A neural network model of flexible grasp movement generation

Jonathan A. Michaels[1,2], Stefan Schaffelhofer[1], Andres Agudelo-Toro[1], Hansjörg Scherberger[1,3,*]

[1]Deutsches Primatenzentrum GmbH, Kellnerweg 4, 37077 Goettingen, Germany
[2]Electrical Engineering Department, Stanford University, Stanford, CA 94305, USA
[3]Faculty of Biology and Psychology, University of Goettingen, 37073 Goettingen, Germany
*Corresponding author. Email: hscherberger@dpz.eu

Abstract

One of the main ways we interact with the world is using our hands. In macaques, the circuit formed by the anterior intraparietal area, the hand area of the ventral premotor cortex, and the primary motor cortex is necessary for transforming visual information into grasping movements. We hypothesized that a recurrent neural network mimicking the multi-area structure of the anatomical circuit and trained to transform visual features into the muscle fiber velocity required to grasp objects would recapitulate neural data in the macaque grasping circuit. While a number of network architectures produced the required kinematics, modular networks with visual input and activity that was encouraged to be biologically realistic best matched neural data and the inter-area differences present in the biological circuit. Network dynamics could be explained by simple rules that also allowed the correct prediction of kinematics and neural responses to novel objects, providing a potential mechanism for flexibly generating grasping movements.

Introduction

Interacting with objects is an essential part of daily life for primates. Grasping is one of our most complex behaviors, requiring the determination of object features and identity, followed by the execution of the correct temporal sequence of precise muscle patterns in the arm and hand necessary to reach and grasp the object. In macaque monkeys, the circuit formed by the anterior intraparietal area (AIP), the hand area (F5) of the ventral premotor cortex, and the hand area of the primary motor cortex (M1) is essential for grasping. These areas share extensive anatomical connections[1], forming a long-range circuit (Figure 1a) where AIP receives the largest amount of visual information, and M1 has the largest output to the brainstem and spinal cord. All three areas have been shown to contain grasp-relevant information well before movement[2–7].

Reversible inactivation of AIP[8,9] or F5[10] results in a selective deficit in pre-shaping the hand during grasping, while M1 lesions lead to profound hand movement deficits[11–13], providing evidence that these areas are required for successful grasping. Additionally, M1 has the largest density of projections directly onto motor neurons for control of the fingers, and precise finger control does not recover after lesion[13]. So far, models of the grasping system have relied on manually tuning the properties of individual neurons to match the assumed role of a given region[14]. No comprehensive model exists of the entire transformation between vision and action, limiting our ability to understand the flexibility of the grasping system.

Goal-driven modeling has emerged as a powerful tool for generating potential neural mechanisms explaining various behaviors[15]. The creation of vast datasets of labeled images (Imagenet[16]) opened the door to studying the computational principles underlying object identification using convolutional neural networks (CNNs), such as Alexnet[17]. Feedforward modeling of the ventral stream using CNNs has led to powerful insights into hierarchy of brain networks[18,19], revealing that subsequent layers of CNNs for object identification align well with brain regions along the ventral stream. Similar approaches have been used in retinal modeling[20], and recent studies incorporating recurrence into CNNs[21,22]. In parallel, advances have been made in understanding motor cortex by modeling it as a dynamical system[23,24] implemented as a recurrent neural network[25–28] (RNN). In these models, preparatory activity sets initial conditions that unfold predictably to control muscles during reaching.

Cortex contains many cytoarchitecturally identifiable areas, but it is unclear what computational benefit is bestowed by this anatomically modular arrangement. While the idea of modular processing is relatively straightforward in a feedforward network, such as a CNN, since information must pass through each subsequent layer, it is unclear what role modules play both in multi-area modular RNNs (mRNNs) and in the distributed cortical grasping circuit during motor control.

In the current work, we bridge the gap between previous work in visual processing and motor control by modeling the entire processing pipeline from visual input to muscle control of the arm and hand. Firstly, we recorded neural activity from AIP, F5, and M1 of two macaque monkeys while they grasped a diverse set of 48 objects. Activity in AIP was best explained by visual features extracted from penultimate layers of Alexnet, a CNN trained to identify objects, M1 activity was best explained by muscle kinematics (i.e. muscle fiber velocity), and F5 was intermediate between the two. Based on these results, we devised a number of neural network architectures to model the function of this circuit. Primarily, we trained an mRNN with sparsely connected modules mimicking cortical areas to use visual features from Alexnet to produce the muscle kinematics required for grasping. Additionally, we trained networks with more homogeneous connectivity, as well as networks with a simple labeled-line code, where each object has a separate input, as opposed to visual features.

