nature neuroscience

Article

https://doi.org/10.1038/s41593-022-01237-9

The geometry of cortical representations of touch in rodents

Received: 17 September 2021

Accepted: 16 November 2022

Published online: 9 January 2023

Check for updates

Ramon Nogueira [1,2,3] ✉, Chris C. Rodgers [2,3,4,5], Randy M. Bruno [2,3,4,6] & Stefano Fusi [1,2,3,4] ✉

Neurons often encode highly heterogeneous non-linear functions of multiple task variables, a signature of a high-dimensional geometry. We studied the representational geometry in the somatosensory cortex of mice trained to report the curvature of objects touched by their whiskers. High-speed videos of the whiskers revealed that the task can be solved by linearly integrating multiple whisker contacts over time. However, the neural activity in somatosensory cortex reflects non-linear integration of spatio-temporal features of the sensory inputs. Although the responses at first appeared disorganized, we identified an interesting structure in the representational geometry: different whisker contacts are disentangled variables represented in approximately, but not fully, orthogonal subspaces of the neural activity space. This geometry allows linear readouts to perform a broad class of tasks of different complexities without compromising the ability to generalize to novel situations.

Making sense of the world often requires the integration of sensory evidence across multiple sources of information. In some situations, this involves only simple operations like linear summation. For example, if we need to determine whether an object is close to our hand or not, we can just move our fingers until any of them touches the object[1,2]. Summing the tactile feedback coming from all fingers and comparing it to a threshold would be sufficient to report whether the object is present or not. In other words, a linear decoder would be sufficient to perform this simple detection task. More difficult tasks, like recognizing the shape of an object by touch, might require more complex decoders to process the stimulus, generate the correct response and generalize to numerous variations of the sensory experience. These tasks might involve active sensing and non-linear integration of the sensory inputs coming from multiple fingers[3,4]. But it is also possible that the way we explore these objects by touch leads to spatio-temporal input patterns that are simpler than expected. For shape recognition, would a linear decoder be sufficient? Would the neural code reflect the difficulty of the task? What kind of neural representations would allow for generalization? To answer these questions, we investigated the problem of shape recognition using recent experimental data[5] in which mice are trained to report whether an object they touch with their whiskers is convex or concave. We found that the task can be solved by simple linear integration of whisker features (linear decoder). A linear decoder is also a good predictor of the decisions of the animals.

We then analyzed the neural activity recorded in somatosensory cortex. We observed that the responses of individual neurons are diverse and seemingly disorganized, as typically observed in cognitive areas[6,7]. In our experiment we could not even observe the somatotopic organization that one would expect from the architecture of the barrel cortex (see also ref. 5). Interestingly, the neural responses are best explained by a process of non-linear spatio-temporal integration, despite the observation that the task can be solved using linear integration.

The lack of organization at the individual neuron level induced us to analyze the representational geometry, which is defined by the set of distances between all the points in the population activity space that represent different sensory stimuli. The geometry is the only aspect of the representation that is preserved across individuals, species,

[1]Center for Theoretical Neuroscience, Columbia University, New York, NY, USA. [2]Zuckerman Mind Brain Behavior Institute, Columbia University, New York, NY, USA. [3]Department of Neuroscience, Columbia University, New York, NY, USA. [4]Kavli Institute for Brain Science, Columbia University, New York, NY, USA. [5]Department of Neurosurgery, Emory University, Atlanta, GA, USA. [6]Department of Physiology, Anatomy and Genetics, University of Oxford, Oxford, UK. ✉e-mail: rn2446@columbia.edu; sf2237@columbia.edu

https://doi.org/10.1038/s41593-022-01237-9

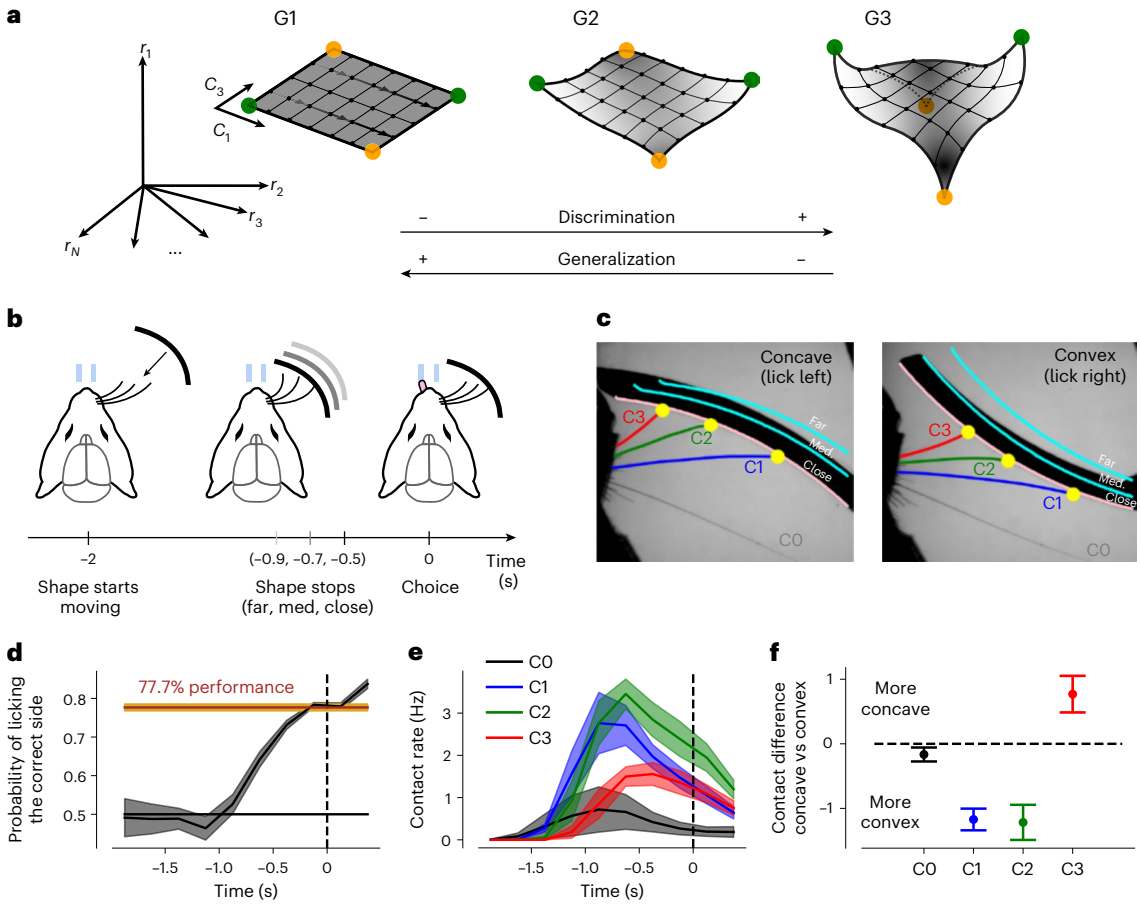


Fig. 1 | Different geometries of the neuronal representations and the whisker-based shape discrimination task. a, Each panel shows the activity space spanned by N orthogonal axes that represent the firing rate of N neurons. Each point in this space corresponds to a pattern of population activity. Low-dimensional representations (G1): the number of contacts C_1 and C_3 are represented along two orthogonal axes. These representations generalize well to unseen experimental conditions (Generalization) but lack flexibility—the ability to discriminate many different groups of points using a linear decoder (Discrimination). A linear decoder trained to report the value of C_1 (high vs low) for a given value of C_3 would generalize to all values of C_3. However, a linear readout would not be able to separate the green from the orange points. They would be separable for high-dimensional representations (G3) which are flexible but generalize poorly. Intermediate geometries (G2) could benefit from the computational properties of both low- and high-dimensional representations. **b**, Animals were presented with either a convex or a concave shape and after two seconds they reported their choice by licking the left (concave) or right (convex) lickpipe. **c**, Whiskers and shape position were monitored with a high-speed camera and an image parsing algorithm[20–22]. Panel adapted from ref. 5. **d**, The probability of making a lick on the correct side increased as a function of time. The choice on a given trial was determined by the side of the first lick after the response window opened ($t = 0$). The mean performance across animals was ~ 78%. **e**, The contact rates of all whiskers (C0, C1, C2 and C3) increased as the shape came within whisking distance (see Extended Data Fig. S1 for time profiles for each shape, and for correct and error trials separately). **f**, Difference in the total number of contacts between concave and convex shapes for all whiskers. C1 and C2 made more contacts for convex shapes, while C3 made more contacts for concave shapes. See Extended Data Fig. S1 for the distribution across animals. Errorbars in (**d**–**f**) correspond to s.e.m. across animals ($n = 10$ mice).

and artificial neural networks (see, for example, ref. 8,9), and different geometries have different computational properties[7,10,11]. It is instructive to discuss a few prototypical geometries (Fig. 1a). The first geometry that we consider (G3) is composed of four points that represent different sensory stimuli and define a relatively high-dimensional object (four points can span three dimensions at most). Throughout the article, when we speak about dimensionality we refer to the embedding dimensionality of the set of points (that is, the minimal number of coordinate axes needed to determine the positions of all points). In the specific example in the figure each point describes the neuronal responses during a sensory experience in which the subject animal explores different objects using whiskers. These sensory experiences are characterized by different values of two variables (for example, they could correspond to low or high number of whisker contacts for two different whiskers). These two variables can be decoded even using a simple linear decoder, which could be implemented by a downstream or a recurrent neuron. But the high dimensionality allows for more: the points can be separated in any possible way into two groups, which might correspond to different required responses. In other words this geometry confers flexibility because a linear readout can be easily trained to perform any binary task without changing the representation. For this geometry individual neurons exhibit responses that are non-linear functions of multiple task relevant variables (non-linear mixed selectivity[6,12]).

An alternative geometry is illustrated in G1 (Fig. 1a). The points representing the four stimuli now define a 2D square, which is lower dimensional than the representation in G3. These representations are called abstract because different variables are represented in orthogonal subspaces (one axis for C_1, whisker 1 contacts, and one axis for C_3, whisker 3 contacts, in the figure), and thanks to this arrangement they have special generalization properties: a linear decoder trained to report the value of one variable defines a coding direction in the neural activity space, and this direction is the same no matter what the values of the other variables are. So a linear decoder can readily generalize to

situations it was never trained on. These representations are called disentangled in the machine learning community[13–15] and they have been observed in several brain areas[7,16,17]. In this scenario individual neurons respond to a single variable or to a linear combination of the task relevant variables (linear mixed selectivity[6,18,19]). These representations allow for generalization at the expense of the flexibility guaranteed by the high-dimensional representations.

Given four experimental conditions, these two geometries are idealized examples of low- and high-dimensional representations, but there are also intermediate scenarios like the one in G2 (Fig. 1a), in which a low-dimensional scaffold is non-linearly distorted. These geometries represent a compromise that could have some of the computational benefits of both high- and low-dimensional representations, and they are the best description of the representations that we observed in somatosensory cortex. The low-dimensional scaffold renders the whisker features disentangled variables, which allows for enhanced generalization. This disentanglement is not a simple consequence of somatotopy, which we actually did not observe. The non-linear distortions make the representations sufficiently high-dimensional to enable a linear readout to perform complex tasks, but they are not so high-dimensional that they compromise the robustness to noise and capacity to generalize. A geometry that balances complexity and generalization, as seen in monkey prefrontal cortex[7], may therefore be a general feature of the cortex.

Results

The whisker-based object discrimination task

Mice were trained on a whisker-based shape discrimination task (Fig. 1b), in which they were asked to report whether an object was concave or convex (see ref. 5 for a detailed description of the experiment). Each trial began with the object moving toward the whiskers. Objects could stop at one of three different distances from the face (far, medium or close). When the response window opened, mice had to make a choice by licking the left lickpipe for concave objects and the right lickpipe for convex objects. The object position and the whiskers were monitored using high-speed video and processed with a deep neural network[20–22] (Fig. 1c). Importantly, mice were free to whisk and lick throughout the trial (2 seconds). On each trial, the choice of the animal was determined by the side of the first lick after the response window opened.

Mice performed the task with a mean accuracy of 77.7% ± 0.9% (s.e.m.) (Fig. 1d). The probability of making a correct lick increased throughout the trial, indicating that mice based their decision on the accumulated sensory evidence gathered by whisking. This implies some form of temporal integration. Most contacts were made between $t = -1.25$ and $t = -0.25$, suggesting that this was the most informative time window (Fig. 1e). The whiskers contacted the two shapes at different rates (Fig. 1f). The contact rate of each whisker followed a similar time profile, and mice made more contacts on correct trials (Extended Data Fig. S1), suggesting that errors resulted from poorer sensory gathering or a lower level of task engagement.

Linear integration is sufficient for object discrimination

We trained a linear decoder to report stimulus shape on a trial-by-trial basis. Its input was the observed spatio-temporal pattern of whisker contacts (Fig. 2a) and the angular position of the whiskers during contacts. We asked whether a linear decoder could use these data to predict the animal's choice, which on error trials differed from the shape's actual identity. Decoding the stimulus can reveal the whisker features that are useful to perform the task, whereas decoding the choice indicates which whisker features are actually used by mice. Both predictions were tested on held-out trials (cross-validation; see Methods).

The most informative set of features comprised all the whisker contacts and angle of contact across time (Fig. 2b, see also ref. 5). When whisker contacts were summed either over time, or across whiskers, the performance decreased, indicating that it is important to read out the full spatio-temporal pattern of whisker contacts and angles. Unsurprisingly, the weights of the classifier trained on contacts summed over time (Fig. 2b inset) reflected the difference in total number of contacts for convex vs concave objects (Fig. 1f). We also observed that the accuracy of the classifiers increased as mice accumulated more evidence (Fig. 2c; see Methods). Other features like force at the base of the whisker (whisker bending) or duration of contacts were not included because we have previously shown that they are less relevant[5].

Would non-linear decoders perform better at determining the shape of the object and the choice of the animal? We trained feed-forward neural networks with non-linear units arranged in multiple layers to predict stimulus and choice. These decoders are more complex, containing more parameters than the linear ones, so they will certainly perform better at classifying the patterns in the training set. However, the cross-validated performance of the non-linear classifier can surpass that of the linear classifier only if non-linear combinations of the features are important. Despite the task being significantly more complex than other whisker-based tasks (for example, pole detection), linear and non-linear decoders performed similarly at classifying stimuli (linear: 90.3% ± 1.2%; best non-linear: 91.3% ± 1.1%) (Fig. 2d; green). A similar result was observed on both correct and error trials (Extended Data Fig. S2a), though for error trials the performance was significantly lower for all decoders. This performance decrease is likely due to the lower number of contacts and overall lower task engagement in error trials (Extended Data Fig. S1). When predicting choice, a similar trend was observed (linear: 72.2% ± 1.6%; best non-linear: 75.6% ± 1.2%) (Fig. 2d; blue), suggesting that animals' decisions were mostly driven by a linear combination of the sensory cues across time and whiskers. On error trials our ability to predict choice was substantially lower, suggesting that these trials were qualitatively different (Extended Data Fig. S2a).

We also fit recurrent neural networks (RNNs) to decode stimulus and choice on a trial-by-trial basis, but the performance was not better than with feed-forward networks (Extended Data Fig. S3a). Finally, we fit the decoders only on trials with a specific stopping position (far, medium, close; Extended Data Fig. S2c) and obtained similar results. Shapes with lower curvatures were harder to discriminate for our decoders (Extended Data Fig. S2d), consistent with the results obtained in behaving mice[5].

Linear and non-linear discrimination in simulated tasks

We wondered whether non-linear decoders could have an advantage for shapes or shape-response associations different from those used in our task. We set up a simulation that reproduced several aspects of the experiment. We simulated the movement of three flexible whiskers that make contacts with the presented shapes (Fig. 3a). We then trained linear and non-linear classifiers to perform different tasks using the simulated spatio-temporal pattern of whisker contacts and angle of contacts. Similar to the behavioral data, linear and non-linear decoders performed equally well at discriminating between the convex and concave shapes used in the actual experiment (Fig. 3b).

Interestingly, non-linearities failed to improve the decoding performance also when we considered shapes with a richer micro-structure and orientations (Fig. 3c,d). Different shape positions and sizes, and shapes with different curvatures, also produced equivalent results (Extended Data Fig. S4). A possible explanation is that the decoders are reading out a spatio-temporal pattern of whisker contacts and angles, which contains information about multiple time steps. Concatenating patterns at different time steps is probably equivalent to projecting the inputs of individual times into a higher dimensional space, which enables a linear decoder to perform as well as a non-linear one. However, it is possible to design simulated tasks that require non-linear combinations of sensory cues (Fig. 3e, see also Extended Data Fig. S4), indicating that behaviorally relevant non-linear tasks exist. These tasks typically involved complex classifications of groups of simple shapes, indicating that the non-linearities are often important when the labels

https://doi.org/10.1038/s41593-022-01237-9

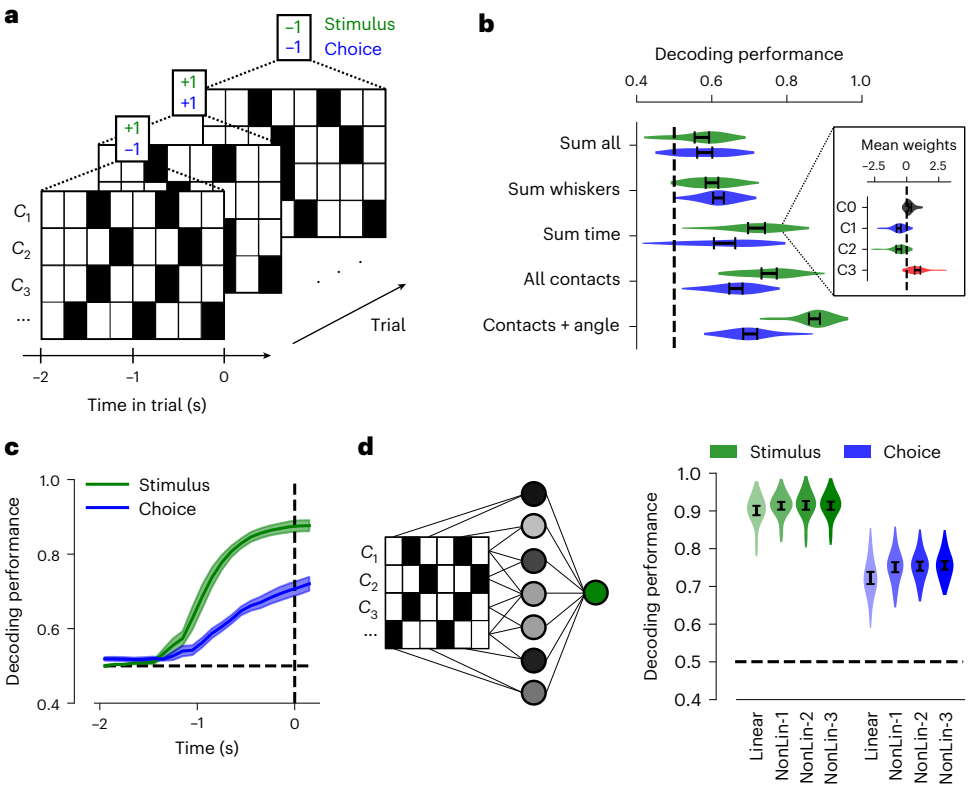


Fig. 2 | The whisker-based shape discrimination task can be solved by linearly integrating whisker contacts across time. a, The spatio-temporal pattern of contacts and angle of contacts across time and whiskers was used to classify shape identity (green) and animal choice (blue) on a trial-by-trial basis. **b**, Stimulus and choice decoding performance when different input features were used. Sum all: the linear decoder used only the sum of whisker contacts across all time bins and whiskers; Sum whiskers: sum across whiskers, but the temporal structure of the sum is retained; Sum time: the decoder considers the vector of the total number of contacts for each whisker; All contacts: the decoder reads out the full spatio-temporal pattern of all contacts from all time bins and whiskers; Contacts+Angle: all contacts and the angle of the whisker at time of contacts, which produces the highest decoding accuracy for both stimulus and choice. Decoding stimulus is expected to be easier than the animal choice because choice depends also on the internal state of the animal. (Inset) The weights obtained by a classifier trained to decode shape identity match the difference in number of contacts between concave and convex shapes (see Fig. 1f). **c**, Decoding performance as a function of time when the decoder reads out the full spatio-temporal pattern of contacts and angle from -2 seconds to the time indicated on the x-axis. The performance increases gradually for both shape and choice, indicating that there is some form of accumulation of evidence. **d**, A multi-layer neural network model is trained to use the full spatio-temporal pattern of contacts and angle of contact to predict the stimulus and the choice of the animal. Non-linear, multi-layer neural networks perform similarly to the linear network with no intermediate layer. Thus, linear integration of the sensory cues is sufficient to predict both stimulus and choice. In all panels stimulus and choice were decorrelated by fitting the classifiers with batches of trials that have an equal number of correct and incorrect trials. Error bars in all panels correspond to s.e.m. across animals ($n = 10$ mice).

assigned to the shapes (that is, their semantics) are complex, rather than the shapes themselves.

