimensions decreased significantly over learning ($M_{stage\ 1} = 0.641\ vs\ M_{stage\ 4} = 0.585, p < 0.05$, one-sided), while the decoding of non-linear dimensions did not ($M_{stage\ 1} = 0.595\ vs\ M_{stage\ 4} = 0.602, p = 0.74$, two-sided; **Supp. Fig. 5g**). The progressive change over learning, not only in the neural dimensionality but also in the type of computations that are represented (linear vs non-linear) conflicts with a static model of the PFC as a high-dimensional nonlinear kernel (e.g., such as that utilised in reservoir computing[15,16]) over long timescales of adaptation.

Changes to neural dimensionality were associated with simultaneous changes in neural geometry, whereby the PFC transformed representations from specific to abstract over learning (**Fig. 2f-h**). This was made evident by cross-generalised decoding employed to identify shared neural representations. We found a decrease in colour and shape cross-generalised decoding over learning (stage 1 vs stage 4). This decrease, confined to the colour-locked period for the colour variable ($cluster\ 1: 0.092 - 0.212s, p < 0.001$; $cluster\ 2: 0.263 - 0.404s, p < 0.001$; **Fig. 2f**), was likely due to a general reduction in colour coding. Similarly, shape cross-generalised decoding decreased during the early shape-locked period ($cluster\ 1: 0.544 - 0.705s, p < 0.001$; **Fig. 2g**). Importantly, we observed a significant increase in cross-generalised XOR decoding over learning ($cluster\ 1: 0.635 - 1.047s, p < 0.001$; **Fig. 2h**). The combination of low colour and shape cross-generalised decoding and high XOR cross-generalised decoding at the end of learning (stage 4, black lines) are characteristic features of the low-dimensional model (**Supp. Fig. 2e**). These results collectively indicate that the task representation underwent a geometrical transition during learning. To further illustrate this transition explicitly, we focused on cross-generalisation of the XOR across different shape widths (broad vs. narrow; **Supp. Fig. 5a**). Training a decoder on XOR classification in narrow shapes and testing it on broad shapes, and vice versa, showed that the XOR representation for both trial types aligned over learning ($cluster\ 1: 0.836 - 1.037s, p < 0.05$; **Supp. Fig. 5c**). While a shared XOR readout could arise from the abstraction of the XOR signal, it may also be attributed to the progressive projection of a reward-prediction signal (or related motor activity, such as licking) into the PFC over learning. Experiment 2 provides further insights into distinguishing between these two potential sources.

We next tested whether changes to neural geometry and dimensionality were reflected in changes to selectivity. We fitted a linear regression to our data, just as we did for our generative models (**Supp. Fig 2a,b**), in which task variables (colour, shape, and colour x shape (XOR)) were used as predictors of each neuron's firing rate. We then examined how selectivity coefficients changed over learning (**Fig. 2i**, learning stage 1 and **Fig. 2j**, learning stage 4). We compared the covariance structure of these selectivity coefficients (**Fig. 2i,j,** bottom right) to the covariances obtained from the random selectivity model (**Fig. 2i**, red contours) and minimal selectivity model (**Fig. 2j**, blue line). At the beginning of learning (stage 1), PFC cells were randomly distributed in selectivity space resembling the high-dimensional model ($p = 0.332$, one-sided, **Fig. 2k** and $p = .99$, one-sided, **Fig. 2l**). However, in late learning (stage 4), selectivity diverged away from randomly mixed selectivity ($p < 0.001$, one-sided, **Fig. 2k**, compare black and red lines) and converged towards the minimal model ($p < 0.001$, one-sided, **Fig. 2l**, compare black and blue lines).

We replicated these decoding, cross-generalised decoding, and selectivity analyses on data sorted by the adaptive trial termination index. Here, the behavioural measure

served as a performance indicator, and sessions were sorted separately for each animal before being pooled into four pseudopopulations across animals (see *Methods, trial termination measure*; **Supp. Fig. 5i-r**). Furthermore, we also performed a re-analysis of an existing dataset[25–27] in which recordings were taken from primate ventral and dorsolateral PFC before and after learning a delayed match-to-sample task that was similar in structure to ours (for details see *Methods, existing lPFC dataset*). This task required active behaviour, and the reward signal was disentangled from the XOR (i.e., match/no match) signal. We found that, as in our data, learning pushed neural activity in the PFC towards a minimal XOR regime (**Supp. Fig. 6**).

Our findings indicate that neural activity in the PFC shifts between two distinct selectivity regimes as learning progresses. Initially, the PFC maximally expanded the representational space by encoding all available variables. Subsequently, after a combination of task variables that predicted the trial's outcome were identified, the PFC reduced its dimensionality over learning towards a minimal model of the task structure. Previous studies on cross-generalisation have suggested that such low-dimensional representations should improve learning of similar problems[7]. This is based on the notion that the classification of new stimuli can be rapidly improved by leveraging already learnt representations, thus resulting in a shared (abstract) representation of task variables between the old and new problems. We tested this prediction in experiment 2.

Generalisation: abstract coding dominates task representations over learning

In experiment 2, we introduced a new colour pair (stimulus set 2) that followed the same shape–outcome associations as the previous colour pair (stimulus set 1) (**Fig. 1d**). Randomly interleaving stimulus set 1 trials and stimulus set 2 trials allowed us to test whether a shared neural representation would be used for both stimulus sets (**Supp. Fig. 1,** far right). Similar to experiment 1, we recorded neural activity from the very first session in which the animals were exposed to the new stimulus set and divided the experimental sessions into four distinct learning stages. To explore learning-induced changes to neural representations, we again employed the decoding and cross-generalised decoding metrics. We next compared these metrics to changes in the structure of neural selectivities. Specifically, for each neuron, we computed its selectivity profiles for each variable (colour, shape, XOR), separately for each stimulus set, and measured their correlation between different sets. Notably, given that the colour was presented before the shape in the experimental setup (**Fig. 1e**), these analyses were conducted for each learning stage in both the colour-locked and shape-locked time windows (**Fig 3** and **Fig. 4**, respectively).

We first examined the propensity of animals to terminate trials when the new stimulus set indicated a lack of reward, similar to the patterns observed in experiment 1. Over

the course of learning with stimulus set 2, animals increasingly terminated non-rewarded trials more frequently than rewarded ones ($r = .52, p < 0.05$, **Supp. Fig. 7r**). Additionally, a facilitation effect was observed when comparing the first three sessions of experiment 2 (stimulus set 2 only) to the first three sessions of experiment 1 (stimulus set 1). Specifically, animals demonstrated significantly more adaptive trial termination early in the learning process with stimulus set 2 compared to stimulus set 1 ($p < 0.01$, **Supp. Fig. 7s**). This early behavioural benefit may be attributed to the utilisation of the previously acquired task representation as a scaffold. Subsequently, we thus investigated how the neural representations of both stimulus sets interacted throughout the learning process.

Hypothetically, PFC could represent the new colours separately from the already learned pair (high-dimensional representation; **Fig 3a**, left), or it could generate an abstract representation (shared across sets) of the colours' contextual meaning (low-dimensional representation; **Fig 3a**, right). Through an exploration of neural activity in the colour-locked period (**Fig. 1e**), we observed that the PFC collapsed the stimulus set dimension (**Fig. 3b**). More specifically, we found that the animals were able to clearly distinguish between the old and new colours during the early stages of learning (**Fig. 3b**, black, 'decoding'; see **Supp. Fig. 7a-e** for temporally resolved decoding). As learning progressed, however, we observed a reduction in stimulus-set decoding ($p < 0.001$, one-sided) and corresponding cross-context decoding of stimulus set ($p < 0.001$, one-sided, **Fig. 3b**, grey, 'cross. decoding'; see **Supp. Fig. 7f-j** for temporally resolved cross. decoding). We also found that at the beginning of learning (stage 1) cells that were selective to colour 1vs3 (e.g. blue vs pink) also tended to be selective to colour 2vs4 (e.g. green vs khaki), i.e., coded for the representation of stimulus set (**Fig. 3c; Supp. Fig. 7k**). PFC cells lost this selectivity pattern over learning ($p < 0.05$, one-sided; **Supp. Fig. 7k**; **Fig. 3d)**. Although context was equally well decodable in early as well as in late learning ($p = 0.99$, two-sided, **Fig. 3e**, black, decoding), its format changed from an item-specific representation into an abstract code (**Fig. 3a**, right). We tested this explicitly by training a linear SVM classifier to decode the context in stimulus set 1 and tested this model on differentiating context in stimulus set 2, and vice-versa (see *Methods, cross-stimulus set generalisation*). We found that the representations of new and old colours became aligned as a function of learning, to allow for a single context readout ($p = 0.01$, one-sided, **Fig. 3e**, grey, 'cross. decoding'). This change in geometry was reflected in a change in neural selectivity, as with the progression of learning, PFC cells exhibited an increasingly positive correlation between selectivity for colour pair 1 and colour pair 2 ($p < 0.001$, one-sided; **Supp. Fig. 7l**; **Fig. 3f,g**). Additionally, colour decoding in set 2 approached the levels of colour decoding in set 1 over learning but this was only a trend-level effect ($p < 0.1$, two-sided; **Fig. 3h**). These changes to neural geometry were accompanied by changes in the dimensionality of representations. More specifically, we found that shattering dimensionality in the colour-locked period significantly decreased over

learning ($p < 0.001$, **Supp. Fig. 7q**). The emergence of a low-dimensional and abstract representation of context before the appearance of the shape suggests that the network occupied a preparatory activity state that permits the immediate, context-appropriate transformation of shape information to reward prediction (XOR representation)[6,34,36,37].

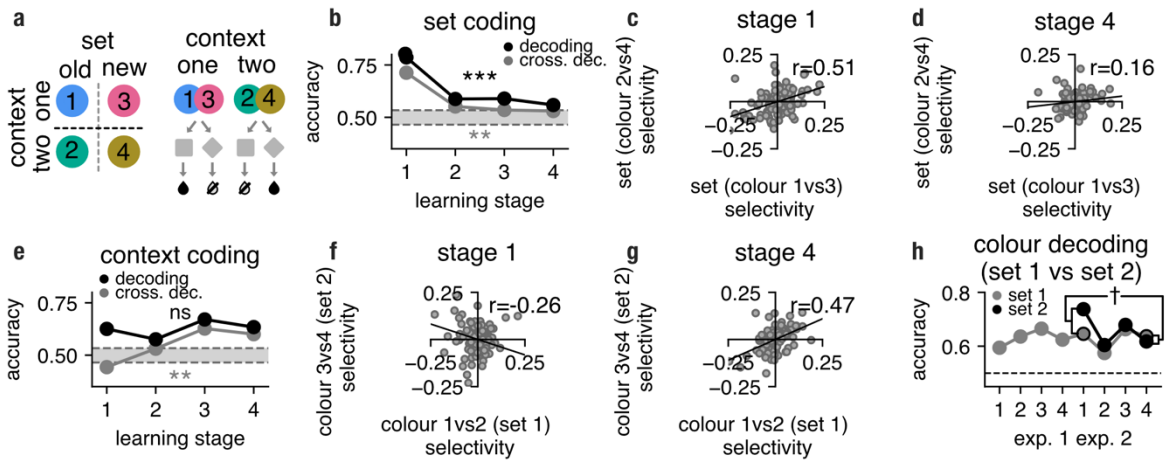


Figure 3. **The PFC aligns new and old colours as a function of learning to allow a single readout of context. a**, New colours (pink and khaki) followed the same shape-reward association as already learnt colours (blue and green; left); this resulted in the colour circle having two dimensions: task relevant (context) and task irrelevant (stimuli set; right). **b**, Decoding (black) and cross-generalised decoding (grey) of set as a function of learning in the colour-locked period (see **Fig. 1e**). The grey shaded areas indicate combined chance-level cross. gen. decoding and chance-level decoding, **c, d**, Selectivity of PFC neurons in stage 1 and stage 4 for set in context 1 and context 2 in the colour-locked period. The line of best fit is shown in black. **e-g**, Analogous to **b-d** but for context coding (f) and context selectivity (g, h). **h**, Decoding of colour separately for stimulus set 1 (grey, 'blue vs green') and stimulus set 2 (black, 'pink vs khaki'). All p-values were calculated from permutation tests (***, p < 0.01; **, p <0.01; *, p < 0.05; †, $p < 0.01$; n.s., not significant). Shading in panels b and e shows the mean ±3 s.d. of chance-level decoding.

We next focused on the neural activity in the shape-locked period, which reflected the state of the PFC when all necessary information for outcome prediction was available (colour and shape). We found that in both early learning (stage 1) and late learning (stage 4), XOR decoding was high and no learning-induced changes to the XOR signal were detected ($p = 0.2$, two-sided, **Fig. 4a**, black, decoding). This suggests that the animals were able to rapidly cross-generalise the XOR representation between stimulus sets as early as in stage 1, consistent with behavioural evidence. We tested this explicitly by training a linear SVM classifier to decode the XOR in stimulus set 1 and tested this model on differentiating the XOR in stimulus set 2, and vice-versa. We observed high scores of cross-generalised XOR decoding in both stage 1 and stage 4, as well as a small increase in this measure as a function of learning ($p = 0.05$, one-sided, **Fig. 4a**, grey, 'cross. decoding'). We also observed a learning-induced increase in the shared selectivity for XOR between set 1 and set 2 ($p < 0.05$, one-sided, **Supp. Fig. 7o; Fig. 4b-c**;). These results suggest that, although the PFC constructed an abstract (shared across sets) representation of the XOR early in learning, it continued to align the set-specific XOR boundaries until this abstract format fully dominated the

XOR representation. No differences between shape decoding were detected across the learning stages ($p = 0.6$, two-sided, **Fig. 4e**, black, decoding). However, both cross-generalised decoding ($p < 0.001$, one-sided, **Fig. 4e**, grey, 'cross. decoding') and selectivity analyses ($p < 0.05$, one-sided, **Supp. Fig. 5m**; **Fig. 4f,g**) showed that the PFC progressively aligned the shape representation between stimulus sets. It is important to note that the same shapes were used in both stimulus set 1 and stimulus set 2. These results suggest that, at the start of the learning process, the PFC encodes shape differently (using a high-dimensional representation) depending on the stimulus set, and that only during late learning (stage 4), the representations converged to a single shared axis (become low-dimensional). We next explored the representation of context (which we already introduced in **Fig. 3**) during the shape-locked period. We found no learning-induced changes to the geometry (decoding: $p = 0.66$, two-sided; cross. decoding: $p = 0.18$, one-sided) or selectivity ($p = 0.15$, one-sided, **Supp. Fig. 7p**; **Fig. 4i-l**). We speculate that the PFC already transformed the context signal in the colour-locked period, so that after this preparatory state was used to guide early shape processing, context information did not serve any relevant function in the late shape-locked period. Finally, we found no differences in XOR decoding as a function of learning in both set 1 and set 2 ($p = 0.93$, one-sided, respectively; **Fig. 4d**). However, the decoding of shape and colour in set 2 approached the levels of shape and colour decoding in set 1 with learning ($p < 0.1$, one-sided and $p < 0.01$, one-sided, respectively; **Fig. 4h, l**).