While all architectures were able to produce the required muscle kinematics, they differed in their ability to explain the neural population dynamics observed in the brain. Networks that received visual features extracted from the CNN matched neural data better than a labeled-line code. Furthermore, models trained with regularizations designed to minimize firing rate and connectivity strength activity matched neural activity recorded in the grasping circuit best. The differences between individual modules in the mRNN paralleled the differences between cortical regions, suggesting that the design of the mRNN model with visual input paralleled the hierarchy observed in the brain. Fixed-point analysis revealed that networks used a simple dynamical strategy to complete the task, which allowed networks trained on a subset of objects to generalize to never-before-seen objects, predicting both the required kinematics and neural population activity. However, this strategy was shared by all networks, regardless of architecture, suggesting that modular computation is not necessary for this task. Together, our results show that modeling the grasping circuit as an mRNN trained to produce muscle kinematics from visual features in a biologically plausible way well matches neural population dynamics and the difference between brain regions, and identifies a simple computational strategy by which these regions may complete this task in tandem.

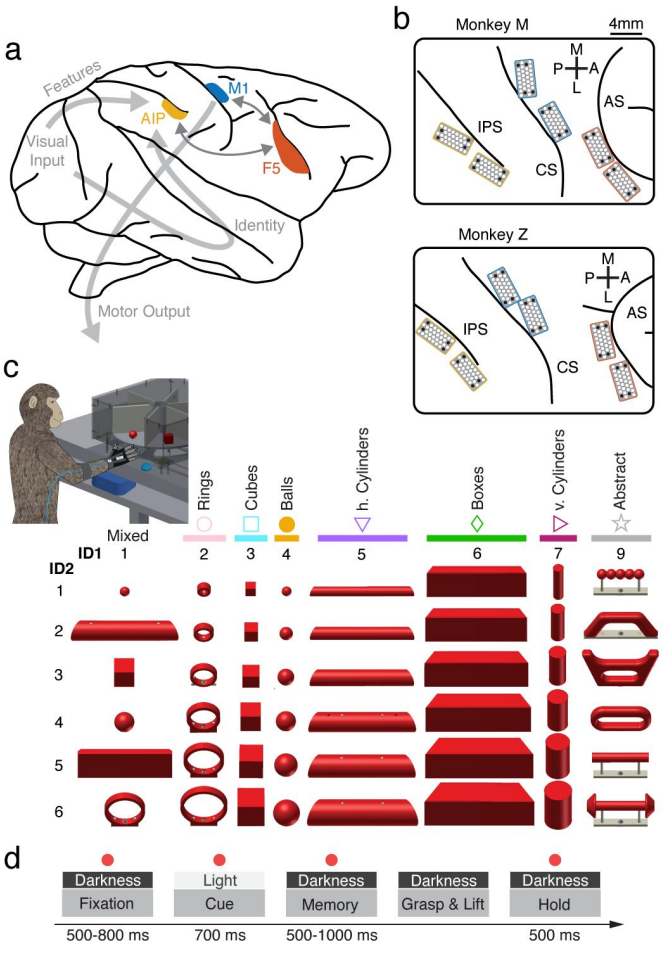

Fig. 1 | Fronto-parietal grasping circuit and experimental design. (a) Simplified brain schematic of the fronto-parietal grasping circuit. Visual information is processed in two parallel streams carrying primarily object features or identity information, both converging on the anterior intraparietal sulcus (AIP). AIP has strong reciprocal connections with the hand area (F5) of the ventral premotor cortex, which has strong reciprocal connections to the hand area of the primary motor cortex (M1). M1 has the majority of subcortical and spinal cord output projections. (b) Location of implanted floating micro-electrode arrays, covering the three desired regions. Arcuate sulcus (AS), central sulcus (CS), intraparietal sulcus (IPS), anterior (A), posterior (P), medial (M), lateral (L). Black dots represent ground and reference electrodes. (c) Monkeys sat in front of a motorized turntable that presented one of six objects to be grasped on any given trial (reproduced from Schaffelhofer et al.[29]). Multiple turntables allowed for a total of 48 objects. Gloves with magnetic sensors allowed full tracking of arm and hand kinematics on single trials. (d) Trials began with visual fixation of a red dot for a variable period. Objects were illuminated temporarily, and monkeys were required to withhold movement until a go cue (blinking of fixation dot) instructed them to grasp and lift the object in darkness. Eye fixation was enforced throughout each trial.

Results

Kinematic and neural activity recorded during a many-object grasping task

We recorded neural activity in the inter-connected anterior intraparietal area (AIP), the hand area (F5) of ventral premotor cortex, and the hand area of motor cortex (M1) using floating micro-electrode arrays (Fig. 1a,b) while two rhesus macaques (monkeys M and Z) performed a delayed grasping task. We presented monkeys with 48 objects composed of shapes of various sizes and orientations on a series of rotating turntables (Fig. 1c). Experimental and behavioral findings have been presented in previous works[2,29]. Monkeys wore a glove that allowed full joint tracking[30] of the arm, hand, and fingers on single trials, and this data was further transformed into muscle space using a previously described musculoskeletal model[31]. On individual trials monkeys had to fixate a red point just under each object, after which the object was illuminated temporarily. Monkeys then waited for a go cue in darkness, after which they reached to, grasped, and lifted the object (Fig. 1d).