Non-linear mixed selectivity in the mouse S1 cortex

To characterize how task variables are represented in the somatosensory cortex (S1), we recorded populations of neurons while mice performed the whisker-based object discrimination task (Fig. 4a). Neurons are predictive of shape identity and animal choice on a trial-by-trial basis as revealed by the performance of a linear classifier (Fig. 4b left panel, see Methods). At the time of the response, shape could be decoded from small ensembles of simultaneously recorded neurons (mean population size of 25.4 cells) with a performance of 56.8% ± 1.7% (green) and the choice of the animal with a performance of 65.4% ± 2.0% (blue). The decoding accuracy for both shape and choice was significantly higher when we grouped together the activity from different recording sessions (pseudopopulations; see ref. 5). Not suprisingly, S1 neurons encoded the number of contacts for each whisker (high or low, Fig. 4b right panel).

To determine whether the neurons also responded to other variables, we trained encoding models that predicted the firing rate of the population of simultaneously recorded neurons from the set of whisking and behavioral variables for all time steps (100 ms) and trials (Fig. 4c). We provided the encoding models with the following regressors: instantaneous whisking contacts and angle, lick side and rate, and trial task variables (current and previous reward, choice and stimulus) (see Methods). Stimulus features like force at the base of the whisker (whisker bending) or contact duration were not included in the model because we have previously shown that they are weakly encoded[5].

We considered four different encoding models that were implemented using feed-forward neural networks: one that implements linear regression, and three feed-forward neural networks with 1-3 hidden layers of rectified linear units (ReLU) and linear output. Note that linear regression can only generate pure and linear mixed selectivity neurons whereas multi-layer networks can respond to non-linear interactions between regressors (non-linear mixed selectivity). We found that the observed neural activity is best explained by the non-linear mixed selectivity encoding model with one hidden layer ($R^2 = 0.111 \pm 0.005$ on held-out data; see Methods) (Fig. 4d; see also Extended Data Fig. S5a,b). The linear model with only pure and linear mixed selectivity was the worst at explaining the neural data ($R^2 = 0.089 \pm 0.005$). Including one

https://doi.org/10.1038/s41593-022-01237-9

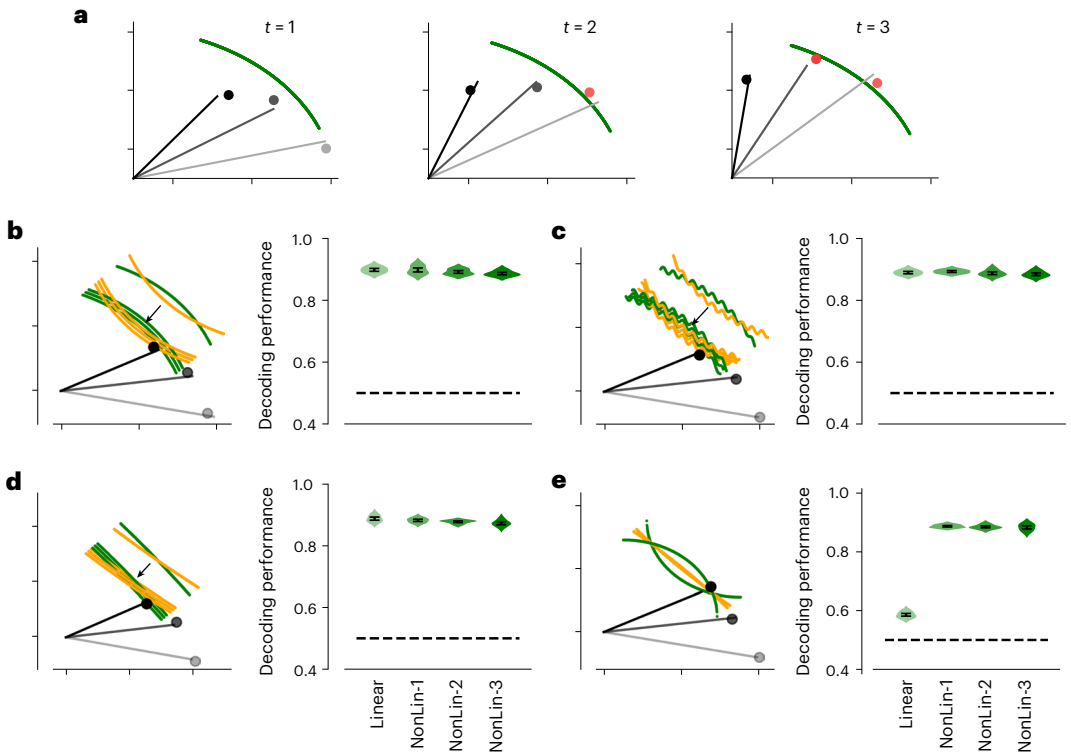


Fig. 3 | Non-linear classifiers outperform linear classifiers on more complex tasks but not on more complex shapes. a, Three snapshots of the simulated whiskers (C1, C2 and C3) making contacts on a moving concave object (green). Red dots correspond to contacts made by a whisker on a given time step (see Methods). **b**, The whisker-based discrimination task of the experiment was simulated with three whiskers and three different stopping locations for concave (green) and convex (orange) shapes (left). Linear and non-linear classifiers performed equally well on the simulated shape discrimination task (right), as in the experiment. **c**, Similar results were found when the same shapes had small wiggles, and when the task was to discriminate rotated flat objects. **d**, Non-linear classifiers performed better when the task required discriminating curved vs flat shapes. Errorbars correspond to s.e.m. across independent simulations ($n = 5$). See also Extended Data Fig. S4.

intermediate layer in the encoding model increased the explanatory power by 24%, a remarkable result given that the non-linearity is not necessary to explain the observed behavior. We also used RNNs as encoding models but they did not perform better than feed-forward encoding models (Extended Data Fig. S3b). The models performed better on correct trials, likely due to the the smaller number of contacts made on the less frequent incorrect trials (Extended Data Fig. S5c). Moreover, they explained better the responses of inhibitory neurons and neurons located in deeper layers of the somatosensory cortex (Extended Data Fig. S5d,e), possibly due to their higher firing rates.

To assess the importance of each regressor, we calculated $\Delta R^2 = R^2_{Full} - R^2_{Reduced}$, which quantifies the loss in prediction power on held-out data when a particular regressor or group of regressors is set to zero (Fig. 5a). We evaluated ΔR^2 for different groups of regressors and found that whisker contacts and continuous whisker angular position were most important variables for explaining the neuronal responses (Fig. 5b; see also Extended Data Fig. S6a). Superficial layers were more strongly driven by contacts than whisker angular position, while deep layers showed the opposite trend (Extended Data Fig. S6b-d). ΔR^2 for the current and previous time steps (time kernel) for the different whisker features showed a recency effect for all whiskers. Population activity was better predicted by C1 and C2 contacts than by C3 contacts, and it was less well predicted by the angular position of C1 than by the other whiskers (Fig. 5c).

What does somatosensory cortex mix?

The recorded neurons displayed a wide range of response properties: while some neurons showed approximate linear mixed selectivity for C1, C2 and C3 contacts (Fig. 6a), others showed sub-linear or XOR-like responses (Fig. 6b).

To assess the non-linear mixing, we used the neural network encoding models because the weights of the intermediate units in these models would be automatically tuned to produce the interaction terms needed to explain the neuronal responses. We determined the contribution of non-linear interactions of specific pairs of variables (or pairs of groups of variables) by setting the two variables (or two groups) to zero and evaluating: 1) ΔR^2, which is the loss in explanatory power for the full non-linear model (see Fig. 5b-d) and relies on interaction terms; 2) ΔR^2_{Linear}, which is the analogous loss for the linear model. We then computed the difference $\Delta R^2 - \Delta R^2_{Linear}$ (Extended Data Fig. S7a), which quantifies the importance of the particular nonlinear interaction that was set to zero. The interaction term is important when this difference is large. We found that the most important interaction was between whisker angular position and contacts (Fig. 6c), followed by interactions between whisker angular position and the other variables. Similar results were observed in both excitatory and inhibitory neurons, and across layers (Extended Data Fig. S7b,c).

We next assessed the importance of the interactions between different whiskers (whisker contacts and angular position) for different time lags. For contacts, the interactions are strongest not for the variables at the current time step, but actually at time-lags of 100ms (Fig. 6d left). This was unexpected given that contacts at the current time steps are those that most affect the neural activity (see Fig. 5c) and could reflect response inhibition between whisker contacts that occur within whisk cycles of 50-100 ms. For the angular position of the whiskers, the strongest interactions were observed in the current time

https://doi.org/10.1038/s41593-022-01237-9

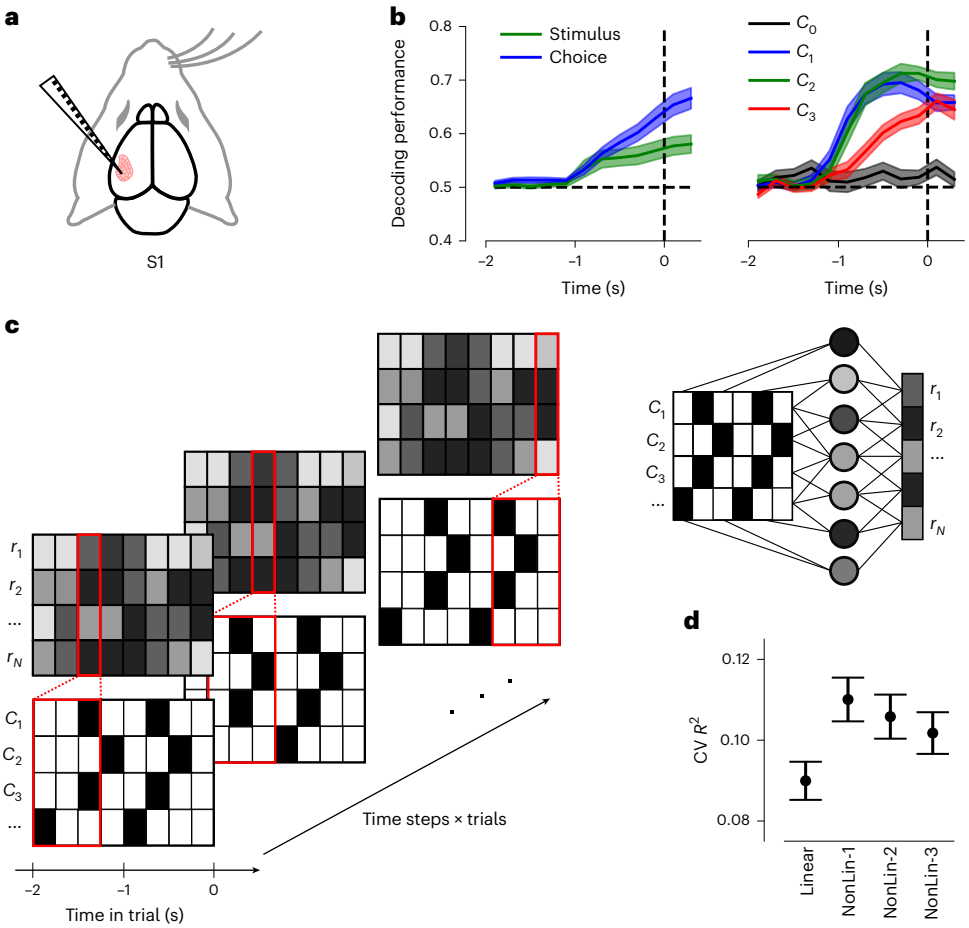


Fig. 4 | Populations of neurons in the mouse somatosensory cortex (S1) exhibit non-linear mixed selectivity for task variables. a, Multiple S1 neurons were simultaneously recorded while mice performed the whisker-based discrimination task described in Fig. 1. **b**, Decoding performance of a linear classifier trained to predict stimulus or choice (left panel), or contacts made by each whisker (right panel) as a function of time. For each timepoint, the decoder read out the neural activity in all the time bins from -2 seconds to the time indicated on the x-axis (simultaneously recorded neurons). As expected, the decoding performance is at chance early in the trial when mice do not make contacts and it reaches ~ 65% at the end of the trial for choice (blue), and ~ 57% for stimulus (green). Whether whiskers made high or low number of contacts on a given trial could be decoded more accurately than shape identity. Errorbars correspond to s.e.m. across simultaneously recorded populations of neurons ($n = 23$ recording sessions). **c**, S1 activity was regressed against task variables. Linear and non-linear neural network models with a different number of intermediate layers were used to reproduce the observed neural activity. **d**, Cross-validated R^2 on populations of S1 neurons for the different encoding models. A fully connected neural network with one hidden layer (NonLin-1) outperforms the linear model and other neural networks with more intermediate layers. Errorbars correspond to s.e.m. across neurons ($n = 584$). See Extended Data Fig. S5a for the distribution of R^2 across neurons.

bin (Fig. 6d right). We also examined non-linear interactions between contacts and whisker position across all whiskers and multiple time steps. Interestingly, the strongest non-linear mixing between angular position and contacts also occurred with a time lag of 100ms (Extended Data Fig. S8; see Discussion for the implications of these observations).

The disentangled geometry of neural representations in S1

To characterize the recorded representational geometry, we studied how it could be used by a downstream linear readout. This is a standard approach in machine learning in which "linear probes" (that is, linear decoders) are used to characterize the representations of hidden layers of deep networks[23]. We trained a linear decoder to perform synthetic classification tasks on recorded neural activity. Different tasks required different geometrical properties. The labels of the synthetic tasks (the outputs of the decoder) were decided on the basis of the observed total number of whisker contacts for two whiskers at a time. The inputs were constructed by combining together the recorded activity of all the neurons from different sessions (same or different animals) and by concatenating the activity vectors of all time bins within a trial (pseudopopulations; see Methods).

We considered complex and easy tasks: for the easy task the desired output was the thresholded weighted sum of the number of contacts for two whiskers (C1 and C3, as in "Easy Task" in Fig. 7a). Weights were random, and each weight vector corresponded to a different implementation of an easy task. In Fig. 7a we illustrate one sample easy task by showing the space of whisker contacts. Each point corresponds to a pair of C1 and C3 contacts, and its color (green and orange) denotes the desired output. The two regions containing different colored points are separated by a line (not shown) whose orientation depends on the random weights. By construction, the easy task is linearly separable. For the complex task, the space of whisker contacts was divided into 4 regions by two orthogonal separating lines, again in random directions. The labels were chosen so that the inputs of diagonally opposed regions require the same desired output ("Complex Task" in Fig. 7a). This task is similar to a XOR task, which is non-linearly separable and requires high-dimensional representations to be solved.

https://doi.org/10.1038/s41593-022-01237-9

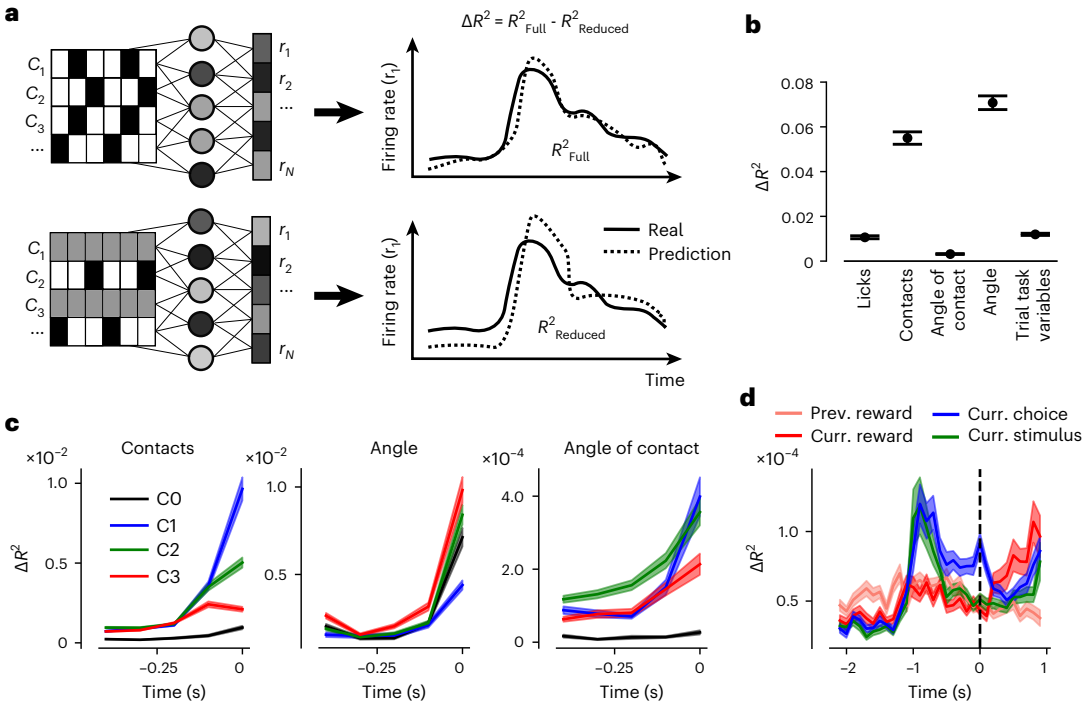


Fig. 5 | Contacts and whisker angular position are the variables that contribute the most to the prediction of S1 population activity. a, We used an encoder model to explain the population's firing rate (r_1,r_2, ..., r_N; simultaneously recorded) as a non-linear function of task variables like whisker C1, C2 and C3 contacts (that is C_1, C_2 and C_3), and calculated R^2_{Full}, the goodness-of-fit of the full model (top). To assess the importance of each regressor, we set the input data for that regressor (or group of regressors) to zero (gray) and assessed the goodness-of-fit of the reduced model, $R^2_{Reduced}$ (bottom). The importance of each regressor can be quantified as the resulting decrease in goodness-of-fit $\Delta R^2 = R^2_{Full} - R^2_{Reduced}$. **b**, Whisker contacts and angular position were the most important groups of regressors on S1 activity as revealed by the decrease in model accuracy ΔR^2 (see Extended Data Fig. S6a for the distributions of R^2_{Full} and $R^2_{Reduced}$ across all recorded neurons). **c**, Decrease in model accuracy (ΔR^2) for whisker contacts (left panel), angular position (central panel) and angle of contact (right panel) for different time lags with respect to current time step. Neural activity was better explained by variables in the current time step. **d**, Previous reward (pale red) was predictive of neuronal activity (ΔR^2) early during the trial whereas the importance of current reward (red) peaked after mice made their choice. Additionally, although current stimulus (green) and choice (blue) followed a similar trend throughout the course of the trial, current choice had a stronger effect on the population's firing at response time ($t = 0$). Similar task variable time profiles have been reported in previous studies in other animals and brain regions[49]. Errorbars in all panels correspond to s.e.m. across neurons ($n = 584$). See Extended Data Fig. S6a for the distribution across neurons.

To benchmark the ability to generalize, we computed the cross-condition generalization performance (CCGP)[7] for all the three whiskers. CCGP was evaluated as a linear decoder's ability to report the number of contacts (high vs low) of one whisker, say C3, for example, for a certain value of the whisker count for a different whisker (for example, when C1 is high). The decoder was trained only on the other values of the different whisker (for example, low C1, see "Generalization" in Fig. 7a). If C1 and C3 contacts are represented in approximately orthogonal subspaces, then the CCGP is high and the C1 and C3 variables are disentangled. The three different idealized geometries of Fig. 1a lead to a different performance on the easy and complex tasks and the generalization benchmark (Fig. 7b).

The linear decoder could perform the easy task with high accuracy (left bar in Fig. 7c). The performance for the complex task was relatively low, but still above chance (central bar in Fig. 7c; see Methods). Interestingly, CCGP was high for all the variables representing the number of contacts of different whiskers (right bars in Fig. 7c).

This means that the coding direction for the number of contacts of each whisker does not depend much on the number of contacts of the other whiskers. In other words, the coding directions for each whisker are approximately parallel to each other when one considers different values of the number of contacts of the other whiskers. The fact that CCGP was high for all the variables excludes the high-dimensional geometry depicted in Fig. 7b (G3). The fact that the performance on the complex task was above chance indicates that the non-linear component of the neural responses is not negligible, and hence that the representational geometry is not compatible with the low-dimensional one illustrated in Fig. 7b (G1). An intermediate geometry, which could be described as a non-linearly distorted low-dimensional scaffold, is the best description of the data (G2).

For these representations the elevated CCGP relies on the linear component of the responses, whereas the distortions needed to solve the complex task require non-linear components. This can be seen directly by generating synthetic neural data using the encoding models described above: the complex task could be better solved when the non-linear component was preserved, whereas linear synthetic representations slightly outperformed non-linear ones on the easy task and CCGP (Fig. 7d).