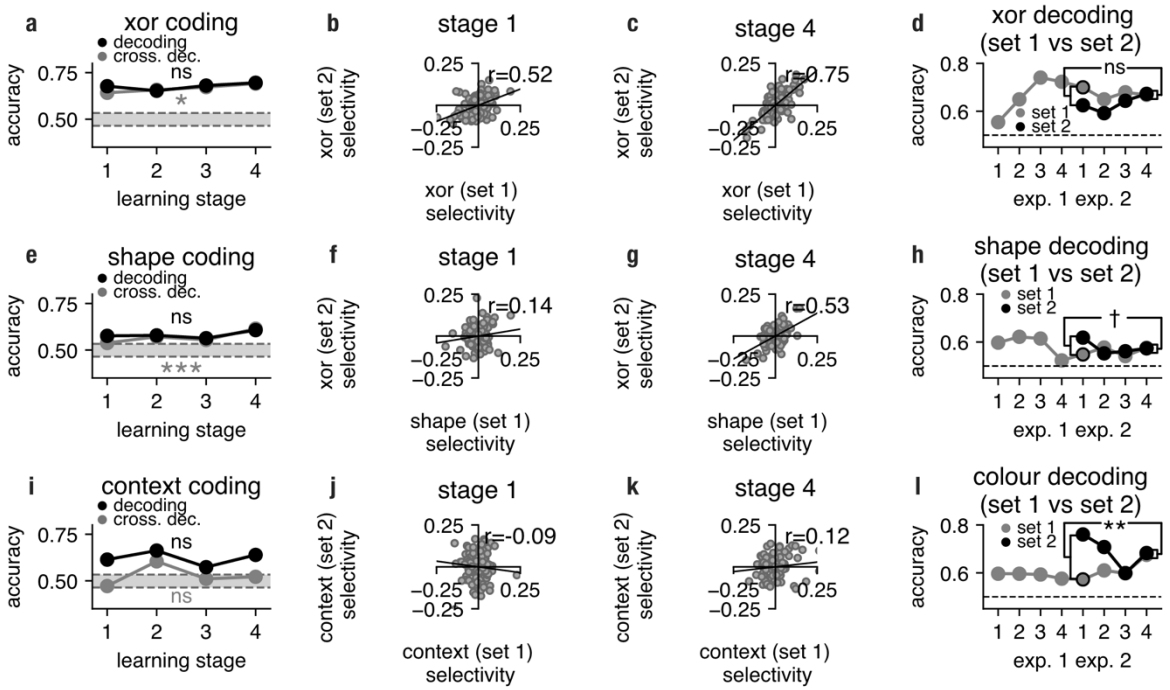


Figure 4. **The PFC utilised abstract stimulus representations when previously unseen stimulus combinations were introduced**. **a**, Decoding (black) and cross-generalised decoding (grey) of XOR as a function of learning in the shape-locked period (see **Fig. 1e**). The grey shaded areas indicate combined chance-level cross. gen. decoding and chance-level decoding, respectively. **b,c**, Selectivity of PFC neurons for XOR in stimulus set 1 and stimulus set 2 in the shape-locked period in stage 1 and

stage 4, respectively. The line of best fit is shown in black. **d**, Decoding of XOR shown separately for stimulus set 1 (grey) and stimulus set 2 (black). **e-l**, Analogous to a-d but for shape (e-h) and context (i-l). All p-values were calculated from permutation tests (***, $p < 0.01$; **, $p < 0.01$; *, $p < 0.05$; †, $p < 0.01$; $n.s.$, not significant). Shading in panels a, e, and i shows the mean ±3 s.d. of chance-level decoding and cross-set decoding.

Discussion

The prefrontal cortex has the capacity to generate both low-dimensional and high-dimensional representations, each of which presents a unique trade-off between generalisability and discriminability. However, the conditions under which each regime is employed currently remain unclear. Our study investigated how the dimensionality and geometry of neural activity changed over learning. We observed that, as learning progressed, neural activity in the PFC transitioned from being high-dimensional with high-discriminability to being low-dimensional and abstract. This transition in the representational strategy was accompanied by a change in population structure, from random and nonlinearly mixed towards minimal and structured. The structured representations that emerged during learning then supported the generalisation of the learned rule to a novel stimulus set.[18,20]

We found that the PFC transitioned from a high- to a low-dimensional regime over multiple days of exposure to a complex task. This corroborates the findings of Hirokawa et al.,[12] who found that neural activity covaried with behaviourally relevant variables, thus occupying a low-dimensional manifold. On the other hand, some studies have suggested that an increase in neural dimensionality is predictive of performance[14]. It is possible that the structure of the task and training provides an explanation for these contrasting findings[18]. In our study, recordings were initiated from the onset of task training, spanning a period of five weeks, with animals experiencing the entire task structure from the first session. In contrast, many other investigations into the primate PFC's involvement in complex cognitive tasks train animals in a fashion that decomposes the task into multiple subcomponents and either builds up task knowledge across training[14] or presents them in a serial (block-wise) manner[7,38]. Additionally, whereas in our study, one source of information (width) was always irrelevant, in other studies, information becomes periodically relevant and irrelevant across multiple blocks, which may promote encoding of currently irrelevant information[36]. These training differences may promote information encoding in a high-dimensional manner that favours discriminability over abstraction.

The acquisition of a low-dimensional representation following the learning of a single XOR rule raises the question of which regime the PFC might adopt when confronted with more complex tasks. Although XOR operations necessitate nonlinear integration and abstraction from sensory input, they can be reduced to simple stimulus-response pairings once the rule has been learnt or when a memory-based strategy is utilised.

However, some tasks are more difficult to decompose, as they require switching between multiple orthogonal or conflicting subtasks. It has been suggested[18,20] that in conditions requiring the performing of multiple tasks in series, a high-dimensional representation could be employed by the PFC in order to maximise flexibility and prevent interference.
Our analytical approach aligns with a growing body of work that emphasises the connection between population structure and neural coding[39]. Specifically, we investigated how diverse forms of single-cell selectivity contribute to the geometry and dimensionality of population-level representations. Our results reveal a negative correlation between neural dimensionality and the emergence of structured selectivity. However, the direction of this relationship may be more nuanced and task-dependent. For instance, a highly structured population with pure selectivity for multiple variables might exhibit higher dimensionality than a randomly mixed population responding to fewer variables, even if the latter is nonlinear.

We cannot rule out the possibility that the shift from a specific to abstract representation of the XOR signal may reflect changes related to reward anticipation or anticipatory behaviours, such as licking. It is possible that an increase of licking behaviour prior to reward onset could potentially drive the enhanced cross-generalisation of the XOR signal over learning. However, this confound cannot account for the same pattern observed with the context variable in experiment 2, where increased context cross-generalisation over learning was observed. During the colour presentation period, the animals lack access to shape information, preventing them from forming a valid reward prediction. Thus, the shift from specific to abstract seen in experiment 2 cannot be attributed to reward anticipation alone or related motor activity.

Our findings suggest that the PFC could employ a multi-phase learning strategy, involving distinct temporal dynamics. Initially, novel tasks could be solved via flexible, reservoir-like dynamics, bypassing the need for immediate synaptic plasticity. As training progresses over longer timescales, the PFC could gradually refine its local connectivity, optimising for performance. This dual-phase approach enables both rapid adaptation and efficient resource allocation, echoing models of the cerebellum-motor cortex interactions, where the cerebellum rapidly drives cortical activity through input control[40]. Similarly, an external region could modulate the PFC's activity on shorter timescales, enabling flexible high-dimensional representations. Over time, the PFC's intrinsic circuitry would consolidate these representations and assume direct task control. Future research could explore the geometry of task representations acquired at different learning stages and the critical role of synaptic plasticity in this process.

The implicit assumption in many experimental paradigms is that the animals are presented with tasks as tabula rasa, devoid of prior knowledge or training. However, it is unlikely that the animal's entire experimental history, including life experience, is

irrelevant to a given task. Our second experiment allowed us to address this issue and explicitly explore the interactions between already learnt and new information. In line with previous predictions[7,17], we found that when a new task instance is added, both the new and old instances were rapidly aligned to common axes and sensory differences between them were collapsed. Notably, different task motifs exhibited distinct generalisation timescales: the XOR representation generalised early in learning, while the context motif required weeks of training to generalise. This disparity likely reflects differences in their structural composition. The XOR rule's use of identical shapes across tasks likely facilitated rapid alignment, leveraging existing neural encoding schemes. In contrast, the context motif's novel colours necessitated additional encoding and adaptation in the prefrontal cortex, slowing generalisation. This suggest that PFC's representational alignment is modulated by the degree of overlap between prior and novel stimuli. Shared features could thus promote efficient transfer of learned representations, while novel features could impose additional encoding demands.

It is perhaps surprising that, given the key role of the PFC in the development and acquisition of structured knowledge, only a few studies have investigated how the structure of PFC representations changes during several training days of an entirely novel task [41,42]. By tracking changes in neural activity across learning, it is possible to identify the biological principles that are required to produce representations supporting higher cognitive functions[43]. Future experiments should extend this paradigm, to track changes in learning even more complex and naturalistic tasks[44]; those that have a compositional structure[45,46] the influence of different learning curricula[47]; and how these representations change within the same individual neurons as opposed to pseudo populations[48–50].

References

1. Gao, P. & Ganguli, S. On simplicity and complexity in the brave new world of large-scale neuroscience. *Current Opinion in Neurobiology* vol. 32 148–155 Preprint at https://doi.org/10.1016/j.conb.2015.04.003 (2015).
2. Langdon, C., Genkin, M. & Engel, T. A. A unifying perspective on neural manifolds and circuits for cognition. *Nat Rev Neurosci* (2023) doi:10.1038/s41583-023-00693-x.
3. Fusi, S., Miller, E. K., Rigotti, M., Karpova, A. & Kiani, R. Why neurons mix: high dimensionality for higher cognition This review comes from a themed issue on Neurobiology of behavior. *Curr Opin Neurobiol* **37**, 66–74 (2016).
4. Badre, D., Bhandari, A., Keglovits, H. & Kikumoto, A. The dimensionality of neural representations for control. *Current Opinion in Behavioral Sciences* vol. 38 20–28 Preprint at https://doi.org/10.1016/j.cobeha.2020.07.002 (2021).
5. Kaufman, M. T. *et al.* The implications of categorical and category-free mixed selectivity on representational geometries. *Current Opinion in Neurobiology* vol. 77 Preprint at https://doi.org/10.1016/j.conb.2022.102644 (2022).
6. Stokes, M. G. *et al.* Dynamic coding for cognitive control in prefrontal cortex. *Neuron* **78**, 364–375 (2013).
7. Bernardi, S. *et al.* The Geometry of Abstraction in the Hippocampus and Prefrontal Cortex. *Cell* **183**, 954-967.e21 (2020).
8. Cromer, J. A., Roy, J. E. & Miller, E. K. Representation of Multiple, Independent Categories in the Primate Prefrontal Cortex. *Neuron* **66**, 796–807 (2010).
9. Roy, J. E., Riesenhuber, M., Poggio, T. & Miller, E. K. Prefrontal Cortex Activity during Flexible Categorization. *Journal of Neuroscience* **30**, 8519–8528 (2010).
10. Flesch, T., Juechems, K., Dumbalska, T., Saxe, A. & Summerfield, C. Orthogonal representations for robust context-dependent task performance in brains and neural networks. *Neuron* **110**, 1258-1270.e11 (2022).
11. Reinert, S., Hübener, M., Bonhoeffer, T. & Goltstein, P. M. Mouse prefrontal cortex represents learned rules for categorization. *Nature* **593**, 411–417 (2021).
12. Hirokawa, J., Vaughan, A., Masset, P., Ott, T. & Kepecs, A. Frontal cortex neuron types categorically encode single decision variables. *Nature* **576**, 446–451 (2019).
13. Duncan, J. An adaptive coding model of neural function in prefrontal cortex. *Nat Rev Neurosci* **2**, 820–829 (2001).
14. Rigotti, M. *et al.* The importance of mixed selectivity in complex cognitive tasks. *Nature* **497**, 585–590 (2013).
15. Maass, W., Natschläger, T. & Markram, H. *Real-Time Computing Without Stable States: A New Framework for Neural Computation Based on Perturbations*. *Neural Computation* vol. 14 (2002).
16. Enel, P., Procyk, E., Quilodran, R. & Dominey, P. F. Reservoir Computing Properties of Neural Dynamics in Prefrontal Cortex. *PLoS Comput Biol* **12**, e1004967 (2016).
17. Barak, O., Rigotti, M. & Fusi, S. The sparseness of mixed selectivity neurons controls the generalization-discrimination trade-off. *Journal of Neuroscience* **33**, 3844–3856 (2013).
18. Musslick, S. & Cohen, J. D. Rationalizing constraints on the capacity for cognitive control. *Trends in Cognitive Sciences* vol. 25 757–775 Preprint at https://doi.org/10.1016/j.tics.2021.06.001 (2021).
19. Ehrlich, D. B. & Murray, J. D. Geometry of neural computation unifies working memory and planning. doi:10.1101/2021.02.01.429156.