We analyzed the data from 10 recording sessions per monkey. On average, each recording session of monkey M consisted of 549±35 (Mean±S.D.) trials, and 153±8, 179±7, and 215±14 single- and multi-units were recorded from AIP, F5, and M1, respectively. On average, each recording session of monkey Z consisted of 490±25 (Mean±S.D.) trials, and 122±10, 137±6, and 126±9 single- and multi-units were recorded from AIP, F5, and M1, respectively.

Previous work using this dataset[2,29] has shown that this circuit is very active during preparation and execution of grasping, containing rich information about the objects, both during presentation and the intervening delay period, and representing temporal information about the kinematic signals required for grasping. Next, we wanted to determine how visual information about grasp targets is used and transformed into the information necessary to execute grasping.

Graded shift from visual to kinematic features in the grasping circuit

The first step in designing a model of the grasping circuit was determining realistic inputs and the structure of the earliest layers. Based on the established role of AIP in grasping and its connectivity to areas containing information about the size, shape, orientation, and identity of objects, we hypothesized that later layers of existing convolutional neural network (CNN) models of the ventral stream may provide potential inputs for AIP. We constructed simulated images of the task from the monkey's perspective (Fig. S1) and fed them into Alexnet[17] (Fig. 2a), a feedforward CNN that was pre-trained to identify objects in ImageNet[16], which contains millions of images. We read out the hidden activity from each of the network layers and compared their ability to explain neural activity in each brain area during the cue period, when the object was visible, using a single-trial, cross-validated regression method similar to previous work in the visual system[32] (Methods).

As predicted, single-trial activity in AIP was better explained by the Alexnet features than in F5 or M1 (Fig. 2b, monkey M - ANOVA, $F = 503.2$, $p < 0.001$, Tukey's HSD $p < 0.001$; Fig. S2a, monkey Z - ANOVA, $F = 96.4$, $p < 0.001$, Tukey's HSD $p < 0.001$). Additionally, the later layers of Alexnet (relu7 layer) predicted activity in AIP better than the early layers (pixel layer, monkey M - ANOVA, $F = 4.1$, $p < 0.001$, Tukey's HSD $p = 0.036$; monkey Z - ANOVA, $F = 6.5$, $p < 0.001$, Tukey's HSD $p = 0.001$), suggesting that Alexnet produced features that were more predictive of neural activity than pure pixel information, and provides realistic inputs for AIP.

To control for differences in firing rate and recording quality between areas, we calculated the internal consistency of each area (i.e. how well single-trial responses correlate across repetitions) and normalized to that value (Methods). A value around 1 would indicate that a set of predictors captures the condition-dependent neural features as well as can be expected given the reliability of the

recorded data. In both monkeys, the above results remained unchanged (Fig. 2b inset, monkey M - ANOVA, $F = 167.4$, $p < 0.001$, Tukey's HSD $p < 0.001$; Fig. S2b, monkey Z - ANOVA, $F = 7.0$, $p = 0.001$, Tukey's HSD $p < 0.04$; relu7 vs. pixel layer, monkey M - ANOVA, $F = 72.8$, $p < 0.001$, Tukey's HSD $p < 0.001$; monkey Z - ANOVA, $F = 106.3$, $p < 0.001$, Tukey's HSD $p < 0.001$). Furthermore, the advanced layers of Alexnet achieved a normalized fit above 0.9 in AIP, suggesting that those visual features well explain neural activity in AIP during the cue period. Visualizing the Euclidean distance between pairs of conditions for both neural data and Alexnet provides additional visual intuition of the similarity between conditions (Fig. 2c) and the difference between Alexnet layers (Fig. S3).

