The non-linear mixed representations in somatosensory cortex support simple and complex tasks

Ramon Nogueira[1,2,3,*], Chris C. Rodgers[2,3,4], Randy M. Bruno[2,3,4], and Stefano Fusi[1,2,3,4,*]

[1]Center for Theoretical Neuroscience, Columbia University, New York, NY 10027
[2]Zuckerman Mind Brain Behavior Institute, Columbia University, New York, NY 10027
[3]Department of Neuroscience, Columbia University, New York, NY 10027
[4]Kavli Institute for Brain Science, Columbia University, New York, NY 10027
[*]Correspondence: rn2446@columbia.edu (R.N.), sf2237@columbia.edu (S.F.)

Adaptive behavior in humans, rodents, and other animals often requires the integration over time of multiple sensory inputs. Here we studied the behavior and the neural activity of mice trained to actively integrate information from different whiskers to report the curvature of an object. The analysis of high speed videos of the whiskers revealed that the task could be solved by integrating linearly the whisker contacts on the object. However, recordings from the mouse barrel cortex revealed that the neural representations are high dimensional as the inputs from multiple whiskers are mixed non-linearly to produce the observed neural activity. The observed representation enables the animal to perform a broad class of significantly more complex tasks, with minimal disruption of the ability to generalize to novel situations in simpler tasks. Simulated recurrent neural networks trained to perform similar tasks reproduced both the behavioral and neuronal experimental observations. Our work suggests that the somatosensory cortex operates in a regime that represents an efficient compromise between generalization, which typically requires pure and linear mixed selectivity representations, and the ability to perform complex discrimination tasks, which is granted by non-linear mixed representations.

Introduction

Making sense of real world sensory inputs often requires the integration of multiple sources of information. In many situations this process of integration involves only simple linear operations. 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 has been detected or not. In other words, a linear decoder would be sufficient to perform this simple detection task. However, recognizing an object by touching it could be a significantly more complex task, which probably involves non-linear integration of the sensory inputs coming from multiple fingers. Here we studied how mice perform a shape recognition task using their whiskers. By analyzing high speed videos of the whisker movements, we discovered that the task can actually be solved by a linear decoder. Moreover, a linear decoder is also the best predictor of the decisions of the animals. However, the neural representations in somatosensory cortex are better explained by a process of non-linear spatio-temporal integration, which typically leads to high dimensional representations. These representations allow a simple downstream readout to perform a broad class of complex tasks without sacrificing the robustness of the representations in simpler tasks, like the one studied in the experiment.

This type of representation constitutes a non-trivial compromise between the flexibility of high dimensional representations, which can be reused in a number of very diverse tasks, and the robustness to noise of low dimensional representations. Indeed, high-dimensional neuronal representations have been shown to enable a linear readout to perform a large number of different tasks [3, 4, 5, 6]. This is a desirable feature in a brain area like prefrontal cortex, which in many complex cognitive tasks is directly involved in the process of decision making. Indeed, for these high dimensional representations, a linear

readout (even by a single neuron) can be easily trained to generate a variety of response patterns, each corresponding to a different task. This can be achieved without modifying the input representations. The disadvantage of these high dimensional representations is that the type of non-linear integration that they entail typically enhances noise and hence can impair the performance in simple tasks that depend on a linear combination of one or very few variables (see e.g. [5]).

For these reasons one would expect that in a sensory area the representations are relatively low dimensional, as they tend to represent only linear combinations of task relevant features. This is one of the reasons why the concept of a receptive field is so popular when studying sensory areas: the concept is actually based on the assumption that the activity can be accurately predicted by computing a linear combination of input features. After all, if these features need to be combined non-linearly to generate complex behaviors, they could be mixed in one of the several downstream areas. Our study shows that this expectation of linearity is not met in the somatosensory cortex of mice. Although this is a sensory area, the neurons respond to rather heterogeneous non-linear combinations of the task relevant features. This is compatible with the recent observations of high dimensional representations in the visual cortex of non-behaving mice [7].

More specifically, this study follows up on a recent study [8] in which mice were trained to perform a whisker-based shape discrimination task while populations of single neurons were recorded in the barrel cortex. On each trial animals were presented with either a concave or a convex shape placed at one of three distances from the animal. By active whisking, monitored by a high speed camera, mice had to report the convexity of the presented object. A linear probe was used to record the activity of populations of single neurons across different layers of the somatosensory cortex. Mice were able to perform the task by comparing the number of contacts across whiskers. In the present study, we trained classifiers to use the spatio-temporal whisking patterns to predict either the convexity of the stimulus or the choice of the animal, and found that linear and non-linear classifiers performed equally well. This indicates that linear integration across time and whiskers is sufficient for solving the task. By comparing linear and non-linear encoders, we could determine and study the linear and non-linear components of the neural responses. Surprisingly, we found that the recorded neural activity was best explained by an encoding model that allowed for non-linear mixed selectivity with respect to task variables. In particular, neuronal populations mixed contacts and angular position of different whiskers, at different times, indicating that the observed activity was the result of non-linear spatio-temporal integration of multiple variables. Indeed, we observed that the shattering dimensionality, a measure of dimensionality of the neural representations that does not depend much on the noise [3, 9], was higher for the recorded neural activity than for the observed spatio-temporal patterns of the whisker inputs, analyzed from the high speed video.

These neural representations allow a linear readout to perform a broad class of complex discrimination tasks. This ability does not come at the cost of a reduced robustness to noise and capacity to generalize, which are usually guaranteed by a lower dimensional representation based on linear mixed selectivity. This observation is similar to what previously seen in prefrontal cortex and hippocampus of monkeys[9]. Our results suggest that the observed neural representations are an efficient compromise between flexibility and robustness, even in a sensory area like somatosensory cortex.

Results

The whisker-based object discrimination task

Mice were trained on a whisker-based shape discrimination task (Fig. 1a), in which they were asked to identify whether a presented object was concave or convex. Each trial began with an object (convex or concave) moving toward the whiskers ($t = -2$ sec). Objects could stop at one of three different distances (far, medium or close), which happened at $t = -0.9$, $t = -0.7$ and $t = -0.5$ seconds, respectively. The uncertainty about distance made the task more difficult, and it encouraged the animals to use all the whiskers at their disposal (C0, C1, C2, and C3). When the response window opened ($t = 0$ seconds), 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, processed with an image parsing algorithm [10, 11, 12] (Fig. 1b, see Methods).

Importantly, mice were free to whisk and lick throughout the course of 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 at $t = 0$ seconds.

Mice performed the task with a mean accuracy of $77.7\% \pm 0.9\%$ (Fig. 1c). The probability of making a lick on the correct lickpipe increased throughout the course of the trial, indicating that mice based their decision on the accumulated sensory evidence gathered by whisking. This implies some form of temporal integration. Likewise, the lick rate also followed a similar trend, which suggests that lick rate could be considered as a proxy for choice confidence (Fig. 1d). As depicted by the temporal profile of whisker contacts, the most informative time window was approximately from $t = -1.25$ to $t = -0.25$ seconds, when most of the contacts were made (Fig. 1e).

The mean difference in total number of contacts between convex and concave objects for whisker C1 was -1.17 ± 0.17, for C2 -1.22 ± 0.27 and for C3 0.77 ± 0.28 (Fig. 1f). Notice that the difference is negative for C1 and C2 and positive for C3. Therefore, by computing the weighted sum of total number of contacts of the three whiskers, it should be possible to discriminate between convex and concave objects. The contact rate of each whisker followed a similar time profile (example mouse, Fig. 1g and all mice in Supplementary Fig. S1). Contact rate for correct trials was higher in the correct trials (example mouse, Fig. 1h and all mice in Supplementary Fig. S2), suggesting that poorer sensory gathering or a lower level of task engagement was the reason for the discrimination errors.

Linear integration is sufficient for object discrimination

In order to understand whether linear integration is actually sufficient to determine the convexity of an object, we tried to predict convexity (stimulus) on a trial-by-trial basis by using a linear decoder that reads out the spatio-temporal pattern of whisker contacts observed during the execution of the task (Fig. 2a). As we did not know whether whisker contacts were the only important variables, we included in the input other variables, like the angular position of the whiskers during contacts. Moreover, we were interested in understanding whether these spatio-temporal patterns could explain the decision of the animal, so we tried to train the decoder also to predict the lick side (choice). Both predictions were tested on held-out trials (cross-validation; see Methods).

By progressively increasing the number of input variables, we realized that the most informative set of stimulus features comprised all the whisker contacts and angle of contact across time (Fig. 2b). See also [8]. Unsurprisingly, the weights of the classifier 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 (see Methods; Fig. 2c). This observation nicely matches the progressively increasing probability of licking on the correct side of Fig. 1c.

Until now we only considered simple linear decoders, which compute a weighted sum of the inputs (whisker contacts and angles at different times) and compare it to a threshold. We then wondered whether a non-linear decoder could better predict the stimulus identity and the choice of the animal. More specifically, we trained two feed-forward neural networks with non-linear units arranged in multiple layers to predict stimulus and choice. These decoders are more complex and they contain more parameters than the linear ones, so they will certainly perform better at classifying the patterns in the training set. However, it is not guaranteed that the cross-validated performance, computed on held-out trials, will actually increase. The cross-validated performance of the non-linear classifier can only surpass the linear classifier if non-linear combinations of the features are important, indicating that the task could be solved more efficiently by combining the inputs variables in a non-linear way.

Despite the complexity of the task, we observed that both linear and nonlinear decoders exhibited a similar performance (linear: $89.0\% \pm 1.5\%$; best non-linear: $90.1\% \pm 1.4\%$) (Fig. 2d; bottom-left; green). A qualitatively equivalent result was observed when trials were split into correct and error (Fig. 2d; bottom-right; stimulus; black and red). The lower overall performance on error trials is likely due to the lower number of contacts and overall lower task engagement (Fig. 1h). When predicting choice on a trial-by-trial basis, a similar trend was observed (linear: $71.4\% \pm 1.6\%$; best non-linear: $75.0\% \pm 1.1\%$) (Fig. 2d; bottom-left; blue), suggesting that animals' decisions were mostly driven by a linear combination across time and whiskers of the sensory cues. However, a very different performance

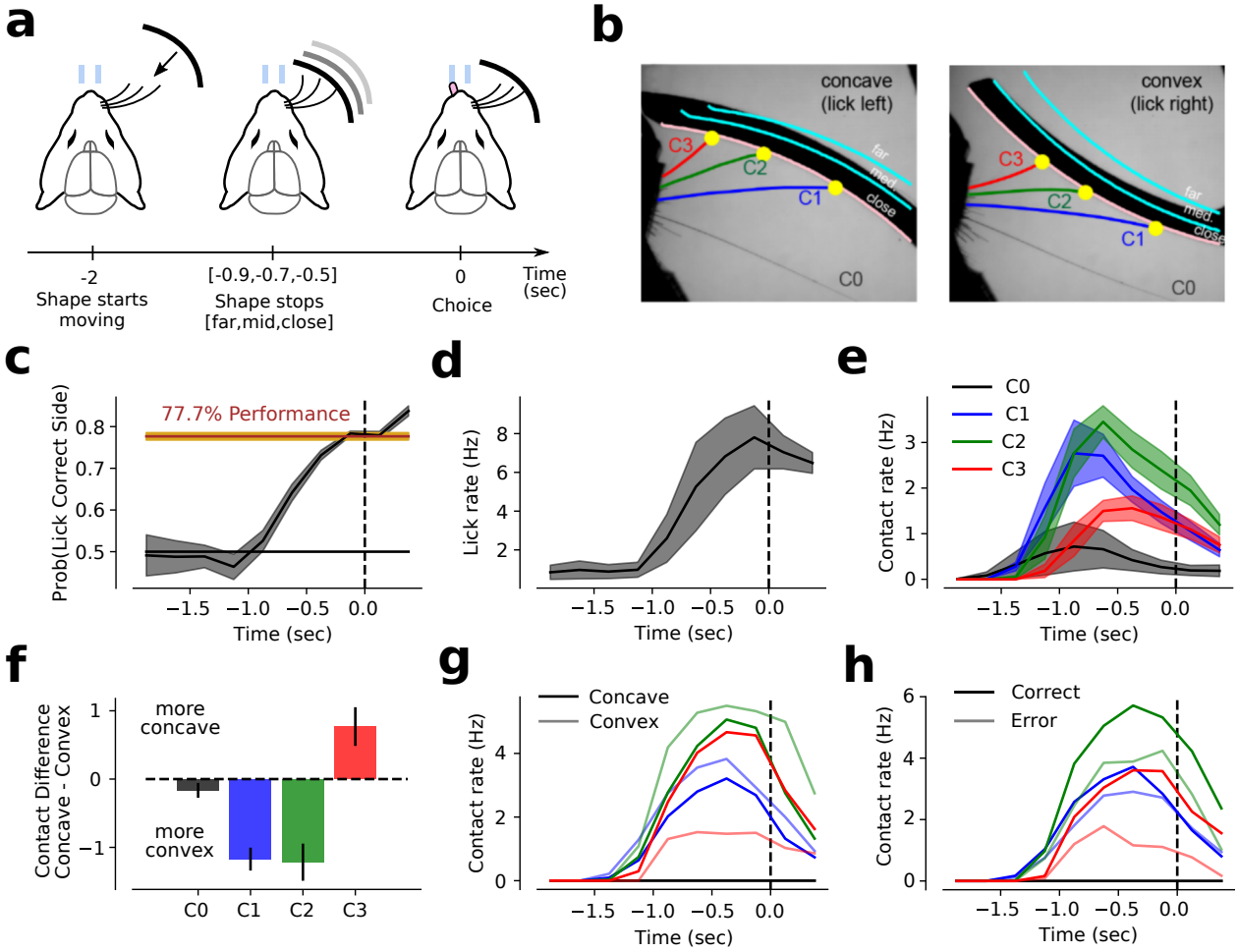

Figure 1: **Mice performed a whisker-based shape discrimination task by accumulating evidence throughout the trial**. (**a**) Animals were presented with either a convex or a concave shape and after two seconds they had to report their choice by licking the left (concave) or right (convex) lickpipe. (**b**) Whiskers and shape position were monitored by combining a high-speed camera with an image parsing algorithm [10, 11, 12]. (**c**) The probability of making a correct choice increased throughout the trial, indicating an accumulation of sensory evidence by active whisking. (**d**) The lick rate increased significantly once the shape was within whisking distance. (**e**) The time profile of contacts was similar across different whiskers. (**f**) C1 and C2 made more contacts for convex shapes, while C3 made more contacts for concave shapes. (**g**) Time profile of whisker contacts for an example animal. As in panel (f), C1 and C2 make more contacts on convex shapes (light colored) while C3 makes more contacts on concave shapes (dark colored). (**h**) Time profile of whisker contacts for correct (dark colored) and error (light colored) trials for an example animal. All whiskers made more contacts on correct than incorrect trials. Panels (c-e) and (g-h) were obtained using a sliding window of 250 milliseconds. Errorbars in (c-f) correspond to s.e.m. across animals.

was observed for choice prediction when trials were split into correct and error (Fig. 2d; bottom-right; choice; black and red), indicating that error trials are qualitatively different. One possible explanation of these effects is that in approximately 60% of the trials animals make very accurate choices that are based in 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 [13, 14]. It is important to note that all models were trained on a balanced dataset in which the the number of correct and error trials was exactly the same. This procedure decorrelates the variables stimulus and choice, which otherwise would be highly correlated.