20. Musslick, S., Saxe, A., Novick, A., Reichman, D. & Cohen, J. D. *Running Head: RATIONAL BOUNDEDNESS OF COGNITIVE CONTROL 1 On the Rational Boundedness of Cognitive Control: Shared Versus Separated Representations*.
21. Aston-Jones, G. & Cohen, J. D. An integrative theory of locus coeruleus-norepinephrine function: Adaptive gain and optimal performance. *Annual Review of Neuroscience* vol. 28 403–450 Preprint at https://doi.org/10.1146/annurev.neuro.28.061604.135709 (2005).
22. Watanabe, M. *Prefrontal Unit Activity During Delayed Conditional Go/No-Go Discrimination in the Monkey. II. Relation to Go and No-Go Responses*. *Brain Research* vol. 382 (1986).
23. di Pellegrino, G. & Wise, S. P. *Visuospatial versus Visuomotor Activity in the Premotor and Prefrontal Cortex of a Primate*. *The Journal of Neuroscience* vol. 13 (1993).
24. Asaad, W. F., Rainer, G. & Miller, E. K. *Neural Activity in the Primate Prefrontal Cortex during Associative Learning). Medial Temporal Structures Critical for Long-Term Memories Are Also Important: Dam-Age to the Hippocampus and/or Subjacent Cortex (Mur*. *Neuron* vol. 21 (1998).
25. Constantinidis, C., Qi, X.-L. & Meyer, T. Single-neuron spike train recordings from macaque prefrontal cortex during a visual working memory task before and after training. *CRCNS. org* (2016).
26. Qi, X. L., Meyer, T., Stanford, T. R. & Constantinidis, C. Changes in prefrontal neuronal activity after learning to perform a spatial working memory task. *Cerebral Cortex* **21**, 2722–2732 (2011).
27. Meyer, T., Qi, X. L., Stanford, T. R. & Constantinidis, C. Stimulus selectivity in dorsal and ventral prefrontal cortex after training in working memory tasks. *Journal of Neuroscience* **31**, 6266–6276 (2011).
28. Raposo, D., Kaufman, M. T. & Churchland, A. K. A category-free neural population supports evolving demands during decision-making. *Nat Neurosci* **17**, 1784–1792 (2014).
29. Gao, P. & Ganguli, S. On simplicity and complexity in the brave new world of large-scale neuroscience. *Current Opinion in Neurobiology* vol. 32 148–155 Preprint at https://doi.org/10.1016/j.conb.2015.04.003 (2015).
30. Sadtler, P. T. *et al.* Neural constraints on learning. *Nature* **512**, 423–426 (2014).
31. Ganguli, S. *et al.* One-Dimensional Dynamics of Attention and Decision Making in LIP. *Neuron* **58**, 15–25 (2008).
32. Chaudhuri, R., Gerçek, B., Pandey, B., Peyrache, A. & Fiete, I. The intrinsic attractor manifold and population dynamics of a canonical cognitive circuit across waking and sleep. *Nat Neurosci* **22**, 1512–1520 (2019).
33. Cueva, C. J. *et al.* Low-dimensional dynamics for working memory and time encoding. *PNAS* **117**, 23021–23032 (2020).
34. Dubreuil, A., Valente, A., Beiran, M., Mastrogiuseppe, F. & Ostojic, S. The role of population structure in computations through neural dynamics. *Nat Neurosci* **25**, 783–794 (2022).
35. Bernardi, S. *et al.* The Geometry of Abstraction in the Hippocampus and Prefrontal Cortex. *Cell* **183**, 954-967.e21 (2020).
36. Mante, V., Sussillo, D., Shenoy, K. V. & Newsome, W. T. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* **503**, 78–84 (2013).
37. Buonomano, D. V. & Maass, W. State-dependent computations: Spatiotemporal processing in cortical networks. *Nature Reviews Neuroscience* vol. 10 113–125 Preprint at https://doi.org/10.1038/nrn2558 (2009).

38. Bartolo, R., Saunders, R. C., Mitz, A. R. & Averbeck, B. B. Dimensionality, information and learning in prefrontal cortex. *PLoS Comput Biol* **16**, (2020).
39. Ostojic, S. & Fusi, S. Computational role of structure in neural activity and connectivity. *Trends in Cognitive Sciences* vol. 28 677–690 Preprint at https://doi.org/10.1016/j.tics.2024.03.003 (2024).
40. Pemberton, J., Chadderton, P. & Costa, R. P. Cerebellar-driven cortical dynamics enable task acquisition, switching and consolidation. *bioRxiv* 2011–2022 (2022).
41. Reinert, S., Hübener, M., Bonhoeffer, T. & Goltstein, P. M. Mouse prefrontal cortex represents learned rules for categorization. *Nature* **593**, 411–417 (2021).
42. Corrigan, B. W. *et al.* Distinct neural codes in primate hippocampus and lateral prefrontal cortex during associative learning in virtual environments. *Neuron* **110**, 2155-2169.e4 (2022).
43. Richards, B. A. *et al.* A deep learning framework for neuroscience. *Nature Neuroscience* vol. 22 1761–1770 Preprint at https://doi.org/10.1038/s41593-019-0520-2 (2019).
44. Hunt, L. T. *et al.* Formalizing planning and information search in naturalistic decision-making. *Nature Neuroscience* vol. 24 1051–1064 Preprint at https://doi.org/10.1038/s41593-021-00866-w (2021).
45. Yang, G. R., Joglekar, M. R., Song, H. F., Newsome, W. T. & Wang, X. J. Task representations in neural networks trained to perform many cognitive tasks. *Nat Neurosci* **22**, 297–306 (2019).
46. Driscoll, L., Shenoy, K. & Sussillo, D. Flexible multitask computation in recurrent networks utilizes shared dynamical motifs 1 2. doi:10.1101/2022.08.15.503870.
47. Dekker, R. B., Otto, F. & Summerfield, C. Curriculum learning for human compositional generalization. (2022) doi:10.1073/pnas.
48. Tolias, A. S. *et al.* Recording chronically from the same neurons in awake, behaving primates. *J Neurophysiol* **98**, 3780–3790 (2007).
49. Steinmetz, N. A. *et al.* Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings. *Science (1979)* **372**, (2021).
50. Zhao, S. *et al.* Tracking neural activity from the same cells during the entire adult life of mice. *Nat Neurosci* (2023) doi:10.1038/s41593-023-01267-x.
51. Tkalčič, M. & Tasič, J. F. *Colour Spaces-Perceptual, Historical and Applicational Background*. http://www.mathworks.com.
52. Maris, E. & Oostenveld, R. Nonparametric statistical testing of EEG- and MEG-data. *J Neurosci Methods* **164**, 177–190 (2007).

Author contributions. M.J.W., D.F.W, J.D., and M.G.S. conceived the study. Ma.K and M.J.W. coded the experimental procedure. J.D. led the experimental recordings, Ma.K., and Mi.K. performed the experimental recordings; M.J.W. and Mi.K performed data pre-processing. J.P.S., M.J.W., and M.G.S. developed the theoretical framework; J.P.S. performed analytical derivations and constructed the optimized networks. J.P.S. and M.J.W. constructed the generative models and designed the analysis methods. M.J.W. performed the data analysis and produced the figures. J.P.S., L.T.H, N.E.M and D.F.W. supervised and reviewed the data analysis. M.J.W., J.P.S., L.T.H, N.E.M, D.F.W. and J.D. interpreted the results and wrote the manuscript. All authors revised the final manuscript.

Acknowledgements. This work was funded by the Wellcome Trust (Sir Henry Wellcome Postdoctoral Fellowship to J.S. (215909/Z/19/Z) and Sir Henry Dale Fellowship L.T.H (208789/Z/17/Z), and award 101092/Z/13/Z to M.Ku., M.B., and J.D.), the Strategic Longer and Larger grant (awarded to L.T.H; BB/W003392/1), the Medical Research Council UK Program (MC_UU_00030/7 to M.Ku., M.B., and J.D.), the Biotechnology and Biological Sciences Research Council (award BB/M010732/1 to M.G.S.), Clarendon Fund and Saven European Scholarship (M.J.W), and the James S. Mc-Donnell Foundation (award 220020405 to M.G.S.). For the purpose of open access, the authors have applied a CC-BY public copyright licence to any author accepted manuscript version arising from this submission. We thank Emilia Piwek and Christopher Summerfield for useful feedback and detailed comments on the manuscript.

Competing interests statement. The authors declare no competing interests.

Data availability, and code availability. All code was custom written in Python using NumPy, SciPy, Matplotlib, Scikit-learn, and Tensorflow libraries. All code will be made available upon peer-reviewed publication. The data that support the findings of this study are available from the corresponding author upon request.

Methods

Data and task

*Animals and task***.** Two adult male rhesus macaques, monkey 1 and monkey 2, were trained in this study. The experiments were conducted in line with the Animals (Scientific Procedures) Act 1986 of the UK and licensed by a Home Office Project License obtained after review by Oxford University's Animal Care and Ethical Review committee. The procedures followed the standards set out in the European Community for the care and use of laboratory animals (EUVD, European Union directive 86/609/EEC). The animals were seated in a sound- and lighting-attenuated experimental booth. Their heads were restrained and faced a 19-inch screen. The centre of the screen was aligned with a neutral eye position. The animals performed a passive object-association task (**Fig. 1d,e**). Importantly, the animals were accustomed to an experimental setting but had no previous exposure to the task or stimuli introduced in this protocol. Neural recordings were collected from the first session as one of the main aims of the study was to capture learning dynamics. In the first experiment, the animals were presented with a colour and a shape, a nonlinear combination of which predicted reward (**Fig. 1d,e**). In experiment 2, a second set of stimuli was additionally introduced to test whether the rule learnt in the first experiment cross-generalised to the new sensory domain (**Fig. 1d**). The colours used in the coloured circles were designed in the CIELab colour space[51]. The L parameter (luminance) was kept constant which ensured that the stimuli were approximately isoluminant; parameters a and b varied with regard to valence but not value which resulted in a circular colour representation[51]. As colours were randomly assigned to conditions for each animal, this circular representation ensured that regardless of which colour pair was assigned to which XOR mapping, the initial colour similarity/dissimilarity within colour pair was kept constant. Additionally, in both experiments, the second object had two features: one relevant for reward prediction (shape, **Fig. 1d**) and one irrelevant (width, **Supp. Fig. 5a**) (for the duration and sequence in which stimuli were presented see **Fig. 1e**) The trial sequence was randomised. All trials with fixation errors were excluded. The dataset contained on average 237.9 ($SD = 23.9$) and 104.8 ($SD = 2.3$) trials for each of the 8 conditions in experiment 1, and 101.0 ($SD = 18.6$) and 54 ($SD = 1.1$) trials per each of the 16 conditions in experiment 2, for monkey 1 and monkey 2, respectively.

*Data acquisition and pre-processing***.** Before the start of the experimental protocol, a titanium head holder with two recording chambers was placed and fixed with stainless steel screws in each animal. The frontal recording chambers were implanted over the lateral prefrontal cortex (lPFC) of the right hemisphere in both animals. Data from a second chamber targeting inferotemporal cortex in the right hemisphere are not considered here. A craniotomy was made beneath each chamber to enable

electrophysiological recording. Recording locations for each animal are shown in **Supplementary Figure 8**. Surgical procedures were carried out under general anaesthesia and were aseptic. A semi-chronic micro-drive system (SC-96, Gray Matter Research) with 1.5 mm interelectrode spacing, interfaced to a multichannel data acquisition system (Cerebus System, Blackrock Microsystems) was used for frontal recordings. Data were recorded over a total of 25 daily sessions in each monkey (monkey 1: 17 sessions in experiment 1 and 8 sessions in experiment 2; monkey 2: 10 sessions in experiment 1 and 15 sessions in experiment 2). The switch to experiment 2 was made after the animal showed a robust reward prediction signal. Notably, electrodes were manually advanced by a minimum of $62.5\ \mu m$ before every session to ensure that activity from new cells was recorded. Neural activity was amplified, filtered ($300\ Hz - 10\ kHz$), and stored for offline pre-processing and analysis. Cluster separation was applied (valley seeking algorithm), and the binary spike train was smoothed using a Gaussian window ($\sigma = 50ms$). We collected spiking activity from 146 and 230 neurons in experiment 1 and from 205 and 151 neurons in experiment 2, for monkey 1 and monkey 2, respectively. Only cells sampled from the ventral and dorsal lateral frontal cortex were included in the data (**Supp. Fig. 8**). No neurons were excluded based on their selectivity profiles. Importantly, as the focus of this study was to track how learning influenced neural geometry and not the magnitude of firing (e.g., repetition suppression effects), we z-scored firing rates of each neuron across the whole session. The obtained firing rate data were then epoched from $200\ ms$ before to $1200\ ms$ after the colour onset. Next, the full set of sessions in each animal were divided up into four learning stages and then sessions in each stage were pooled across animals, e.g., the first learning stage was comprised of first 5 sessions from monkey 1 and first 3 sessions from monkey 2. We found that four learning stages were sufficient to observe learning-induced effects. All subsequent analyses were implemented in Python using custom-written code and run on combined data (monkey 1 and 2). Two types of analyses were used in this study: (1) timepoint-resolved, where a specific method was applied to every time point in the epoch to track how representations evolved in trial time, and (2) time-averaged, where a method was run on time-averaged data (e.g., $[\mathrm{t}_{400ms}, \mathrm{t}_{500ms}]$, colour-locked or shape-locked) in the time window preceding the shape display or trial outcome. In the former time window, we examined the neural geometry when only the colour information is known (**Fig. 1e**; colour-locked), whereas just before outcome onset (**Fig. 1e**, shape-locked), we examined whether neural geometry reflected the colour and the shape and their combination (XOR) before the animals received feedback about the value of the trial.

Adaptive trial termination. To assess learning, we measured the proportion of trial terminations through fixation breaking in both rewarded and non-rewarded trials. Specifically, for each session, we counted the number of trial terminations in rewarded and non-rewarded trials when fixation breaking occurred during the shape-locked period (**Fig. 1e**), a phase where all necessary information for outcome prediction is

available but the reward is not yet delivered. These counts were then normalised by dividing by the total number of fixation errors recorded in the session. The adaptive trial termination measure was computed by dividing the normalised non-reward trial count by the normalised rewarded trial count for each session separately. We next divided sessions into four learning stages and fitted a linear regression model with learning stage as the predictor of the adaptive trial termination. To estimate the p-value, we employed a permutation approach, randomising the session-to-learning stage association ($n = 10{,}000$ permutations).