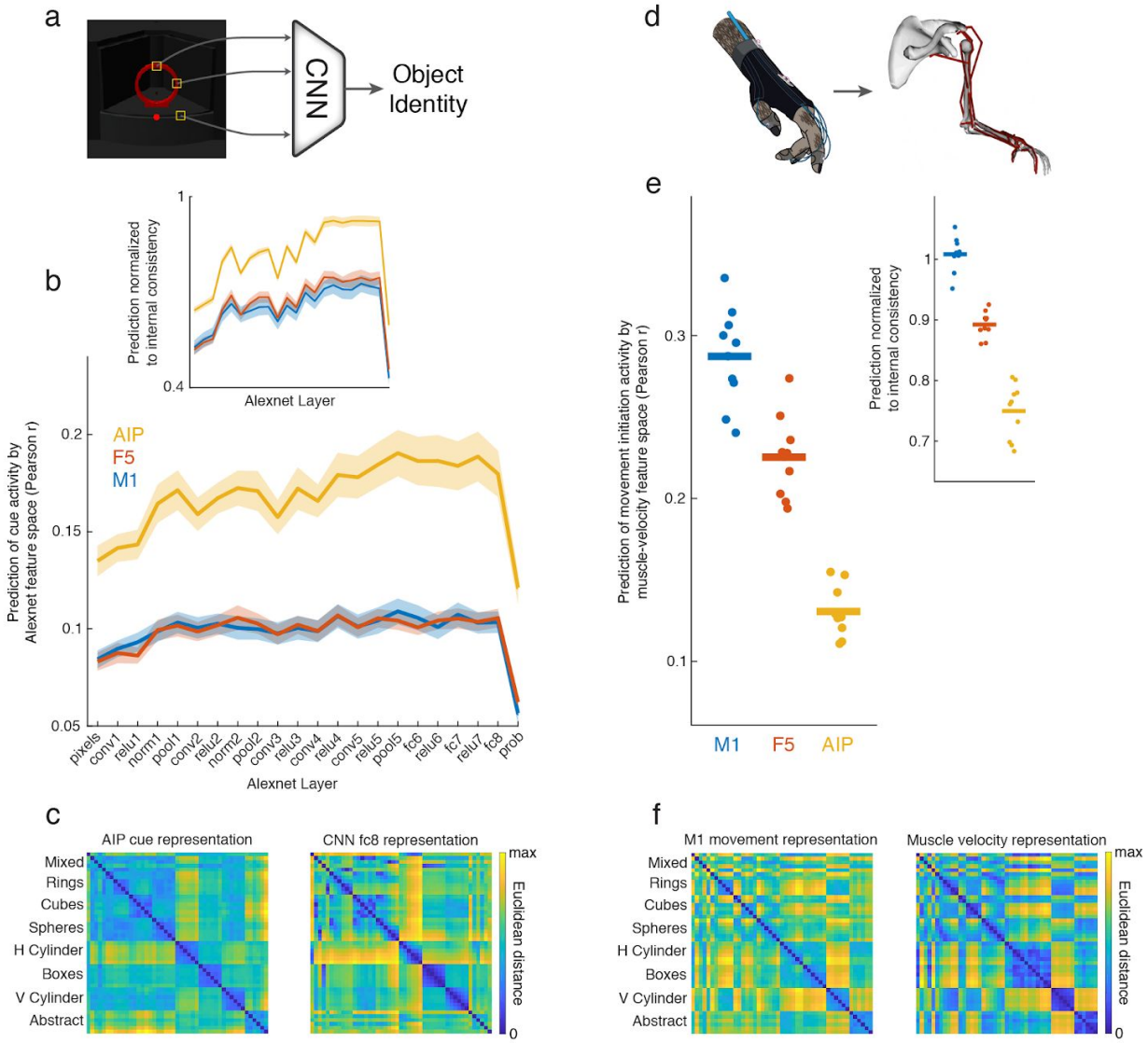

Fig. 2 | Graded shift from visual to kinematic features in the fronto-parietal grasping circuit of monkey M. (a) Simulated images of all objects were fed through a convolutional neural network (CNN) pre-trained to extract object identity (Alexnet). (b) The representation of all objects in each layer of the CNN (first 20 principal components) was regressed against the single-trial neural activity of each unit during the cue period, when the object was visible, and the median fit was taken over all units within one recording session. Solid line and error surfaces represent the mean and s.e.m.over all recording sessions of monkey M. (b - inset) To ensure that results were not due to varying signal quality or firing rate between areas, insets shows regression results normalized to the median internal consistency of each area

(i.e. half of trials correlated with the other half condition-wise). (c) Example Euclidean distance between neural representations of each object in AIP during the cue period and in the fc8 layer of the CNN (session M9). (d) Joint angles (27 DOF) were transformed into muscle length space (50 DOF) using a musculoskeletal model. For visualization purposes, not all muscles are shown. (e) The mean muscle velocity of all grasping conditions during movement initiation (200 ms before to 200 ms after movement onset) was regressed against the single-trial neural activity of each unit during the same time period. Each point represents one recording session of monkey M. (e - inset) Same normalization procedure as in (b - inset). (f) As in (c), but comparing the movement initiation representation in M1 to the muscle velocity representation in the same time window (session M3).