Disentanglement is not a simple consequence of somatotopy. One possible explanation for the elevated CCGP is that there are segregated populations of neurons, each encoding the number of contacts for one whisker. This specialized representation might reflect the somatotopic architecture of the barrel cortex. However, this is not the case because we analyzed separately the populations of neurons in different columns and we found that: 1) C1, C2 and C3 contacts were encoded in all the columns (Extended Data Figs. S6e-h, S9), as observed in ref. 5; 2) CCGP is elevated for all the whiskers in all the columns (inset in Fig. 7d; see also Extended Data Fig. S9).

https://doi.org/10.1038/s41593-022-01237-9

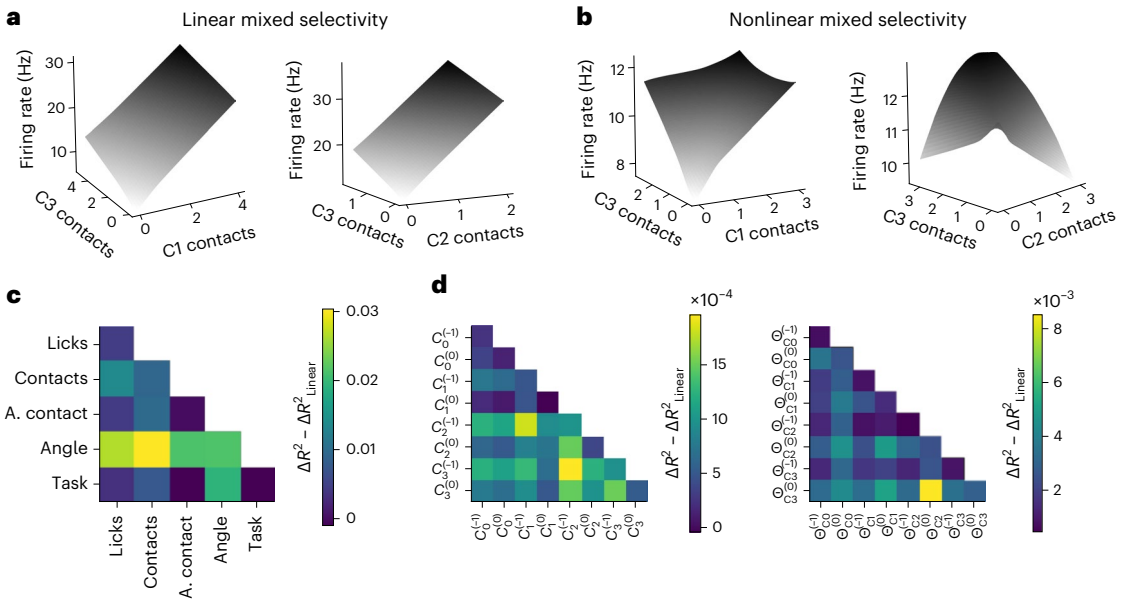


Fig. 6 | Neurons show non-linear mixed selectivity, mostly originating from non-linear interactions between contacts and whisker angular position. a,**b**, Tuning curves of example recorded neurons that exhibit approximately linear mixed selectivity (a) and non-linear mixed selectivity (b) for C1, C2 and C3 whisker contacts. All tuning curves were obtained from the best encoding model (non-linear with one hidden layer, see Fig. 4d). **c**, Non-linear mixed selectivity ($\Delta R^2 - \Delta R^2_{Linear}$) for the interaction between the different groups of task variables. The interaction between the groups whisker contacts and angular position was the most important non-linear contribution to the encoding model. Diagonal elements account for the strength of the non-linearity between a particular task variable and the activity of the neuronal population. **d**, Non-linear mixed selectivity contribution for different time steps and whisker contacts (left) and angular position (right). The interaction between contacts made by whiskers in the previous time step (100 ms time lag) (for example, $C_3^{(-1)}$ vs $C_2^{(-1)}$) had more explanatory power on S1 population activity than the interaction on the current time step (0 ms time lag) (for example, $C_3^{(0)}$ vs $C_2^{(0)}$). For angular position, the interaction between whiskers at the current time step (for example, $\Theta_{C3}^{(0)}$ vs $\Theta_{C2}^{(0)}$) had more explanatory power in S1 than on previous time steps.

Task difficulty modulates RNNs' representational geometry

As seen in the previous sections, the task is linear but the neural representations contain a non-linear component. There are at least two possible explanations. First, S1 is employed in multiple tasks, some of which may be complex enough to require a non-linear component. When trained on these complex tasks, S1 could also perform easy tasks, though the generalization performance might be reduced. Second, the non-linearities may be unavoidable even when the task is easy and does not require them. For example, it is possible that the type of temporal integration needed to perform the task requires some form of non-linearities as in echo state machines or liquid state machines[24]. The results thus far cannot exclude the second possibility because the models that we used to predict the stimulus, the choice of the animal and the neural activity use inputs from different time steps that are concatenated together. This is a possible way of implementing temporal integration, but it is not clear whether and how it can be realized by a neural circuit, like a recurrent neural network.

To answer these questions we simulated a recurrent neural network (RNN) trained to perform tasks with different levels of difficulty that are similar to the shape discrimination task (Fig. 8a). Notice that the RNNs were only required to generate the correct response in the artificial tasks, and not to reproduce the neural data as in ref. 25,26.

The synthetic tasks had the following structure: each trial lasted 30 time steps, and at each time step we fed the RNN with a vector containing three binary variables (each representing contacts made by one of three whiskers). Each binary variable was random (an independent Bernoulli process) with either a high or low success rate λ, for a total of 8 different conditions. The desired outputs defined two tasks with different levels of difficulty: easy (Fig. 8b,d,f), and complex (Fig. 8c,e,g) (see Methods). The easy task could be solved by linearly mixing information across input channels whereas the complex task required non-linear mixing. RNNs were trained on either the easy or the complex task to determine the input, recurrent, and readout weights (see Extended Data Figs. S10 for results on a third very complex task).

We then froze the trained network and studied the geometry of the neural representations using the same approach adopted for real data: we trained a linear readout to perform tasks with two different levels of complexity. We finally asked how well the representations learned for one task could be used to perform the other task. Specifically, we trained each network on either the easy or the complex task, froze the recurrent and input weights, and then optimized linear readouts to use those representations to perform different tasks (easy, easy other, and complex task) (Fig. 8d,e; see Methods). To study the generalization properties of the neural representations we estimated the CCGP for different dichotomies (ways of dividing the points into two groups) that correspond to different easy tasks (CCGP easy, CCGP other in Fig. 8f,g; see Methods).

Networks trained on an easy task produced a high discrimination and generalization performance (high CCGP) for the easy task only (dark brown; Fig. 8d,f). The complex task could not be performed indicating that in this case the non-linearities are weak and the dimensionality of the representations is relatively low. These results show that the non-linearities are not necessary to perform temporal integration.

Training a network on the complex task produced representations that allowed a linear readout to perform both the easy and complex tasks (Fig. 8e) and interestingly the easy task could be performed better than the task the network was trained on. This flexibility is provided by the non-linear components of the responses, as shown in Extended Data Fig. S10e. Importantly, CCGP was well above chance for the variable defined by the easy task: this means that the non-linearities needed to perform the complex task have a relatively low impact on the ability to generalize. In other words, training on the complex task leads to

https://doi.org/10.1038/s41593-022-01237-9

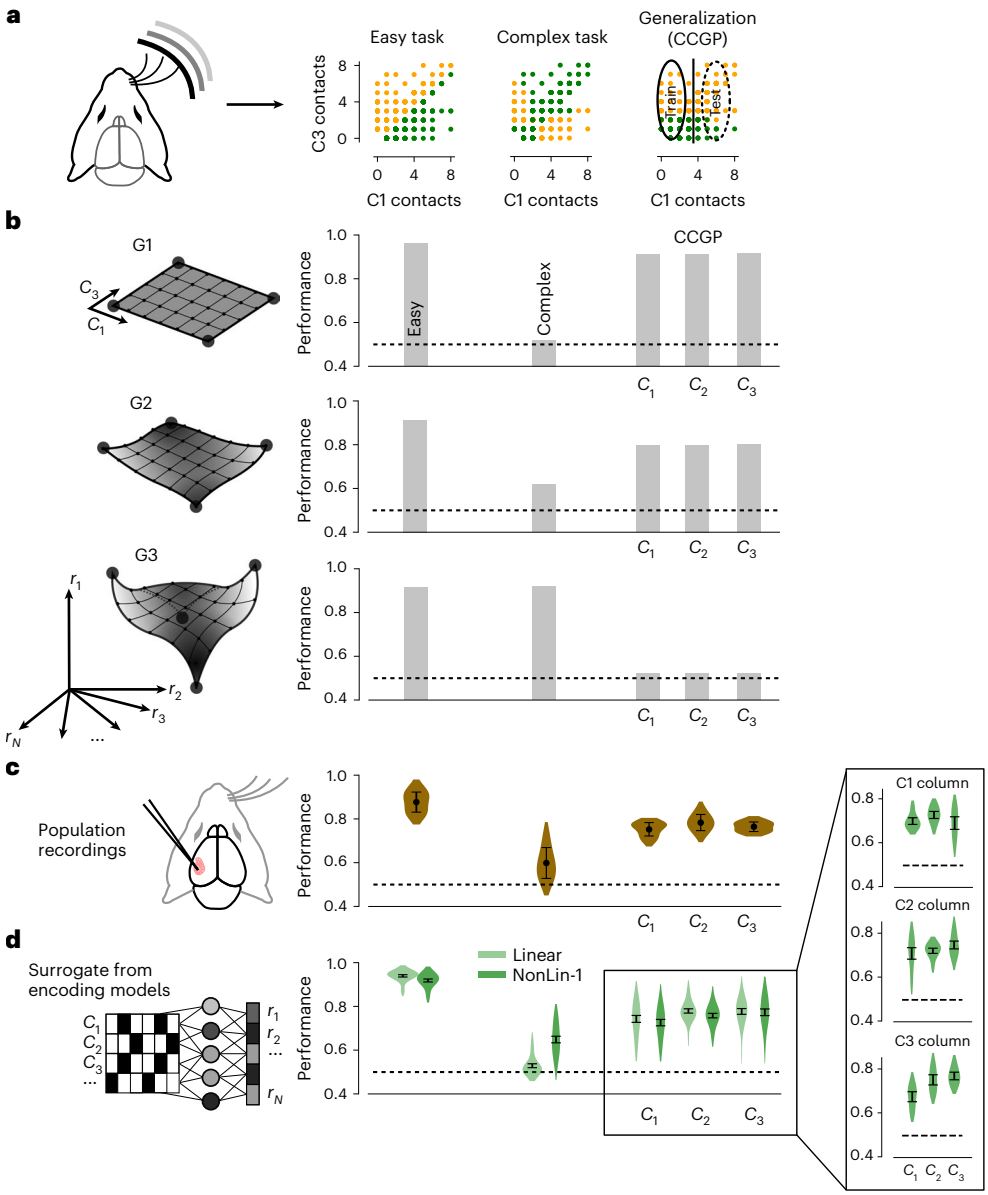


Fig. 7 | The geometry of representations observed in S1. a, Synthetic tasks and tests used to probe the geometry. Easy task: the linear classifier should output 1 (orange) if the (random) weighted sum of C_1 and C_3 is larger than a threshold, and 0 otherwise (green). The two classes are linearly separable. The values of C_1 and C_3 (or other pairs of whisker contacts; see Methods) are taken from the experiment. Complex task: the (C_1,C_3) space is divided into 4 regions by two orthogonal random directions. The two classes are colored in orange and green. The task is not linearly separable. Generalization test (example for C_3): a linear decoder is trained to discriminate between high and low C_3 on low C_1 (Train), and it is tested on high C_1 (Test). CCGP for C_1 and C_2 were evaluated equivalently (see Methods). **b**, Expected performance on the synthetic tasks for the three different idealized geometries of Fig. 1a. For low-dimensional representations (G1), both the easy task performance and generalization (CCGP) are high. The performance is poor for complex tasks. High-dimensional representations (G3) allow for high performance in complex tasks, but generalization is poor. Intermediate geometries (G2) benefit from the computational properties of both low- and high-dimensional representations. **c**,**d**, Performance on the easy task, complex task, and generalization benchmark (CCGP) when we used the real neural data (c) (pseudosimultaneous recordings) or the surrogate data generated by the linear (pale green) and non-linear (one hidden layer; darker green) encoding models (d) (fit on simultaneously recorded populations). For the easy task, performance was high for real (c) and surrogate data (d). For the complex task, the performance was above chance for real and surrogate data, except for the linear surrogate representations. For the generalization benchmark, CCGP is high for all whisker contact variables (C_1, C_2, and C_3) for the real and the surrogate data across all columns in S1 (inset; see also Extended Data Fig. S9). Errorbars in (c) correspond to the standard deviation across cross-validation iterations ($n = 10$) (see Methods), whereas in (d) they correspond to s.e.m. across populations of simultaneously recorded neurons ($n = 23$ recording sessions).

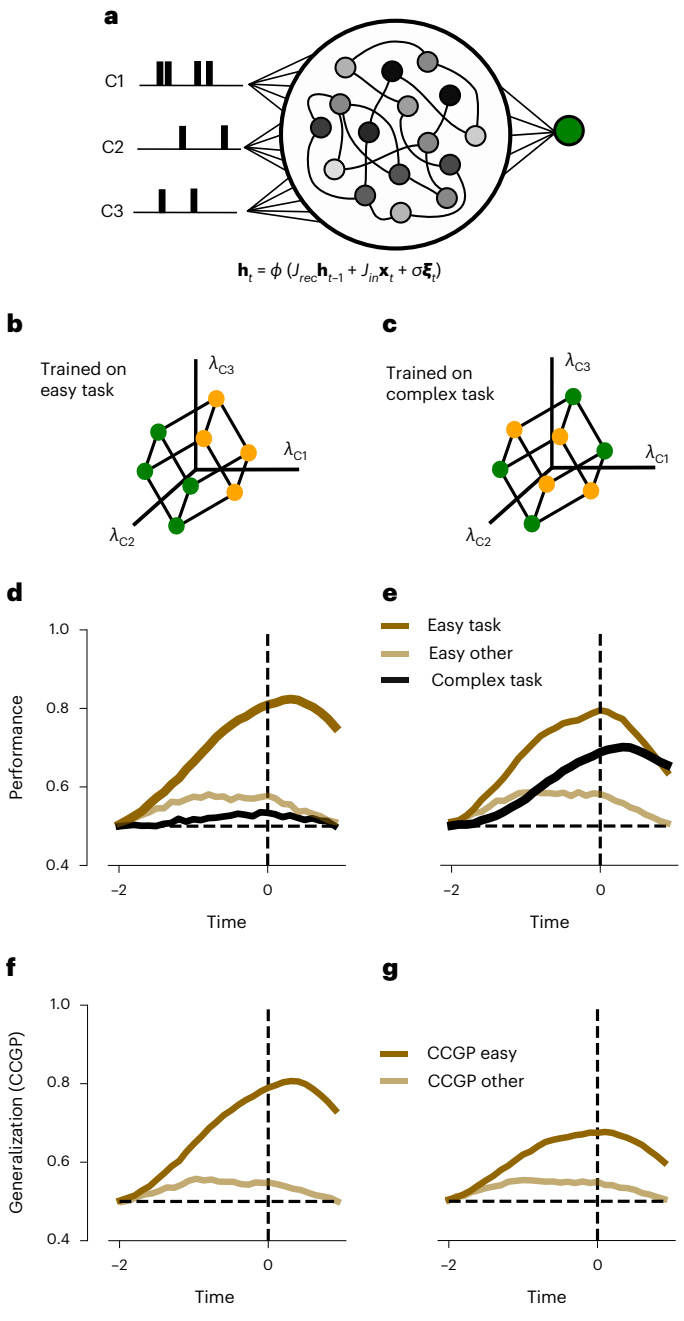
non-linear components in the neural responses and allows the network to perform a broad class of different tasks, with a modest cost in terms of generalization (see also Extended Data Fig. S10). This is probably why we observed a non-linear component in the neural responses of S1, despite the fact that the task does not require them.

Discussion

The neural responses in somatosensory cortex are diverse and seemingly disorganized non-linear functions of multiple variables describing whisker features (non-linear mixed selectivity). The mixed variables characterize multiple whiskers at different times, leading to

https://doi.org/10.1038/s41593-022-01237-9

an interesting form of spatio-temporal mixed selectivity. Strikingly, the non-linearity is observed despite the fact that the task can be solved by linearly integrating the same task-relevant variables. The non-linear responses appear to lack any evident organization - even somatotopy (see also ref. 5). However, an interesting organization appeared when we analyzed the representational geometry: whisker contacts, which are important features for performing the shape discrimination task, are represented in subspaces that are approximately orthogonal. This geometry is not a simple reflection of the anatomical organization of the somatosensory cortex. Indeed, the representational geometry in three different columns is similar, with neurons that respond to features of all the three whiskers used by the animal. This kind of factorized or disentangled representation has been observed in the prefrontal cortex[7,27], hippocampus[7,17], infero-temporal cortex and perirhinal cortex[16,19,28], and motor cortex of monkeys[29]. It is known to have important computational properties for generalization[13,14].



Why non-linear neuronal responses when the task is linear?

As the non-linearity detracts from robustness to noise, why are the representations non-linear even if the shape discrimination task is linear? We showed that the observed geometry actually represents a good compromise between the ability of a linear readout to perform complex discrimination tasks - which is typical of high-dimensional representations - and the robustness to noise and the ability to generalize to novel situations - which is typical of low-dimensional representations. This interesting compromise can be reproduced in a simulation of a RNN trained to perform several different tasks that are similar to the shape discrimination task. In these simulations we also observed that the non-linear component of the representations is progressively more important in more complex tasks, and that its cost in terms of noise robustness and generalization is relatively small. This suggests that the somatosensory cortex operates in an interesting regime that is probably the result of training on a variety of tasks and allows for flexibility and generalization.

Encoding models to characterize the population response

Previous studies showed that the dimensionality of neural representations can be maximal (monkey PFC[6]), very high (rodent visual cortex[30]), or as high as it can be given the task's structure[31]. More recently, in ref. 7 the authors showed that representations can have the maximal dimensionality required to separate all possible groups of stimuli with a linear decoder (shattering dimensionality) and, at the same time, exhibit a low-dimensional scaffold which allows for cross-condition generalization. These studies focused on computationally relevant properties of the representational geometry, ignoring the detailed information about the individual neurons' response. Other studies looked more closely at the components of responses that are important for characterizing this geometry, focusing on the two important ingredients for getting high dimensionality: mixing and diversity[32,33]. Sometimes the responses of individual neurons can be well described by linear mixed selectivity[16,18,19], indicating that the representations are low-dimensional, or disentangled[14].

Here we adopted a new approach to characterize both collective properties of the representations and the dynamic response of individual neurons. Using the neural network encoding models we could characterize the response of a population of simultaneously

Fig. 8 | The geometry of representations in RNNs is modulated by the difficulty of the task. a, We simulated and analyzed the representations of a recurrent neural network model (RNN) performing a synthetic task that is similar to the shape discrimination task. A set of noisy and fully connected ReLU units receive input from three independent channels (C1, C2 and C3). The state of the network at a given time step ($\mathbf{h}_t$) is determined by its state at the previous time step ($\mathbf{h}_{t-1}$), the input at the current time step ($\mathbf{x}_t$) and independent Gaussian noise ($\boldsymbol{\xi}_t$) (see Methods). **b**, The easy task is a linear integration task across input channels. RNNs in panels (d),(f) were trained on the easy task. **c**, The complex task consists of a non-linear integration task with respect to two input channels (XOR). RNNs in panels (e),(g) were trained on the complex task (see Methods). **d**, The performance on the easy task (dark brown) as a function of time within a trial for RNNs trained on the easy task. Performance increased as a function of elapsed time in all networks and tasks. A readout unit fails at performing an orthogonal easy (easy other; pale brown), and the complex task (black). **e**, Probability correct on the easy (dark brown), orthogonal easy (easy other; pale brown), and complex tasks as a function of time within a trial for RNNs trained on the complex task. **f**, Generalization ability, estimated as the cross-condition generalization performance (CCGP), is high for the easy task and low for the orthogonal other easy task for RNNs trained on the easy task. **g**, Generalization (CCGP) for RNNs trained on the complex task is overall lower than for RNNs trained on the easy task. See Extended Data Fig. S10 for results also on a very complex task. Different levels of input noise produced qualitatively equivalent results for all networks, tasks and generalization (Extended Data Fig. S10). For all panels, the performance curves correspond to the mean across random realizations of input patterns and tasks ($n = 50$) (see Methods).

recorded neurons. This is more than reproducing the responses of all the individual neurons, because the encoding models can also capture the correlations between the activities of different neurons, which can affect the geometry of the representations and their consequences on information encoding and behavior (see, for example, ref. 34–38). This approach was motivated by the fact that the task involves active sensing, which is closer to natural behavior but more difficult to analyze. In contrast to the aforementioned monkey experiments[32], the trial temporal structure is highly variable as it depends on how the animal moves the whiskers. Because the sensory input in this kind of behavior involves a larger set of variables which are continuous, we used a more unbiased approach to identify those variables that could be important to predict the animal's behavior and neural activity. We started from this larger set of variables that characterize complex spatio-temporal patterns and let the encoding model find those that are most important.

Limitations on assessing non-linearity of mixed selectivity

One important issue is that mixed selectivity is always defined with respect to a set of variables[39] and this is one of the limitations of all the analyses that focus on the responses of individual neurons. For example, neurons could respond to a non-linear combination of two variables x and y, say xy. If one considers $z = xy$ as an additional variable, then the non-linearity disappears, as all the neurons can be described as a linear combination of x, y, z. It is possible that for a different choice of variables, the observed mixed selectivity would be less non-linear than what we observed. However, the choice of the variables was actually dictated by the analysis of the whisker features and interestingly, a linear combination of these features is sufficient to predict shape and choice, but not the neural activity. Even when we considered additional variables (for example, whisker angles were used to predict the activity but not the stimulus or the behavior), we still needed non-linear interactions. This is significant because these additional variables could be related to non-linear interactions between other variables. Nevertheless, a linear encoding model that has access to all these variables still performs worse than a non-linear one.