Constantinidis et al. 2016 dataset. We also used an existing dataset of electrophysiological recordings[25] which have been described in detail previously[26,27]. In brief, neural activity was recorded from the ventral and dorsal lateral PFC (similar to the areas targeted in this study) in four rhesus monkeys who performed a feature match-to-sample task. More specifically, the animals were required to report after a delay period whether the shape of the first stimulus was the same as the shape of the second stimulus. Note that a match/no-match rule is equivalent to an XOR rule. Importantly, neural activity was recorded before the animals were exposed to the task rule (passive viewing) and after they had learned the rule. As both correct match and correct no-match trials were rewarded, the match/no-match signal was not confounded with a reward prediction signal. To test whether neural activity was pushed towards a minimal regime in such experimental conditions we employed the same decoding and selectivity measures as used in the analysis of our dataset (see **Fig 2**). We examined neural data averaged across the presentation of the second stimulus and the subsequent delay period ($[\mathrm{t}_{0ms}, \mathrm{t}_{2000ms}]$; stimulus 2-locked). Furthermore, neural activity was analysed for all stimulus pairs combined. For the 8 stimuli, we paired them into 4 sets of pairs and performed our analyses separately on each pair of stimuli (and averaged results over all 4 pairs) so that chance decoding was the same as in our dataset (i.e., 50%).

Models

Multiple linear regression. We can model the firing rate $\boldsymbol{r}$ of a neuron (either from our generative models or our data) at a given time point as a linear combination of the three main task variables: colour, shape and the interaction between colour and shape (the XOR term):

$$\boldsymbol{r} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1}$$

where $\boldsymbol{r}$ is a vector of 1x K dimensionality containing the time-averaged firing rates for K trials; $\mathbf{X}$ is the design matrix of dimensionality $\mathrm{K \, x \, D}$ where rows correspond to the K trials and columns correspond to the value of the D task variables such as colour, shape and XOR ($\mathrm{D} = 3$) in each trial. $\boldsymbol{\beta}$ is a D-by-1 vector populated with the coefficients for each of the task variable estimated for the nth neuron. Finally, $\boldsymbol{\epsilon}$ contains K residuals. The $\boldsymbol{\beta}$ vector specifies the coordinates of the nth neuron in the selectivity space spanned by D task variables (**Fig. 1b,c** and **Fig. 2i,j**). That is, every neuron can be represented as a point in a space where each axis corresponds to the cell's selectivity for a task variable.

Generative models. Neural selectivity can be defined by the matrix $\mathbf{S}_{data} = (\mathrm{S}_{nd})_{1 \leq n \leq N, 1 \leq d \leq D}$, where the nth row corresponds to the nth unit, and columns correspond to the regression coefficients for the three task variables colour, shape, and colour x shape (XOR) that form the axes of the considered space.[5] This cloud of points is then centred by removing the mean ($\sum_n S_{nd} = 0$, for each of the D task variables). Here, we explored two types of selectivity distributions and their representational properties. Firstly, we examined a random mixed selectivity model in which selectivities are captured by a spherical multivariate Gaussian distribution $\mathbf{S}_{random} \sim \mathcal{N}_d(\mathbf{0}, \sigma^2 \mathbf{I}_{\boldsymbol{d}})$. In such a model, all variables can be decoded equally well from the population resulting in a high-dimensional representation and there is poor cross-generalisation between variables. The second selectivity model we examine results from a system performing the XOR task while being constrained to exhibit low overall firing rates (i.e., a form of metabolic cost). We derived analytically that maximising XOR decodability while minimising such a metabolic cost results in units being selective only to the interaction term (colour x shape; XOR) and having no selectivity to the linear terms (colour or shape; **Supp. Materials Section 1**). A matrix describing the selectivity of such a population can be thus formulated as $\mathbf{S}_{minimal} \sim \mathcal{N}_d\big(\mathbf{0}, \sigma^2 \mathrm{diag}(0,0,1)\big)$, where the covariance matrix is an diagonal matrix with two first diagonal terms equal to zero and the third equal to one. We call this the minimal XOR model. Importantly, to allow comparisons between the observed selectivity and model selectivity (minimal or random), the generative models were constructed using parameters derived from the data. Specifically, we used the mean value of diagonal entries of the covariance matrix $\widetilde{\mathbf{\Sigma}}_{data}$ estimated from $\mathbf{S}_{data}$ to set the value of the variance parameter σ^2 in both $\mathbf{S}_{random}$ and $\mathbf{S}_{minimal}$ (note that this ensures that both models have the same total variance). Furthermore, we showed that multiple linear regression was able to recover the underlying minimal and random models from artificially generated firing rates under various levels of noise (**Supp. Fig. 11**).

Optimised feedforward networks. We used $N = 400$ units in these networks and their firing rates were described by eq. 1 with σ = 2. The output $\boldsymbol{z}$ of these networks was given by a softmax readout

$$\boldsymbol{z} = \mathrm{Softmax}(\mathbf{W}_{\mathrm{out}}\boldsymbol{r} + \boldsymbol{b}), \tag{2}$$

where $\mathbf{W}_{\mathrm{out}}$ are the two sets of readout weights (connecting the hidden layer to the readout unit 1 (XOR == 0) and weights connecting the hidden layer to readout unit 2 (XOR == 1)) and $\boldsymbol{b}$ is the readout bias. We optimized these networks with back-propagation using a canonical cross-entropy cost function

$$\mathcal{L} = \mathcal{H}(\boldsymbol{p}, \boldsymbol{z}) + \frac{\lambda}{2N}\|\boldsymbol{r}\|_2^2, \tag{3}$$

where the first part of eq. 3 denotes the cross-entropy loss $\mathcal{H}(\boldsymbol{p}, \boldsymbol{z})$ between the true probabilities of reward $\boldsymbol{p}$ (which were equal to 0 or 1, depending upon the stimuli for that trial) and the model's readout probabilities $\boldsymbol{z}$ and the second term corresponds to a metabolic cost on all firing rates. Before training, the values of the $\boldsymbol{\beta}$'s were drawn randomly from a Gaussian distribution with 0 mean and variance 0.05, and elements of $\mathbf{W}_{\mathrm{out}}$ and $\boldsymbol{b}$ were set to 0. We trained two types of models: (1) with no regularisation ($\lambda = 0$) and (2) with high regularisation ($\lambda = 20$; **Supp. Fig. 3a** and **1b**, respectively). In line with our predictions, selectivity to task variables in models trained with high regularisation converged on the minimal selectivity regime whereas models trained with no regularisation produced randomly mixed selectivity (**Supp. Fig. 3c,d**).

Analysis methods

Decoding. To test what information was represented by the observed neural population as a function of learning, we employed linear SVM decoding[7,14]. In contrast to the regression analysis, where the regression coefficients was estimated for every neuron separately, decoding analyses were run on pseudo-populations. This approach enhanced statistical power, reducing the likelihood of Type II errors, and mitigated the impact of session-specific sampling bias. Given the varying number of trials per session (both within and between animals), a sliding window method was employed to utilise all available data. Specifically, each session was divided into three windows, each matching the size of the session with the fewest trials ($n = 801$). Neural activities ($trials\ x\ neuron\ x\ time$) for the first 801 trials in each session during learning stage 1 were combined along the neuron axis to form a pseudo-population. This procedure was repeated for the middle and final 801 trials across each learning stage. This resulted in a matrix $\mathbf{X} = (\mathrm{X}_{knt})_{1\leq k\leq K, 1\leq n\leq N, 1\leq t\leq T}$ for each of the time windows and each of the four learning stages, where the first dimension corresponds to K trials, the second dimension corresponds to N neurons (combined from two animals), and the last dimension corresponds to T time points. Then binary SVM classifiers were used to decode the task variables (colour, shape, width and XOR in experiment 1 and

context, stimulus set, shape, width and XOR in experiment 2) at each time point for temporally resolved decoding (**Fig. 1b-d**; **Supp. Fig. 5b,j-m** and **Supp. Fig. 7a-e**) and from time-averaged firing rates ($[t_{-100ms}, t_{0ms}]$; colour-locked, **Fig. 3b,e,** outcome-locked, **Fig. 4a,e,j**). An equivalent decoding procedure was used when analysing generative models (**Supp. Fig. 2d**). Decoding was performed in a cross-validated way where K trials were split randomly into set 1 and set 2, with each containing 50% of trials. The decoder was fitted using the set 1 and tested on set 2. The procedure was then repeated using set 2 as the training set and set 1 as the test set. Both decoding scores were then averaged. This procedure was repeated 10 times for different random splits of trials in sets 1 and 2, and these 10 resulting scores were then averaged. The decoding results were averaged over three time windows.

Shattering dimensionality. To estimate shattering dimensionality[7,14] in experiment 1(**Fig. 2e**) and 2 (**Supp. Fig. 7q**), we used the same decoding approach as described above except that we averaged decoding scores over all 35 possible dichotomies that could be theoretically represented given the task structure (i.e., three linear variables form a cube in state space that can be dissected into two sets of 4 vertices in 35 possible ways; see **Supp Fig. 10** for illustration). We estimated the decodability of each of these dichotomies and tracked their mean decoding accuracy as a function of learning. To assess learning-induced changes to the linear and nonlinear components of neural dimensionality separately, we split the dichotomies into linear ($n = 7$) and nonlinear decoding axes ($n = 28$) and tracked their mean decoding accuracy as a function of learning (**Supp. Fig. 5f**). Exploring the theoretically maximal dimensionality of the neural representations in experiment 2 using shattered dimensionality, where four linear variables (colour, stimulus set, shape, width) were used, resulted in the identification of 6435 theoretically possible binary decoding problems. We tracked the mean of these dimensions over time (**Supp. Fig. 10q**). Due to the computational expense of analysing a high number of dimensions, the statistical comparison between stage 1 and stage 4 was conducted on time-averaged activities.

Cross-generalised decoding. To examine the neural geometry of the task variables we used cross-generalised SVM decoding[35]. In contrast to a typical cross-validation procedure, the testing happens not only on trials that were previously not seen but also on trials that correspond to different conditions. To achieve a high cross-generalisation score, it is therefore not sufficient to generalise across trial-wise noise but also to generalise across conditions[7]. Specifically, the labels of eight unique conditions ($2\ colours\ x\ 2\ shapes\ x\ 2\ widths$; see **Supplementary Fig. 11a**) were split into two sets of four labels each, depending on the tested variable (e.g., colour 1 labels vs colour 2 labels; see **Supplementary Fig. 11b**). Each subset was further divided into a training set and a testing set (colour 1 vs colour 2 training set and colour 1 vs colour 2 testing set). A decoder was trained on the training set and then tested on the testing set, and vice versa; the two scores were then averaged. We identified 36

possible train-test splits (see **Supplementary Fig. 11c**), and the cross-generalised decoding score was obtained by averaging these scores. This method was used to determine whether the format of a task variable is abstract, meaning the variable is encoded in the same format as a function of the remaining task variables. Note that some of the train-test splits correspond to decoding a task variable as a function of a single other variable (e.g., decoding colour as a function of shape; see **Supplementary Fig. 11c**, shaded geometries), while others examine the decoding of the variable as a function of a mix of variables.

Cross-stimulus set generalisation (decoding and selectivity analyses). Cross-generalised decoding performed for experiment 2 data (both run on time-averaged firing rates and temporally resolved; **Fig. 3b,e**, **Fig. 4a,e,i** and **Supp. Fig. 10f,h-j**, respectively) differed in one aspect from the algorithm described in the *Cross-generalised decoding* section. As the aim of the analysis described here was to identify the neural format of the main task variables used across stimulus sets, only one splitting variable was used (i.e., stimulus set) to obtain cross-generalisation scores for the task-relevant variables (context, shape, and XOR). This reduced the possible cross-generalisation decoding axes to four possible binary decoding problems (e.g., when performing cross-generalised decoding for the colour variable we can: (1) train on differentiating colour 1 from colour 2 in stimulus set 1 and test on differentiating colour 3 from colour 4 in stimulus set 2, (2) train on differentiating colour 3 from colour 4 in stimulus set 2 and test on differentiating colour 1 from colour 2 in stimulus set 1, (3) train on differentiating colour 1 from colour 3 and test on differentiating colour 2 from colour 4, and (4) train on differentiating colour 2 from colour 4 and test on differentiating colour 1 from colour 3; these four decoding scores were then averaged). Using this procedure, we explored the cross-stimulus set generalisation potential of the colour, shape, width and XOR variables. Additionally, to test how selectivity of PFC cells changed as a function of learning in experiment 2 we employed a Pearson correlation metric. Specifically, we compared how similar the colour, shape and XOR coefficients in stimulus set 1 are to coefficients for the same variables in stimulus set 2 (**Fig. 3c,d,f,g** and **Fig. 4b,c,f,g,j,k**), which yielded three correlation scores for each of the main task variables (**Supp. Fig. 10k-o**). This was done for each of the four learning stages to explore whether selectivity for stimulus set 1 aligns with selectivity for stimulus set 2 as a function of learning, consistent with a shared abstract representation.