Having established that later layers of a CNN trained to identify objects provide natural inputs to AIP, the next step was to determine reasonable outputs of the grasping circuit. As mentioned previously, monkeys wore a tracking glove[30] that allowed the extraction of 27 degrees of freedom of movement information, almost completely capturing reach to grasp movement trajectories. The joint angle signal was further transformed into a 50-dimensional muscle space using a musculoskeletal model of the primate arm and hand[33] (Fig. 2d), allowing detailed access to muscle kinematics in the hand that would be very difficult to obtain using single muscle recording techniques. While this model does not give us direct access to muscle force or activity, it provides a kinematic signal that bears many similarities to muscle activity. We opted to analyze the muscle velocity signal, since it is invariant to starting hand posture. Similar to the analysis of visual features, we used a 50-dimensional muscle velocity signal to predict single unit activity around movement onset (200 ms before to 200 ms after movement onset). Activity in M1 was better predicted by muscle features than F5 or AIP (Fig. 2e, monkey M - ANOVA, $F = 110.0$, $p < 0.001$, Tukey's HSD $p < 0.001$; Fig. S2d, monkey Z - ANOVA, $F = 89.2$, $p < 0.001$, Tukey's HSD $p < 0.001$), showing the opposite correspondance with cortical regions as the visual feature analysis. Furthermore, the results did not change when controlling for internal consistency (Fig. 2e inset, monkey M - ANOVA, $F = 154.0$, $p < 0.001$, Tukey's HSD $p < 0.001$; Fig. S2e, monkey Z - ANOVA, $F = 68.7$, $p < 0.001$, Tukey's HSD $p < 0.001$), with normalized fits around 1, suggesting that M1 data is predicted as well as possible by the muscle velocity signal, also supported by the similarity in feature space between M1 and muscle velocity (Fig. 2f).

Together, these results strongly suggest a visuomotor gradient from AIP to F5 to M1 that transforms visual features of objects into muscle kinematic signals. However, these analyses only provide snapshots in time and cannot explain the temporal evolution of neural population activity nor the computational mechanisms required to complete the task.

A modular recurrent neural network model of vision to hand action

To build a comprehensive model of the grasping circuit incorporating temporal dynamics, we devised a modular recurrent neural network (mRNN) inspired by the above results and the known anatomical connectivity of the grasping circuit (Methods). The model consisted of three interconnected stages designed to reproduce the muscle dynamics necessary to grasp objects (Fig. 3a). The visual input consisted of an 8-dimensional visual feature signal consisting of the first 8 principal components (92% variance explained) of the features in one of the layers of Alexnet (fc8) that best fit AIP activity while viewing the simulated images. This visual signal entered the input module, a fully-connected RNN that relayed information to the intermediate module through a flat layer of 8 neurons (no recurrence), providing sparsity in the connections between modules. Similarly, the intermediate module projected to the output module through a flat layer, and feedback connections existed for each of the feedforward connections. In order to match kinematic timing, all three modules received a hold signal that cued movements 200 ms before desired movement onset, which was approximately when the monkey's hand lifted off of a handrest button. The output module was most directly responsible for generating the 50-dimensional muscle velocity signal required to grasp each object up to 400 ms into

movement and to suppress movement earlier in the trial. Figure 3b shows inputs for an example trial, including the 8-dimensional visual cue signal and the one dimensional hold signal. During the fixation, memory, and movement periods only the fixation point was presented, while during the go cue the fixation point disappeared for 100 ms.