Mixed selectivity and the architecture of the neural circuit

Our analysis showed that non-linear interactions are an important component of the neuronal responses. The strongest interactions are between the variables representing the angular position of the whiskers at the current time and the whisker contacts that occurred in the preceding 100 ms time step. The interactions between contacts of different whiskers affect the current neural activity if the contacts happened in the preceding time bin. Although the interactions are delayed, the information about the whisker contacts is not, and the strongest contribution to the neural activity comes from the contacts of the current time step. This means that the first information to arrive in the somatosensory cortex is more linear, and the interaction terms affect the neural activity with a delay of the order of 100 ms. We speculate that the whisker contact information arrives first from segregated inputs containing information about separate whiskers. The interaction terms appear later and could originate from some non-linear recurrent neural circuit which might be local, within somatosensory cortex, or long-range, involving other areas such as secondary somatosensory, motor, frontal cortex or secondary thalamic nuclei (for example, ref. 40–42). For the information about whisker position (expressed as angles in our analysis) the dominant interaction terms are instead between angles at the current time. It is possible that this information is already non-linearly mixed in other brain areas (downstream, like motor cortex, or upstream like primary thalamus and brainstem[43,44]).

New analytical methods for naturalistic experiments

Our general framework for analyzing behavioral and electrophysiological data is particularly valuable in experiments in which the animals perform natural tasks, which are becoming increasingly popular[45–48]. Fitting neural networks to predict stimulus identity and animal choice from features extracted from high-speed videos is useful to identify the most important variables to perform the task successfully and the behavioral strategy actually followed by the animal, especially for naturalistic behavior, in which we have limited control over the strategies adopted by the animals. Moreover, using neural networks to fit neuronal activity from the recorded task variables can be understood as an unbiased multi-dimensional generalization of a population tuning curve. Even though the tuning information is implicitly contained in the architecture and weights of the encoding model, it can still provide crucial insights about the coding properties and geometrical structure of the recorded neuronal population. In our case the animals actively sample the objects by moving the whiskers, and this can greatly complicate the study of the geometry of the neural representations. For example, some of the quantities used in the past to characterize the representational geometry like the shattering dimensionality require lengthy calculations, involving a number of operations that scales exponentially with the number of experimental conditions. This becomes prohibitive in an experiment like the one we analyzed where we need to consider complex spatio-temporal patterns to characterize the sensory input. Our method can still inform us about the geometry of the representations (it considers the activity of a population of neurons), but with a more favorable scaling. For all these reasons we believe that the method we propose here can be applied to a number of more natural tasks which are becoming progressively more feasible in the neuroscience community.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41593-022-01237-9.

References

1. Johansson, R. S. & Flanagan, J. R. Coding and use of tactile signals from the fingertips in object manipulation tasks. *Nat. Rev. Neurosci.* **10**, 345–359 (2009).
2. Bensmaia, S. J., Tyler, D. J. & Micera, S. Restoration of sensory information via bionic hands. Nature Biomedical Engineering1-13 (2020).
3. Davidson, P. W. Haptic judgments of curvature by blind and sighted humans. *J. Exp. Psychol.* **93**, 43 (1972).
4. Lederman, S. J. & Klatzky, R. L. Hand movements: A window into haptic object recognition. *Cogn. Psychol.* **19**, 342–368 (1987).
5. Rodgers, C. C. et al. Sensorimotor strategies and neuronal representations for shape discrimination. *Neuron* **109**, 2308–2325 (2021).
6. Rigotti, M. et al. The importance of mixed selectivity in complex cognitive tasks. *Nature* **497**, 585–590 (2013).
7. Bernardi, S. et al. The geometry of abstraction in the hippocampus and prefrontal cortex. *Cell* **183**, 954–967 (2020).
8. Haxby, J. V. et al. A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron* **72**, 404–416 (2011).
9. Guntupalli, J. S. et al. A model of representational spaces in human cortex. *Cereb. cortex* **26**, 2919–2934 (2016).
10. Chung, S. & Abbott, L. Neural population geometry: An approach for understanding biological and artificial neural networks. *Curr. Opin. Neurobiol.* **70**, 137–144 (2021).
11. Saxena, S. & Cunningham, J. P. Towards the neural population doctrine. *Curr. Opin. Neurobiol.* **55**, 103–111 (2019).

12. Fusi, S., Miller, E. K. & Rigotti, M. Why neurons mix: high dimensionality for higher cognition. *Curr. Opin. Neurobiol.* **37**, 66–74 (2016).
13. Bengio, Y., Courville, A. & Vincent, P. Representation learning: A review and new perspectives. *IEEE Trans. pattern Anal. Mach. Intell.* **35**, 1798–1828 (2013).
14. Higgins, I. et al. *β*-VAE: Learning basic visual concepts with a constrained variational framework. International Conference on Learning Representations (ICLR) (2017).
15. Higgins, I., Racanière, S. & Rezende, D. Symmetry-based representations for artificial and biological general intelligence. Frontiers in Computational Neuroscience28 (2022).
16. Higgins, I. et al. Unsupervised deep learning identifies semantic disentanglement in single inferotemporal neurons. *arXiv preprint arXiv:2006.14304* (2020).
17. Boyle, L., Posani, L., Irfan, S., Siegelbaum, S. A. & Fusi, S. The geometry of hippocampal ca2 representations enables abstract coding of social familiarity and identity. *bioRxiv* (2022).
18. Raposo, D., Kaufman, M. T. & Churchland, A. K. A category-free neural population supports evolving demands during decision-making. *Nat. Neurosci.* **17**, 1784 (2014).
19. Chang, L. & Tsao, D. Y. The code for facial identity in the primate brain. *Cell* **169**, 1013–1028 (2017).
20. Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M. & Schiele, B. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *European Conference on Computer Vision*, 34-50 (Springer, 2016).
21. Pishchulin, L. et al. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4929-4937 (2016).
22. Mathis, A. et al. Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. *Nat. Neurosci.* **21**, 1281–1289 (2018).
23. Alain, G. & Bengio, Y. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644* (2016).
24. Buonomano, D. V. & Maass, W. State-dependent computations: spatiotemporal processing in cortical networks. *Nat. Rev. Neurosci.* **10**, 113–125 (2009).
25. Yamins, D. L. et al. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl Acad. Sci.* **111**, 8619–8624 (2014).
26. Yamins, D. L. & DiCarlo, J. J. Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* **19**, 356–365 (2016).
27. Panichello, M. F. & Buschman, T. J. Shared mechanisms underlie the control of working memory and attention. *Nature* **592**, 601–605 (2021).
28. She, L., Benna, M. K., Shi, Y., Fusi, S. & Tsao, D. Y. The neural code for face memory. *bioRxiv* (2021).
29. Elsayed, G. F., Lara, A. H., Kaufman, M. T., Churchland, M. M. & Cunningham, J. P. Reorganization between preparatory and movement population responses in motor cortex. *Nat. Commun.* **7**, 1–15 (2016).
30. Stringer, C., Pachitariu, M., Steinmetz, N., Carandini, M. & Harris, K. D. High-dimensional geometry of population responses in visual cortex. *Nature* **571**, 361–365 (2019).
31. Gao, P. & Ganguli, S. On simplicity and complexity in the brave new world of large-scale neuroscience. *Curr. Opin. Neurobiol.* **32**, 148–155 (2015).
32. Lindsay, G. W., Rigotti, M., Warden, M. R., Miller, E. K. & Fusi, S. Hebbian learning in a random network captures selectivity properties of the prefrontal cortex. *J. Neurosci.* **37**, 11021–11036 (2017).
33. Dang, W., Jaffe, R. J., Qi, X.-L. & Constantinidis, C. Emergence of non-linear mixed selectivity in prefrontal cortex after training. Journal of Neuroscience (2021).
34. Meshulam, L., Gauthier, J. L., Brody, C. D., Tank, D. W. & Bialek, W. Collective behavior of place and non-place neurons in the hippocampal network. *Neuron* **96**, 1178–1191 (2017).
35. Nogueira, R. et al. The effects of population tuning and trial-by-trial variability on information encoding and behavior. *J. Neurosci.* **40**, 1066–1083 (2020).
36. Stefanini, F. et al. A distributed neural code in the dentate gyrus and in ca1. *Neuron* **107**, 703–716 (2020).
37. Valente, M. et al. Correlations enhance the behavioral readout of neural population activity in association cortex. *Nat. Neurosci.* **24**, 975–986 (2021).
38. Frost, N. A., Haggart, A. & Sohal, V. S. Dynamic patterns of correlated activity in the prefrontal cortex encode information about social behavior. *PLoS Biol.* **19**, e3001235 (2021).
39. Hirokawa, J., Vaughan, A., Masset, P., Ott, T. & Kepecs, A. Frontal cortex neuron types categorically encode single decision variables. *Nature* **576**, 446–451 (2019).
40. Zhang, W. & Bruno, R. M. High-order thalamic inputs to primary somatosensory cortex are stronger and longer lasting than cortical inputs. *Elife* **8**, e44158 (2019).
41. Manita, S. et al. A top-down cortical circuit for accurate sensory perception. *Neuron* **86**, 1304–1316 (2015).
42. Banerjee, A. et al. Value-guided remapping of sensory cortex by lateral orbitofrontal cortex. *Nature* **585**, 245–250 (2020).
43. Moore, J. D., Mercer Lindsay, N., Deschênes, M. & Kleinfeld, D. Vibrissa self-motion and touch are reliably encoded along the same somatosensory pathway from brainstem through thalamus. *PLoS Biol.* **13**, e1002253 (2015).
44. Ranganathan, G. N. et al. Active dendritic integration and mixed neocortical network representations during an adaptive sensing behavior. *Nat. Neurosci.* **21**, 1583–1590 (2018).
45. Gulli, R. A. et al. Context-dependent representations of objects and space in the primate hippocampus during virtual navigation. *Nat. Neurosci.* **23**, 103–112 (2020).
46. Roussy, M. et al. Ketamine disrupts naturalistic coding of working memory in primate lateral prefrontal cortex networks. Molecular Psychiatry1-16 (2021).
47. Nelson, M. E. & MacIver, M. A. Sensory acquisition in active sensing systems. *J. Comp. Physiol. A* **192**, 573–586 (2006).
48. Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A. & Poeppel, D. Neuroscience needs behavior: correcting a reductionist bias. *Neuron* **93**, 480–490 (2017).
49. Nogueira, R. et al. Lateral orbitofrontal cortex anticipates choices and integrates prior with current information. *Nat. Commun.* **8**, 1–13 (2017).

Methods

Behavioral task and recordings

This experiment has been described in detail[5]. Here we provide a brief summary of the behavioral setup and data acquisition. Mice in our colony were continuously backcrossed to C57BL/6J wild-type mice from Jackson Laboratories, and all mice reported here were bred in-house. They were kept on a 12 hour, non-reversed light cycle and were typically tested during the day. We used males and females arbitrarily and in roughly equal proportion (6 females and 4 males).

Ten head-fixed mice were trained to perform a shape discrimination task in the dark by making contacts with whiskers C0, C1, C2 and C3 (Fig. 1). On each trial, either a concave or convex shape (custom designed and 3D-printed) was moved within reach of the mouse's whiskers with a linear actuator. All trials started at $t = -2$ seconds when the shapes started moving. Shapes were moved with the same speed in all trials and they could stop at three different locations: far, medium and close, which occurred at $t = -0.9$, -0.7 and -0.5 seconds, respectively. Including three different final positions was important to prevent animals from using simpler strategies based on distance to the shape and to force them to integrate contacts across whiskers and time to perform the discrimination task. All trials had a fixed duration of 2 seconds. At $t = 0$ the response window opened and mice had to report their choice by licking either on the left or right lickpipe for concave and convex shapes, respectively. Licks were monitored by infrared beams or capacitive touch sensors. Even though mice were free to lick throughout the trial, the choice on each trial was determined by the side of the first lick after the response window opened ($t = 0$). The behavioral performance for each animal was determined as the percentage of correct choices.

Whisker and shape position were recorded with a high-speed camera (200 frames/second). Whisker tracking was based on a modified version of 'pose-tensorflow' package[20,21], which is the 'feature detector' network used in the first version of DeepLabCut[22]. The network was trained to track eight equally spaced joints per whisker. Whisker contacts were identified when the distance between the tip of a particular whisker and the edge of the shape was smaller than 10 pixels. Angular position was defined as the angle of the line between the tip and the base of each whisker.

Populations of individual neurons (single units) were simultaneously recorded in mouse somatosensory cortex (S1) during the whisker-based shape discrimination task (Fig. 4). Mice were implanted with a custom-designed stainless steel headplate between postnatal day 90 and 180. We removed the scalp and fascia covering the dorsal surface of the skull and positioned the headplate over the skull and affixed it. To permit electrophysiological recording we used a dental drill to thin the cement and skull over S1, rendering it optically transparent, and coated it with cyanoacrylate glue. We used intrinsic optical signal imaging to locate the cortical columns of the barrel field corresponding to the whiskers on the face. We then used a scalpel to cut a small craniotomy directly over the columns of interest. Between recording sessions, the craniotomy was sealed with silicone gel. To record neural activity, we head-fixed the mouse in the behavioral arena. We lowered an electrode array using a motorized micromanipulator. We used an OpenEphys acquisition system with two digital headstages to record 64 channels of neural data at 30 kHz at the widest possible bandwidth (1 Hz to 7.5 kHz). We used KiloSort[50] to detect spikes and to assign them to putative single units. We identified inhibitory neurons from their waveform half-width, that is the time between maximum negativity and return to baseline on the channel where this waveform had highest power. Neurons with a half-width below 0.3 ms were deemed narrow-spiking and putatively inhibitory. We measured the laminar location of each neuron based on the manipulator depth and the channel on which the waveform had greatest RMS power.

A total of 584 neurons were recorded from 23 sessions that included 7 different mice. The mean number of simultaneously recorded neurons was 25.4. From these 584 neurons, 68 were recorded in layer 2/3, 157 in layer 4, 249 in layer 5 and 96 in layer 6. Also, from the total number of neurons 16% were categorized as inhibitory and 84% as excitatory neurons. All experiments were conducted under the supervision and approval of the Columbia University Institutional Animal Care and Use Committee.

Decoding Behavior

On each trial, we built a matrix that contained behaviorally relevant variables through time. In the following, we will refer to this matrix as the spatio-temporal whisking pattern gathered by the behaving mice. We used 20 time bins per feature after dividing 2 seconds into time bins of 100 ms. For whiskers C0, C1, C2 and C3 we included number of contacts and angle of contact (Fig. 2a), since these were shown to be the most informative whisker features for both decoding shape and lick side[5]. In main text and figures, we will use C0, C1, C2 and C3 when referring to whisker identity, and C_0, C_1, C_2 and C_3 when referring to the contacts made by each of these whiskers. The total amount of features on each trial was 160, 8 whisker features (contacts and angle of contacts for each whisker) times 20 time bins. All features for each individual session were normalized to null mean and unit standard deviation. For each mouse, we concatenated all recording sessions into a single super-session, which significantly increased the number of trials used to fit each model. Trials that did not register any lick within the first 500 ms after response time ($t = 0$) were discarded from the analysis. In total we used 10 mice, with a mean of 1266 trials per mouse (super-sessions). All analysis were performed with custom written python and pytorch scripts.

We decoded the identity of the presented shape (stimulus; green) or lick side (choice; blue) on a trial-by-trial basis. In Fig. 2b,c the model was trained after balancing correct and incorrect trials and the quantity to be decoded (stimulus or choice). For instance, when the decoder was trained to predict stimulus identity, we randomly sampled (without replacement) trials from the train set such that correct, incorrect, concave shape and convex shape trials were equally populated. By balancing correct and incorrect trials we ensured that stimulus and choice were uncorrelated. Otherwise, information about choice would have been artificially boosted by stimulus information. We refer to this balancing as decorrelation, and it was repeated 10 times. In Fig. 2b,c the data was split into train, test and validation (2 nested KFold, $k = 4$) in order to optimize the $l2$ regularization strength over the range $[10^{-7}, 10^{3}]$ (20 steps log-evenly spaced). The reported decoding performances corresponds to the mean across cross-validations and decorrelations on the validation set after optimizing regularization strength on the test set. In Fig. 2b,c we used logistic regression (sklearn).

For Fig. 2b we gradually increased the complexity of the behavioral features to decode stimulus and choice by considering: sum of all contacts across time and whiskers (Sum all), sum of all contacts across whiskers (Sum whisker), sum all contacts across time (Sum time), all contacts across whiskers and time (All contacts) and all contacts and angles of contact across whiskers and time (Contacts + Angle). The inset in 2b corresponds to the weights of the classifier trained after summing contacts across time. Information about stimulus and choice across time was calculated by linearly decoding the cumulative number of features (contacts and angle of contacts) up to that particular time (Fig. 2c).

We analyzed the complexity of the whisker-based shape discrimination task by decoding the spatio-temporal whisking pattern with different decoding models (multilayer feedforward networks with 0, 1, 2 or 3 hidden layers of 100 ReLU units). In the following, because a feedforward network with 0 hidden layers is equivalent to a linear classifier, we will use these two terms synonymously. The models in Fig. 2d were trained and tested following the same steps than for Figs. 2b,c. However, instead of using logistic regression (sklearn) we fit the feedforward networks with stochastic gradient descent (batch size 64, 100 epochs) on pytorch, where the optimal learning rate η was

obtained following the same procedure than for the regularization strength ([10^{-7}, 1], 20 steps log-evenly spaced). We used *cross-entropy* loss and ADAM optimizer. The reported decoding performances on correct and error trials (Extended Data Fig. S2a) correspond to the mean performance on the validation set (see above) after splitting trials into correct and error. In Extended Data Fig. S3c we followed the exact same procedure using 20 and 200 units in the hidden layers, which produced equivalent results to Fig. 2d. We also used recurrent neural networks (RNNs) as decoding models in Extended Data Fig. S3a with 5, 10, 60 and 100 recurrent units. Instead of fitting the classifier with the whole spatio-temporal pattern of whisker contacts and angles of contacts (8 features × 20 time steps = 160) on each trial, the input to the RNNs on each time step on each trial consisted of the vector of contacts and angle of contact for whiskers C0, C1, C2 and C3 on that time step and trial. We used backpropagation through time to fit the RNNs (ADAM). The optimal learning rate and regularization was obtained in the same way as for the feedforward decoding models. RNNs as decoding models showed similar or lower decoding performance than feedforward decoding models.

In Extended Data Fig. S2c we performed the same analysis as in Fig. 2d but we only used trials that corresponded to the three different stopping distances separately ("Far", "Medium", and "Close"). We also analyzed datasets where three mice were presented with flatter shapes instead of the standard ones used in whisker-based discrimination task (Extended Data Fig. S2d). One of these mice was also part of the pool of 10 mice used in all the presented behavioral analysis whereas the other two were only used to analyze the flatter version of the task. Mice and decoding models had lower performances for flatter shapes (Extended Data Fig. S2d, see also ref. 5).

In order to discard the possibility that the results in Fig. 2d were a consequence of a low number of trials, linear and non-linear decoders were also trained on synthetic tasks of whisker contacts with different levels of difficulty (Extended Data Fig. S2b). We created two *ad-hoc* tasks from the spatio-temporal whisking patterns gathered by the animals, the easy and the very complex tasks. Mice were never trained on these tasks, they correspond to tasks that have been defined on the whisker contact space for whiskers C1, C2 and C3 *a posteriori*. For each mouse we first summed contacts through time on each trial (total contact space). The easy task was defined by splitting the trials in the super-session into two linearly separable classes on the total contact space (orange vs green in Extended Data Fig. S2b). For the very complex task, trials were split into two non-linearly separable classes (3D-parity) also on the total contact space. In both tasks the contact space was first transformed by a unitary random rotation. Importantly, for both the easy and the very complex task, the two classes were equally populated. This was achieved by adding Gaussian noise (standard deviation of 0.1) on each whisker total counts on each trial so that a median split was uniquely defined. Therefore, each trial on a super-session was assigned either to class 1 or class 2 for the easy task, and either to class 1 or class 2 for the complex task. For each mouse, the easy or very complex task were performed by reading out the feature matrix that contained whiskers C0, C1, C2 and C3 contacts across time (20 time bins of 0.1 seconds, 80 features in total). The procedure for fitting the different models was the same as in Fig. 2d, with the only difference that we did not need to balance correct and error trials. The $l2$ regularization strength and the learning rate η exploration intervals were [10^{-6}, 1] and [10^{-4}, 1] respectively, log-evenly spaced in 10 steps. Unsurprisingly, linear and non-linear decoders performed equally well on simple integration tasks, whereas non-linear decoders were necessary to perform complex tasks that require non-linear integration of sensory evidence (Extended Data Fig. S2b).

Simulation of the whisker discrimination task

In order to gain additional insights of the whisker-based discrimination task, we built a simulation of the experiment and analyzed the simulated data like we did for the real data. On each trial we simulated the movement of three whiskers C1, C2 and C3. The size of the simulated box was 12x12 (a.u.), and the base of all three whiskers was placed at the origin of the two axes (0,0) (Fig. 3a). Like in the real experiment, each trial consisted of 2 seconds, and we simulated time steps of 0.1 seconds (20 steps in total), which was also the same time window used to analyze the real experiments. The angular position of whisker C1 (longest whisker, pale grey) at the beginning of each trial $\phi_{0,C1}$ was sampled from a von Mises distribution ($\mu = 0$, $\kappa = 10$) and the position of C2 (middle; grey) and C3 (shortest; black) were determined by adding 1/3 and 2/3 of a radian with respect to the position of C1, respectively. The angular position of whisker *Ci* on time step *t* was $\phi_{t,Ci} = \sin(\omega t + \phi_{0,C1} + i/3)$ where ω was drawn from a gaussian distribution ($\mu = 3$, $\sigma = 0.1$) on each trial. Hence, the position on the (x, y) plane of whisker *Ci* on each time step was $(l_x, l_y)_{t,Ci} = (L_i \cos(\phi_{t,Ci}), L_i \sin(\phi_{t,Ci}))$, where $L = \{10, 8.5, 7\}$ were the lengths of the three whiskers.