Measuring similarity between selectivity distributions. To test the observed neural population for the presence of random mixed or minimal selectivity (**Fig. 1b,c,** and **Fig. 2i,j**), we firstly obtained regression coefficients for the three variables of interest (colour, shape and XOR; eq. 1) and constructed the selectivity space $\mathbf{S}_{data}$. To assess the similarity of $\mathbf{S}_{data}$ to $\mathbf{S}_{minimal}$ and $\mathbf{S}_{random}$, we computed the covariance matrix of $\mathbf{S}_{data}$ ($\widetilde{\mathbf{\Sigma}}_{data}$) as well as the covariance matrices of the expected random and minimal

distributions given $\mathbf{S}_{data}$ ($\widetilde{\mathbf{\Sigma}}_{minimal}$ and $\widetilde{\mathbf{\Sigma}}_{random}$; see *Generative models)*. Finally, we calculated the normalised distance of the observed selectivity from model random selectivity

$$d_{i_{from\,random}}(\widetilde{\mathbf{\Sigma}}) = \frac{\|\widetilde{\mathbf{\Sigma}}-\widetilde{\mathbf{\Sigma}}_{i_{random}}\|_2 - \mathbb{E}(\|\widetilde{\mathbf{\Sigma}}_{i_{random}}-\widetilde{\mathbf{\Sigma}}_{i_{random}}\|_2)}{\mathbb{E}(\|\widetilde{\mathbf{\Sigma}}_{i_{random}}-\widetilde{\mathbf{\Sigma}}_{i_{minimal}}\|_2) - \mathbb{E}(\|\widetilde{\mathbf{\Sigma}}_{i_{random}}-\widetilde{\mathbf{\Sigma}}_{i_{random}}\|_2)}, \tag{4}$$

and the normalised distance of the observed selectivity to minimal selectivity

$$d_{i_{from\,minimal}}(\widetilde{\mathbf{\Sigma}}) = \frac{\|\widetilde{\mathbf{\Sigma}}-\widetilde{\mathbf{\Sigma}}_{i_{minimal}}\|_2 - \mathbb{E}(\|\widetilde{\mathbf{\Sigma}}_{i_{minimal}}-\widetilde{\mathbf{\Sigma}}_{i_{minimal}}\|_2)}{\mathbb{E}(\|\widetilde{\mathbf{\Sigma}}_{i_{random}}-\widetilde{\mathbf{\Sigma}}_{i_{minimal}}\|_2) - \mathbb{E}(\|\widetilde{\mathbf{\Sigma}}_{i_{minimal}}-\widetilde{\mathbf{\Sigma}}_{i_{minimal}}\|_2)}. \tag{5}$$

where the subscript i denotes a random draw and the expectations were computed over 1000 draws. From both the denominators and numerators, the distance within each of the models was subtracted to centre the measure around 0. More specifically, $\mathbb{E}(\|\widetilde{\mathbf{\Sigma}}_{i_{random}} - \widetilde{\mathbf{\Sigma}}_{i_{random}}\|_2)$(the expected difference between two different randomly drawn selectivity distributions from the random model) was, for example, subtracted from the denominator and numerator of $d_{from\,random}$ to account for within model distance. Additionally, both $d_{from\,random}$ and $d_{from\,minimal}$ were normalised by the distance between selectivities generated using both generative models ($\mathbb{E}(\|\widetilde{\mathbf{\Sigma}}_{i_{random}} - \widetilde{\mathbf{\Sigma}}_{i_{minimal}}\|_2)$) which resulted in the metrics being bounded between 0 and 1 (when $\|\widetilde{\mathbf{\Sigma}}_{data} - \widetilde{\mathbf{\Sigma}}_{minimal}\|_2$ is equal or smaller than $\|\widetilde{\mathbf{\Sigma}}_{random} - \widetilde{\mathbf{\Sigma}}_{minimal}\|_2$). This was done to allow for a comparison of similarity estimates across learning stages. The Euclidean distance metric was chosen as the main analysis tool in this study based on simulations in which we generated different proportions of random and minimal selectivity across a single population (from 0% minimal and 100% of random to 100% minimal and 0% random) and compared the precision with which multiple metrics recovered the true proportions. We compared the Euclidean distance metric to the PAIRs metric, which has been used previously in the literature[12,28], and to the symmetric Kullback–Leibler divergence estimate (KL divergence) which benefits from a strong theoretical basis and is assumption-agnostic. We found that, compared to the KL divergence and PAIRs metrics, the Euclidean distance measure yielded the highest precision of tracking learning-induced changes to neural selectivity (**Supp. Fig. 9a,b**). Specifically, our simulations showed that both the KL divergence and PAIRs can be used to precisely identify extreme selectivity regimes (either strong random selectivity or strong minimal selectivity) but fail at identifying intermediate selectivity regimes showing a strong bias towards random mixed selectivity (**Supp. Fig. 9a,b**). As the focus of this study was to track learning dynamics, a metric that allows to identify a broad range of selectivity regimes was chosen for the final analysis. Nonetheless, the results from experiment 1 (**Fig. 1k,l**) were broadly replicated using the symmetric KL divergence estimate (**Supp. Fig. 9c,d**) and PAIRs (**Supp. Fig. 9e,f**).

Principal Component Analysis. PCA was used as a measure of neural dimensionality in experiment 1(**Supp. Fig. 5e**). Firstly, pseudo populations were constructed for each learning stage using the same procedure as described in the *Decoding* section. Then, firing rates were averaged in the time window preceding the outcome presentation ($[\mathrm{t}_{400ms}, \mathrm{t}_{500ms}]$, shape-locked;). Next, principal components were run on condition averages. This was done separately for each learning stage. To compute how the variance explained (ratio) by the first PC changed as a function of learning, trials were randomly split into test and train 10 times; PCA was fitted then on train trials and the test trial firing rates were projected onto them to compute variance explained. The results from 10 random splits were then averaged. Note that width 1 and width 2 trials were pooled together. The null distribution for the permutation test was computed by randomly shuffling neurons between stage 1 and stage 4, and repeating the described PCA procedure ($n = 500$).

Statistical testing

Decoding and cross. gen. decoding. Throughout the study, we employed non-parametric permutation tests to test statistical significance within each learning stage and between learning stages (learning-induced effects). Two types of null distributions were thus constructed: (1) for statistical testing in time-resolved and time-specific decoding analyses the labels describing the trial dimension (k) of the pseudo-population matrix $\mathbf{X} = (\mathrm{X}_{knt})_{1\leq k\leq K, 1\leq n\leq N, 1\leq t\leq T}$ were randomly permuted 500 times; (2) to test for learning-induced effects, the matrices $\mathbf{X}_{stage\ 1}$ and $\mathbf{X}_{stage\ 4}$ were concatenated along the neuron dimension (n) and then 500 new $\mathbf{X}'_{stage\ 1}$ and $\mathbf{X}'_{stage\ 4}$ matrices were generated by randomly assigning neurons to either $\mathbf{X}'_{stage\ 1}$ or $\mathbf{X}'_{stage\ 4}$. One-sided tests were used when testing the predictions of the minimal model and two-sided tests were used when no differences were expected. Additionally, to control for time-related family-wise errors a cluster-based permutation correction was added to the time-resolved decoding[52].

Selectivity measures. To test whether observed selectivity was dissimilar to the random selectivity regime and similar to minimal selectivity regime 1000 random and minimal models were generated using data-derived parameters for each learning stage. Next, the $d_{rand.from\ rand.}$ and $d_{rand.from\ min.}$ distances were computed for 1000 randomly generated models according to eq. 4 and eq. 5 (with $\widetilde{\mathbf{\Sigma}}_{random}$ as input) to serve as null distributions for both comparisons. Note that the observed selectivity was compared to random model selectivity when analysing the data's similarity to random (**Fig. 2k**) as well as minimal selectivity (**Fig. 2l**). Furthermore, as in experiment 2 we tested whether selectivity for task variables was similar in stimulus set 1 to variables

in stimulus set 2 and whether this selectivity alignment changed over learning, two null distributions were thus constructed: (1) statistically significant selectivity alignment was assessed by comparing the observed correlation to a distribution ($n = 500$) of correlations obtained after shuffling one of the selectivity vectors; (2) learning-induced effects in selectivity alignment were assessed by comparing the observed difference in alignment between stage 1 and stage 4 to a distribution of differences computed after randomly shuffling neurons between stage 1 and 4.

Learning shapes neural geometry in the prefrontal cortex

Supplementary materials

Section 1.1: How to set selectivity parameters for optimal XOR decoding

We assume that neural activities $\boldsymbol{x}$ of $\boldsymbol{N}$ neurons are given by the following regression model

$$\boldsymbol{x} = \boldsymbol{\beta}_0 + \delta_{c,c_1}\boldsymbol{\beta}_C + \delta_{s,s_1}\boldsymbol{\beta}_S + \delta_{c,c_1}\delta_{s,s_1}\boldsymbol{\beta}_I + \boldsymbol{\eta} \tag{1}$$

Where $\boldsymbol{\beta}_C, \boldsymbol{\beta}_S, \boldsymbol{\beta}_I$ are the regression 'coefficients' for colour, shape, and the interaction term, respectively, $\delta_{c,c_1} = 1$ if the colour c is colour 1 and -1 otherwise (same for δ_{s,s_1} for shapes s), and $\boldsymbol{\eta} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_1)$. Therefore,

$$\boldsymbol{x}|r_1 \sim \mathcal{N}(\boldsymbol{\mu}_{r_1}, \boldsymbol{\Sigma}_1) \tag{2}$$

$$\boldsymbol{x}|r_2 \sim \mathcal{N}(\boldsymbol{\mu}_{r_2}, \boldsymbol{\Sigma}_2). \tag{3}$$

Where r_1 is one XOR condition (i.e., $c = 1, s = 1$ or $c = 2, s = 2$) and r_2 is the other XOR condition (i.e., $c = 1, s = 2$ or $c = 2, s = 1$) and $\boldsymbol{\Sigma}_2$ is some noise covariance matrix. Note that with the inclusion of the interaction term, it is sufficient to separate the two XOR conditions. We now calculate $\boldsymbol{\mu}_{r_1}$ and $\boldsymbol{\mu}_{r_2}$:

$$\begin{aligned}\boldsymbol{\mu}_{r_2} &= \frac{\boldsymbol{\beta}_0 + \boldsymbol{\beta}_C + \boldsymbol{\beta}_S + \boldsymbol{\beta}_I}{2} + \frac{\boldsymbol{\beta}_0 - \boldsymbol{\beta}_C - \boldsymbol{\beta}_S + \boldsymbol{\beta}_I}{2} \\ &= \boldsymbol{\beta}_0 + \boldsymbol{\beta}_I\end{aligned} \tag{4}$$

and,

$$\begin{aligned}\boldsymbol{\mu}_{r_2} &= \frac{\boldsymbol{\beta}_0 - \boldsymbol{\beta}_C + \boldsymbol{\beta}_S - \boldsymbol{\beta}_I}{2} + \frac{\boldsymbol{\beta}_0 + \boldsymbol{\beta}_C - \boldsymbol{\beta}_S - \boldsymbol{\beta}_I}{2} \\ &= \boldsymbol{\beta}_0 - \boldsymbol{\beta}_I\end{aligned} \tag{5}$$

Therefore,

$$\boldsymbol{x}|r_1 \sim \mathcal{N}(\boldsymbol{\beta}_0 + \boldsymbol{\beta}_I, \boldsymbol{\Sigma}_2) \tag{6}$$

$$\boldsymbol{x}|r_2 \sim \mathcal{N}(\boldsymbol{\beta}_0 - \boldsymbol{\beta}_I, \boldsymbol{\Sigma}_2). \tag{7}$$

Therefore, we need $\boldsymbol{\beta}_I > 0$ to be able to separate the two XOR conditions.

Section 1.2: An energy cost on unnecessary neural activity

If we also consider minimising the squared norm of neural activity for each condition (i.e., an energy cost), we have

$$\mathbb{E}\left[\left\|\boldsymbol{x}_{c,s}\right\|^2\right] = \mathbb{E}\left[\boldsymbol{x}_{c,s}\right]^{\mathrm{T}} \mathbb{E}\left[\boldsymbol{x}_{c,s}\right] + \mathrm{Tr}(\boldsymbol{\Sigma}_1) \tag{8}$$

where the subscripts c and s correspond to colour and shape indices, respectively. Therefore,

$$\mathbb{E}\left[\left\|\boldsymbol{x}_{1,2}\right\|^2\right] = (\boldsymbol{\beta}_0 - \boldsymbol{\beta}_C + \boldsymbol{\beta}_S - \boldsymbol{\beta}_I)^{\mathrm{T}} (\boldsymbol{\beta}_0 - \boldsymbol{\beta}_C + \boldsymbol{\beta}_S - \boldsymbol{\beta}_I) + \mathrm{Tr}(\boldsymbol{\Sigma}_1) \tag{9}$$

$$\mathbb{E}\left[\left\|\boldsymbol{x}_{2,1}\right\|^2\right] = (\boldsymbol{\beta}_0 + \boldsymbol{\beta}_C - \boldsymbol{\beta}_S - \boldsymbol{\beta}_I)^{\mathrm{T}} (\boldsymbol{\beta}_0 + \boldsymbol{\beta}_C - \boldsymbol{\beta}_S - \boldsymbol{\beta}_I) + \mathrm{Tr}(\boldsymbol{\Sigma}_1) \tag{10}$$

$$\mathbb{E}\left[\left\|\boldsymbol{x}_{1,1}\right\|^2\right] = (\boldsymbol{\beta}_0 - \boldsymbol{\beta}_C - \boldsymbol{\beta}_S + \boldsymbol{\beta}_I)^{\mathrm{T}} (\boldsymbol{\beta}_0 - \boldsymbol{\beta}_C - \boldsymbol{\beta}_S + \boldsymbol{\beta}_I) + \mathrm{Tr}(\boldsymbol{\Sigma}_1) \tag{11}$$

$$\mathbb{E}\left[\left\|\boldsymbol{x}_{2,2}\right\|^2\right] = (\boldsymbol{\beta}_0 + \boldsymbol{\beta}_C + \boldsymbol{\beta}_S + \boldsymbol{\beta}_I)^{\mathrm{T}} (\boldsymbol{\beta}_0 + \boldsymbol{\beta}_C + \boldsymbol{\beta}_S + \boldsymbol{\beta}_I) + \mathrm{Tr}(\boldsymbol{\Sigma}_1) \tag{12}$$

Therefore, the total mean energy cost m is given by

$$m = \frac{1}{4} \sum_{c=1,2} \sum_{s=1,2} \mathbb{E}\left[\left\|\boldsymbol{x}_{c,s}\right\|^2\right] \tag{13}$$

$$= \boldsymbol{\beta}_0{}^{\mathrm{T}}\boldsymbol{\beta}_0 + \boldsymbol{\beta}_C{}^{\mathrm{T}}\boldsymbol{\beta}_C + \boldsymbol{\beta}_S{}^{\mathrm{T}}\boldsymbol{\beta}_S + \boldsymbol{\beta}_I{}^{\mathrm{T}}\boldsymbol{\beta}_I + \mathrm{Tr}(\boldsymbol{\Sigma}_1) \tag{14}$$

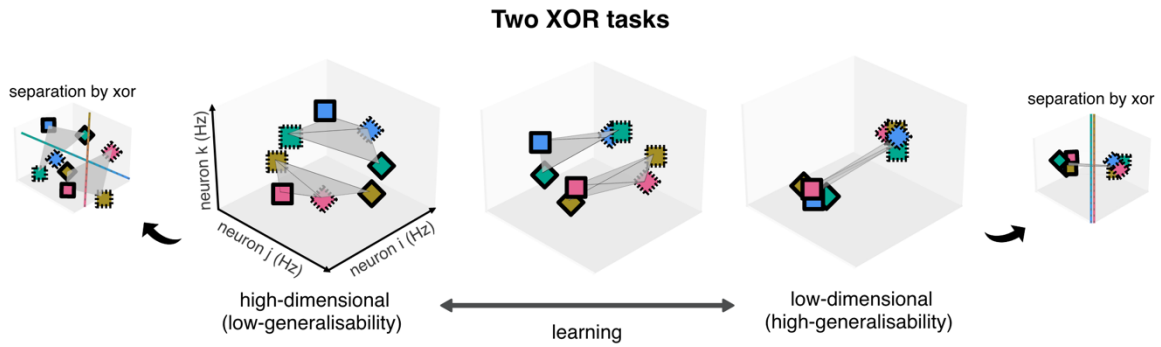
$$= \|\boldsymbol{\beta}_0\|^2 + \|\boldsymbol{\beta}_C\|^2 + \|\boldsymbol{\beta}_S\|^2 + \|\boldsymbol{\beta}_I\|^2 + \mathrm{Tr}(\boldsymbol{\Sigma}_1) \tag{15}$$