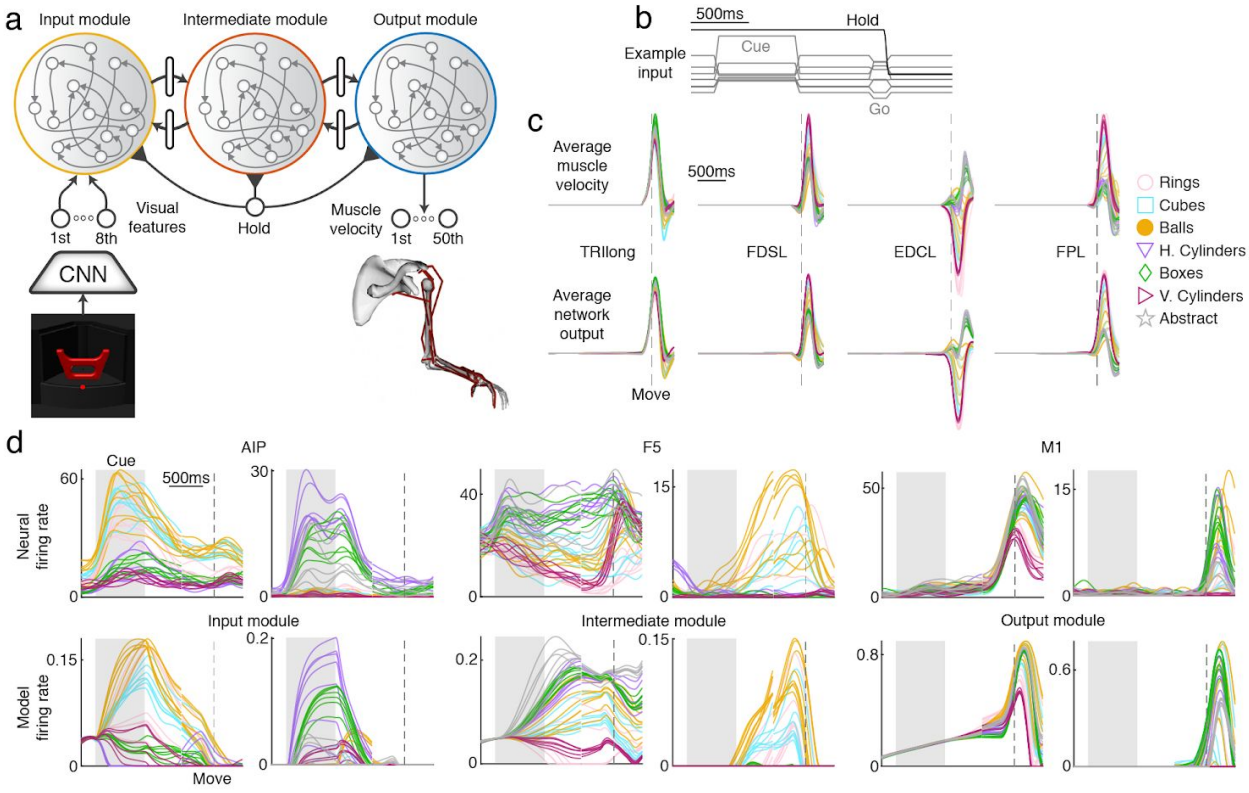

Fig. 3 | Modular recurrent neural network model of the fronto-parietal grasping circuit. (a) Schematic of neural network model. Visual features of each object (first 8 principal components of fc8 layer) are fed into an input module, which is reciprocally connected to an intermediate module through a flat layer of 8 units, which is similarly connected to an output module. The output module must recapitulate the muscle velocity for every object grasped by the monkey. Every module received a hold signal that is released 200 ms prior to movement onset. (b) Example input for an exemplary trial. (c) Average muscle velocity for four example muscles (TRIlong - triceps long head, FDSL - flexor digitorum superficialis digit 5, EDCL - extensor digitorum communis digit 5, FPL - flexor pollicis longus) showing recorded kinematics and network output (session M2). (d) Two example units from each pair of modules and brain regions showing similar properties and highlighting common features of each area. Traces were aligned to two events, cue onset and movement onset, and concatenated together. The shaded gray area represents the cue period, while the dashed line represents movement onset. For (c) and (d), the multiple traces for each type of object represent the different sizes within a turntable.

We used an optimization procedure (Methods) to train one network per monkey to recapitulate all the behavior of that monkey. Each network was trained to reproduce the average muscle velocity of a random set of 4% of all successful trials completed by each monkey, holding out the other 96% of trials. It is crucial to emphasize that no neural data was used in any training procedure, allowing us to compare the neural dynamics of the recorded data to the internal dynamics of our model. While we hypothesized that training networks to reproduce the behavior of the task would lead the internal neural dynamics of the model to match recorded data, it's unclear how

additional constraints on the training procedure may affect this result. Therefore, in addition to the constraint of muscle kinematics, we wanted to test to what extent regularizations that encourage biological neural activity might increase the similarity to neural data. Networks were trained in two stages. The objective of initial training was to reproduce the desired muscle kinematics, while during the second stage we included two additional constraints: 1) a penalty on the mean firing rate of each unit, encouraging units to keep as low firing rate as possible, and 2) a penalty on the squared magnitude of the input and output weights (Methods). To minimize the effect of network initialization on our results, 5 randomly initialized networks were trained separately, and all analysis results for each dataset were averaged across randomizations.

Trained networks were successfully able to reproduce the desired muscle kinematics (Fig. 3c), achieving on average 2.9% normalized error after the first training stage and 3.0% after biological regularization (3.6% normalized error for both training stages in monkey Z). In addition to successful recapitulation of muscle kinematics, networks were also able to suppress output before the movement period and maintain an internal representation of the task conditions in the absence of a visual cue (Fig. 3c).

mRNN model reproduces single unit and population level neural dynamics

To gain an initial intuition of how the hidden state of the regularized mRNN compares to neural data, we plotted the average firing rates of 6 example units that showcase the similarities between the modules and the brain regions of interest (Fig. 3d). Units in AIP and the input module were often characterized by large responses to the visual cue that were either partially maintained through the memory period into movement, or decayed rapidly after the disappearance of the stimulus. Units in F5 and the intermediate module often showed sustained responses throughout the trial that were sensitive to time within the trial. M1 and output module units showed the largest response during movement itself, but often had stable or ramping activity earlier in the trial.