Shapes with different curvatures were modelled as circle segments with different radius. In Fig. 3a-c the shapes had a radius of $R = 11$, whereas in Fig. 3e (and Extended Data Fig. S4g) we used $R = 6$ (green) and $R = 50$ (orange) (see also Extended Data Fig. S4f for $R = \{6, 11, 22, 50\}$). On each time step the shape moved by 0.2 units towards the origin, and three different stopping times were used after the beginning of the trial: 9, 10 and 11 time steps. The wiggles on Fig. 3c were obtained by adding a sinusoidal function of amplitude 0.2 and frequency 10 to the shape in the y-axis direction.

In order to incorporate the flexibility of the real whiskers in our models, on each time step the tip of each whisker was obtained by adding gaussian noise ($\sigma = 0.3$) to the position of the rigid whisker (solid circles) (Fig. 3a). On each time step, we evaluated whether the noisy tip of the whisker made contact with the shape (red dot when contact occurred). Following the same approach we used for the real data, on each trial we constructed a matrix where each column corresponded to a time step (20 columns) and each row corresponded to contacts (0 no contact, 1 contact) and angle of contact for each whisker (6 rows in total). Each simulation consisted of 2000 trials. We also fit linear and non-linear classifiers (multilayer perceptrons with 0, 1, 2 and 3 hidden layers) to predict whether a particular trial corresponded to a concave or a convex shape or other variants of the task. For Fig. 3e and Extended Data Fig. S4g,h, the shapes were static throughout the trial.

We also simulated variants of the experiment in which we modified the size and position of the shapes, among others. In particular, we used more distant stopping locations (7, 11 and 13 time steps) (Extended Data Fig. S4a); smaller shapes (Extended Data Fig. S4b); further shapes (2 a.u. further) (Extended Data Fig. S4c); closer shapes (1 a.u. closer) (Extended Data Fig. S4d); and flatter shapes ($R = 24$) (Extended Data Fig. S4e). We also used rotated and flat shapes ($\theta = 0.1$ and $R = 100$) (Fig. 3d). In all cases linear and non-linear decoders produced qualitatively equivalent results, although performances were overall lower for flatter, further and smaller shapes.

Encoding Models

On each trial we built a matrix that contained all the experimental variables that we considered could affect the firing rate of S1 populations. We analyzed the time interval $t = [-2.1, 1.0]$ seconds in time bins of 100 ms, which spanned from the beginning of the trial to one second after response window opened (31 time steps per trial). The experimental variables used in the encoding models were: contacts, angle of contact and angular position of whiskers C0, C1, C2 and C3; lick side and lick rate; current and previous reward, stimulus, shape position and choice. We will refer to whisker and lick variables as *continuous-variables* and previous and current reward, stimulus, position and choice as *trial-variables*. Other features like force at the base of the whisker (whisker bending) or contact duration were not included in the model because they have been shown to be weakly encoded in S1 during the whisker-based shape discrimination task (see[5]). For each recording

session we concatenated all the time steps across all trials (Fig. 4c). On each time step S1 population activity was regressed against the current *continuous-variables* and up to five time steps backwards in time (500 ms = 5 steps × 100 ms). *Trial-variables* were arranged as indicator variables throughout the length of the trial. Population activity was regressed using a total of 70 *continuous-variables* (70 = 14 variables × 5 time steps) plus 248 *trial-variables* (248 = 8 variables × 31 time steps). Both neuronal activity and regressors were normalized to null mean and unit variance. Trials that did not register any lick within the first 500 ms after response time ($t = 0$) were discarded from the analysis. In total we used 23 recording sessions from 7 different mice, with 25.4 mean number of simultaneously recorded neurons and 4883 mean number of effective trials used to fit the models (trials × time steps).

We analyzed the encoding properties of populations of neurons in mouse S1 by regressing the neuronal activity against the experimental variables described above. Similar to behavior, we used different encoding models with different levels of flexibility (multilayer feedforward networks with 0, 1, 2 or 3 hidden layers of 100 ReLU units). Similar to classification, an encoding model with 0 hidden layers is equivalent to a linear regression. We fit the encoding models by minimizing the mean-squared-error (MSE-loss) between the predicted and the real firing rate (stochastic gradient descent, batch size 64, 100 epochs; Fig. 4). To validate our results with a different loss function, Poisson-loss was also used to fit the models, which produced qualitatively equivalent results (Extended Data Fig. S5f,g). The linear model can only implement pure and linear mixed selectivity, while encoding models that include at least one hidden layer can implement non-linear mixed selectivity[6,12]. On each recording session, models were fit by splitting the data into train, test and validation (2 nested KFold, $k = 4$). The partition was performed based on the real trials of the experiment so that time steps from the same trial were always grouped in the same partition. Otherwise, due to the correlation between the neuronal activity on consecutive time steps, performances on the validation set could have been artificially boosted. The optimal regularization strength $l2$ and learning rate η were obtained by identifying the values that produced the highest performance on the test set over the ranges $[10^{-7}, 10^{2}]$ and $[10^{-7}, 10^{-1}]$ respectively (20 steps log-evenly spaced). As goodness-of-fit for the different encoding models, we used the metric $R^2 = 1 - Loss/Variance$. The reported R^2 corresponded to the mean across cross-validations on the validation set after optimizing regularization strength and learning rate on the test set. As expected, when all models were tested on training data, more parameters entailed better firing rate prediction (Extended Data Fig. S5b). All the encoding models were implemented in pytorch and optimized with the ADAM algorithm. The reported performance on correct and error trials correspond to the mean *CV* R^2 on the validation set (see above) after splitting trials into correct and error (Extended Data Fig. S5c). Errorbars in Fig. 4 correspond to s.e.m. across recorded neurons. In Extended Data Fig. S3d, we followed the exact same procedure with 20 and 200 units in the hidden layers as encoding models, which produced equivalent results to Fig. 4. We also used recurrent neural networks (RNNs) as encoding models in Extended Data Fig. S3b with 5, 10, 60 and 100 recurrent units. Instead of fitting the encoding model with the current and the last 500ms of the whisking spatio-temporal and licking pattern, and task variables, the input to the RNNs on each time step consisted of the value of all these variables on the current time step. We used backpropagation through time to fit the RNNs (ADAM). The optimal learning rate and regularization was obtained in the same way as for the feedforward encoding models. RNNs as encoding models produced lower R^2 than feedforward encoding models.

In order to evaluate the individual contributions of regressor (or group of regressors) x_i to the predictability of the population's firing rate, we evaluated the quantity $\Delta R^2 = R^2_{Full} - R^2_{Reduced}$ (Fig. 5a), where R^2_{Full} corresponds to the performance of the full model and $R^2_{Reduced}$ corresponds to the performance of the model when regressor (or group of regressors) x_i is set to zero. For instance, in Fig. 5b ΔR^2 for *Contacts* was calculated by setting to zero the variables C_0, C_1, C_2, and C_3 for the current and up to five time steps in the past (4 whiskers × 5 time steps = 20 regressors). This method is preferred over re-training the whole model without regressor x_i because of the correlations between regressors x_i and x_j, so that we make sure that the reported contribution takes into account the correlation with the rest of regressors.

For each pair of regressors (or pairs of groups of regressors) x_i and x_j, we evaluated the pure non-linear interaction (contribution) to the encoding model by evaluating $\Delta R^2 - \Delta R^2_{Linear}$ (Fig. 6 and Extended Data Fig. S7a). Here ΔR^2 corresponds to the loss in predictive power for the non-linear model when both x_i and x_j are set to zero and ΔR^2_{Linear} is the equivalent for the linear encoding model. Because non-linear models also include the linear terms, subtracting the contribution from the pure linear model was necessary in order to isolate pure non-linear interactions.

Similar to the synthetic tasks presented in Extended Data Fig. S2b, we also created two synthetic tasks based on whisker contacts: the easy and the complex tasks. Additionally, to benchmark the ability to generalize, we evaluated the cross-condition generalization performance (CCGP). The easy and the complex tasks corresponded to a linear and an XOR task with respect to the contacts of pairs of whiskers, respectively, whereas the CCGP tested how well a linear classifier trained to perform a simple discrimination task on a set of trials would generalize to an unseen set of trials. Given that the encoding models were fit using 5 time steps (100 ms × 5 times steps = 500 ms) for all the *continuous-variables*, for each time step we first summed the number of contacts across the current and previous four time steps for each whisker independently. Gaussian white noise was also introduced in all whisker contacts to obtain a well defined median to create the different tasks (standard deviation of 10^{-3}). All tasks were defined as 2D tasks on the summed number of contacts across 5 time steps, so they were constructed from the three different pairs that could be built from the set {C1,C2,C3}: (C1,C2), (C1,C3) and (C2,C3). Importantly, all the time steps in which no whisker contacts were registered for the sum across these 5 time steps were discarded from this analysis. On the easy task, the coloring of the different regions in the whisker contact space (for example, C_1 vs C_3) was defined by a linear boundary, whereas for the complex task it corresponded to an XOR task (Fig. 7a). In both cases the task boundaries were obtained by performing a random unitary rotation on the whisker contact space and splitting each dimension with respect to the median. For the generalization benchmark (CCGP), the process was slightly different. By splitting all the trials into low and high number of contacts for each whisker, we created four different conditions. Cross-condition generalization performance (CCGP)[7] was evaluated as the performance of a linear classifier to discriminate between low and high number of contacts for whisker i when trained only on low contacts for whisker j and tested on high contacts on whisker j. For instance, for the (C_1,C_3) pair, a linear classifier was trained to discriminate between low and high number of C1 contacts using only trials of low C3 contacts and tested on high number of C3 contacts (and viceversa). CCGP for whisker C1 corresponded to the mean across training on C2 low and testing on C2 high, training on C2 high and testing on C2 low, and the same process conditioning on whisker C3. CCGPs for whiskers C2 and C3 were evaluated equivalently but conditioning on C1,C3 and C1,C2, respectively.

Once the three tasks were defined for each time step, we generated surrogate representations for each encoding model by introducing the pair of whisker contact variables into the different encoding models. This procedure was only performed on the validation partition. For instance, for the pair (C1,C3) we generated surrogate activity on each time step by introducing in the different encoding models only the experimental variables contacts C1 and C3 for the current and previous four time steps. From these surrogate representations, linear classifiers (cross-validated logistic-regression) were fit to perform all three tasks. The reported performance in Fig. 7d corresponds to the

mean performance across cross-validations of the encoding models and pairs of regressors.

To evaluate whether the performance on the easy and complex tasks were significantly above chance, we compared them with their null distributions. For each element of the null distribution we shuffled the class labels for each pattern of surrogate activity and fit linear classifiers to perform the easy and complex tasks as described above. Each element of the null distribution corresponded to the mean across cross-validations of the encoding models and pairs of regressors. The null distribution was obtained by repeating this process 1000 times. To evaluate whether the reported CCGPs for the surrogate patterns were significantly above chance, we compared it with the null distribution. We followed a similar procedure to that described in ref. 7. In short, each experimental condition was randomly rotated in the surrogate activity space by shuffling each trial with respect to the identity of the neurons. The same random shuffle was used for all trials in a given condition. This procedure destroys the geometrical structure of the representation but approximately maintains the distance between the different conditions. The null distribution was obtained by repeating this process 1000 times. Performance on the easy and complex tasks and CCGP in Fig. 7d are significantly above from chance ($P < 0.05$), besides for the complex task using surrogate populations generated with the linear encoding models.

We also evaluated contact information for the different whiskers and columns by decoding whether a set of trials corresponded to a high or low number of contacts for each whisker using neurons recorded from only a particular column of S1 (Extended Data Fig. S9a). CCGPs for the different columns and whiskers was evaluated following the same process described above (Extended Data Fig. S9a). Given that different population sizes were recorded for the different columns of the S1, in order to compare information across columns, all the performances in Extended Data Figs. S9a were obtained using 10 neurons, which is the smallest number of simultaneously recorded neurons across all columns. Errorbars in Fig. 7 and Extended Data Fig. S9 correspond to s.e.m. across recording sessions.

Population Decoding

Populations of mouse S1 neurons were recorded during the whisker-based shape discrimination task. Linear classifiers were fit to predict different experimental variables on a trial-by-trial basis. Information about a particular variable for a given time step was calculated using the entire population activity from the beginning of the trial to that particular moment (Fig. 4b). Time bins of 200 ms were used and population activity was normalized to null mean and unit variance. Trials that did not register any lick within the first 500 ms after response time ($t = 0$) were discarded from the analysis. The mean number of simultaneously recorded neurons and trials per session was 25.4 and 157.5, respectively. In total 23 recording sessions from 7 different mice were analyzed. In all panels the data was split into train, test and validation (2 nested KFold, $k = 4$) in order to optimize the $l2$ regularization strength over the range $[10^{-4}, 10^{4}]$ (10 steps log-evenly spaced). In all cases, logistic regression was used as our linear classification model (sklearn).

Shape identity (stimulus; green) and lick side (choice; blue) were predicted on each trial by reading out the population activity (left panel in Fig. 4b). Similar to decoding from the spatio-temporal pattern of whisker features, the classifiers were trained after balancing correct and incorrect trials and the quantity to be decoded (stimulus or choice). We refer to this balancing as decorrelation, and it was repeated 10 times. The reported decoding performances correspond to the mean across cross-validations and decorrelations on the validation set after optimizing regularization strength on the test set. From population activity we also decoded whether a particular trial corresponded to a high or low number of contacts for the different whiskers (right panel in Fig. 4b). For each whisker we summed the total number of contacts made up to a particular point in time and labeled each trial according to whether it was below or above the median number of contacts. Gaussian noise was added in all trials (standard deviation of 0.1) to obtain a unique median. The reported decoding performances correspond to the mean across cross-validations on the validation set after optimizing regularization strength on the test set.

Populations of recorded neurons were also used to perform the easy and the complex tasks and the generalization benchmark (Fig. 7c) (see previous section). From all the neuronal recordings, pseudopopulations of neurons were constructed and linear classifiers (logistic regression) were fit to perform these three tasks. To define the easy and complex tasks and the generalization benchmark (CCGP) on each recording session, we first summed the number of contacts throughout the entire trial (2 sec). Similar to the equivalent analysis on surrogate representations (see previous section), all three analysis were defined with respect to pairs of whisker contacts variables: (C_1,C_2), (C_1,C_3) and (C_2,C_3). For the easy and complex tasks, a random unitary rotation was performed on the whisker contact space for a given pair and all those trials that did not include whisker contacts were discarded from the analysis. Four experimental conditions corresponding to low and high number of contacts for two whisker variables were defined. From each experimental condition, 200 trials were sub-sampled with replacement for both the train and test set. The simultaneously recorded activity of S1 neurons across a particular trial was flattened with respect to the time axis (200ms time bins; 10 time bins per trial). For a given recording session we constructed the train and test matrices with dimensions 800 (200 trials per condition × 4 experimental conditions) and number of neurons × 10 time bins. It is important to note that with this procedure the train and test matrix did not share any trials, which would artificially boost the estimated performance for the different tasks. For each recording session we repeated this procedure and stacked the different train and test matrices along the dimension of neurons. A total of 584 neurons were recorded across all sessions, which produced a train and a test matrix with 5840 columns (584 × 10 time bins). Two different linear classifiers were fit on the train matrix and tested on the test matrix to perform the easy and the complex task, respectively. The reported performances on the easy and complex tasks in Fig. 7c corresponds to the mean across pairs of whiskers and 10 iterations of this process. To evaluate whether the performance on the easy and complex tasks were significantly above chance, we compared them to the null distribution. For each element of the null distribution we shuffled the class labels for each recorded pattern, built the train and test pseudopopulation matrices, and fit linear classifiers to perform the easy and complex tasks as described above. Each element of the null distribution corresponded to the mean across pairs of whiskers and 10 iterations of this process with the same shuffled labels. The null distribution was obtained by repeating this process 1000 times. Performance on the easy and complex tasks in Fig. 7c are significantly above chance ($P < 0.05$).

In order to evaluate the generalization properties of the recorded neurons (CCGP), we proceeded in a similar way but we worked on the original whisker contact space instead (no unitary rotation). Also, given that the cross-validation is performed across conditions when evaluating the CCGP, only one matrix of pseudopopulation activity was constructed by sub-sampling with replacement from each experimental condition (200 trials per condition). Similarly, the reported CCGP in Fig 7c corresponds to the mean across 10 iterations of this process. To evaluate whether the reported CCGPs for the real recordings were significantly above chance, for each panel we constructed a null-hypothesis distribution and evaluated the probability of obtaining the real CCGP when sampling from it. We followed the same procedure described in[7]. In short, each experimental condition was randomly rotated in the activity space by shuffling each trial with respect to the identity of the neurons. The same random shuffle was used for all trials in a given condition. This procedure destroys the geometrical structure of the

representation but approximately maintains the distance between the different conditions. We performed 1000 iterations and computed the probability of obtaining the real CCGPs. In Extended Data Fig. S9b, we used pseudopopulations of neurons from specific columns to decode high or low number of contacts for whiskers *C*1, *C*2, and *C*3 (top), as well as the generalization benchmark (CCGP) (bottom). In both cases, we followed the same procedures described above.

Recurrent Neural Networks

Recurrent neural networks (RNNs) were trained to perform a task similar to the whisker-based shape discrimination task. The recurrent network consisted of 60 ReLU units whose activity at time t ($\mathbf{h}_t$) was determined by the following equation:

$$\mathbf{h}_t = \phi\left(J_{rec}\mathbf{h}_{t-1} + J_{in}\mathbf{x}_t + \sigma\boldsymbol{\xi}_t\right), \quad (1)$$

where $\phi()$ is the ReLU non-linearity, $\boldsymbol{\xi}_t$ is independent and unitary Gaussian noise on each time step and σ is the strength of this noise ($\sigma = 1$ in all our units).

The stimulus $\mathbf{x}_t$ consisted of three channels that on each time step could be either 0 or 1, an artificial analogy of whiskers C1, C2 and C3 making contacts or not. On each trial, each input channel corresponded to a random realization of a *Bernoulli* process (T time steps) with two possible underlying mean values λ_{low} or λ_{high}. This made a total of 8 different experimental conditions (2 conditions per channel and 3 channels) (Fig. 8). From these 8 experimental conditions we defined three different tasks, the easy, the complex and the very complex task. For Fig. 8 we only show the easy and complex tasks, see Extended Data Fig. S10 for results on the very complex task. For all tasks, the input information was transformed by an unitary random rotation (same rotation in all time steps and trials). We fit a different RNN for each task, and in each RNN input, recurrent and output weights were trained. The easy task was defined as a task that linearly separated the 8 experimental conditions into 2 groups of 4 (left panel in Extended Data Fig. S10); the complex task corresponded to an XOR with respect to C1 and C2 (middle panel in Extended Data Fig. S10); and the very complex task was defined as a 3D-parity with respect to all channels C1, C2 and C3 simultaneously (right panel in Extended Data Fig. S10). Given 8 experimental conditions, there were 3 different easy tasks: separation with respect to the C1 axis only (easy task 1); C2 axis only (easy task 2); and C3 axis only (easy task 3). There were also 3 different complex tasks: separation with respect to a 2D-XOR on (C1,C2) (complex task 1); on (C1,C3) (complex task 2); and on (C2,C3) (complex task 3). There was only one very complex task, a 3D-Parity task with respect to all channels. In Fig. 8, easy task, easy other, and complex task corresponded to easy task 1, the mean across easy tasks 2 and 3, and complex task 1, respectively.

To recreate the experimental conditions, inputs lasted for 20 time steps but a random delay of $\Delta t = [0, 9]$ time steps was introduced at the beginning of each trial. All networks were trained to make a decision at $T = 20$. The three networks were trained on datasets of 400 trials per experimental condition and for all channels $\lambda_{low} = 0$ and $\lambda_{high} = 1$. We used *cross-entropy* as loss function, the $l2$ regularization strength was set to 10^{-3} and the learning rate $\eta = 0.005$. We used the ADAM optimizer, batches of 20 trials and as many epochs as necessary to reach 10^{-3} error on the loss function (~ 10 epochs for the easy task, ~ 20 for the complex, and ~ 50 epochs for the very complex task). Once trained, networks were tested on 40 trials per experimental condition and for all channels $\lambda_{low} = 0.35$ and $\lambda_{high} = 0.65$ for the easy task, $\lambda_{low} = 0.3$ and $\lambda_{high} = 0.7$ for the complex task and $\lambda_{low} = 0.23$ and $\lambda_{high} = 0.77$ for the very complex task.