We then investigated what other tasks could theoretically be performed using the same spatio-temporal patterns of inputs that we constructed by analyzing high speed videos. Clearly, the task of the experiment can be performed by a linear decoder, as just shown. We now describe a broad class of other "easy" tasks that can also be performed without pre-processing the inputs, and another class of "complex" tasks that require a different representation of the inputs. We constructed these two classes of tasks by choosing a certain output (virtual motor response) for each recorded spatio-temporal input. In particular, for the easy task (Fig. 2e) the desired output was chosen on the basis of a randomly weighted sum of C1, C2 and C3 contacts. When this sum was larger than a threshold, then the output would be left, otherwise it would be right (see Methods). This class of tasks is linearly separable by design and it can be solved by a linear decoder. For the complex task, the input regions belonging to the two responses were defined by a parity problem (a 3D-XOR, which is a non-linearly separable task; see Methods). We then trained both linear and non-linear decoders to perform these tasks. Unsurprisingly, both the linear and nonlinear decoders had a similar performance for the easy task, while only the non-linear decoders could perform above chance on the complex task. This simple exercise shows that there are tasks based on the observed spatio-temporal inputs that would require a more complex readout. We speculate that at least some natural sensory computations may belong to this constructed set of tasks.

Nonlinear mixed selectivity in the mouse barrel cortex

To characterize how task variables are represented in the somatosensory cortex (S1) of behaving mice, populations of neurons were simultaneously recorded in barrel cortex while mice performed the whisker-based object discrimination task (Fig. 3a). Barrel cortex neurons were predictive of shape identity and animal choice on a trial-by-trial basis as revealed by the performance of linear classifiers (Fig. 3b left panel, see Methods). At response time shape category could be decoded with a performance of $56.8\% \pm 1.7\%$ (green) and the animal's choice with a performance of $65.4\% \pm 2.0\%$ (blue). Both shape and choice could be reliably decoded from neuronal ensembles in which we grouped together the activity from different recording sessions (pseudopopulations)[8]. As expected, populations of barrel cortex neurons encoded information about whisker contacts (Fig. 3b right panel), where the task was to predict whether a particular activity pattern corresponded to a trial with high or low number of contacts (see Methods). Information in C1 and C2 peaked earlier than in C3, and overall they were more strongly encoded than C3.

Given that populations of neurons in S1 could be responsive to multiple task variables, we also trained encoding models that predicted the firing rate of the recorded population from the set of whisking and behavioral variables for all time steps (100 ms) and trials (Fig. 3c). In particular, we used as regressors of the encoding models instantaneous task variables like whisking contacts and angle, lick side and rate, as well as trial task variables like current and previous reward, choice and stimulus (see Methods). Fitting encoding models is useful when it is unclear which variables are modulating the activity of a population of neurons. Moreover, having a mapping between task variables and neural activity not only is informative about the neural code but also helps to construct denoised neural representations.

We considered four different encoding models that were implemented using feed-forward neural networks with 0, 1, 2 and 3 hidden layers of rectified linear units (ReLu). The neural network with 0 hidden layers is equivalent to a linear encoding model (linear regression), which can generate only pure and linear mixed selectivity neurons. In contrast, the 1, 2, and 3-layer neural nets can generate neuronal responses that depend on nonlinear interactions between the different regressors (non-linear

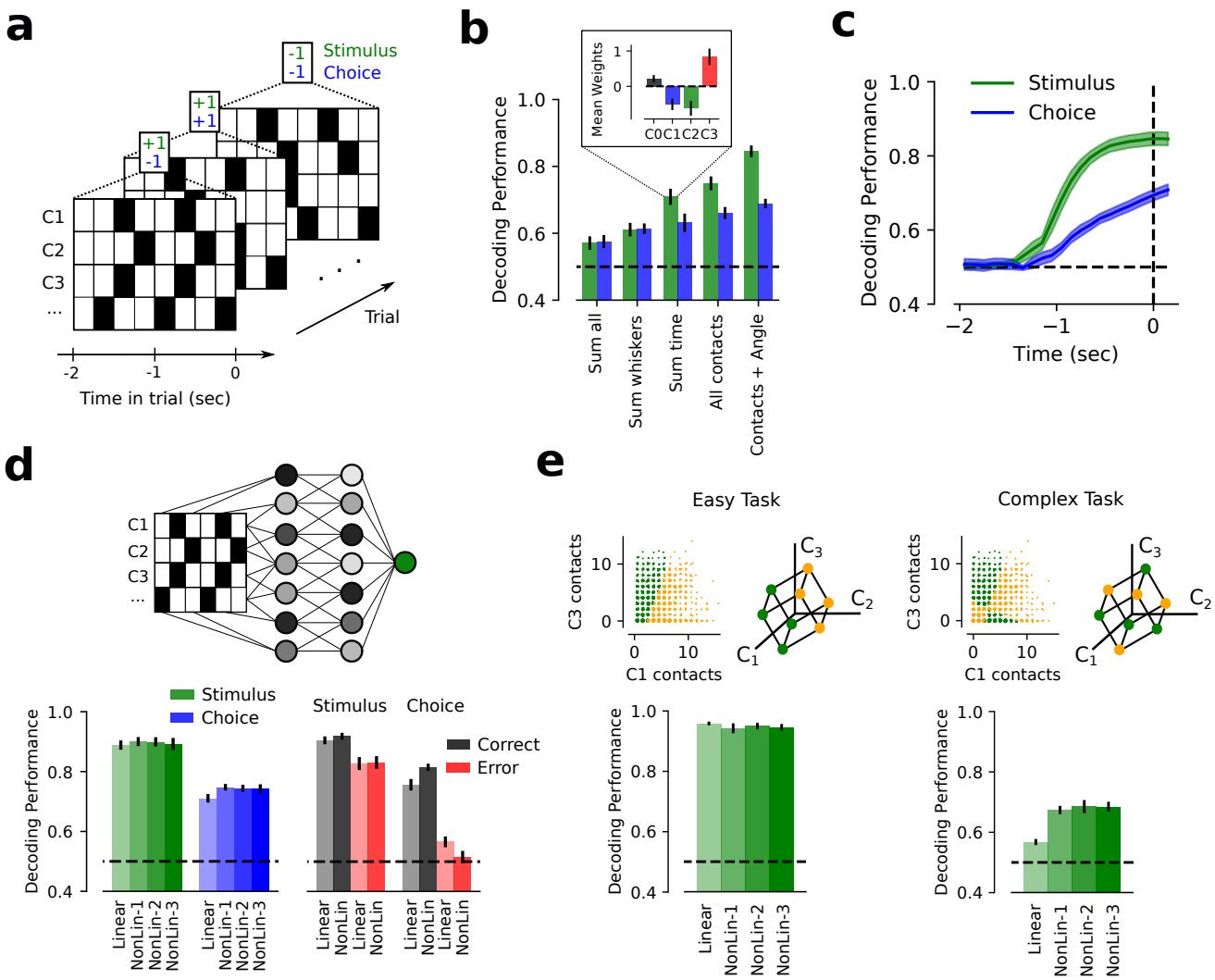

Figure 2: **The whisker-based shape discrimination task can be solved by linearly integrating sensory cues across time and whiskers**. (**a**) The recorded pattern of contacts across time and whiskers was used to classify shape identity (green) and animal choice (blue) on a trial-by-trial basis. (**b**) As input features to the classifiers we used: sum contacts across all time bins and whiskers ("Sum all"); sum across whiskers ("Sum whiskers"); sum across time ("Sum time"); all contacts from all time bins and whiskers ("All contacts"); and all contacts and angle of contacts ("Contacts + Angle"). Using all sensory cues across time and whiskers yields maximal performance for decoding shape identity and animal choice. (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**) When using the cumulative number of time bins, classifier performance increases as a function of time throughout the trial for both shape identity and animal choice. (**d**) Non-linear sensory cue integration across time and whiskers (fully connected neural networks) does not substantially improve classification performance compared to linear cue integration for either shape identity (stimulus; green) or animal lick side (choice; blue) (left panel). The same trend is observed for stimulus decoding when splitting between correct and error trials. On the contrary, correct trials conveyed much more information about animals' choice than incorrect trials (right panel). (**e**) Virtual easy (linearly separable; left panel) and complex (non-linearly separable, 3D-parity; right panel) tasks. Non-linear cue integration is only advantageous when the task itself requires complex sensory integration across time and whiskers. Error bars in all panels correspond to s.e.m. across mice.

mixed selectivity). We found that the activity of populations of barrel cortex neurons was best explained by a nonlinear mixed selectivity encoding model (1-layer neural network; $R^2 = 0.111 \pm 0.005$ on held-out data; see Methods) (Fig. 3d). 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 intermediate layers in the encoding model produced an increase of 23.7% in explanatory power. As expected, when all models were tested on train data, more parameters entailed better firing rate prediction (Supplementary Fig. S3). Models tested on correct trials showed better performance than those tested on incorrect trials, likely due to the reduced number of trials and the smaller number of contacts made on mistakes (see Fig. 1h).

The encoding models showed a higher performance for inhibitory neurons and neurons located in deeper layers of the somatosensory cortex (Supplementary Fig. S4), possibly due to their higher firing rates. Importantly, all of these results were qualitatively equivalent when a Poisson-loss function was used instead (Supplementary Fig. S5).

To assess the importance of each regressor, we calculated $\Delta R^2 = \Delta R^2_{Full} - \Delta R^2_{Reduced}$, which represents the loss in prediction power on held-out data when a particular regressor or group of regressors is set to zero (Fig. 4a, see Methods). Whisker contacts and continuous whisker angular position were the two most important variables for explaining the neuronal responses (Fig. 4b). Interestingly, superficial layers (2/3) were more strongly driven by sensory (contacts) than motion variables (whisker position), while deep layers (5 and 6) showed the opposite trend (Supplementary Fig. S6). As expected, the time kernel for whisker features showed a recency effect for all whiskers. Population activity was better predicted by contacts of C1 and C2 than by C3, whereas it was less well predicted by the angular position of C1 than by the other whiskers (Fig. 4c). In agreement with [8], we also found a deviation from classic somatotopy: C1 contacts were more strongly represented than C2 contacts in the C2 column, and than C3 contacts in the C3 column (Supplementary Fig. S7). The previous reward R_{-1} had the strongest effect early in the trial, while the current reward R_0 peaked after the response window opened and the animal made its choice (Fig. 4d). Although the current stimulus S_0 (shape category) and choice C_0 followed a similar trend throughout the course of the trial, C_0 had a stronger effect on the neural activity at the beginning of the response window. Similar task variable time profiles have been reported in previous studies in other animals and brain regions [6].

What does somatosensory cortex mix?

Different geometries of neural representations have different computational advantages. High dimensional representations, which require non-linear mixed selectivity, allow a linear readout like a downstream neuron to implement a large number of input-output functions. In other words, the readout can shatter the input space in a multitude of different ways, each corresponding to a different task that can be performed. The number of possible ways that a linear readout can separate the inputs divided by the total number of possible input-output functions is called the shattering dimensionality [3, 9], and it is high in the presence of non linear mixed selectivity. In contrast, low dimensional representations, which can be obtained with highly specialized neurons (pure selectivity) or with linear mixed selectivity, allow for better generalization to novel situations and are typically more robust to noise [5, 9].

The recorded neurons displayed a wide range of response properties: while some neurons showed linear mixed selectivity for C1, C2 and C3 contacts (Fig. 5a), others showed sublinear or XOR-like responses (Fig. 5b). To study non-linear mixed selectivity we could have fitted a linear model with interaction terms for each neuron. Instead, we decided to fit the neural network encoding model to all neurons simultaneously, so that we could determine which interaction terms were shared by multiple neurons from an agnostic approach. The state of activation of each unit in the intermediate layer is a weighted sum of the inputs passed through a non-linearity. Thanks to the non-linearity, the units' responses contain interaction terms. The training procedure that finds the weights of the intermediate units, tends to find automatically the interaction terms that are needed to reproduce the neuronal activity. These terms are then available to all the neurons whose activity we intend to reproduce.