To minimise m while keeping $\boldsymbol{\beta}_I > 0$, which we need for performance, we can set $\boldsymbol{\beta}_0 = \boldsymbol{\beta}_C = \boldsymbol{\beta}_S = \mathbf{0}$ which gives

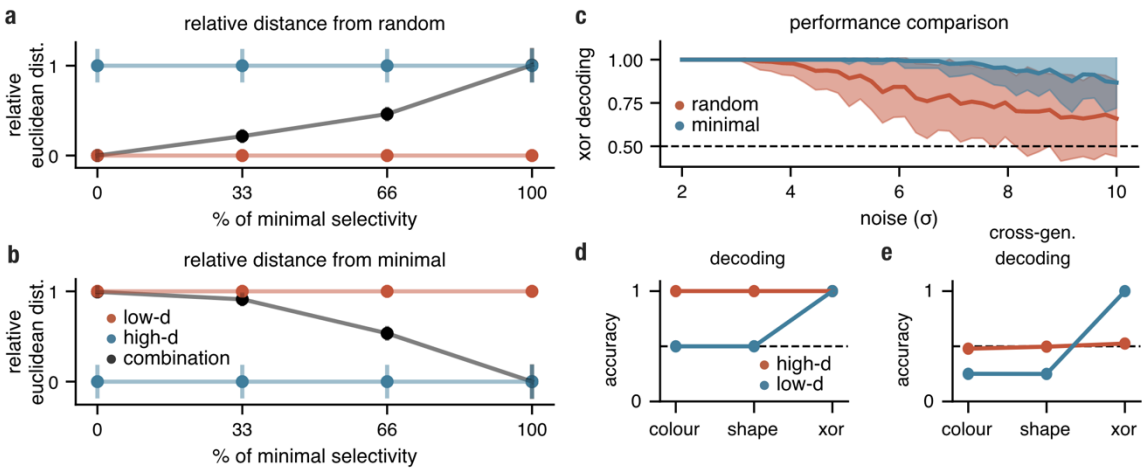
$$m = \|\boldsymbol{\beta}_I\|^2 + \mathrm{Tr}(\boldsymbol{\Sigma}_1)$$

(16)

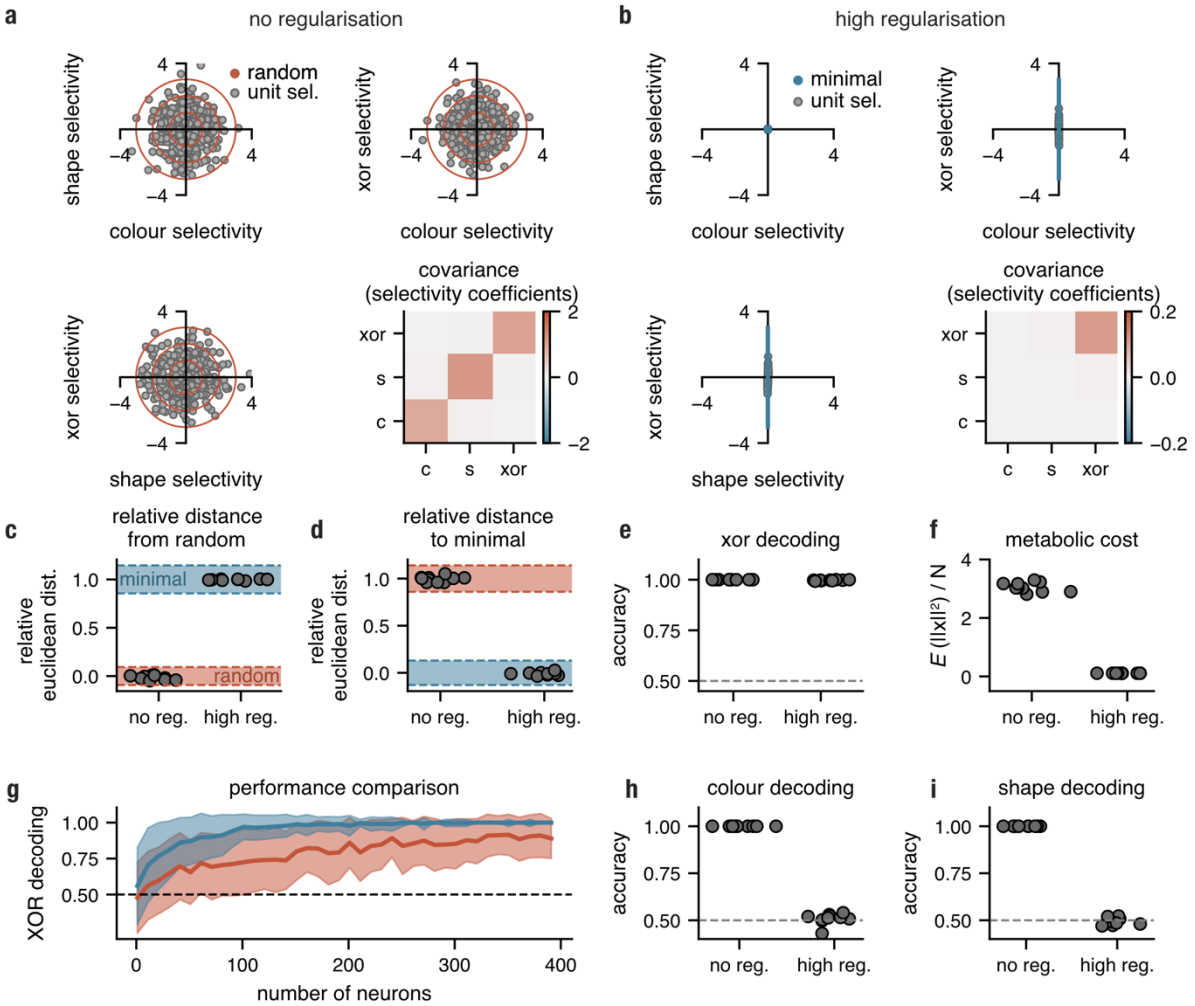
Section 2: supplementary figures



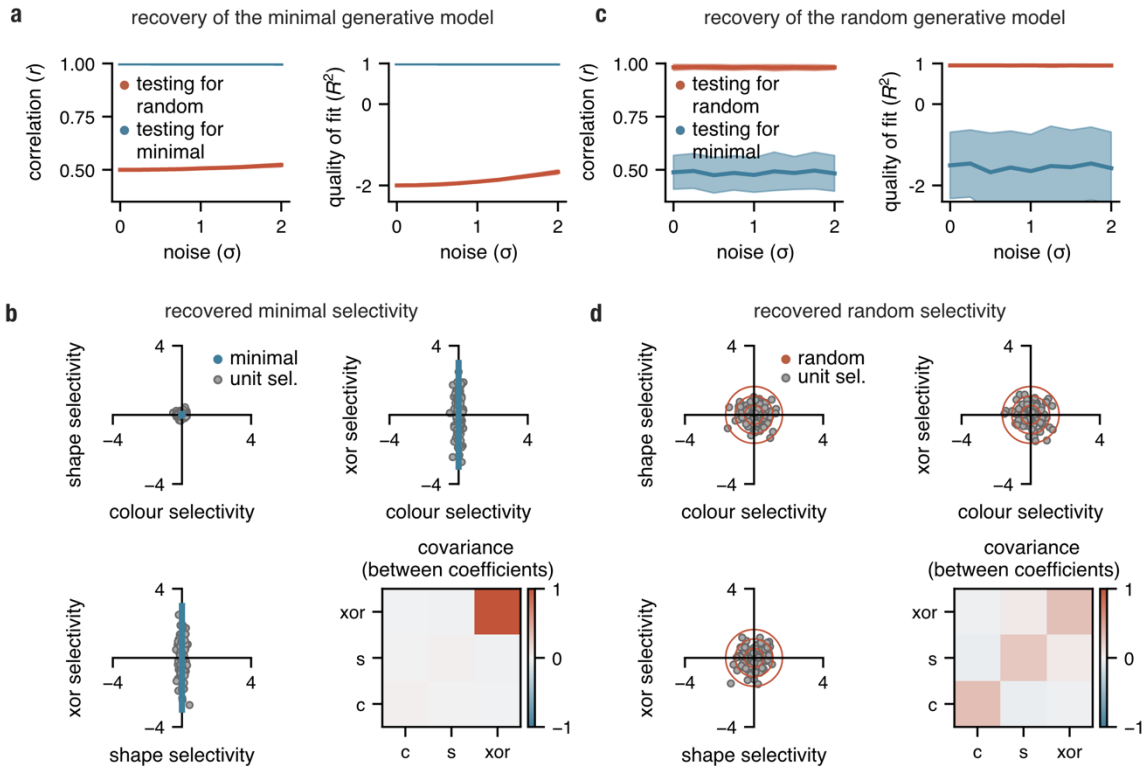
Supplementary figure 1. Different solutions to the discriminability-generalisability trade-off when a new stimulus set is being learnt. Low-dimensional representations enable high generalisability. When the task representation is high-dimensional (left), it is not trivial that the XOR discriminant from one task (blue and green shapes) would correctly differentiate the XOR feature in the second task (pink and khaki); i.e., aligning both the new axis to the old one is constrained if high discriminability is to be maintained. When the neural code is low-dimensional, both tasks can be easily aligned to a common axis enabling a shared XOR discriminant that generalises across tasks.



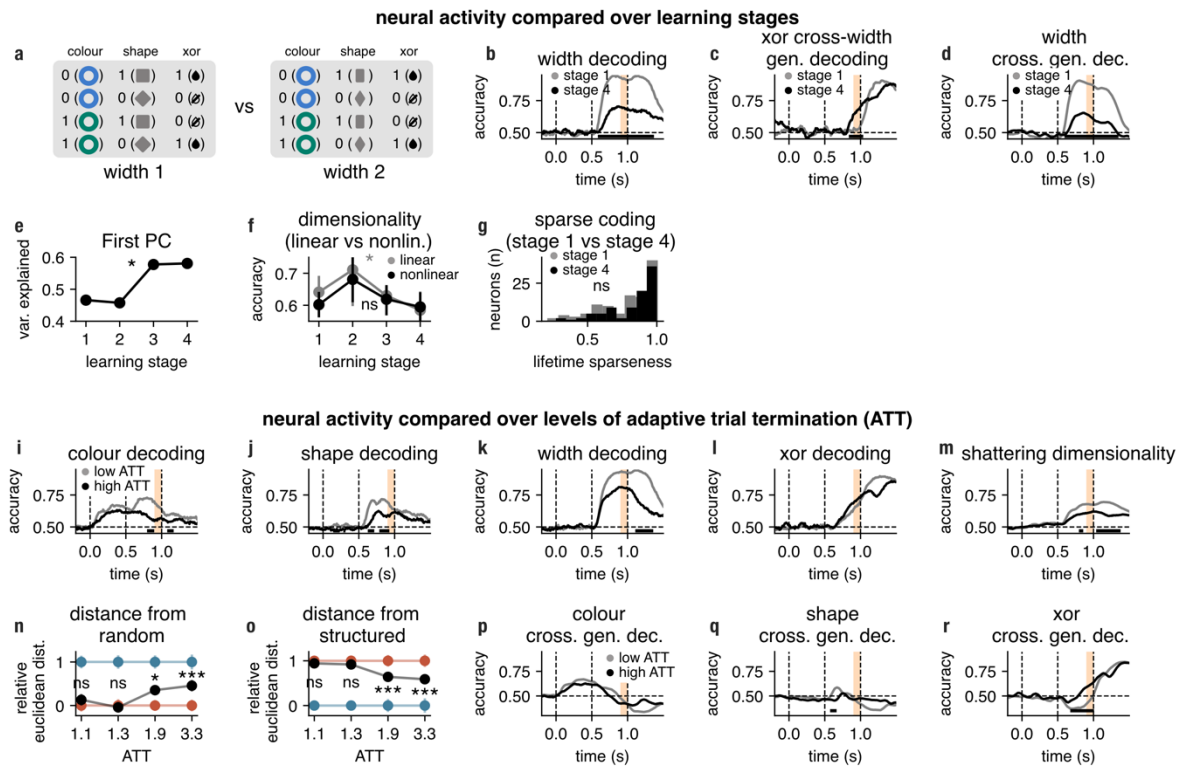
Supplementary figure 2. Predictions from random mixed and minimal selectivity models. **a**, Relative distance between the covariance matrices of either the random (red), minimal (blue) or the model with varying proportions of minimal selectivity (black) to the random selectivity model. **b**, Same as panel **c** but the distance is calculated relative to the minimal selectivity model. Red and blue error bars show standard deviations ($\pm 1\ s.d.$ over 1000 randomly drawn models; see *Methods*, *measuring similarity between selectivity distributions*) of the relative Euclidean distance between the covariance matrix of the random, minimal model (with matched total variance to the data) and the covariance matrix expected from random selectivity; black error bars show the standard deviation of the relative distance between the surrogate covariance (with varying proportions of minimal selectivity) and random covariance ($\pm 1\ s.d.$ over 1000 random models). **c,** XOR decoding as a function of noise (σ) in the neural activities for the random and minimal selectivity models. Dashed grey line shows chance-level decoding. **d**, Mean (over 100 models) decoding of task variables for the random (red) and minimal (blue) models. Dashed grey line shows chance-level decoding. **e**, Mean (over 100 models) cross-generalised decoding of task variables for the random (red) and minimal (blue) models. Dashed grey line shows chance-level decoding.