While these example units are useful insights into both the simulation and the neural data, a proper characterization requires a full analysis of the neural population dynamics. To compare the population dynamics of neural and simulated data we used canonical correlation analysis (CCA), a commonly used dimensionality reduction technique[25] for finding linear combinations of the units in each population to produce a low-dimensional set of correlated dimensions (Methods). Before performing CCA, we separately reduced the dimensionality of trial-averaged neural and simulated data across all brain regions and modules to 12 principal components (PCs) in order to restrict the analysis to dimensions of high variance. On average, 12 PCs captured 90% of the variance across all brain regions of monkey M and 94% across all brain regions of monkey Z. We found a very high similarity between the population dynamics of the regularized mRNN model and the neural data for monkey M (Fig. 4) and monkey Z (Fig. S4), having an average canonical correlation (CC) of 0.68 over all 12 orthogonal dimensions (0.66 mean CC for monkey Z). In agreement with previous work[34], the dimension of highest correlation (CV 1) captured the signal typically most dominant in recordings of motor or premotor cortex, a condition-independent signal showing strong modulation shortly before movement initiation. Subsequent dimensions captured various aspects of the neural activity, including complex movement dynamics specific to each movement, as well as sustained memory period activity, cue selectivity, and temporal dynamics throughout the memory period. Importantly, correlations remained high across many orthogonal dimensions, with the average CC remaining above 0.8 for the first 7 dimensions.

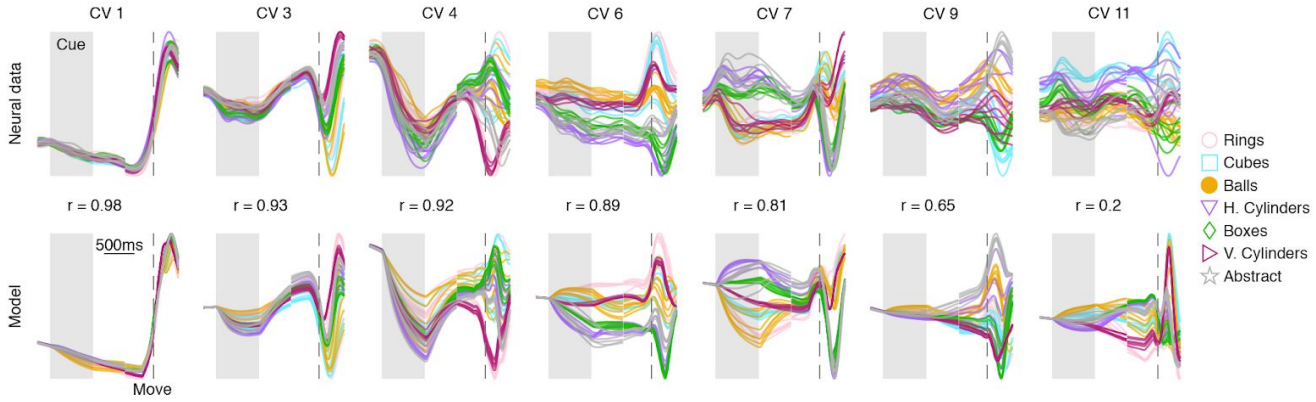

Fig. 4 | Regularized mRNN with visual input matches recorded neural data in monkey M. (a) Example canonical variables (CVs) from canonical correlation analysis (first 12 principal components) between neural and simulated data across all brain regions and modules (session M9), showing r-value for each dimension. The y-axis represents the arbitrary units of the canonical correlation, scaled separately for each CV. There are multiple traces for each type of object, representing the different sizes or types within a turntable.

Biological regularizations improve match to neural data across multiple network architectures

The regularized mRNN model with visual input achieved very high similarity to neural data, providing a model of how the brain may generate grasping movements. However, it is important to characterize how that fit may be affected by the presence or absence of biological regularizations, as well as examine whether or not other network architectures could achieve similar results (monkey M - Fig. 5a; monkey Z - Fig. S5a). Firstly, we designed 3 other modular architectures, one that did not include feedback connections between modules (Feedforward), one that had no bottleneck between modules (No-bottleneck) and one that did not use visual input generated by a CNN (Labeled-line). The purpose of the Labeled-line architecture was to test if the CNN input was necessary to achieve the best fit to neural data, or if the network could learn an equivalent representation simply by being trained on muscle kinematics and regularization.The input for this network consisted of a labeled-line code (also known as “one-hot”), where each condition was cued by a separate dimension that was 1 for that condition and 0 for all others. Secondly, we designed two additional architectures that did not have a modular design. The first was a fully-connected network (Homogeneous) that contained the same number of units as the modular networks. The second was a network that contained the same sparsity in connectivity as the mRNN (Sparse), but no specific structure, to test if sparsity itself was enough to induce high fits to neural data. All of these alternative network structures were trained successfully, achieving normalized errors in the range of 1.1-3.6%.