Once the network was trained, we determined the contribution of nonlinear interactions of specific

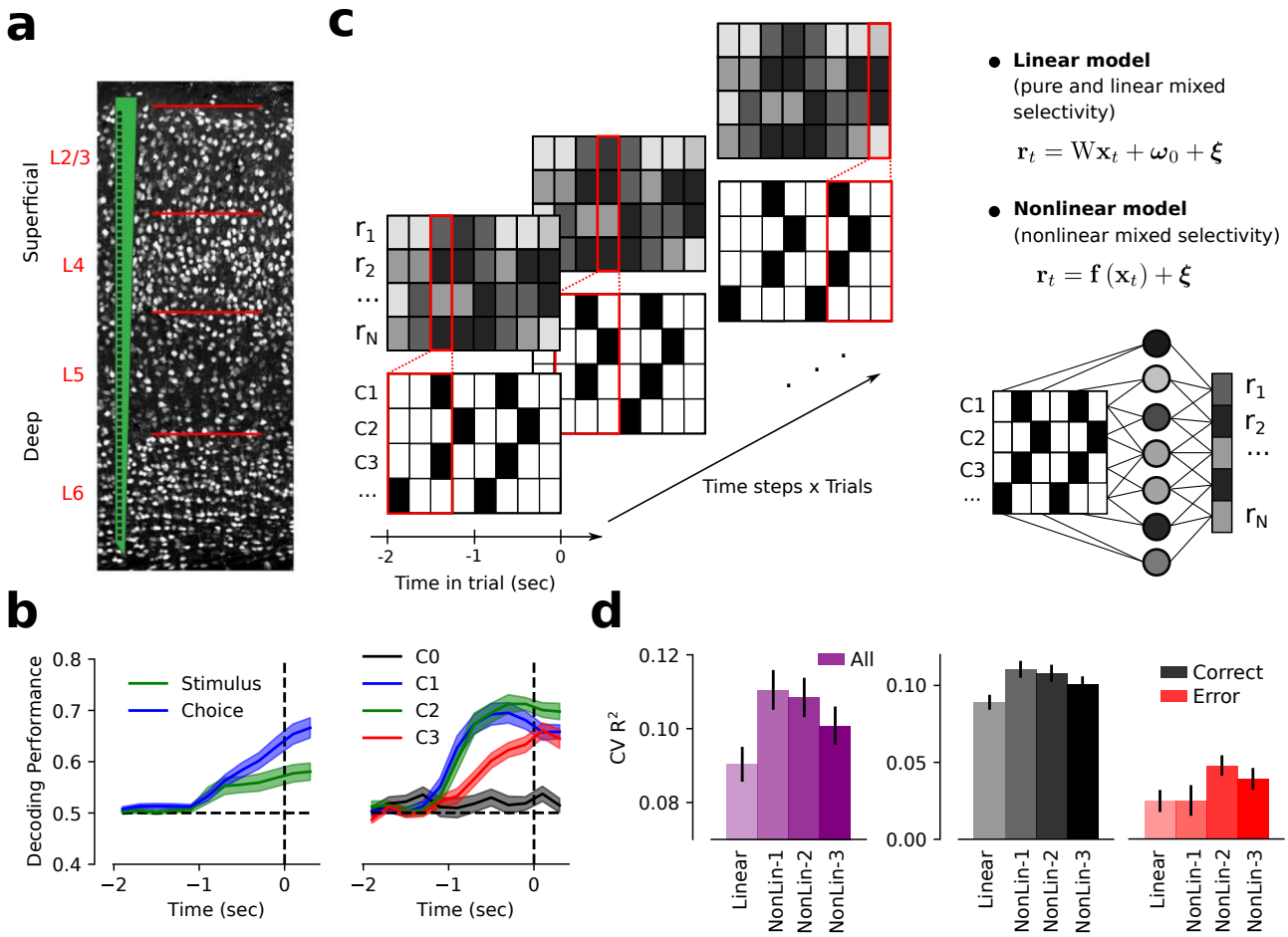

Figure 3: **Populations of neurons in mouse barrel cortex exhibit non-linear mixed selectivity for task variables**. (**a**) Populations of barrel cortex neurons were simultaneously recorded while mice performed a whisker-based discrimination task. (**b**) The performance of classifiers using neural data from different timepoints to predict stimulus or choice (left panel) or contacts made by each whisker (right panel). As expected, decoding performance is at random early in the trial when mice do not make contacts and it reaches 65% at the end of the trial. The identity of the whisker making contact could be decoded more accurately than the shape identity. (**c**) Barrel cortex activity was regressed against task variables. A linear model and non-linear models with different levels of flexibility were used to fit neural firing rate. (**d**) A fully connected neural network with one hidden layer (NonLin-1) outperforms the linear model and other neural networks with higher degrees of flexibility (left panel). Overall, encoding models explained S1 activity better on correct than error trials (right panel). Errorbars correspond to s.e.m. across neurons.

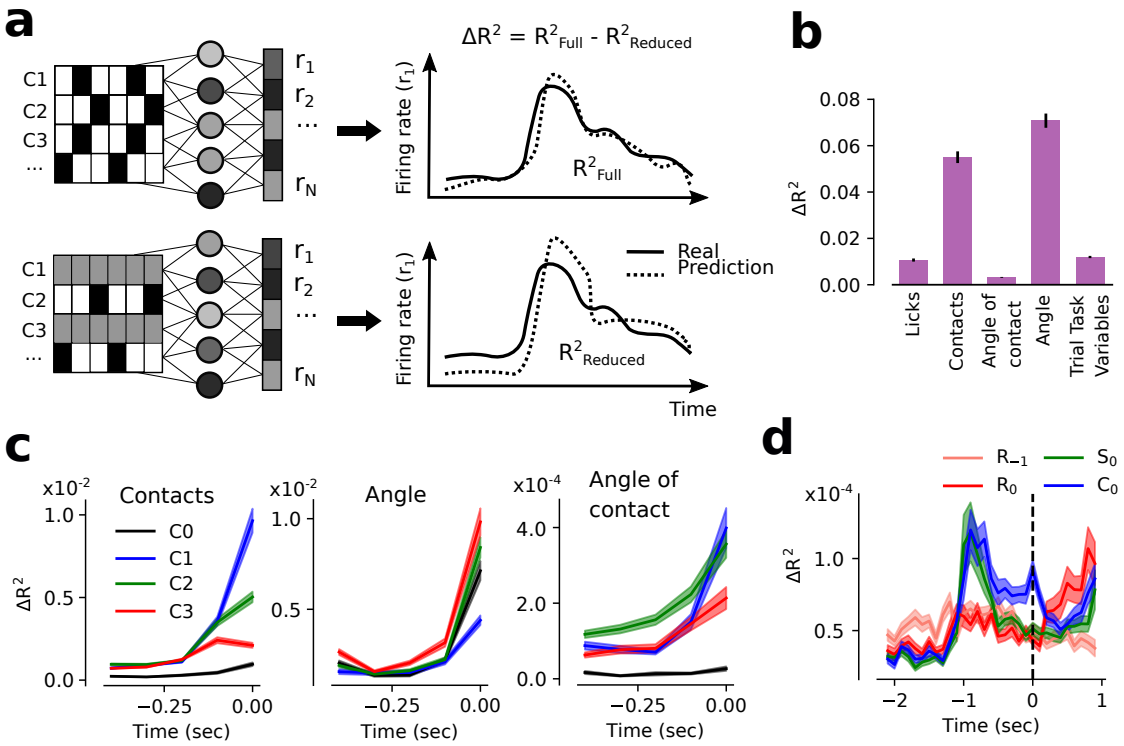

Figure 4: **Contacts and whisker angular position are the strongest predictors of S1 population activity**. (**a**) We fit an encoder model to explain the population's firing rate (r_1,r_2, ..., r_N) as a nonlinear function of task variables like whisker C1, C2 and C3 contacts, and calculated R^2_{Full}, the goodness-of-fit of the full model (Top panels). 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 panels). In this way we quantified the importance of each regressor 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 factors on barrel cortex activity as revealed by the decrease in model accuracy ΔR^2. (**c**) The decrease in model accuracy ΔR^2 for whisker contacts, angle of contact and angular position revealed that spikes were better explained by the most recent contacts. (**d**) Previous reward R_{-1} was encoded by the population early during the trial whereas current reward R_0 peaked after mice made their choice. Additionally, although current stimulus S_0 and choice C_0 followed a similar trend throughout the course of the trial, C_0 had a stronger effect on firing rate just as the response window opened at t = 0. Errorbars in all panels correspond to s.e.m. across neurons.

pairs of variables by setting the two variables to zero and evaluating both ΔR^2, which is the loss in explanatory power for the full non-linear model (Fig. 4b,c,d) and ΔR^2_{Linear}, which the analogous loss for the linear model. We then computed the difference $\Delta R^2 - \Delta R^2_{Linear}$ (see Methods and Supplementary Fig. S8). If this difference is close to zero, then the interaction term under consideration is not important, because the loss is the same whether we use the full non-linear model, which can compute the interaction term, or the linear model, which cannot. If the difference is large, then the interaction term is important.

We found that the interaction between whisker contacts and angular position was the most important one ($\Delta R^2 - \Delta R^2_{Linear} = 0.030$; Wilcoxon signed-rank test, $P < 0.001$; Fig. 5c). The interactions between whisker angular position and the rest of the variables were also strong. No significant deviations from this trend were observed when excitatory and inhibitory neurons and cortex layers were analyzed separately (Supplementary Fig. S9). We next assessed nonlinear mixed selectivity between different whiskers and at different timesteps, separately for contacts and for whisker position. For contacts, nonlinear mixed selectivity across whiskers was strongest for contacts that occurred slightly in the past (100ms), whereas for whisker angular position nonlinear mixed selectivity was strongest in the current time bin (Fig. 5d). Finally, we jointly examined nonlinear mixed selectivity between contacts and whisker position across all whiskers and at multiple timesteps. Interestingly, the strongest nonlinear mixing between angular position and contacts also occurred 100ms in the past (Supplementary Fig. S10). This time lag suggests that information about whisker contacts might be more strongly affected by network lateral connections or top-down signals than whisker angular position. Finally, similar to how neurons in C2 and C3 showed an unexpected response to C1 contacts (Supplementary Fig. S7), we observed that they also responded to non-linear interactions with the C1 whisker, regardless of their anatomical location (Supplementary Fig. S11).

What are the observed representations good for?

The representations generated by the non-linear neural network with one hidden layer turned out to be the best description of the observed representations. What type of tasks is it possible to perform when these representations are passed to a downstream non-linear readout? And in particular, what are the advantages and disadvantages of the non-linear components of the neuronal responses? To answer these questions we trained a linear readout to perform synthetic tasks when reading the neural representations generated by the non-linear model (see also Fig. 6a). We compared its performance to a readout that has access to the neural representations of the linear encoder, which are the best approximation of the observed representations when the neurons are allowed to have only linear components and no interaction terms. The difference in performance, which of course depends on the task, can highlight the computational advantages of linear vs non-linear representations in different situations.

Our approach based on encoders can separate the effects of linear and non-linear components, similarly to the method introduced in [3]. A significant advantage of this new approach is that it provides us with a full pattern of neural activity from the different encoding models for each individual trial.

Similar to Fig. 2e, we defined an easy and a complex task (left and middle columns in Fig. 6b) on the whisker contact variables (see Methods). As before, the easy task was designed to be solvable by a linear readout, while the complex task was designed not to be solvable by a linear readout from the whisker contacts. For the easy task, which does not require non-linearity, linear representations performed slightly better than non-linear representations. On the other hand, the small loss in the easy tasks is compensated by a significant advantage of non-linear representations in the complex task. Indeed, for complex tasks, the performance is at chance for linear representations. Finally, linear representations allow for a stronger form of generalization, cross-condition generalization performance (CCGP), which is a signature of a process of abstraction [9]. Not surprisingly, CCGP was higher for linear representations, which are lower dimensional. However, once again, as in the case of simple tasks, the advantage was small (right column in Fig. 6b; see Methods). Shattering dimensionality was also evaluated as in [3, 9], and the encoding model with one hidden layer yielded the highest dimensionality (Fig. 6c, see Methods). Finally, the ability to perform complex tasks and retain a

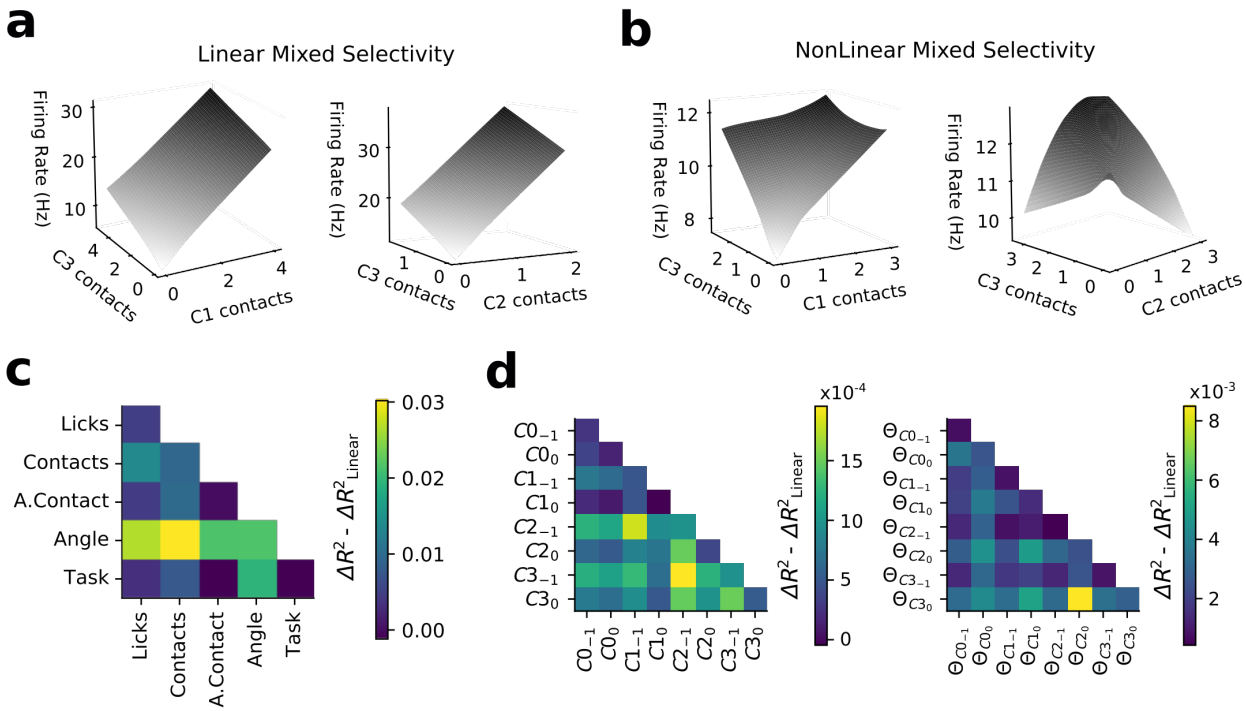

Figure 5: **Neurons show nonlinear mixed selectivity, especially for interactions between contacts and whisker angular position**. (**a-b**) Multidimensional tuning curves of example neurons that exhibit linear mixed selectivity (a) and non-linear mixed selectivity (b) for C1, C2 and C3 whisker contacts. All tuning curves were obtained from the best encoding model (non-linear with one hidden layer, see Fig. 3d). (**c**) Nonlinear mixed selectivity ($\Delta R^2 - \Delta R^2_{Linear}$) for the interaction between the different task variable groups. The interaction between whisker contacts and angular position was the most important nonlinear contribution to the encoding model. (**d**) Nonlinear mixed selectivity contribution for different time steps and whisker contacts (left) and angular position (right). The strongest interaction for contacts between whiskers occurs at the previous time step, while for angular position it occurs at the current time step.