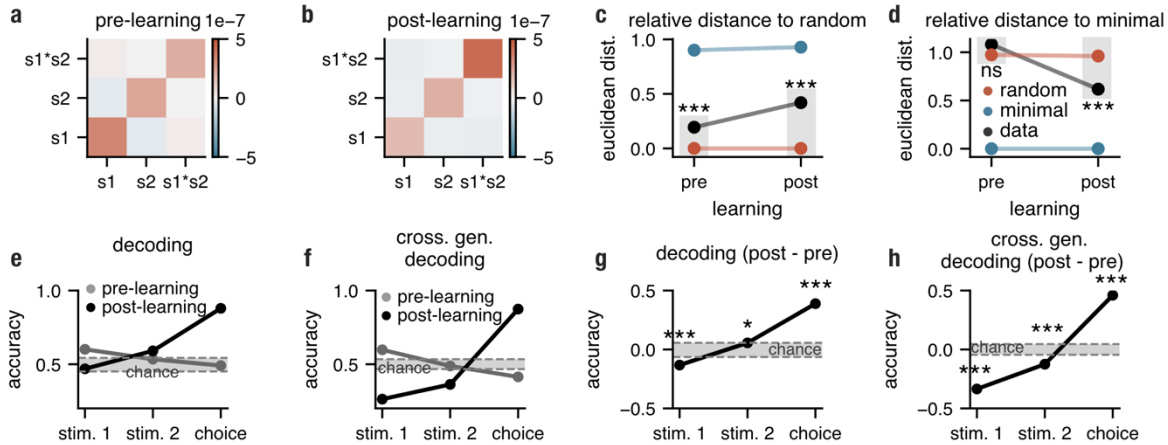
Supplementary figure 3. Optimised feedforward networks converge to the minimal XOR selectivity model. Twenty feedforward networks were trained (10 with high levels of regularisation and 10 with no regularisation) to perform the XOR task. **a, b,** Selectivity observed in no and high regularisation networks, respectively; models of random (red ellipses) and minimal selectivity (blue ellipses) well approximated the observed selectivity. **c, d**, After training, low regularisation networks converged on a random mixed selectivity regime and high regularisation networks on a minimal XOR regime. **e**, post-training XOR decoding (linear SVM) for both no and high regularisation models. **f**, No regularisation models exhibited substantially lower metabolic cost (cf. **Supp. materials** eq. 8). **g**, Comparison of XOR decoding obtained from minimal and random generative models as a function of population size. **h**, **i**, Colour and shape decoding (linear SVM) for no and high regularisation models, respectively.



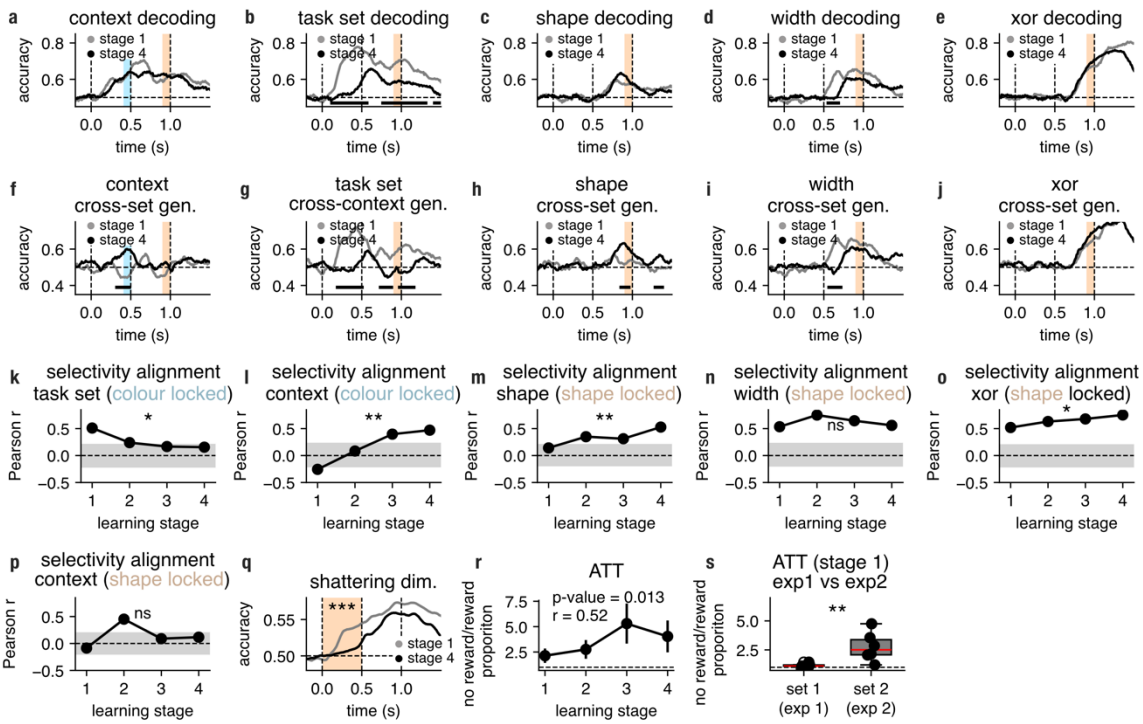
Supplementary figure 4. Linear regression recovers underlying generative models. a, Pearson correlation and R^2 computed between the covariance matrix of recovered selectivity and true underlying selectivity (when minimal generative model was used) for different levels of noise. **b**, Selectivity coefficients obtained after running a linear regression plotted for each unit in selectivity space (for $\sigma = 2$); minimal model overlaid in blue; covariance matrix computed between the recovered selectivity coefficients. **c**,**d**, Analogous to a,b but when the random model was used to generate data; random model overlaid in red. Shaded areas in a and c indicate the mean $\pm 1\ s.d.$ computed over 100 different initialisations.



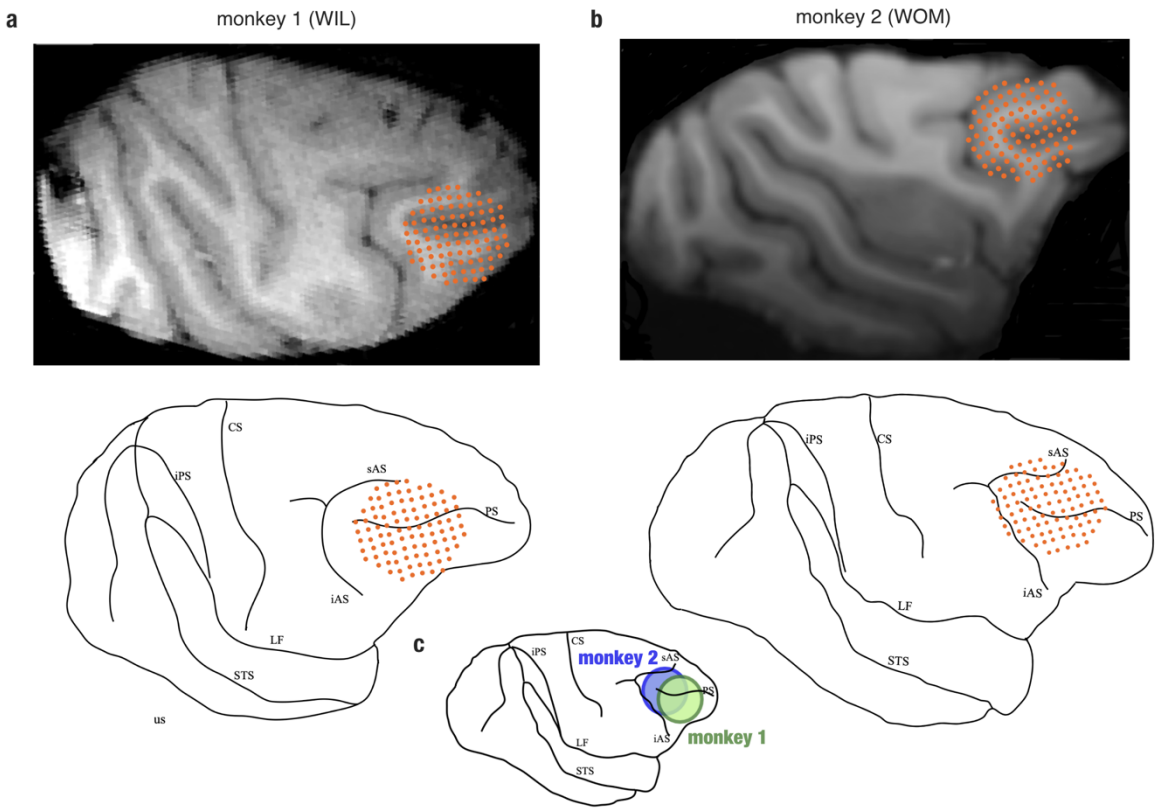
Supplementary figure 5. Decoding of task variables as a function of learning in experiment 1. **a**, schematic of narrow and broad shape trials. This feature was not predictive of reward. **b**, Temporally resolved linear SVM decoding of width; horizontal bars indicate statistical significance; dashed black bars indicate significant differences in coding between stage 1 and stage 4; dashed grey bars indicate trend-level significance in stage 1 vs stage 4 decoding. The pale orange area indicates the time window for which all subsequent decoding analyses were run. Vertical three dashed lines show the onset of the colour, shape and the outcome, respectively. **c**, Temporally resolved cross-width generalised decoding of XOR in stage 1 and stage 4. A SVM decoder was train to classify XOR in broad shape trials and tested in narrow shape trials, and vice versa. Above chance decoding indicates that the same XOR representation was used in both type of trials. **d**, Time resolved cross-generalised linear decoding of width for learning stages 1 (grey) and 4 (black). **e**, variance explained (ratio) by the first principal component plotted as a function of learning (see *Methods, principal component analysis* for details). **f**, Mean $\pm 1\ s.d.$ linear decoding of all linear dichotomies (grey line) and non-linear dichotomies (black line). **g**, Distributions of the lifetime sparseness index observed in stage 1 (grey) and stage 4 (black). A Kolmogorov–Smirnov test revealed no differences between these distributions, $D = 0.15$, $p = .18$. **h**, Session-wise correlation between mean XOR decoding in the pre-reward period (shaded orange areas in previous panels) and the trial termination measure. **i**, Mean and normalised norm of firing rates plotted as a function of time for stage 1 and stage 4; horizontal bars indicate a statistically significant difference between these stages. **j-s,** Analogues to Figure 2 but run on sessions sorted using adaptive trial termination. All p-values were calculated from permutation tests (***, $p < 0.01$; **, $p < 0.01$; *, $p < 0.05$; †, $p < 0.01$; $n.s.$, not significant).



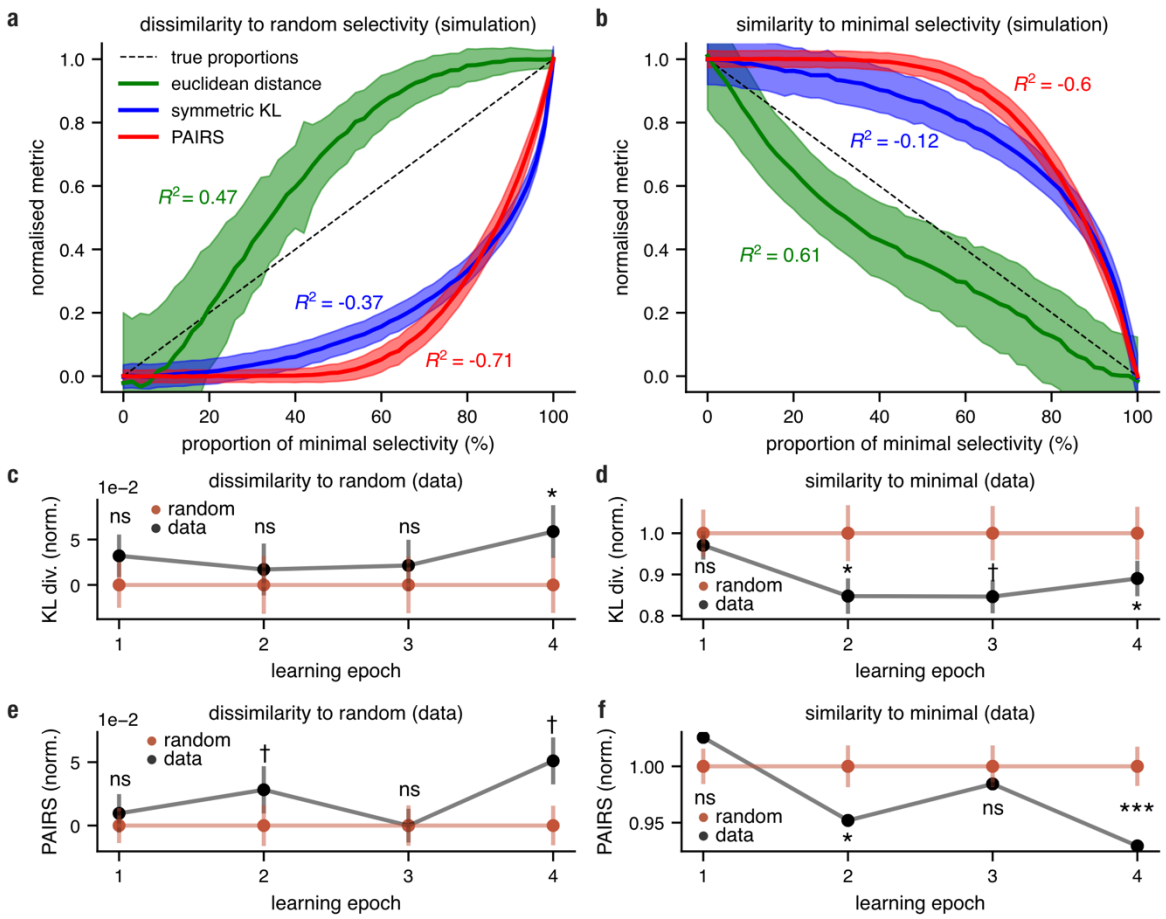
Supplementary figure 6. The re-analysis of Constantinidis et al.[25–27] dataset. We employed the same analysis methods as in **Fig 2,** and **Fig 3** to test whether the PFC activity reported in Constantinidis et al. converged on a minimal XOR model. **a**, the covariance matrix describing relations between the selectivity for stimulus 1, stimulus 2 and their interaction (XOR) in the pre-learning phase of the experiment. **b**, Same as panel a but post learning. **c**, Relative Euclidean distance between the covariance matrix of selectivity coefficients and the covariance matrix expected from random selectivity (with matched total variance) plotted as a function of learning (*Methods*, *measuring similarity between selectivity distributions*). **d**, Same as panel **c** but we show the relative distance from the covariance matrix expected from minimal selectivity (with matched total variance). **e**, Decoding of task variables for pre- and post-learning stages. **f**, Cross-generalised decoding of task variables plotted as a function of learning. **g, h**, Learning-induced accuracy differences in decoding and cross-generalised decoding, respectively. Shaded areas in **e-h** illustrate chance-level decoding obtained by shuffling trial labels (for details see *Methods, statistical testing*). All p-values were calculated from permutation tests (***, $p < 0.01$; **, $p < 0.01$; *, $p < 0.05$; n.s., not significant).