For each network, input, recurrent and output weights were learned. Additionally, for each network, recurrent and input weights were frozen and readout weights for the other tasks were also trained on the activity of the artificial units (logistic regression). For instance, for the network trained on the easy task (easy task 1) in Fig. 8b, all learnable weights were optimized for the easy task using backpropagation through time (dark brown). However, additional readout weights on the artificial units' activity were also trained for the orthogonal other easy task (easy tasks 2 and 3; pale brown), and the complex task (complex task 1; black). These additional readout weights were trained on the train set at decision time ($T = 20$) and tested on the test set on all time steps. In Extended Data Fig. S10 we show the performance curves for additional readout weights when trained on all tasks (easy and complex tasks 1,2,3 and very complex task). For Figs. 8c,e,g all weights were trained to perform complex task 1 and additional readout weights on the artificial units were trained to perform the the easy, and the orthogonal easy (easy other) tasks.

We also evaluated the ability of each network to generalize to unseen experimental conditions by means of the cross-condition generalization performance (CCGP). A very similar procedure to Fig. 7 was used to evaluate CCGP for the three different RNNs. For instance, in Fig. 8f, a linear classifier was trained to perform the easy task (easy task 1; dark brown) by reading out the activity of the artificial units. The classifier was trained on the set of trials defined by easy task 2 = +1 and tested on the set of trials that defined easy task 2 = 0 (and vice-versa). The same procedure was followed for the set of trials defined by easy task 3 = +1 and tested on trials defined by easy task 3 = 0 (and vice-versa). The reported CCGP was the mean across these four procedures. For the rest of CCGPs, the same train-test procedure was followed as defined by the rest of orthogonal easy tasks (see Extended Data Fig. S10).

For each panel in Fig. 8, and Extended Data Fig. S10 we trained and tested 50 different networks and reported the mean performance across test sets. Each network was trained on a different random realization of the input and rotation. In Extended Data Fig. S10 the low, medium and high noise levels corresponded to ($\lambda_{low} = 0.23, \lambda_{high} = 0.77$), ($\lambda_{low} = 0.3, \lambda_{high} = 0.7$) and ($\lambda_{low} = 0.35, \lambda_{high} = 0.65$), respectively.

We analyzed the complexity of the different tasks in the same way that we analyzed the complexity of the whisker-based shape discrimination task (see Fig. 2). For all trained networks, we used different classifiers with different levels of flexibility (multi-layer feedforward networks with 0, 1, 2 or 3 hidden layers of 100 ReLU units). These classifiers were trained to predict the output of the easy, the complex and the very complex task on a trial-by-trial basis by reading out the spatio-temporal pattern that was used as input to the networks (Extended Data Fig. S10d). The $l2$ regularization strength and the learning parameter η were optimized over the ranges $[10^{-8}, 10^{2}]$ and $[10^{-6}, 10^{0}]$ respectively (10 steps log-evenly spaced). The errorbars for each panel in Extended Data Fig. S10d correspond to the s.e.m. across 4 different network instances.

The encoding properties of the artificial units on each network were also analyzed in the same way we analyzed the population activity of mouse S1 neurons (Extended Data Fig. S10e), by fitting encoding models of feedforward networks of 0, 1, 2 and 3 hidden layers. In this case, the $l2$ regularization strength and the learning parameter η were optimized over the ranges $[10^{-8}, 10^{2}]$ and $[10^{-6}, 10^{0}]$ respectively (10 steps log-evenly spaced). The errorbars for each panel in Extended Data Fig. S10e correspond to the s.e.m. across 240 neurons (4 networks × 60 units).

RNNs with this architecture were also used to decode stimulus and choice on a trial-by-trial basis from the spatio-temporal whisking patterns as well as to predict S1 population activity. In particular, we used networks of 5, 10, 60 and 100 noise-less units, regularization strength of 0.001, and learning rate of 0.001. These RNNs produced decoding and encoding performances that were lower or similar to their feed-forward counterparts (Extended Data Fig. S3).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The datasets analyzed in this study have been deposited to Zenodo: https://doi.org/10.5281/zenodo.4743837.

Code availability

The code used for data acquisition and pre-processing is available at https://github.com/cxrodgers/Rodgers2021. The code used for all analyses in this study is available at https://github.com/ramonnogueira/TheGeometryOfS1.

References

50. Pachitariu, M., Steinmetz, N. A., Kadir, S. N., Carandini, M. & Harris, K. D. Fast and accurate spike sorting of high-channel count probes with kilosort. *Adv. neural Inf. Process. Syst.* **29**, 4448–4456 (2016).
51. Ashwood, Z. C. et al. Mice alternate between discrete strategies during perceptual decision-making. *Nat. Neurosci.* **25**, 201–212 (2022).
52. Calhoun, A. J., Pillow, J. W. & Murthy, M. Unsupervised identification of the internal states that shape natural behavior. *Nat. Neurosci.* **22**, 2040–2049 (2019).

Acknowledgements

We would like to thank the members of the Center for Theoretical Neuroscience, Marcus K. Benna and Mattia Rigotti for all their insightful comments and suggestions. Support was provided by NINDS/NIH (R01NS094659, R01NS069679, F32NS096819, and U01NS099726); NSF 1707398 (Neuronex); the Gatsby Charitable Foundation (GAT3708); the Simons Foundation; the Swartz Foundation; Northrop Grumman; a Kavli Institute for Brain Science postdoctoral fellowship (to CR); and a Brain and Behavior Research Foundation Young Investigator Award (to CR).

Author contributions

RN and SF conceived the project and the analytic approach. CR developed the behavior, videography and performed the electrophysiological recordings. RN analyzed the data. RN, CR, RB, and SF decided how to interpret the results. RN and SF wrote, and CR and RB edited the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

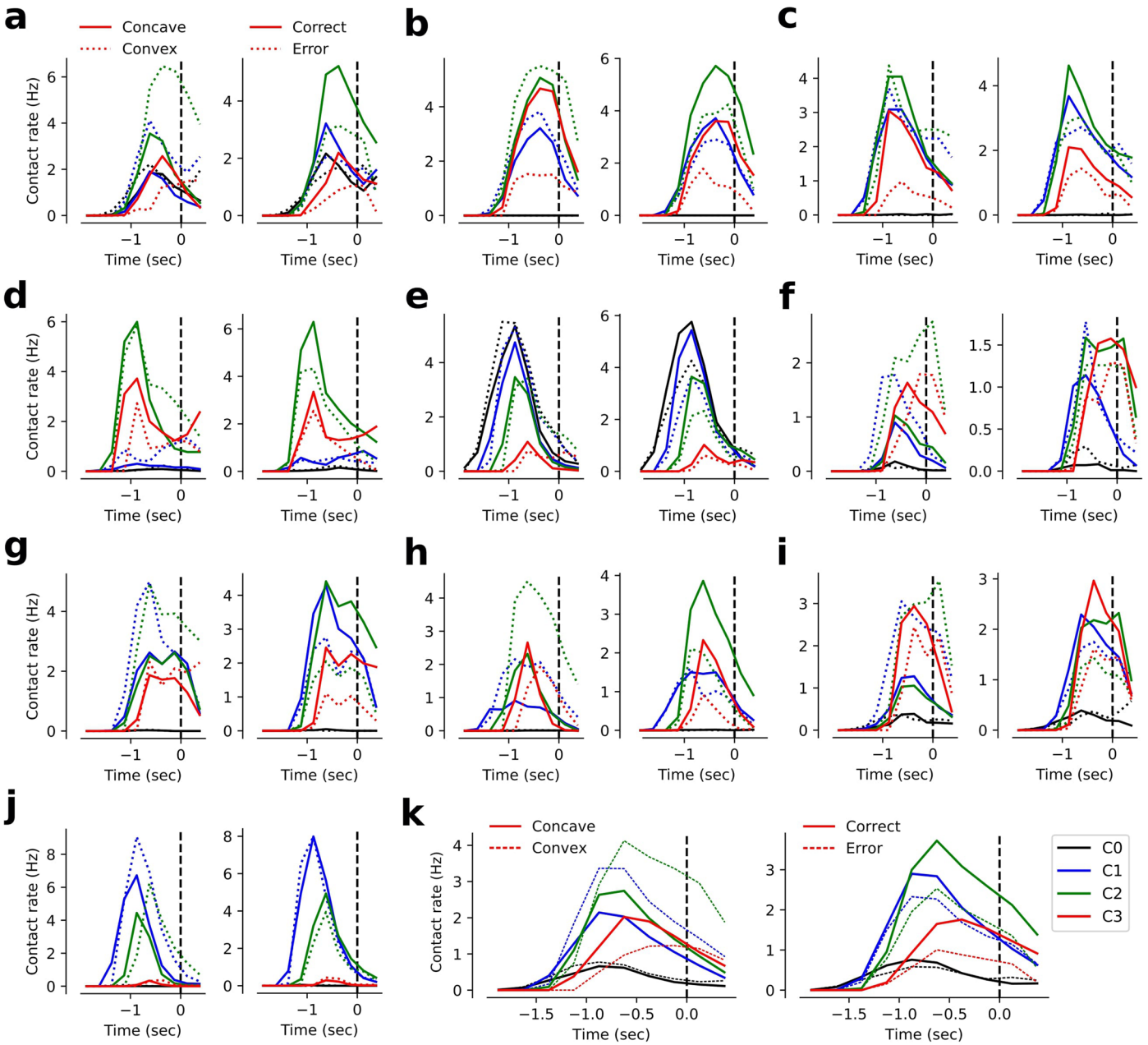
Extended data is available for this paper at https://doi.org/10.1038/s41593-022-01237-9.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41593-022-01237-9.

Correspondence and requests for materials should be addressed to Ramon Nogueira or Stefano Fusi.

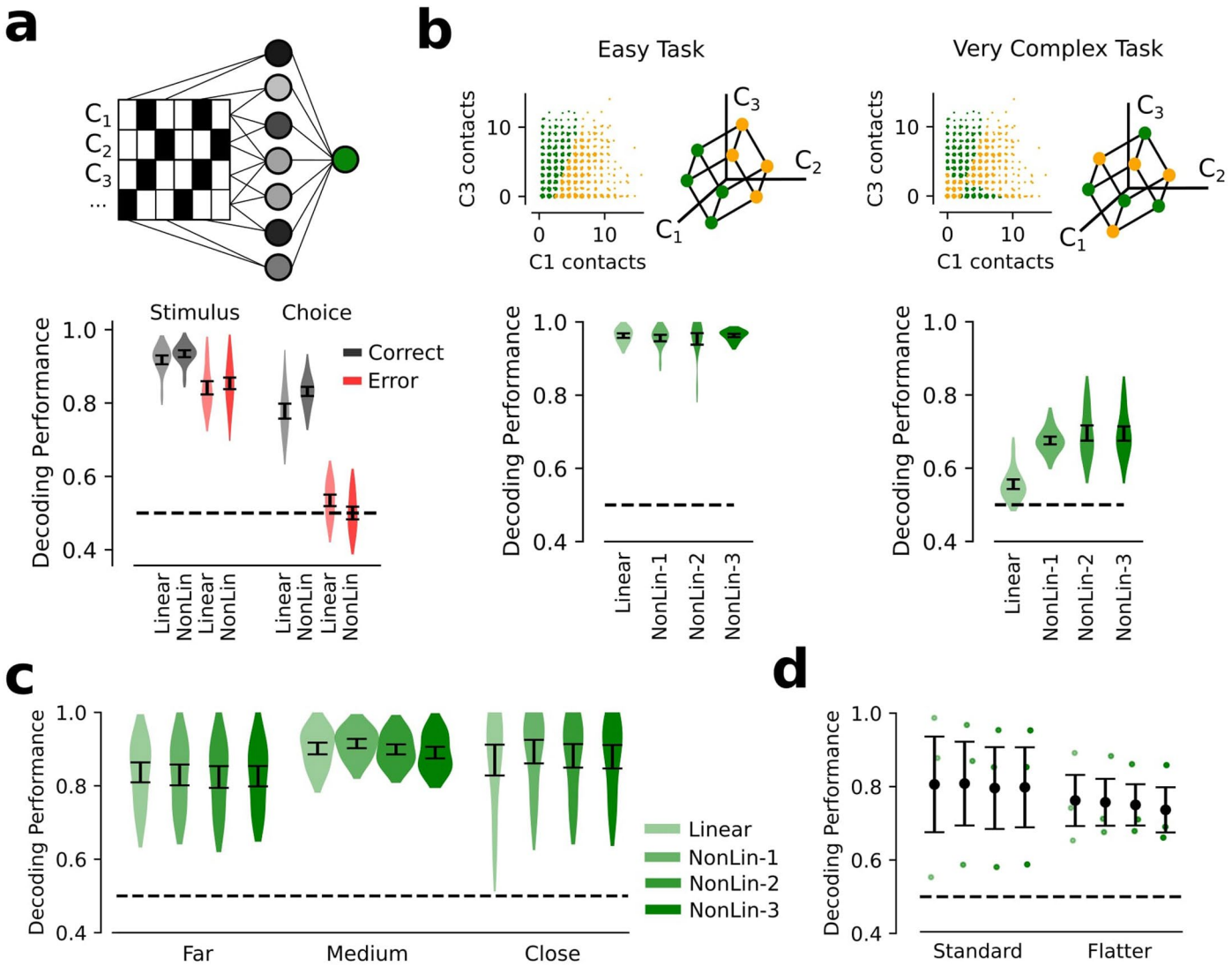
Peer review information *Nature Neuroscience* thanks the anonymous reviewers for their contribution to the peer review of this work.

https://doi.org/10.1038/s41593-022-01237-9



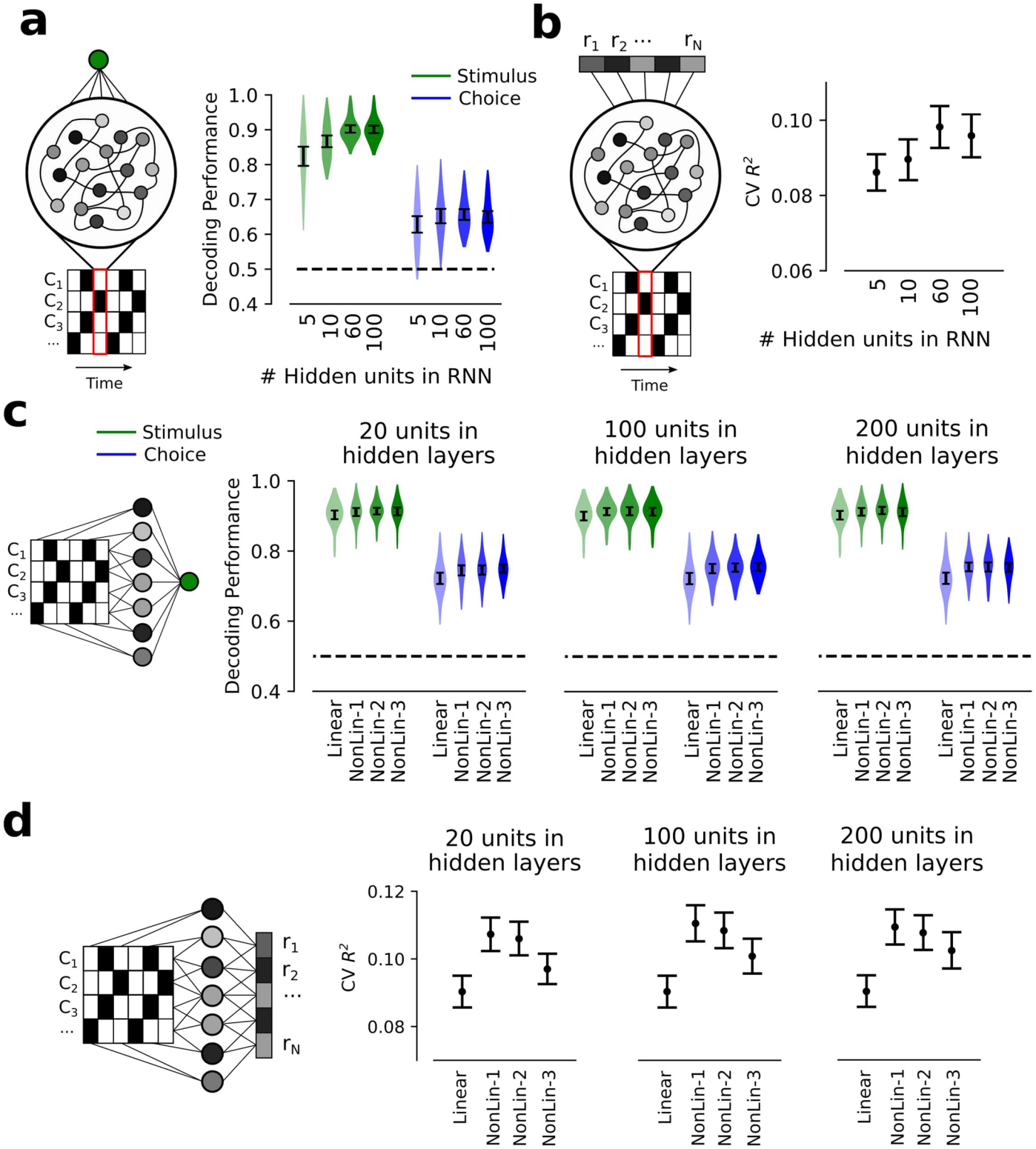
Extended Data Fig. 1 | Contact rate for convex and concave shapes and for correct and error trials for all mice. Contact rate (y-axis) as a function of time throughout the trial (x-axis) separately for convex and concave shapes (left), and for correct and error trials (right). Contacts were higher for convex than concave shapes for whisker C1 and C2, whereas whisker C3 showed the opposite trend. Contacts were higher for correct than error trials for all whiskers and animals. (**a-j**) Results for all mice. (**k**) Mean results across animals (n = 10 mice).

https://doi.org/10.1038/s41593-022-01237-9



Extended Data Fig. 2 | Linear and non-linear decoders performed similarly for correct and error trials, different stopping locations, and flatter shapes. (**a**) A multi-layer feedforward network model is trained to use the full spatio-temporal pattern of contacts and angle of contacts to predict the stimulus and the choice of the animal on a trial-by-trial basis (see Fig. 2). Models were trained using all trials and tested on correct (black) and incorrect (red) trials. Only linear and non-linear models with one hidden layer are shown. Stimulus decoding (left) produced higher decoding performance for correct trials than errors, probably due to the higher number of contacts made by mice on correct trials. On the contrary, correct trials conveyed much more information about animals' choice than incorrect trials (right). One possible explanation of these effects is that in approximately 60% of the trials animals make very accurate choices that are based on properly sampled sensory cues. In the other 40% of the trials, animals still sample information properly but their choice is inaccurate and based on a hidden variable we do not have access to[51,52]. (**b**) Decoding performance (y-axis) for the different decoders (x-axis) for the easy (linearly separable; left panel) and very complex (non-linearly separable, 3D-parity; right panel) tasks (see Methods). Non-linear cue integration is only advantageous when the task itself requires complex sensory integration across time and whiskers. (**c**) Linear and non-linear classifiers performed equally well on the shape discrimination task from the real spatio-temporal pattern of whisker contacts when conditioned on trials corresponding to far, medium and close stopping locations. Error bars in panels (a-c) correspond to s.e.m. across mice (n = 10). (**d**) Similar to behavior (see[5]), flatter shapes were more difficult to discriminate than the standard ones from the real spatio-temporal pattern of whisker contacts. Errorbars correspond to s.e.m. across mice that were presented both standard and flatter shapes (n = 3).

https://doi.org/10.1038/s41593-022-01237-9

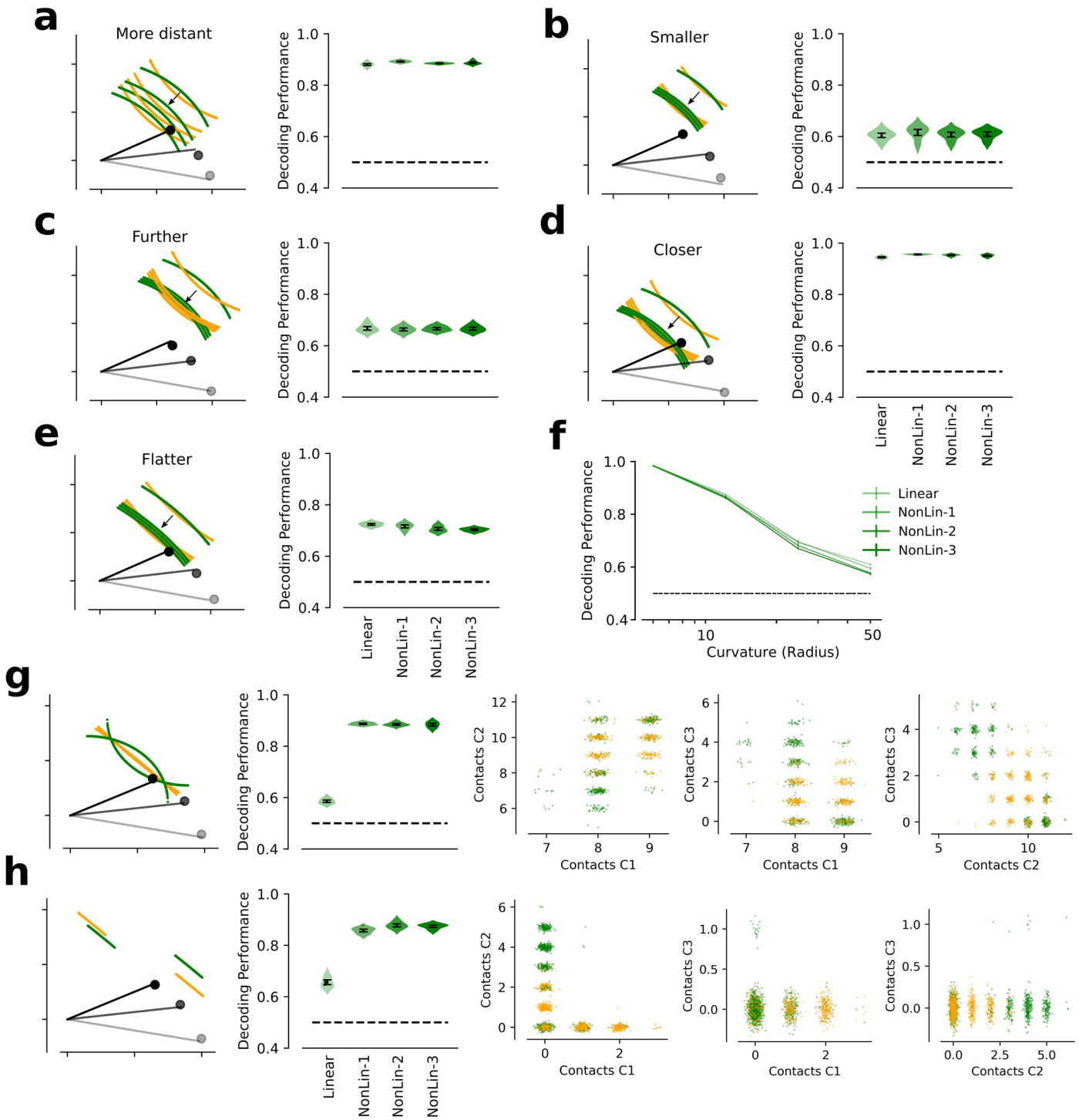


Extended Data Fig. 3 | See next page for caption.