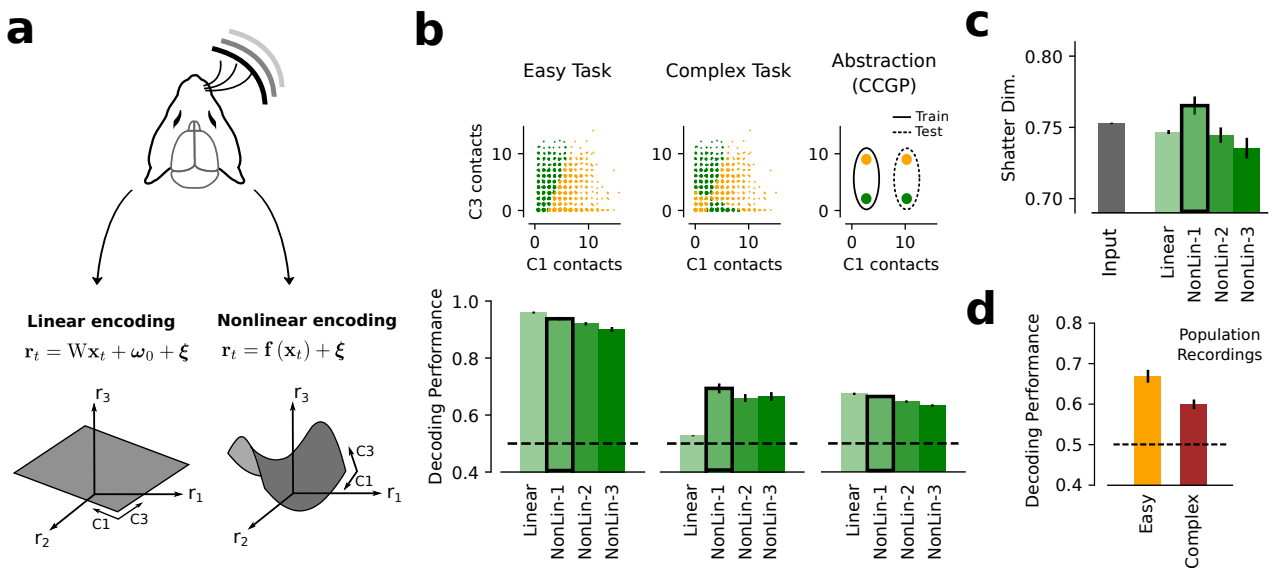

Figure 6: **Nonlinear sensory encoding in the barrel cortex allows for complex task performance while maintaining generalization properties**. (**a**) By combining the pattern of contacts sampled by the animals with the different encoding models (linear and non-linear), we generated surrogate neuronal activity that was used as input to decoders performing different tasks. (**b**) While surrogate neuronal activity generated with a linear encoding model shows high performance for easy task and abstraction, it is at chance level for the complex task (see Fig. 2e and Methods). Non-linear encoding models perform slightly worse at easy tasks and abstraction but well above chance for complex tasks. (**c**) Shattering dimensionality is the highest for denoised neuronal representations created with the non-linear encoding model with only one hidden layer. (**d**) Decoding performance was above chance on both the easy (orange) and the complex tasks (XOR; brown) when using the actual neuronal representations. The two tasks were defined on the basis of the total number of contacts at the choice period (see Methods). Errorbars in all panels correspond to s.e.m. across populations of simultaneously recorded neurons (recording sessions).

high performance on simple tasks is preserved if we use the actual representations instead of the representations generated by the non-linear encoding model, which are basically denoised non-linear representations (Fig. 6d, see Methods).

Overall these results suggest that, even though the whisker-based discrimination task can be solved by a linear combination of the contacts, S1 representations are efficiently tuned for a broader class of tasks, which can be significantly more complex. Importantly, the very same representations that allow a linear readout to perform a large number of complex tasks, are almost as robust to noise as the linear representations optimized for simpler tasks. So the observed non-linear representations allow a simple linear readout to perform rather complex tasks without compromising its ability to generalize to novel situations, or to be robust to noise.

The activity of recurrent neural networks trained to perform complex tasks is similar to the recorded activity

In the previous section we have seen that the mouse barrel cortex encodes the sensory stimuli by non-linearly combining the task variables even when such complex mixing is not required for the performance of the discrimination task. We next asked whether such high dimensionality is an unavoidable property of neural networks or whether it actually depends upon the complexity of the task. To answer this question we trained recurrent neural networks (RNNs) to perform simple and complex tasks that are similar to the shape discrimination task (Fig. 7a). We decide to use synthetic tasks to have complete control over their complexity. For simple tasks linear integration was sufficient, whereas for complex tasks non-linearity was required. Importantly, the RNNs were required to generate the correct response in the artificial tasks, and not to reproduce the neural data [15, 16].

On each trial, a three dimensional binary variable (representing contacts made by each of three whiskers) was fed into the RNN (60 ReLu noisy units). The trial consisted of a total of 30 time steps, and the input was only delivered during the first 20 time steps. The different input channels were Bernoulli processes with either a high or a low success rate λ, which made a total of 8 different experimental conditions. These are stochastic processes that generate sequences of binary variables that are supposed to mimic whisker contacts. Two different tasks were defined, the easy task (Fig. 7b) and the complex task (Fig. 7c), which consisted of a linearly and non-linearly separable task of the input space, respectively. Analogously to the whisker-based discrimination task, these two tasks require the accumulation of sensory cues over time. The easy task could be solved by linearly mixing information across input channels whereas the complex task required nonlinear mixing of sensory cues by the artificial units. For both networks, a read-out unit was trained to perform the task by linearly reading out the activity of the recurrent network and its performance increased as a function of elapsed time in both cases (Fig. 7b,c).

Similar to how we analyzed mouse behavior and S1 populations of neurons, we fitted classifiers and encoding models with different levels of complexity to both the input and the activity of the RNNs. We found that both the linear and non-linear classifiers performed equally well on the easy task by reading out from the input channels, while only non-linear classifiers performed above chance for the complex task (Fig. 8a). The activity of the neurons of the RNNs was also fit with different encoding models and we found that the advantage of including non-linear mixed selectivity was larger for the RNNs trained on the complex task (Fig. 8b). Indeed, for the RNN trained on the easy artificial task, non-linear encoding models produce a marginal improvement with respect to the linear one, which demonstrates that high dimensional representations are not an unavoidable property of neural networks. These results indicate that while the whisker-based discrimination task can be understood as a linear integration task through time and space, neuronal representations in the somatosensory cortex are better described under a non-linear mixed selectivity paradigm that has been optimized for complex behavioral tasks that require complex sensory combination across input channels.

Discussion

Characterizing the properties of neural representations in sensory brain areas during perceptual decision-making is an important step towards understanding the relation between brain activity and behavior. In this study, mice were trained to discriminate concave and convex shapes by actively scanning them with their whiskers. Mice learnt the task well and their performance grew monotonically with elapsed time, indicating that their choice was based on the accumulation of sensory evidence. The shape and the choice of the animal could be predicted using a simple linear decoder that reads the spatio-temporal whisking patterns and contacts. From this analysis we concluded that the whisker-based shape discrimination task is relatively simple, as it can be described as a linear integration task over time and whiskers. In contrast to the simplicity of the task, we observed that the activity of neurons in the somatosensory cortex is best described by a non-linear encoding model. Individual neurons exhibit diverse non-linear mixed selectivity to the whisker contacts and angles at different times. These complex neural representations are not needed to perform the task the animals have been trained on. This complexity is certainly important in more difficult tasks that require non-linear mixing but it could disrupt the performance in simple tasks, as non-linearity can amplify noise. However, the recorded representations actually represent a good compromise between the ability 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. Finally, recurrent neural networks (RNNs) trained to perform a simple and a complex equivalent of the whisker-based discrimination task reproduced the behavioral and electrophysiological results.

Previous studies showed that the dimensionality of neural representations can be maximal (monkey PFC [3]), very high (rodent visual cortex [7]), or as high as it can be given the structure of the task [17]. More recently, in [9] the authors showed that representations can have the maximal shattering dimensionality and, at the same time, a low dimensional scaffold which allows for cross-condition

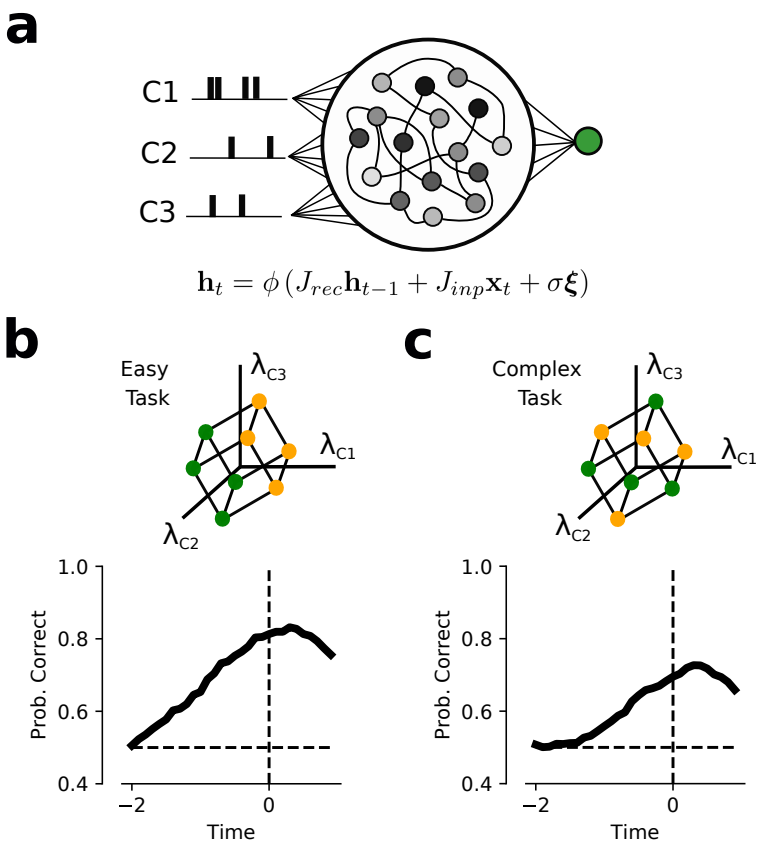

Figure 7: **Recurrent neural networks (RNNs) were trained on an artificial analog of the whisker-based shape discrimination task**. (**a**) A set of noisy and fully connected ReLu units receive input from three independent Bernoulli processes (C1, C2 and C3). On each trial, each channel corresponded to a Bernoulli instance drawn from with either a low or a high success rate parameter λ. (**b**) The easy task consists of a linear integration task across input channels (top panel). The performance of the neural network increases monotonically with elapsed time (bottom panel). (**c**) The complex task consists of a nonlinear integration task across input channels (3D-parity; top panel). The performance of the neural network also increases monotonically with elapsed time (bottom panel).

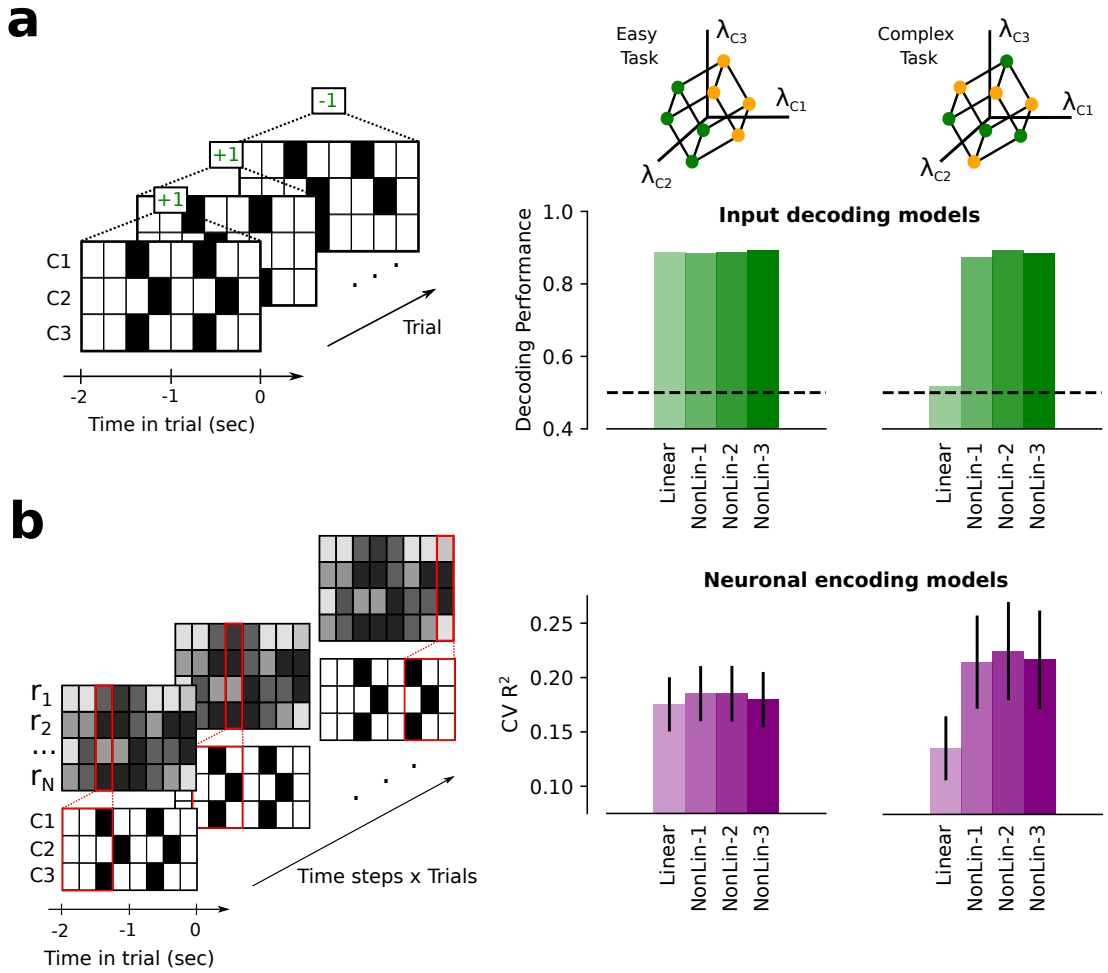

Figure 8: **The RNNs trained to perform the easy and the complex tasks reproduce the behavioral and electrophysiological results obtained in the whisker discrimination task**. (**a**) Linear and non-linear classification models that read-out from the input space perform equally well on the easy task. On the contrary, only non-linear classifiers that allow for complex cue combination perform above chance on the complex task. The behavioral results obtained on the whisker-based discrimination task are aligned with the easy task (see Fig. 2). (**b**) The encoding properties of the artificial units differ qualitatively between the easy and the complex RNNs. An RNN trained to perform a task that requires non-linear integration of sensory cues is better explained by an encoding model with non-linear mixed selectivity, while a non-linear encoding scheme provide little additional explanatory power for the easy task RNN. In contrast to (a), the encoding properties of the barrel cortex are aligned with those of an RNN trained to perform a complex task (see Fig. 3).