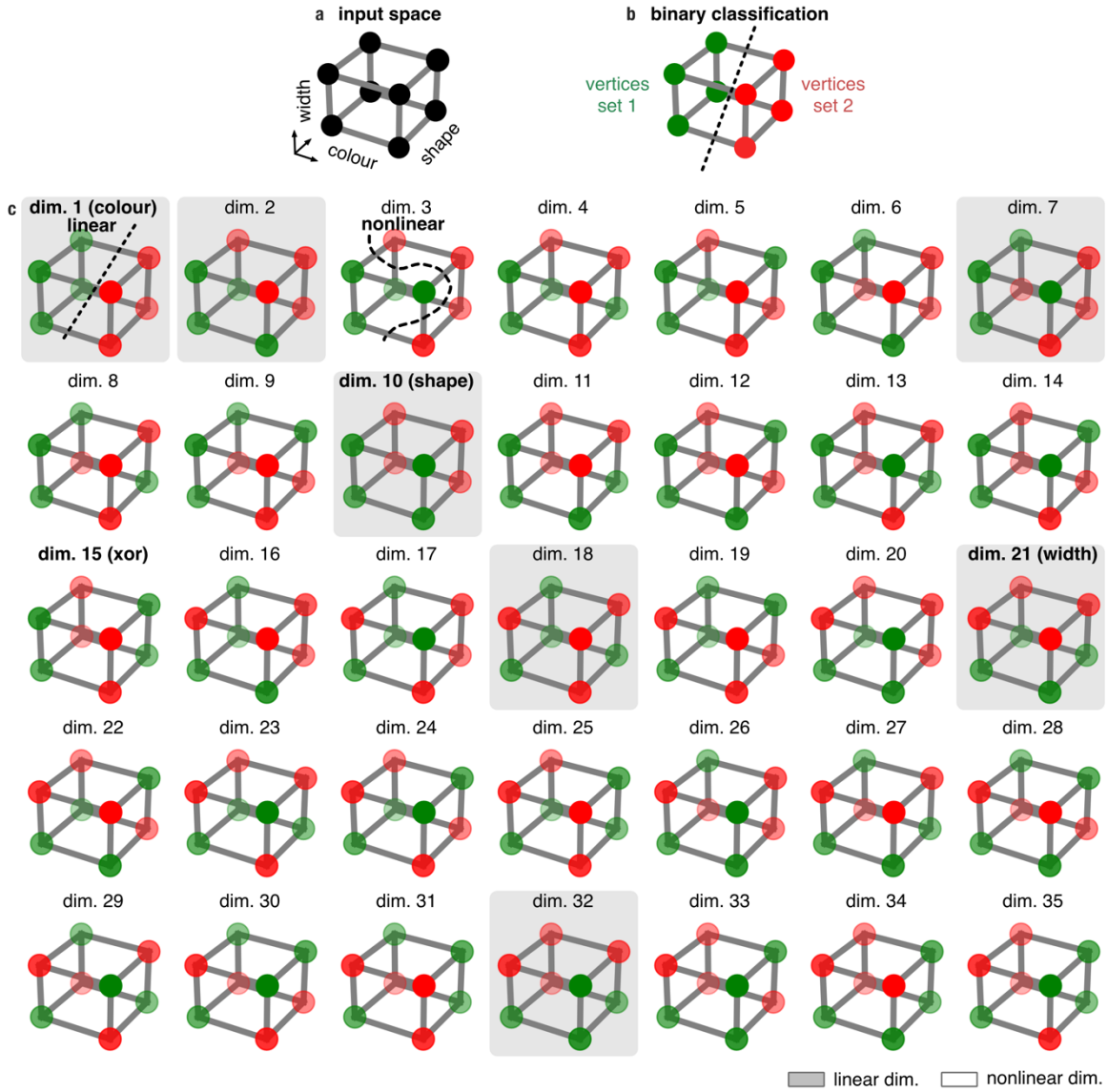
Supplementary figure 7. **a-e**, Temporally resolved linear SVM decoding for main task variables in experiment 2. Turquoise shading in the far left panel indicates the time window in which subsequent selectivity analyses were performed for context and stimulus set (cf. panels k-l); pale orange shading indicates when the remaining selectivity analyses were performed (cf. panels m-p). Horizontal dotted lines represent chance-level decoding whereas vertical dotted lines indicate the onset of the colour, shape and the trial outcome. The black bars indicate significant differences in decoding between stage 1 and stage 4, respectively. **f-j**, Temporally resolved cross-generalised decoding of main task variables (trained on set 1 and tested on set 2 (and vice-versa) for context, shape, width and XOR); cross-generalised stimulus set decoding was obtained by training on set1 vs set2 in context 1 and testing on context 2 (and vice versa) **k**, Correlation between selectivity coefficients for stimulus set (set 1 vs set 2) from context 1 and context 2 in late the colour-locked period (cf. turquoise in a) plotted as a function of learning; shaded areas indicate chance-level correlations between two random selectivity vectors obtained after shuffling trial labels. **l**, Correlation between selectivity coefficients for context (context 1 vs context 2) in stimulus set 1 and stimulus set 2 in late the colour-locked period plotted as a function of learning. **m-o**, Correlations of selectivity coefficients obtained from stimulus set 1 and stimulus set 2 for shape, width and XOR, respectively. **p**, Analogous to k but computed for the late target-locked period (cf. pale orange in a). **q**, Time resolved neural dimensionality as measured by shattering dimensionality (mean over all 6435 possible task dichotomies) in stage 1 and stage 4; the pale orange shaded area indicates the time window in which the statistical comparison between stage 1 and stage 4 was computed on averaged neural data. **r**, Analogous to Fig. 2a but computed for stimuli set 2 trials. **s**, Comparison of the ATT observed in the first 3 sessions (per animal) in experiment 1 (when only stimulus set 1 was presented) and 3 first sessions (per animal) in experiment 2 for stimuli set 2 trials only. All p-values were calculated from permutation tests (***, $p < 0.01$; **, $p < 0.01$; *, $p < 0.05$; †, $p < 0.01$; n.s., not significant).



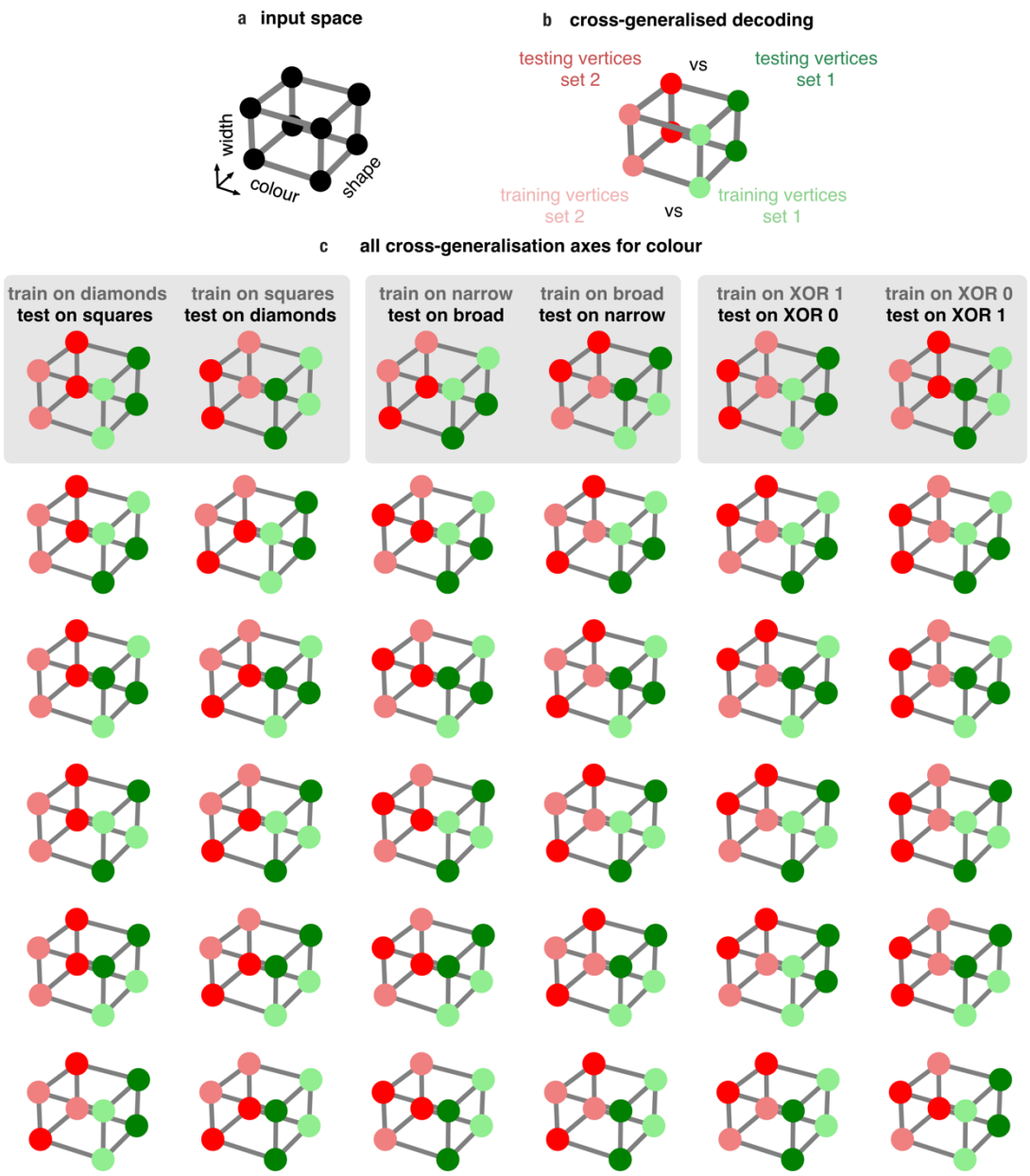
Supplementary figure 8. Electrode locations in monkey 1 (a) and monkey 2 (b) and their comparison (c).



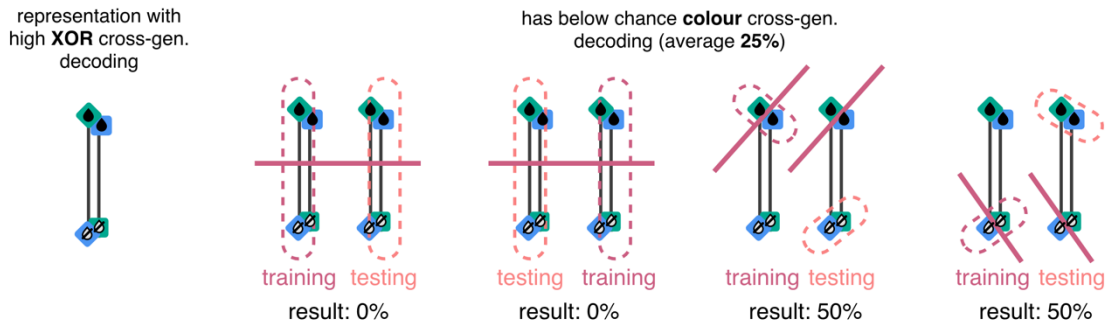
Supplementary figure 9. The precision of Euclidean distance, symmetric KL estimate and PAIRS metrics in tracking learning-induced changes to neural selectivity. a, Relative Euclidean distance between the covariance matrix of selectivity coefficients obtained from a simulated mixed population (random-minimal) and the covariance matrix expected from pure random selectivity plotted as a function of minimal selectivity proportions (0-100%; for details see *methods*, *measuring similarity between selectivity distributions*). Coloured annotations indicate mean R^2 values computed between the true proportions (dotted lines) and estimated proportions (coloured bold lines). **b**, Same as panel **a** but we show the relative distance from the covariance matrix expected from minimal selectivity; shaded areas illustrate mean $\pm 1\ s.d.$ for each of the metrics computed from 1000 randomly drawn selectivity models. **c, d** Selectivity results from experiment 1(**Fig. 3d, e**) computed using symmetric KL divergence estimate. **e, f** Selectivity results from experiment 1(**Fig. 3d, e**) computed using the PAIRS metric.



Supplementary figure 10. Schematic depiction of shattering dimensionality in the XOR task in experiment 1. a. Colour, shape, and width form a cube in the input space, with each of the 8 vertices representing a unique combination of these input variables. **b**. The vertices can be divided into two equally sized groups of four vertices each, representing a binary classification problem. **c**. All 35 theoretically possible binary problems (red vs. green vertices) are obtained by randomly splitting the vertices into two equally sized groups. Dimensions 1, 10, 15, and 21 correspond to colour, shape, XOR, and width, respectively. The grey and white backgrounds indicate whether a dimension is linear (can be split by a plane, e.g. dimension 1) or nonlinear (cannot be split by a plane, e.g. dimension 3), respectively. Colour intensity corresponds to spatial depth, with solid colours signifying vertices closer to the viewer and pale colours signifying those farther away.



Supplementary figure 11. Schematic depiction of cross-generalised decoding of the variables. **a**. Colour (blue vs green), shape (diamond vs square), and width (narrow vs broad) form a cube in the input space, with each of the 8 vertices representing a unique combination of these input variables. **b**, To test whether a variable is encoded in an abstract format relative to a single variable, two binary linear classification problem were defined. Firstly, a classifier was trained on differentiating that variable (e.g. colour) only on a subset of trials (e.g., diamond shape trials). Next, this classifier was tested on the remaining subset of trials (e.g., square shape trials). For example, a high score obtained from such a decoding procedure indicates that the same representation of colour was used as a function of different levels of shape, a hallmark of abstract coding. **c**, To test whether a variable (e.g., colour) has an abstract format relative to all remaining task variables, this procedure needs to be repeated for all possible train and test splits (when colour 1 is always on the left and colour 2 is always on the right side of the cube). These 36 scores are then averaged to obtain cross-generalised decoding of colour. Some of the test-train splits correspond to task variables like shape, width and XOR (grey background) while others represent a mixture of input variables.



Supplementary figure 12. Relationship between high XOR cross-gen. decoding and below chance colour cross-gen. decoding. The clustering of all rewarded trials (XOR == True) and non-rewarded trials (XOR == False) on opposite sides of an axis (representing the abstract XOR) results in high cross-gen. decoding for the XOR. Consequently, colour and shape decoding exhibit below-chance cross-gen. scores. This occurs because some cross-gen axes are inverted, leading to 0% accuracy scores, while others assign a single class label to all examples when testing, resulting in a 50% score.