https://doi.org/10.1038/s41593-022-01237-9

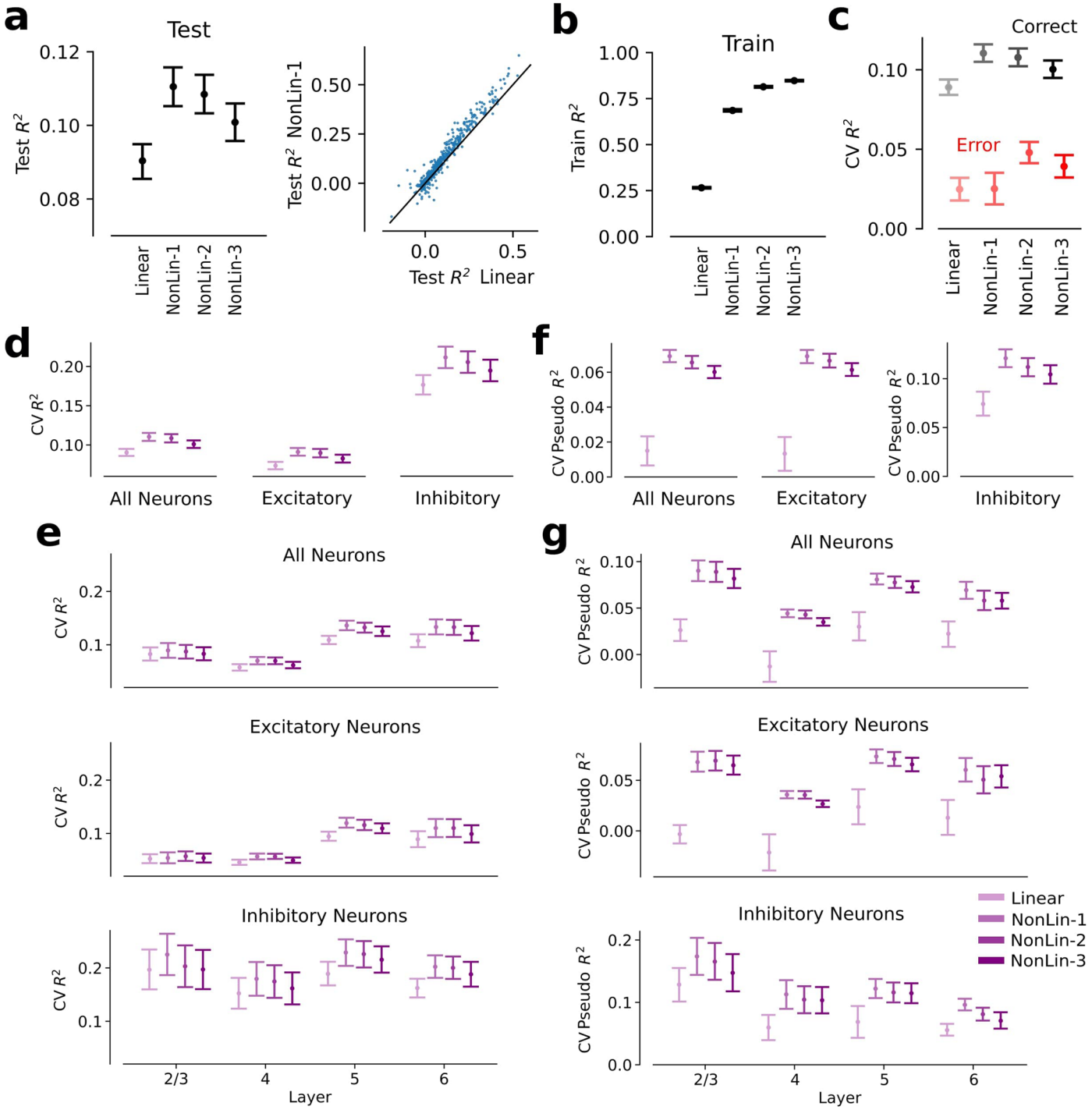
Extended Data Fig. 3 | Different feedforward and recurrent architectures were used to decode behavior and fit the S1 encoding models. (**a**) Stimulus (green) and choice (blue) were decoded on a trial-by-trial basis from the spatio-temporal pattern of whisking contacts using RNNs with different number of hidden units. On each time step and trial the input to the RNN decoder was the number of contacts and angle of contact for all the whiskers. Errorbars correspond to s.e.m. across mice (n = 10). Feed-forward networks performed in general better than RNNs for decoding behavior (see Fig. 2). (**b**) S1 population activity was regressed against whisker and task variables (encoding model) using RNNs with different number of hidden units. Errorbars correspond to s.e.m. across neurons (n = 584). See Extended Data Fig. S5a for the distribution of CV R^2 across neurons for the linear and the best non-linear (feed-forward with one hidden layer of hundred units) encoding models. Feed-forward networks performed in general better than RNNs on explaining S1 population activity (see Fig. 4). (**c**) Linear and non-linear feed-forward networks with different number of units (artificial neurons) in the hidden layers are equally good at predicting the presented shape (stimulus; green) and animal's choice (choice; blue) on a trial-by-trial basis from the spatio-temporal pattern of whisker contacts and angle of contacts (see Fig. 2). Errorbars correspond to s.e.m. across mice (n = 10). (**d**) The profile of explanatory power for S1 population activity across models is qualitatively equivalent for 20, 100 (see Fig. 4), and 200 units in the hidden layers of the feed-forward encoding models. Errorbars correspond to s.e.m. across neurons (n = 584). See Extended Data Fig. S5a for the distribution of CV R^2 across neurons for the linear and the best non-linear (feed-forward with one hidden layer of hundred units) encoding models.

https://doi.org/10.1038/s41593-022-01237-9



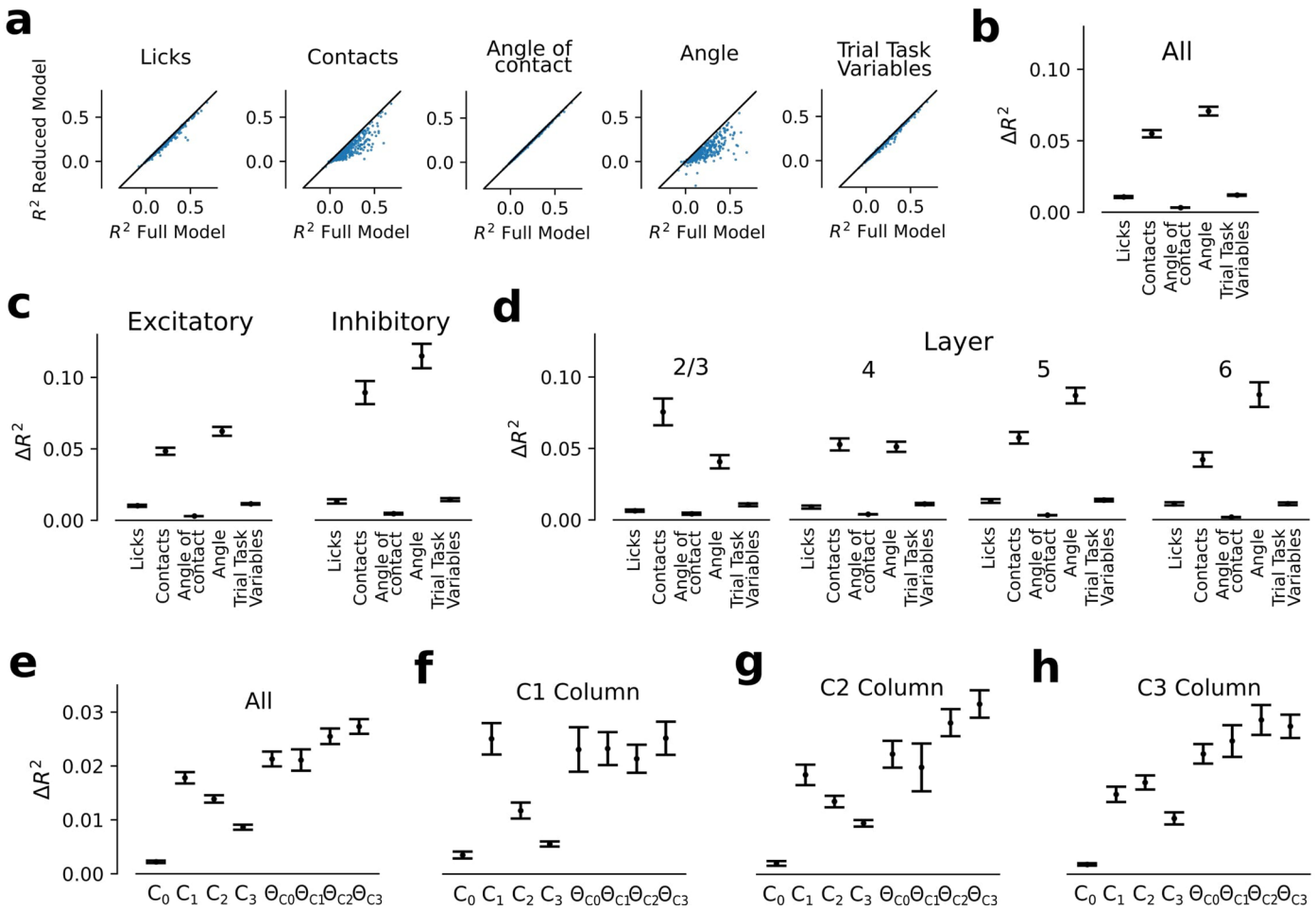
Extended Data Fig. 4 | Linear and non-linear classifiers perform equally well on other shapes used in the simulated whisker discrimination task. (**a-e**) Linear and non-linear classifiers performed equally well from the spatio-temporal pattern of simulated whisker contacts and angles of contact when different general parameters of the shapes were used in the simulations: more distant stopping locations (a); smaller shapes (b); further shapes (c); closer shapes (d); and flatter shapes (e) (see Methods). (**f**) Flatter shapes (larger radius) were more difficult to discriminate than more curved shapes on a simulation of the whisker-based discrimination task. (**g**) Curvature discrimination task: discriminate between curved (green) or flat (orange) shapes (first panel). Non-linear classifiers substantially outperform linear ones on the simulated curvature discrimination task when the full spatio-temporal pattern of contacts and angle of contacts is used to discriminate between flatter (orange) and more curved (green) shapes (second panel). Total number of contacts for the pair (C1, C2) (third panel), (C1, C3) (fourth panel) and (C2, C3) (fifth panel). The boundary between the two categories is non-linear. (**h**) Example of another simulated non-linear discrimination task (green vs orange bars). Non-linear classifiers substantially outperform linear ones on this task when the full spatio-temporal pattern of contacts and angle of contacts is used. The boundary between the two categories is non-linear, especially for C1 vs C2. Errorbars in all panels correspond to s.e.m. across independent simulations (n = 5).

https://doi.org/10.1038/s41593-022-01237-9



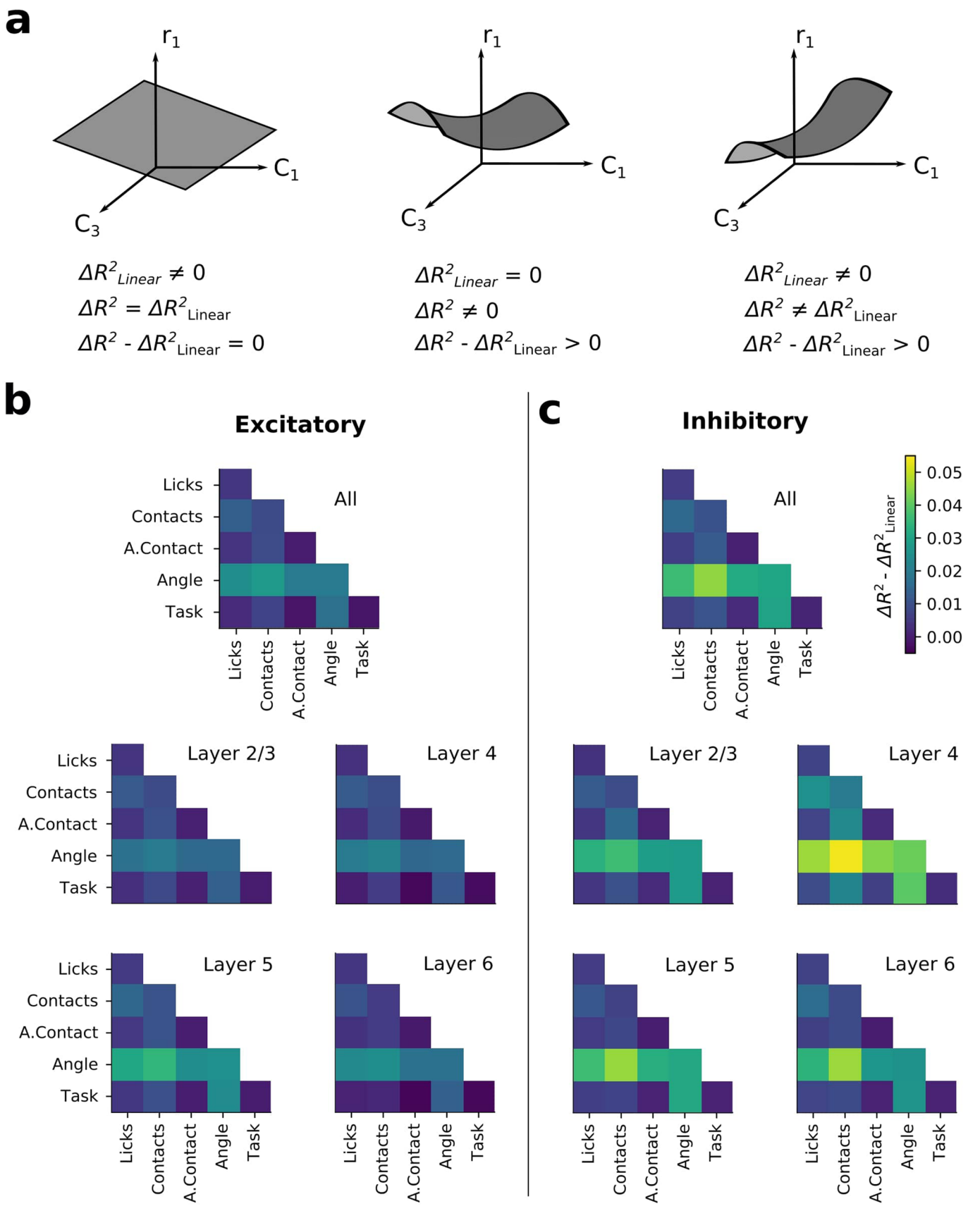
Extended Data Fig. 5 | Mean goodness-of-fit across neurons, correct and incorrect trials, neuronal types, and S1 layers. (**a**) Mean goodness-of-fit across neurons (y-axis; R^2) for the different encoding models (x-axis) on held-out data (left), and the distribution of R^2 across neurons for the non-linear (one hidden layer) (y-axis) and the linear encoding models (x-axis) (right). All models were fit on simultaneously recorded populations. For held-out data, the best model is a feed-forward fully connected network with only one hidden layer (NonLin-1). (**b**) Mean goodness-of-fit across neurons (y-axis; R^2) for the different encoding models (x-axis) on the data used for training. For the train data, the more complex the model (more parameters), the better the prediction. (**c**) Mean CV R^2 (y-axis) across S1 neurons for the different encoding models (x-axis) when evaluated in correct (black) and incorrect trials (red). Encoding models explained S1 activity better on correct than error trials. Errorbars in panels (a-c) correspond to s.e.m. across neurons (n = 584). (**d**) Mean CV R^2 across neurons for the different encoding models (x-axis) on held-out data for all neurons (left), only excitatory (middle) and only inhibitory neurons (right). (**e**) Mean CV R^2 across neurons for the different encoding models on held-out data for neurons across layers for all (top), excitatory (middle) and inhibitory neurons (bottom). (**f-g**) Encoding fits when using Poisson loss instead of mean squared error (MSE). The y-axis shows the Poisson-loss equivalent of the R^2, the Pseudo-R^2. The Pseudo-R^2 is calculated as 1 - PLoss/Variance, where PLoss is the negative Log-likelihood of the Poisson model. All models were fit on simultaneously recorded populations. Errorbars in all panels correspond to s.e.m. across neurons (all n = 584; excitatory n = 491; inhibitory n = 93; layer 2/3 n = 68; layer 4 n = 157; layer 5 n = 249; layer 6 n = 96).

https://doi.org/10.1038/s41593-022-01237-9



Extended Data Fig. 6 | Contribution of the different regressors to S1 activity across neuronal types, layers, and columns in S1. (**a**) Distribution of R^2 for the full model (x-axis; R^2 Full Model) vs R^2 for the different ablations (y-axis; R^2 Reduced Model) across all recorded neurons. The R^2 for the different ablations is calculated by testing the model on held-out trials where those features (or groups of features) have been set to zero. (**b**) Mean ΔR^2 across all neurons (y-axis; $\Delta R^2 = R^2_{Full} - R^2_{Reduced}$) for the different groups of variables. Whisker contacts (Contacts) and continuous angle (Angle) are the most important groups of regressors in predicting S1 activity. Errorbars correspond to s.e.m. across neurons (n = 584). (**c**) ΔR^2 separately for excitatory and inhibitory neurons. Inhibitory populations show a higher ΔR^2 because their R^2 is overall higher (see Extended Data Fig. S5e). (**d**) ΔR^2 across layers of the somatosensory cortex. Whisker contacts have a stronger effect on superficial layers, while whisker angle has a stronger effect on deeper layers. (**e**) The metric ΔR^2 reveals that neurons in S1 do not strictly obey somatotopy during the whisker-based discrimination task (see also[5]). The contribution to the encoding model's performance (y-axis; ΔR^2) for the groups of variables associated with whiskers' contacts (C_0, C_1, C_2 and C_3 for all time steps) and angular position (θ_{C0}, θ_{C1}, θ_{C2}, and θ_{C3} for all time steps) for all neurons. (**f-h**) Whisker and angular position encoding strength for C1 column (b), C2 column (c) and C3 column (d). While C1 contacts is the strongest driver in C1 column, C2 and C3 columns are not dominated by C2 and C3 contacts, respectively. Errorbars in all panels correspond to s.e.m. across neurons (all n = 584; excitatory n = 491; inhibitory n = 93; layer 2/3 n = 68; layer 4 n = 157; layer 5 n = 249; layer 6 n = 96; C1 column n = 117; C2 column n = 208; C3 column n = 204).