generalization. All these studies focused on computationally relevant properties of the geometry of the representations of neural populations, ignoring the detailed information about the response of individual neurons. Other studies looked more closely at the components of neuronal responses that are important for high dimensionality: mixing and diversity [18]. Here we adopted a new approach to characterize both collective properties of the representations and the dynamic response of individual neurons. Indeed, using the neural network encoding models we could characterize the response of an entire population of neurons. This is more than reproducing the responses of all the individual neurons of the population, because the encoding models can capture also the correlations between the activities of different neurons, which can be important to determine the geometry of the representations and their effects on information and behavior [19]. This approach was motivated by the fact that the task involves some form of active sensing, which is closer to natural behavior but terribly difficult to analyze. In contrast to the monkey experiments mentioned above [18], the trial temporal structure is highly variable as it depends on the way the animal decides to move the whiskers. In our case we did not have a small set of discrete variables that could characterize the sensory input. So we used a more unbiased approach to identify those variables that could be important to predict the behavior and the neural activity of the animal. We started from a much larger set of variables that characterize complex spatio-temporal patterns, and we let the encoding model find those that are most important.

One important issue is that non-linear mixed selectivity is always defined with respect to a set of variables[20]. These variables typically characterize the sensory input and the motor output. Also in our case non-linearities are defined with respect to a set of variables, and this could be one of the limitations of our analysis. However, it is important to stress that we used the same variables to characterize the task and the neural activity. The main result is that a linear combination of these variables is the best predictor of the stimulus identity and the choice of the animal, but not of the neural activity. Even when we considered additional variables (e.g. 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.

Recently, it has been shown that neural network models (Deep Convolutional Networks, DCN) provide us with the best description of the neuronal responses in the primate's visual system [15, 16, 21]. All these models, which are constructed by training the networks to perform a classification task, include non-linearities which certainly play a fundamental role [22, 23, 24]. These results showed also that the classical concept of a receptive field, which is inherently linear, must be revisited. Interestingly, this approach has even been used to produce synthetic visual stimuli that are able to maximally activate real neurons [25, 26]. Similarly, accurate models of the somatosensory system could be used to synthesize artificial somatosensory stimuli to drive neuronal activity or even behavior. We believe this study represents a first step in this direction.

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 in the field [27, 28, 29, 30]. 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 geometry of neural representations like the shattering dimensionality can 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 without incurring such unfavorable 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.

Methods

Behavioral task and recordings

This experiment has been described in detail in [8]. Here we provide a brief summary of the behavioral setup and data acquisition.

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. 1a). 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 ($t = 0$) (Fig. 1c,d).

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 [11, 10], which is the 'feature detector' network used in the first version of DeepLabCut [12]. 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 barrel cortex during the whisker-based shape discrimination task. 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 barrel cortex, 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 [31] to detect spikes and to assign them to putative single units. We identified inhibitory neurons from their waveform half-width, i.e. 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. Recordings were performend in layers 2/3, 4, 5 and 6 of the primary somatosensory cortex (S1). Neurons in layer 1 or the cortical subplate were discarded from this analysis because they were difficult to sort and showed variable properties across mice.

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 of Behavior

On each trial, we built a matrix that contained behaviorally relevant features 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 [8]. The total amount of features on each trial was 160, 8 whisker features times 20 time bins. 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. Each feature was normalized to null mean and unit standard deviation. Trials that did not registered 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,d, 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,d, 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 both Fig. 2b and d, 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 perceptrons with 0, 1, 2 or 3 hidden layers of 100 ReLu units). In the following, we use multilayer perceptron with 0 hidden layers and linear classifier 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 fitted the multilayer perceptrons 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 correspond to the mean performance on the validation set (see above) after splitting trials into correct and error.

In panel Fig. 2e we created two *ad-hoc* tasks from the spatio-temporal whisking patterns gathered by the animals, the easy and the complex tasks. Mice were never trained on these tasks, they correspond to tasks that have been defined on the whisker contact space *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. For the complex task, trials were split into two non-linearly separable classes (XOR) 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 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. For each mouse, the easy or 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. Errorbars in panels Fig. 2b,c,d,e correspond to the standard error of the mean (s.e.m.) across mice (super-sessions).

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*. For each recording session we concatenated all the time steps across trials (Fig. 3b). 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 $\times$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 $\times$ 5 time steps) plus 248 *trial-variables* (248 = 8 variables $\times$ 31 time steps). Both neuronal activity and regressors were normalized to null mean and unit variance. Trials that did not registered 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 (across layers 2/3, 4, 5 and 6) and 4883 mean number of effective trials used to fit the models (trials $\times$ time steps).

We analyzed the encoding properties of populations of neurons in the mouse barrel cortex by regressing the neuronal activity against the experimental variables described above. Similarly 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). We fitted 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. 3). To validate our results with a different loss function, Poisson-loss was also used to fit the models (Supplementary Fig. S5). Importantly, in our setup the encoding model with 0 hidden layers (linear model) is equivalent to a linear regression. 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 [3, 5]. 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 $[10^{-7}, 2]$ and $[10^{-7}, -1]$ 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. All the encoding models were implemented in pytorch and optimized with the ADAM algorithm. The reported decoding performances on correct and error trials correspond to the mean performance on the validation set (see above) after splitting trials into correct and error (Fig. 3c). Errorbars in Fig. 3 correspond to s.e.m. across recorded neurons.

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. 4a), 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 x_i is set to zero. 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. 5). 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.

We also generated surrogate neuronal data on each time step by introducing some of the experimental variables used as regressors into the different encoding models (Fig. 6a). This procedure was only performed on the validation partition. Similarly to the behavioral case, we defined two different virtual tasks, the easy task and the complex task, which corresponded to a linear task and a XOR on the input variables (Fig. 6b). We used C1, C2 and C3 contacts for the current time step to create these virtual tasks. In order to have a well defined median for all these three variables, Gaussian white noise was introduced in these regressors before creating the surrogate neuronal activity (standard deviation of 1). Both the easy and the complex task were defined as 2D tasks, 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}. For instance, for {C1,C3} we generated surrogate activity by introducing in the different encoding models only the experimental variables contacts C1 and C3 for the current time step. On the easy task the coloring of the different regions in the C1-C3 space was defined by a linear boundary, whereas for the complex task it corresponded to a XOR task. In both cases the task boundaries were obtained by performing a random unitary rotation on the C1-C3 space and splitting based on the medians. From the surrogate neuronal data, linear classifiers (cross validated logistic regression) were fitted to perform these two virtual tasks. The reported performance in Fig. 6b corresponds to the mean performance across cross-validations and the three pairs of regressors used to define the task. We also evaluated cross-condition generalization performance (CCGP; Abstraction) on the surrogate data in a similar way to [9]. For instance, for the pair of regressors C1 and C3, we added Gaussian noise and split all trials based on their medians. Four experimental conditions and three different dichotomies were obtained. Abstraction was evaluated as the performance of a linear classifier trained on two experimental conditions and tested on the other two. The reported values values of Abstraction in Fig. 6b correspond to the mean across the the different cross-validations, dichotomies (not including the xor-dichotomy) and pairs of variables. To evaluate shattering dimensionality, similarly to performance on the easy and complex tasks (Fig. 6b), we split all trials into high and low contacts for C1, C2 and C3 from their median values after adding white noise to obtain a uniquely defined median (standard deviation of 1). We obtained eight different experimental conditions ($2^3 = 8$) from two different values (high or low) of three variables (C1, C2 and C3). For 100 iterations we randomly colored the eight different experimental conditions and fitted a linear classifier (cross-validated Logistic Regression; sklearn) on the input variables (black) and on the surrogate neuronal activity generated from the four different encoding models (green). The reported values of shatter dimensionality correspond to the mean across cross-validations, colorings and recording sessions. Errorbars in all Fig. 6 panels correspond to s.e.m. across recording sessions.

Population Decoding

Populations of mouse barrel cortex neurons were recorded during the whisker-based shape discrimination task. Linear classifiers were fitted 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 (Figs. 3b and 6d). Time bins of 200 ms were used and population activity was normalized to null mean and unit variance. Trials that did not registered 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 (Fig. 3b). Similarly 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 (Fig. 3b). 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.

We defined an Easy and a Complex task from the pattern of contacts and predicted the category label by reading-out population activity on a trial-by-trial basis (Fig. 6d). The two tasks were defined as a linearly separable task (easy task) and a XOR on the contacts space (complex task). We used whiskers C1, C2 and C3 and constructed 3 sub-tasks for both the easy and the complex task from the 3 pairs {C1,C2}, {C1,C3} and {C2,C3}. For example, for the pair {C1,C3}, on each time step trials were split into high or low number of contacts for each whisker by summing the number of contacts each whisker made up to that point. As before, Gaussian noise was added (s.d. of 0.1) to obtain a unique median. A total of 4 different experimental conditions were obtained with this method. The contacts space for this particular pair was transformed by a unitary rotation and the easy and the complex tasks were defined by a linear boundary and an XOR from these 4 experimental conditions, respectively. The reported decoding performances correspond to the mean across the 3 sub-tasks (C1 vs C2, C1 vs C3 and C2 vs C3) and across cross-validations on the validation set after optimizing regularization strength on the test set.

Recurrent Neural Networks

In order to further validate the results obtained at both the behavioral and electrophysiological levels, recurrent neural networks were trained to perform an artificial analogous of 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:

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

where $\phi()$ is the ReLu non-linearity, $\boldsymbol{\xi}$ is independent and unitary Gaussian noise 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 (3 channels $\times$ 2 conditions per channel) (Fig. 7a). From these 8 experimental conditions two different tasks were defined, the easy and the complex task. For both tasks, the input information was transformed by a unitary rotation (same rotation in all time steps). The easy task was defined as a task that linearly separated the 8 experimental conditions into 2 groups of 4 (Fig. 7b), while the Complex task was defined as a 3D-XOR (3D-parity) on the rotated stimulus space (Fig. 7c). 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. The networks were trained to make a decision at $T = 20$ in all trials. The two 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^{-10} and the learning rate $\eta = 0.005$. We used ADAM as optimizer, batches of 20 trials and as many epochs as necessary to reach 10^{-3} error on the loss function ($\sim$ 10 epochs for the easy task and $\sim$ 50 epochs for the 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 and $\lambda_{low} = 0.23$ and $\lambda_{high} = 0.77$ for the complex task.

We analyzed the complexity of the easy and complex tasks in the same way we analyzed the complexity of the whisker-based shape discrimination task. For both networks, we used different classifiers with different levels of flexibility (multi-layer perceptrons with 0, 1, 2 or 3 hidden layers of 100 ReLu units). These classifiers were trained to predict the output of the easy and the complex task on a trial-by-trial basis by reading-out the spatio-temporal pattern that was used as input to the networks (Fig. 8a). The $l2$ regularization strength and the learning parameter η were optimized over the ranges $[10^{-6}, 1]$ and $[10^{-5}, 1]$ respectively. .

The encoding properties of the easy and complex task units were also analyzed in the same way we analyzed the population activity of mouse barrel cortex neurons (Fig. 8b). In this case, the $l2$ regularization strength and the learning parameter η were optimized over the ranges $[10^{-5}, 10^{2}]$ and $[10^{-6}, 1]$ respectively.

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); NeuroNex (DBI-1707398); the Gatsby Charitable Foundation (GAT3419); 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 wrote and CR, RB, and SF edited the manuscript.

Competing interests

The authors declare no competing interests.

References

[1] Johansson, R. S. & Flanagan, J. R. Coding and use of tactile signals from the fingertips in object manipulation tasks. *Nature Reviews Neuroscience* **10**, 345–359 (2009).

[2] Bensmaia, S. J., Tyler, D. J. & Micera, S. Restoration of sensory information via bionic hands. *Nature Biomedical Engineering* 1–13 (2020).

[3] Rigotti, M. *et al.* The importance of mixed selectivity in complex cognitive tasks. *Nature* **497**, 585 (2013).

[4] Barak, O., Rigotti, M. & Fusi, S. The sparseness of mixed selectivity neurons controls the generalization-discrimination trade-off. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **33**, 3844–3856 (2013).

[5] Fusi, S., Miller, E. K. & Rigotti, M. Why neurons mix: high dimensionality for higher cognition. *Current opinion in neurobiology* **37**, 66–74 (2016).

[6] Nogueira, R. *et al.* Lateral orbitofrontal cortex anticipates choices and integrates prior with current information. *Nature Communications* **8**, 1–13 (2017).

[7] Stringer, C., Pachitariu, M., Steinmetz, N., Carandini, M. & Harris, K. D. High-dimensional geometry of population responses in visual cortex. *bioRxiv* 374090 (2018).

[8] Rodgers, C. C. *et al.* The sensorimotor strategies and neuronal representations of tactile shape discrimination in mice. *BioRxiv* (2020).

[9] Bernardi, S. *et al.* The geometry of abstraction in the hippocampus and prefrontal cortex. *Cell* (2020).

[10] 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).

[11] 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).

[12] Mathis, A. *et al.* Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. *Nature neuroscience* **21**, 1281–1289 (2018).

[13] Ashwood, Z. C. *et al.* Mice alternate between discrete strategies during perceptual decision-making. *bioRxiv* (2020).

[14] Calhoun, A. J., Pillow, J. W. & Murthy, M. Unsupervised identification of the internal states that shape natural behavior. *Nature neuroscience* **22**, 2040–2049 (2019).

[15] Yamins, D. L. *et al.* Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences* **111**, 8619–8624 (2014).

[16] Yamins, D. L. & DiCarlo, J. J. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience* **19**, 356–365 (2016).

[17] Gao, P. & Ganguli, S. On simplicity and complexity in the brave new world of large-scale neuroscience. *Current opinion in neurobiology* **32**, 148–155 (2015).

[18] 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. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **37**, 11021–11036 (2017).

[19] Nogueira, R. *et al.* The effects of population tuning and trial-by-trial variability on information encoding and behavior. *Journal of Neuroscience* **40**, 1066–1083 (2020).

[20] Hirokawa, J., Vaughan, A., Masset, P., Ott, T. & Kepecs, A. Frontal cortex neuron types categorically encode single decision variables. *Nature* **576**, 446–451 (2019).

[21] Kar, K., Kubilius, J., Schmidt, K., Issa, E. B. & DiCarlo, J. J. Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature neuroscience* **22**, 974–983 (2019).

[22] Ramirez, A. *et al.* Spatiotemporal receptive fields of barrel cortex revealed by reverse correlation of synaptic input. *Nature neuroscience* **17**, 866–875 (2014).

[23] Peron, S. P., Freeman, J., Iyer, V., Guo, C. & Svoboda, K. A cellular resolution map of barrel cortex activity during tactile behavior. *Neuron* **86**, 783–799 (2015).

[24] Petersen, R. S. *et al.* Diverse and temporally precise kinetic feature selectivity in the vpm thalamic nucleus. *Neuron* **60**, 890–903 (2008).

[25] Bashivan, P., Kar, K. & DiCarlo, J. J. Neural population control via deep image synthesis. *Science* **364** (2019).

[26] Xiao, W. & Kreiman, G. Xdream: Finding preferred stimuli for visual neurons using generative networks and gradient-free optimization. *PLoS computational biology* **16**, e1007973 (2020).

[27] Gulli, R. A. *et al.* Context-dependent representations of objects and space in the primate hippocampus during virtual navigation. *Nature neuroscience* **23**, 103–112 (2020).

[28] Roussy, M. *et al.* Naturalistic coding of working memory in primate prefrontal cortex. *bioRxiv* (2020).

[29] Nelson, M. E. & MacIver, M. A. Sensory acquisition in active sensing systems. *Journal of Comparative Physiology A* **192**, 573–586 (2006).

[30] 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).

[31] 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. *Advances in neural information processing systems* **29**, 4448–4456 (2016).

Supplementary Information

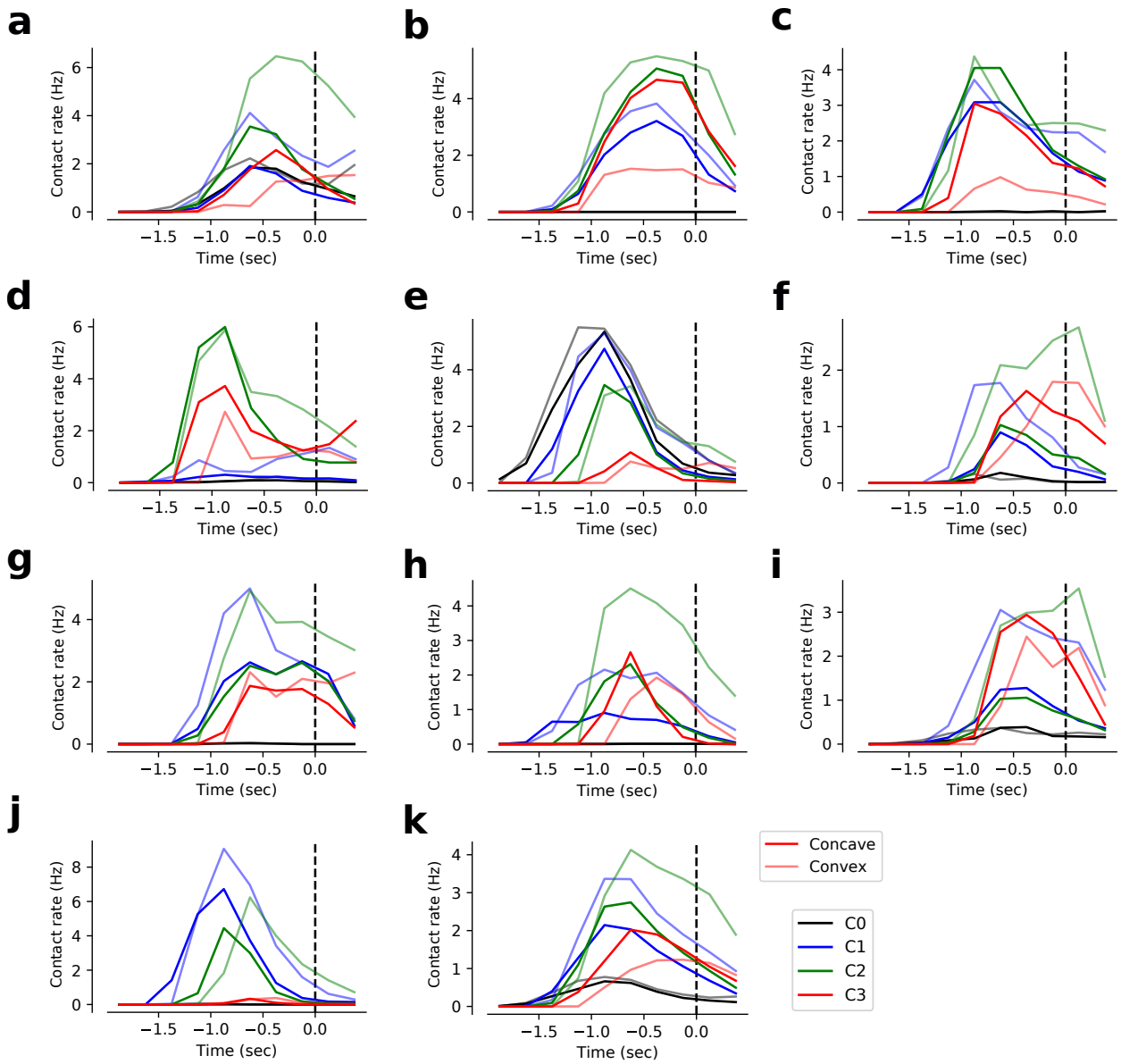

Figure S1: Contact rate was higher for convex than concave shapes in C1 and C2, whereas C3 showed the opposite trend. (**a-j**) Results for all the mice. (**k**) mean results across mice.

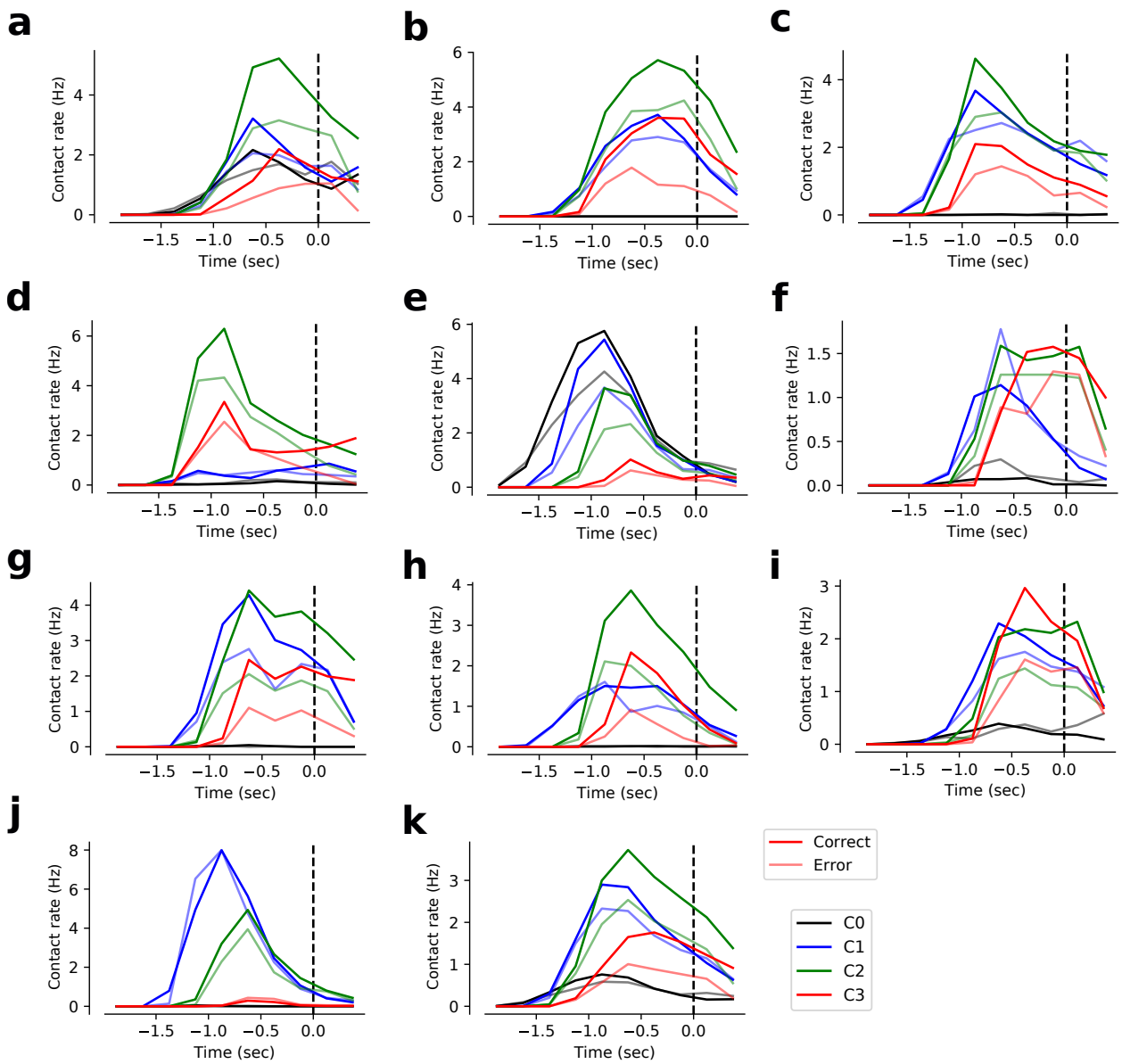

Figure S2: Contact rate was higher for correct than error trials for all whiskers and animals, (**a-j**) Results for all the mice. (**k**) mean results across mice.

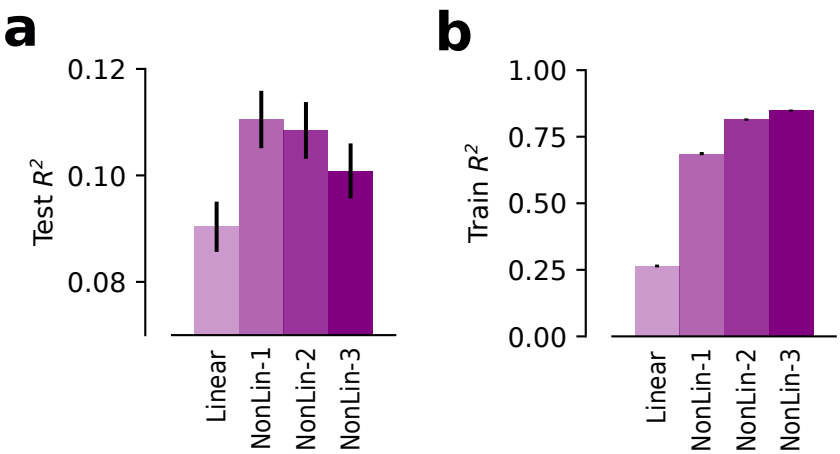

Figure S3: (**a**) The best encoding model when tested on held-out data is a feedforward fully connected network with only one hidden layer (NonLin-1). (**b**) As expected, when tested on the train set, the more complex the model (more parameters), the better the prediction.