https://doi.org/10.1038/s41593-022-01237-9

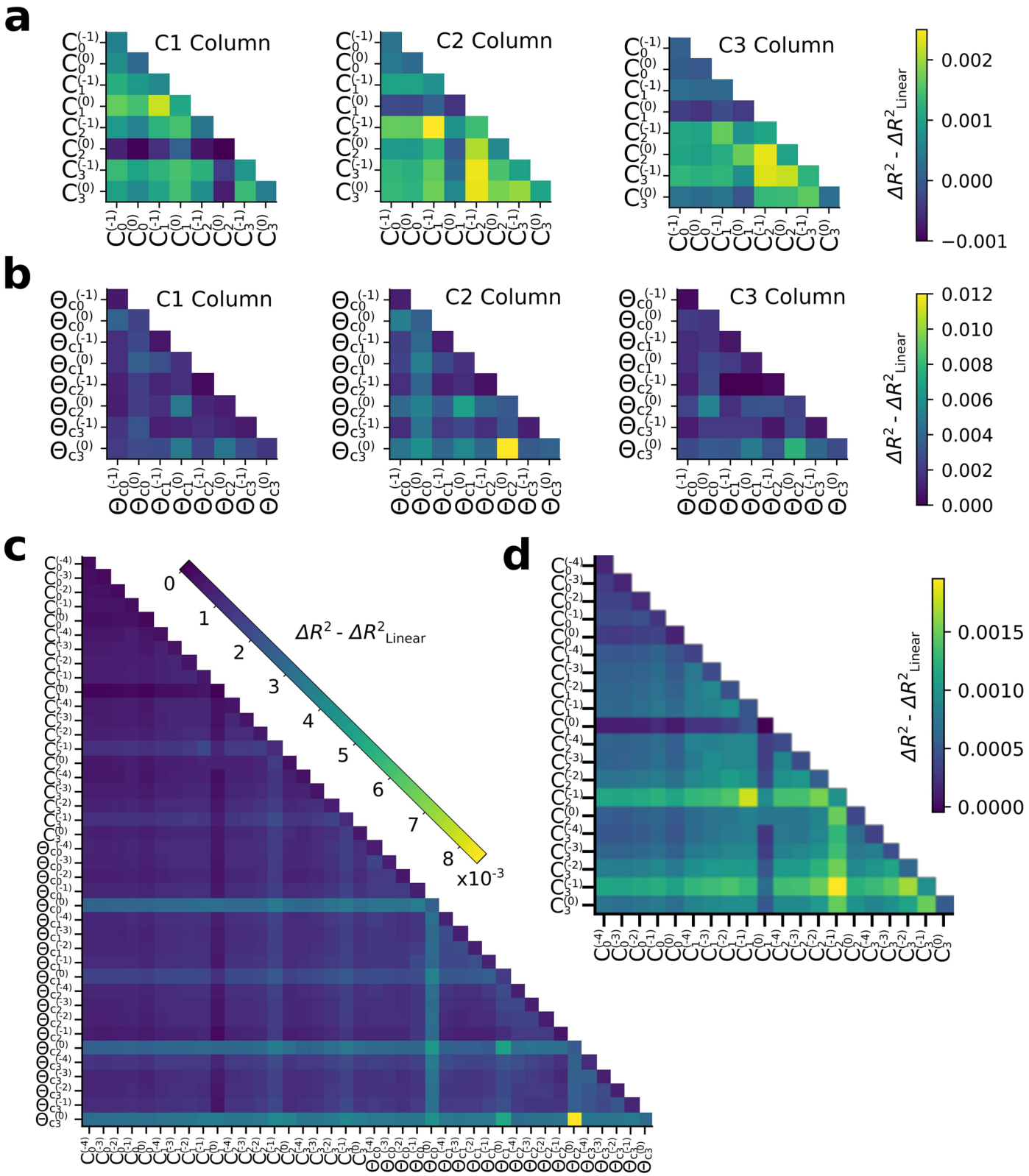


Extended Data Fig. 7 | See next page for caption.

https://doi.org/10.1038/s41593-022-01237-9

Extended Data Fig. 7 | Non-linear mixed selectivity across neuronal types and cortical layers for the different groups of task variables. (**a**) Different coding scenarios would produce different values for $\Delta R^2 - \Delta R^2_{Linear}$. Here we show different tuning schemes with respect to C_1 and C_3 for a fictional neuron (r_1). The metric $\Delta R^2 - \Delta R^2_{Linear}$ was used to evaluate to what extend the pure non-linear terms were important to predict the population's firing rate. (Left) If the relationship between neuronal activity and encoding variables is linear, $\Delta R^2_{Linear} \neq 0, \Delta R^2 = \Delta R^2_{Linear}$ and therefore $\Delta R^2 - \Delta R^2_{Linear} = 0$. (Middle) If the relationship between neuronal activity and encoding variables is purely non-linear, $\Delta R^2_{Linear} = 0, \Delta R^2 \neq 0$, and $\Delta R^2 - \Delta R^2_{Linear} > 0$. (Right) If the encoding model is composed of both linear and non-linear components, $\Delta R^2_{Linear} \neq 0$, $\Delta R^2 \neq \Delta R^2_{Linear}$, and $\Delta R^2 - \Delta R^2_{Linear} > 0$. (**b-c**) Pure non-linear mixed selectivity contribution ($\Delta R^2 - \Delta R^2_{Linear}$) for the interaction between the different blocks of variables across neuronal types and S1 layers. Results were qualitatively equivalent for the excitatory (a) and the inhibitory (b) populations.

https://doi.org/10.1038/s41593-022-01237-9

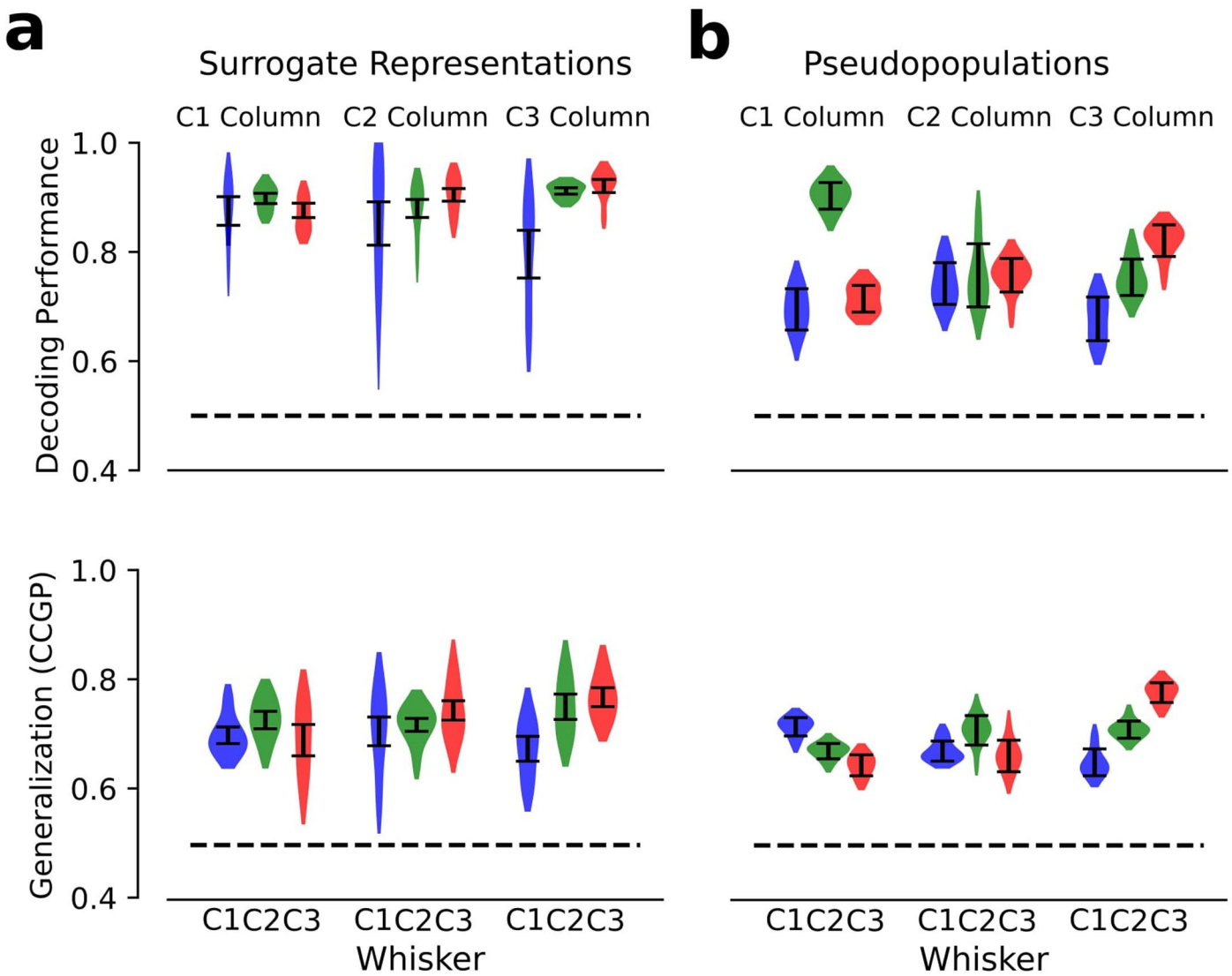


Extended Data Fig. 8 | See next page for caption.

https://doi.org/10.1038/s41593-022-01237-9

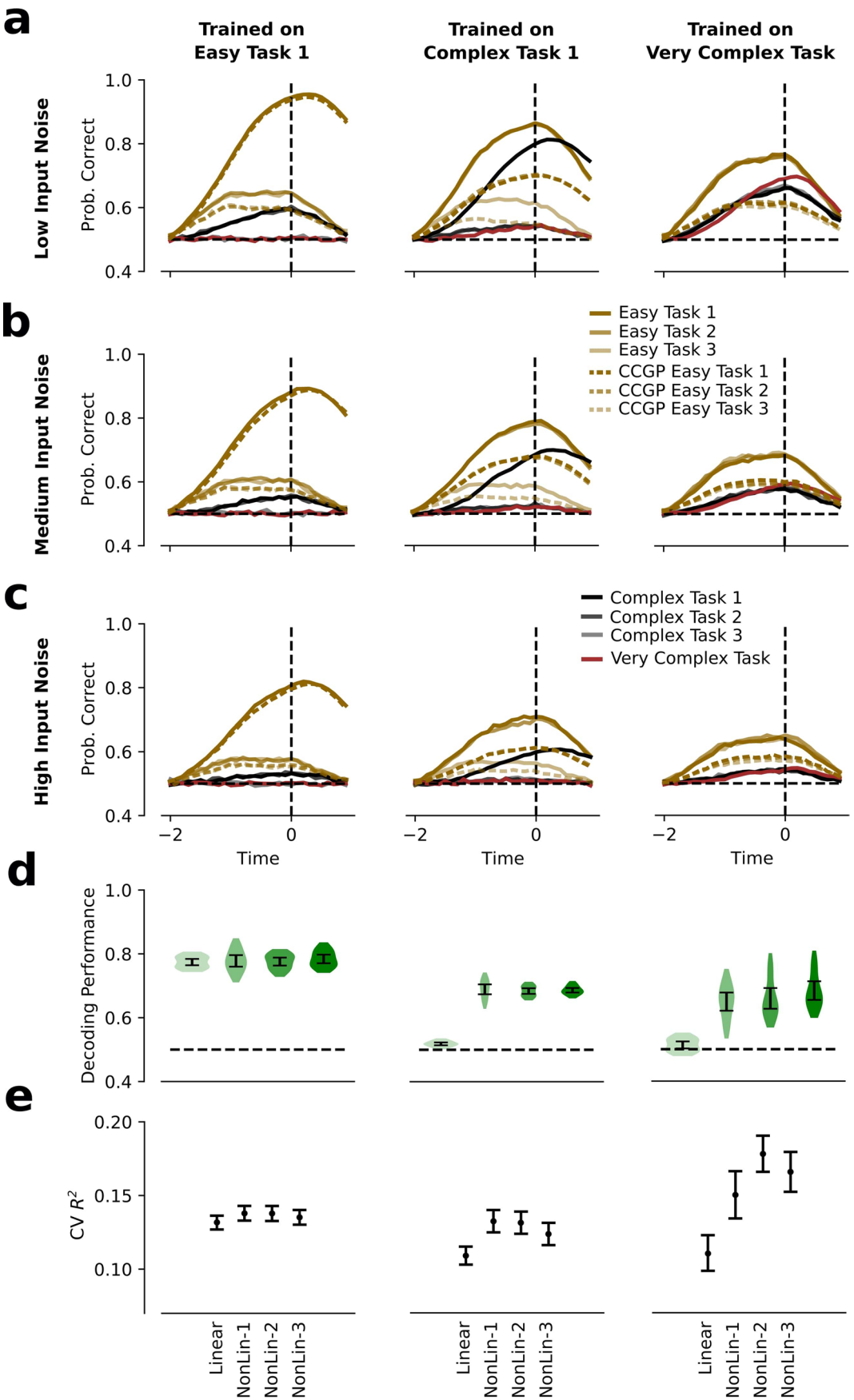
Extended Data Fig. 8 | Non-linear mixed selectivity for whisker contacts and angular position for the different time steps and cortical columns in S1. (**a**) Pure non-linear mixed selectivity ($\Delta R^2 - \Delta R^2_{Linear}$) contribution for the interaction between contacts for the different time steps (time lags) and whiskers separately by columnar location (mean across neurons). The strongest interactions occur at the time lags of 100 ms (1 time step). Even though C1 column shows that C1 terms have the strongest interaction, C2 and C3 columns present a more heterogeneous interaction pattern. (**b**) Equivalent plot for the interaction between angular position for the different time steps (time lags) and whiskers. The strongest interactions occurs at time lags of 0ms. All columns present strong interactions terms with the rest of whiskers. (**c**) Pure non-linear mixed selectivity contribution ($\Delta R^2 - \Delta R^2_{Linear}$) for the interaction between contacts and angular position for the different time steps (time lags) and whiskers (single regressors in the encoding models) (mean across neurons). The strongest non-linear contribution in whisker angular position occurs on the current time step for all whiskers. The strongest non-linear contribution for the interaction between contacts and angular position occurs on time lags between angular position (and neuronal activity) and contacts of 100 ms (1 time step). (**d**) Equivalent plot for the interaction between contacts for the different time steps (time lags) and whiskers. The strongest non-linear contribution in whisker contacts occurs on time lags between neuronal activity and contacts of 100 ms (1 time step) for all whiskers.

https://doi.org/10.1038/s41593-022-01237-9



Extended Data Fig. 9 | Different whiskers are represented across S1 columns in approximately orthogonal sub-spaces. (**a**) Top: Decoding performance on whether the sum of whisker contacts across the current and previous four time steps corresponded to a high or a low number of contacts with respect to the median for each different whisker (decode high vs low number of contacts for each whisker) (C1 blue; C2 green; C3 red) and columns (see Methods). Information about whisker contacts for all whiskers was present in all columns. Decoding Performance was evaluated on surrogate activity generated by the best encoding model (NonLin-1) for each recording session. In each recording session, activity from only one column was recorded. In order to compare information across columns, surrogate activity was generated for 10 neurons, which corresponded to the smallest number of simultaneously recorded neurons across recording sessions (columns). Bottom: Qualitatively equivalent results were found when the CCGP was evaluated (see Methods). All columns encode information about all whiskers in approximately orthogonal spaces. Errorbars in all panels correspond to s.e.m. across recording sessions (C1 column n = 6; C2 column n = 9; C3 column n = 6). (**b**) Top: Decoding performance on whether the total number of contacts in a trial (2 sec.) corresponded to a high or a low number of contacts with respect to the median for each different whisker (decode high vs low number of contacts for each whisker) (C1 blue; C2 green; C3 red) and columns using pseudopopulations of neurons (see Methods). All bars are significantly above chance. Bottom: CCGP with respect to the total number of contacts for the different whiskers and columns. Even though there seems to be some somatotopic structure on CCGP for pseudopopulations, the differences are small and all bars are significantly above chance. Errorbars correspond to the standard deviation across cross-validation iterations (n = 10).

https://doi.org/10.1038/s41593-022-01237-9



Extended Data Fig. 10 | See next page for caption.

https://doi.org/10.1038/s41593-022-01237-9

Extended Data Fig. 10 | The geometry of representations and encoding properties in RNNs change for the easy, the complex, and the very complex tasks. (**a**) Probability of correct response (y-axis) as a function of time (x-axis) when the RNN was trained on the easy task 1 (left panel), the complex task 1 (central panel) and the very complex task (right panel). RNNs were trained on low input noise ($\lambda_{low} = 0.23$ and $\lambda_{high} = 0.77$) (see Methods). For each network, additional readout weights on the activity of the artificial units were trained to perform the rest of the tasks (solid lines). While an RNN trained on easy task 1 produced the best performance for easy task 1, the neuronal representations were not well suited for the rest of tasks (left). When an RNN was trained on the complex 1 (central) and the very complex (right) tasks, it produced representations that allowed the performance of many different tasks. This came at the expense of losing performance for the easy task 1. Generalization performance, defined as cross-condition generalization performance (CCGP), was also tested for the three easy tasks (dashed lines) when the RNN was trained on the easy task 1 (left panel), the complex task 1 (central panel) and the very complex task (right panel) (see Methods). While abstraction (CCGP) is high for easy task 1 when the RNN was trained on the easy task 1, it is low for easy tasks 2 and 3. Since complex task 1 is defined as the 2D-XOR between C1 and C2, CCGP was higher for easy task 1 and 2 (C1 and C2) than for easy task 3 (C3). For the RNN trained on the very complex task, CCGP was significantly above chance for all easy-task variables. (**b-c**) The same qualitative results were obtained when medium (b; $\lambda_{low} = 0.3$ and $\lambda_{high} = 0.7$) and high (c; $\lambda_{low} = 0.35$ and $\lambda_{high} = 0.65$) noise levels were used instead. For all panels in (a-c) the performance curves correspond to the mean across random realizations of input patterns and tasks (n = 50) (see Methods). (**d**) Similar to Fig. 2d, linear and non-linear classification models that read out the input (x-axis), were trained to perform the easy (left panel; easy task 1), the complex (central panel; complex task 1) and the very complex tasks (right panel). On the easy task, both linear and non-linear classifiers performed equally well, as shown by decoding performances (y-axis) of the different models. On the contrary, only non-linear classifiers that allow for complex cue combination, performed above chance on the complex (central) and very complex (right) tasks. The behavioral results obtained on the whisker-based discrimination task (see Fig. 2) are aligned with the easy task (left panel). In all panels errorbars correspond to the s.e.m. across different network realizations (n = 4). (**e**) Similar to Fig. 4d, CV R^2 on explaining artificial units activity (y-axis), is plotted against the different encoding models (x-axis). RNNs trained to perform tasks that require non-linear integration of sensory cues are better explained by an encoding model that allows for non-linear mixed selectivity (central and right panels), while a non-linear encoding scheme provide marginal additional explanatory power for the easy task RNN (left panel). The higher the complexity of the trained task, the higher the advantage of non-linear encoding models on explaining the activity of the artificial units. In contrast to (d), the encoding properties of S1 are aligned with those of RNNs trained to perform tasks that require non-linear combination of sensory evidence (see Fig. 4). Errorbars correspond to the s.e.m. across artificial units (n = 240; 4 networks × 60 units).

nature portfolio

Corresponding author(s): Stefano Fusi and Ramon Nogueira

Last updated by author(s): Nov 5, 2022

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Custom made code was used. It is available at https://github.com/cxrodgers/Rodgers2021
Data analysis	Custom made code in python and pytorch was used. It is available at https://github.com/ramonnogueira/TheGeometryOfS1

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The datasets analyzed in this study have been deposited to Zenodo: https://doi.org/10.5281/zenodo.4743837.
See also Rodgers 2022 for a detailed description of this dataset: Rodgers, C.C. A detailed behavioral, videographic, and neural dataset on object recognition in mice. Sci Data 9, 620 (2022). https://doi.org/10.1038/s41597-022-01728-1

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	*Use the terms sex (biological attribute) and gender (shaped by social and cultural circumstances) carefully in order to avoid confusing both terms. Indicate if findings apply to only one sex or gender; describe whether sex and gender were considered in study design whether sex and/or gender was determined based on self-reporting or assigned and methods used. Provide in the source data disaggregated sex and gender data where this information has been collected, and consent has been obtained for sharing of individual-level data; provide overall numbers in this Reporting Summary. Please state if this information has not been collected. Report sex- and gender-based analyses where performed, justify reasons for lack of sex- and gender-based analysis.*
Population characteristics	*Describe the covariate-relevant population characteristics of the human research participants (e.g. age, genotypic information, past and current diagnosis and treatment categories). If you filled out the behavioural & social sciences study design questions and have nothing to add here, write "See above."*
Recruitment	*Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.*
Ethics oversight	*Identify the organization(s) that approved the study protocol.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	A total of 584 neurons were recorded in this study and used to analyze the geometry of representations in S1. 10 mice were used to analyze the shape discrimination task and behavior. The results of the encoding and decoding models were robust across neurons and mice.
Data exclusions	Mice that did not learn the task well were discarded from all analysis (3 discarded)
Replication	The reported performance of the decoding and encoding models correspond to the performance on validation trials, which were not used for training the models. See the Methods section for a detailed description of how cross-validation was implemented in the analysis of this manuscript.
Randomization	When training the decoding end encoding models, data was randomly split into train, test and validation (cross-validation). See the Methods section for a detailed description of how cross-validation was implemented.
Blinding	Decoding and encoding models were cross-validated. Trials were randomly assigned to train, test and validation groups. See the Methods section for a detailed description of how cross-validation was implemented in the analysis of this manuscript.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☐	☒ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals	Mice C57BL/6J, bred at Columbia University from Jackson lines. 10 mice in total. Mice began behavioral training between postnatal days 90 and 180.
Wild animals	No wild animals were used in the study
Reporting on sex	We used males and females arbitrarily and in roughly equal proportion (6 females and 4 males). Female mice typically weighed less than male mice and drank correspondingly less water, but we adjusted the reward size based on weight to achieve roughly equal trial counts. Because we observed no other differences, we pooled the data from both sexes.
Field-collected samples	No field collected samples were used in this study
Ethics oversight	All experiments were conducted under the supervision and approval of the Columbia University Institutional Animal Care and Use Committee

Note that full information on the approval of the study protocol must also be provided in the manuscript.