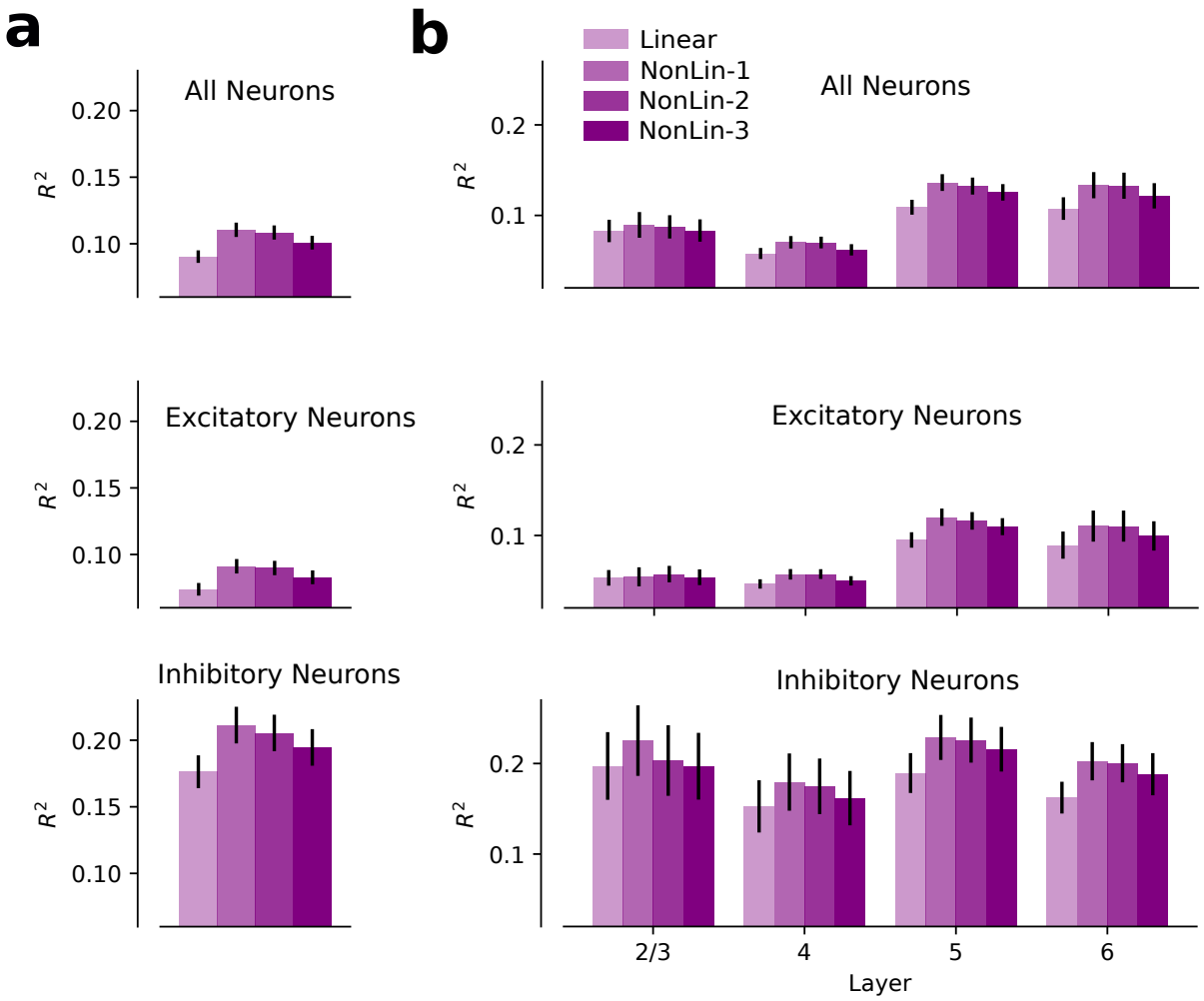

Figure S4: The encoding models explain neurons firing rate best for inhibitory neurons and deeper layers. (**a**) Performance of the different encoding models on held-out data for all neurons (top), only excitatory (middle) and only inhibitory neurons (bottom). (**b**) Performance of the different encoding models on held-out data for neurons across layers for all (top), excitatory (middle) and inhibitory neurons (bottom).

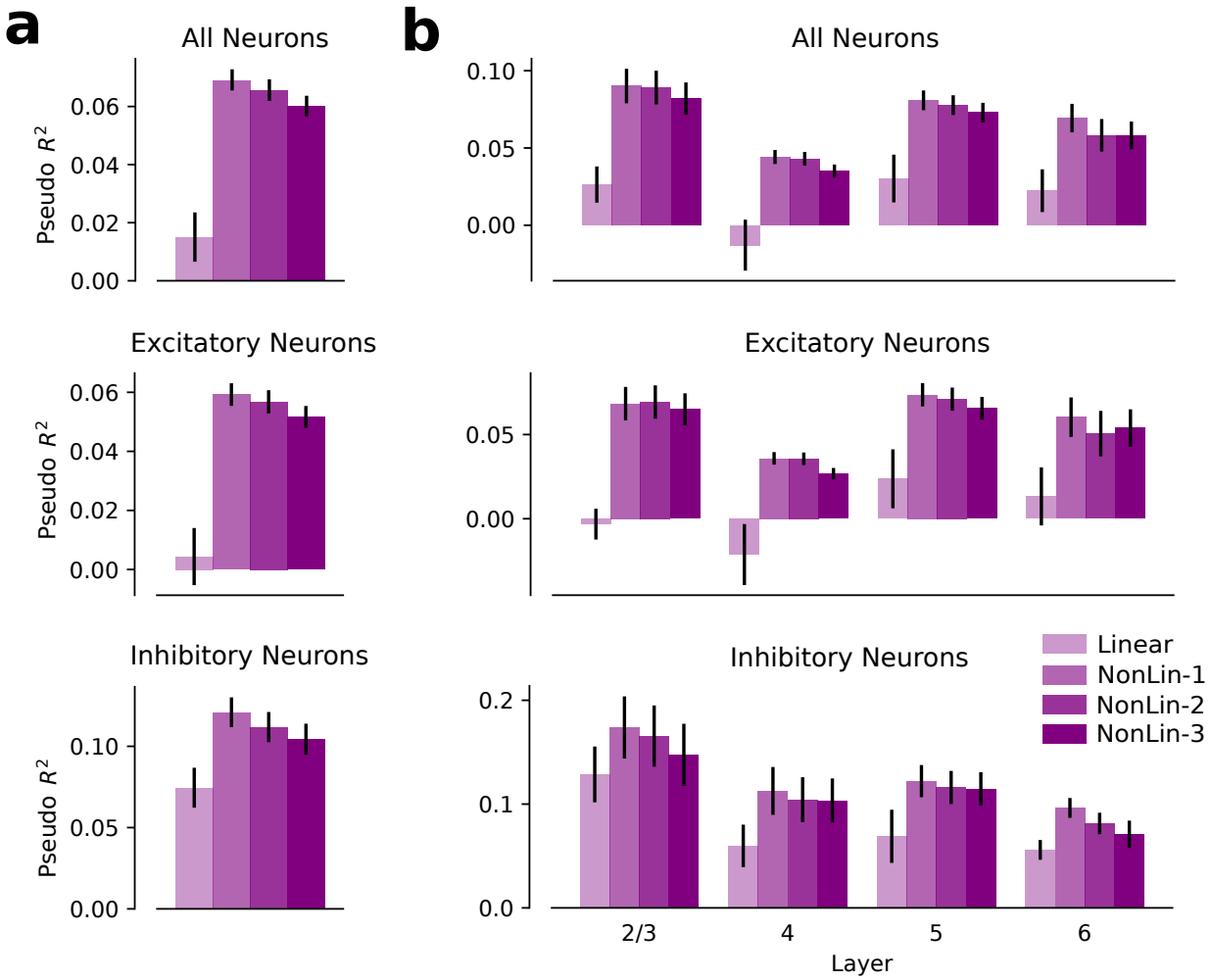

Figure S5: Encoding fits when using Poisson loss instead of mean squared error (MSE). The y-axis shows the Poisson-loss equivalent of the R^2, the Pseudo-R^2. The Pseudo-R^2 is calculated as 1 - LLH/Variance, where LLH is the Log-likelihood of the Poisson model. (**a-b**) see Figure S4.

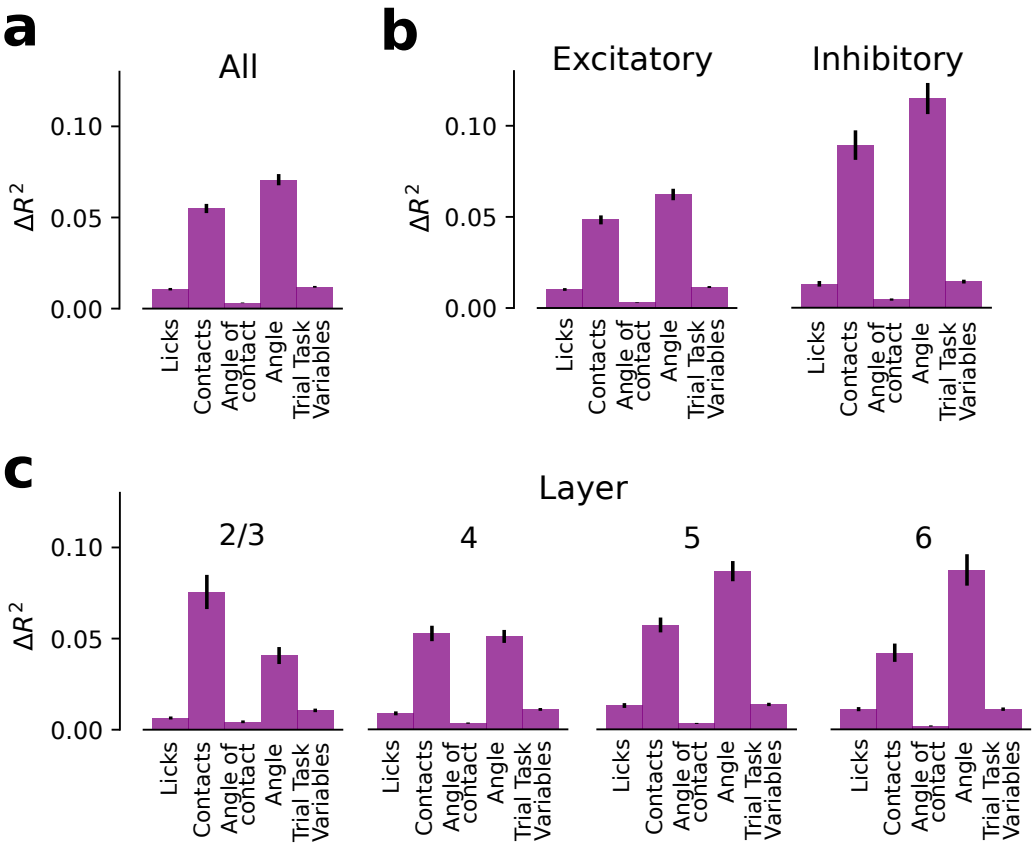

Figure S6: Difference between R^2 of the full model and the R^2 for the different ablations (ΔR^2). The R^2 for the different ablations is calculated by testing the model on trials where those features have been set to zero. (**a**) ΔR^2 for all neurons. Whisker contacts and angle are the most important features in predicting barrel cortex activity. (**b**) ΔR^2 excitatory and inhibitory neurons. Inhibitory populations show a higher ΔR^2 because their R^2 is overall higher (see Fig. S4). (**c**) ΔR^2 across layers of the barrel cortex. Whisker contacts have a stronger effect on superficial layers (2/3), while whisker angle has a stronger effect on deeper layers.

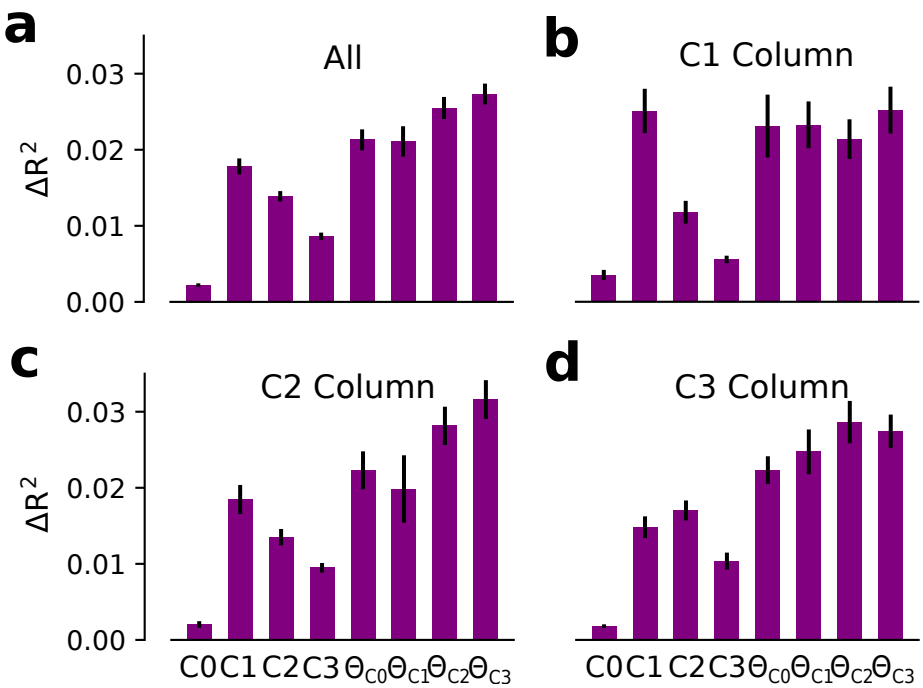

Figure S7: The metric ΔR^2 reveals that somatotopy is not present in the barrel cortex during a whisker-based discrimination task. (**a**) The contribution to the encoding model's performance for whiskers' contacts (C0, C1, C2 and C3) and angular position (θ_{C0},θ_{C1},θ_{C2}, and θ_{C3}) for all neurons. (**b-c**) whisker and angular position encoding strength for C1 barrel (b), C2 barrel (c) and C3 barrel (d). While C1 contacts is the strongest driver in C1 barrel, C2 and C3 barrels are not dominated by C2 and C3 contacts, respectively. Errorbars in all panels correspond to s.e.m. across neurons.

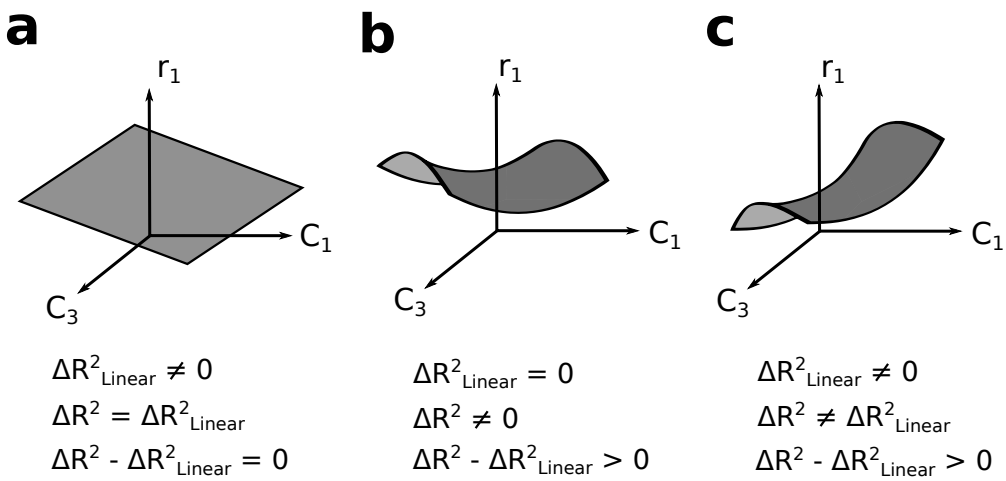

Figure S8: (**a**) The metric $\Delta R^2 - \Delta R^2_{Linear}$ was used to evaluate to what extend the pure nonlinear terms were important to predict the population's firing rate. If the relationship between neuronal activity and encoding variables is linear, $\Delta R^2 = \Delta R^2_{Linear}$ and therefore $\Delta R^2 - \Delta R^2_{Linear} = 0$ (**b**) If the relationship between neuronal activity and encoding variables is purely non-linear, $\Delta R^2_{Linear} = 0$, $\Delta R^2_{Linear} \neq 0$ and $\Delta R^2 - \Delta R^2_{Linear} > 0$. (**c**) If the encoding model is composed by both linear and nonlinear components, $\Delta R^2_{Linear} \neq 0$, $\Delta R^2 \neq 0$ and $\Delta R^2 - \Delta R^2_{Linear} > 0$.

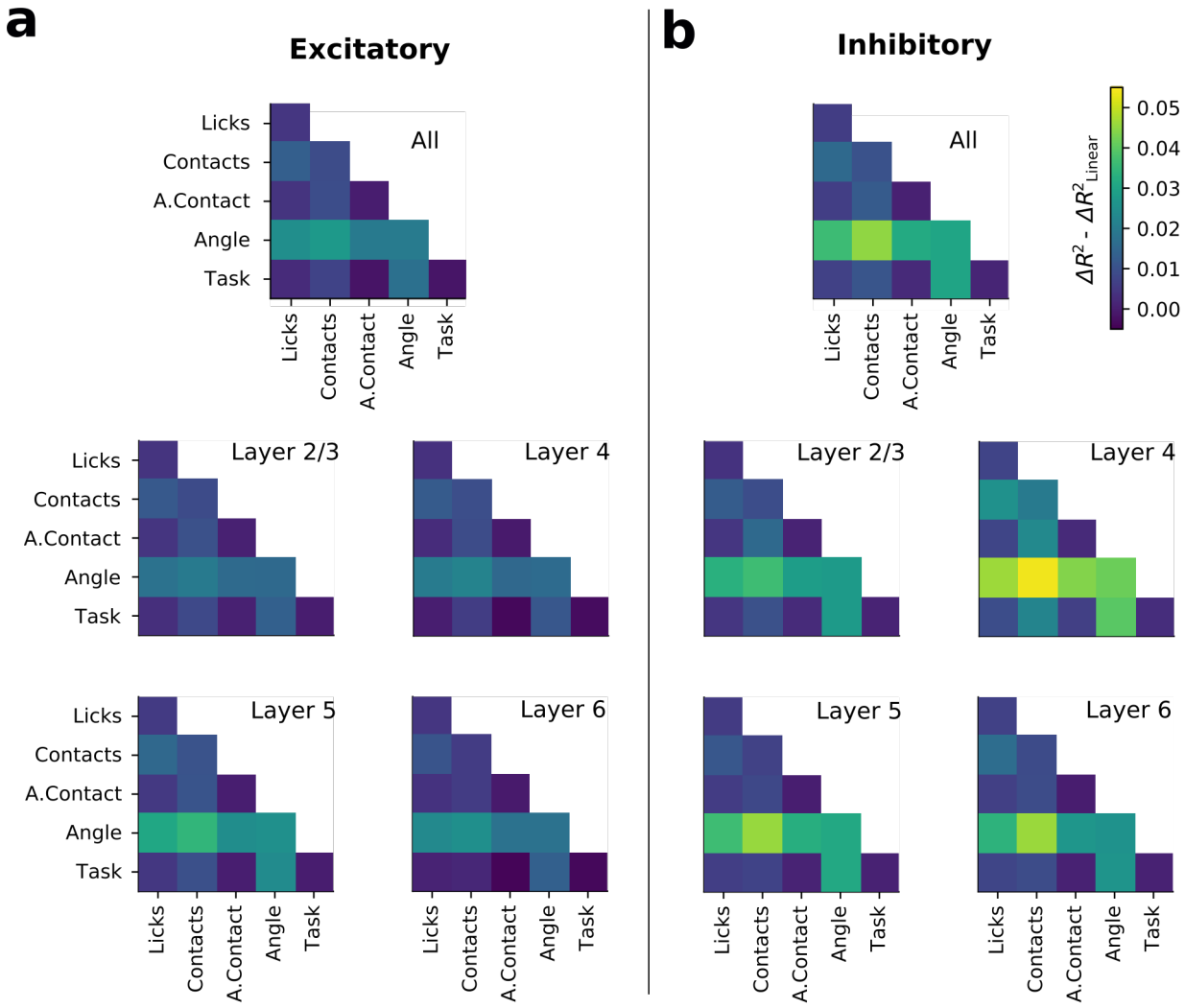

Figure S9: Pure nonlinear mixed selectivity contribution $\Delta R^2 - \Delta R^2_{Linear}$ for the interaction between the different blocks of variables across neuronal types and S1 layers. Results were qualitatively equivalent for the different populations.

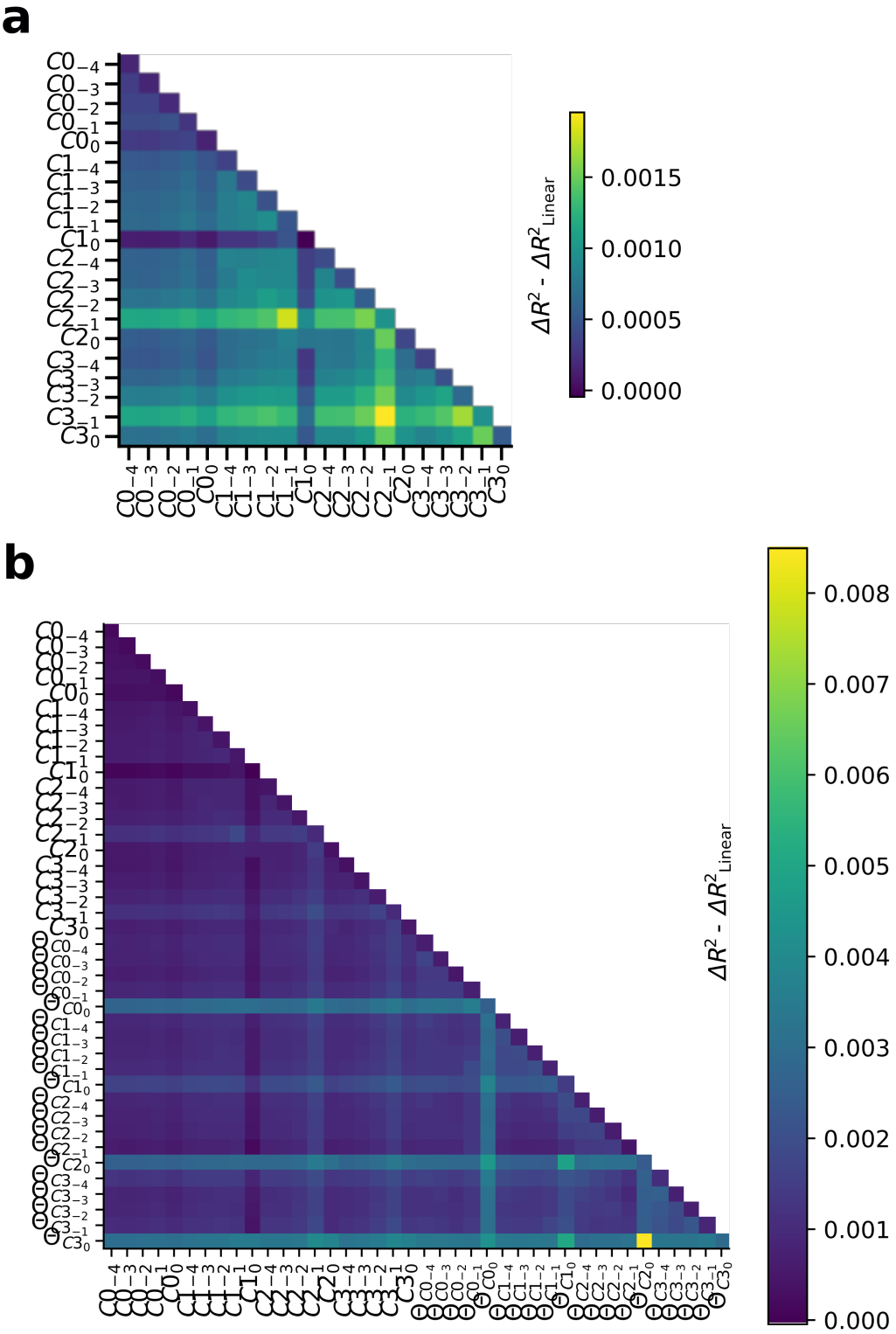

Figure S10: (**a**) Pure nonlinear mixed selectivity contribution for the interaction between contacts for the different time steps and whiskers. The strongest nonlinear contribution in whisker contacts occurs on the previous time step for all whiskers. (**b**) Pure nonlinear mixed selectivity contribution for the interaction between contacts and angular position for the different time steps and whiskers. The strongest nonlinear contribution in whisker angular position occurs on the current time step for all whiskers. The strongest nonlinear contribution for the interaction between contacts and angular position occurs also at the previous time step.

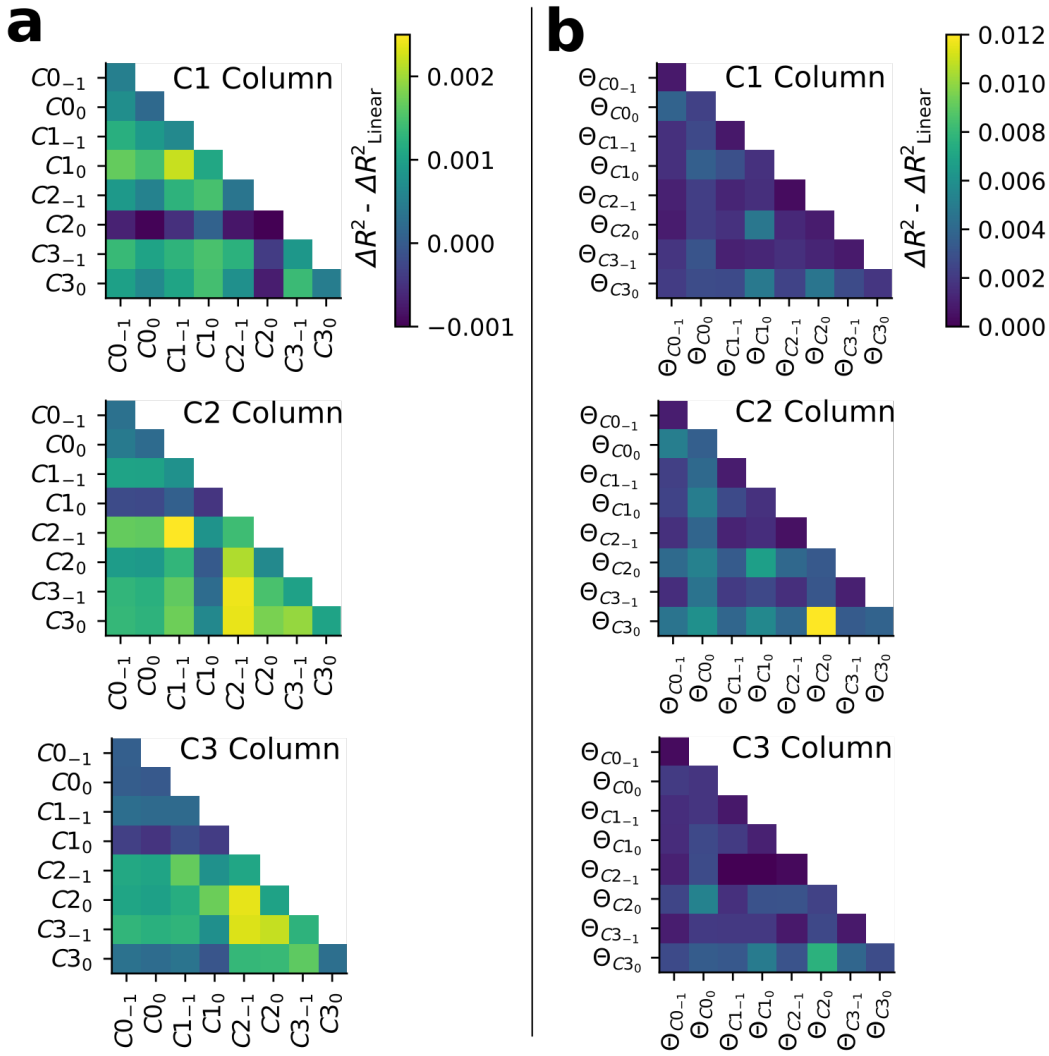

Figure S11: (**a**) Pure nonlinear mixed selectivity contribution for the interaction between contacts for the different time steps and whiskers. The strongest interactions occur at the previous time step. Even though C1 column shows that C1 terms have the strongest interaction, C2 and C3 columns present a more heterogeneous interaction pattern. (**b**) Pure nonlinear mixed selectivity contribution for the interaction between angular position for the different time steps and whiskers. The strongest interactions occurs at the current time step. All columns present strong interactions terms with the rest of whiskers.