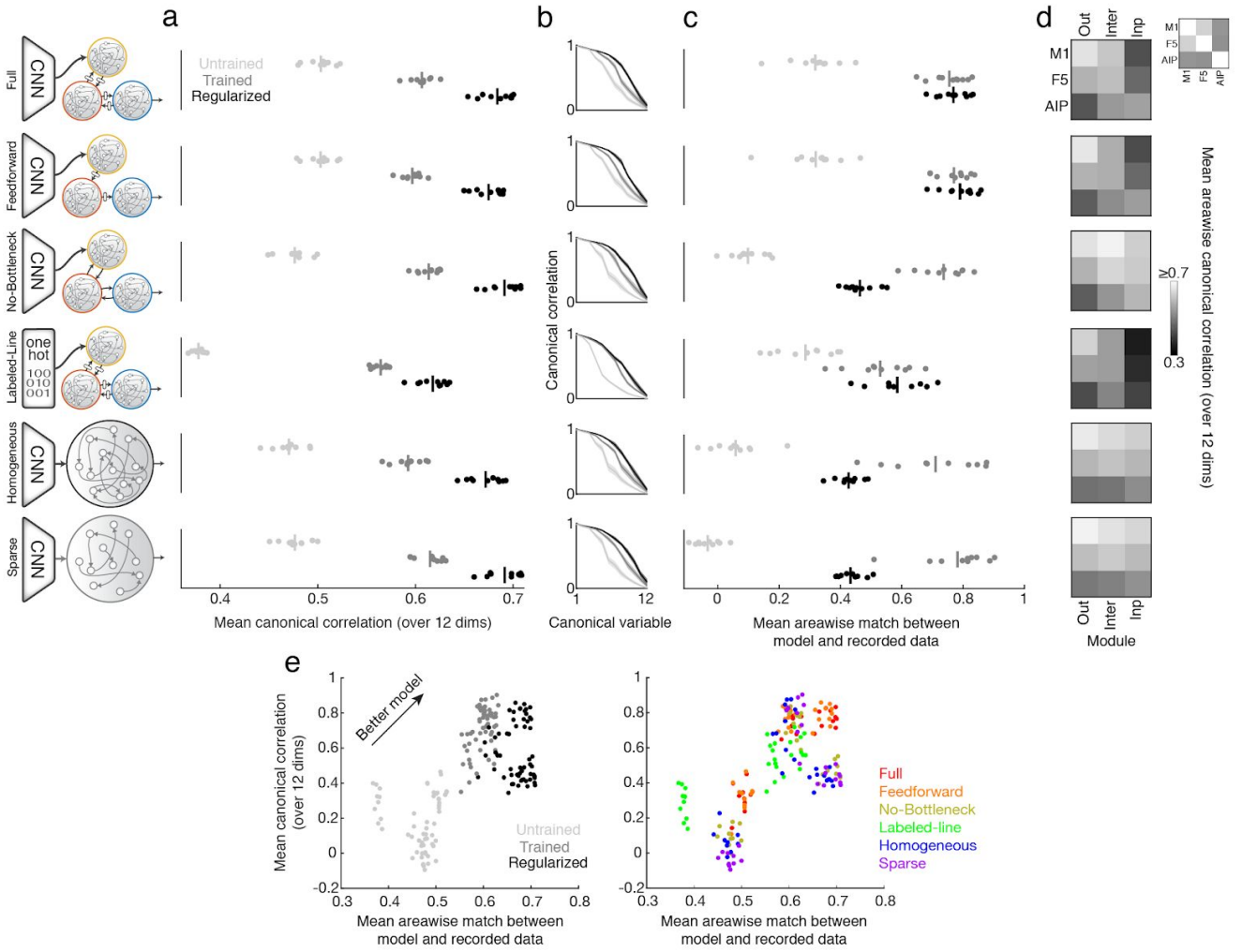

Fig. 5 | Temporal features of regularized neural network model match recorded neural data and align with corresponding brain regions in monkey M. (a) Results of canonical correlation analysis across all sessions of monkey M using 6 model architectures for untrained networks, networks trained to produce kinematics, and networks with additional regularizations. Vertical bars represent the mean, and each dot represents a single session. (a - first) Full model with CNN input and three modules. (a - second) same as first model, but with only feedforward connections. (a - third) Same as Full model, but with no flat layer bottleneck between modules. (a - fourth) Three module design receiving a labeled-line input (one-hot), where each condition is represented by a separate input dimension. (a - fifth) A homogeneous, fully-connected module receiving CNN input. (a - sixth) A single, sparsely-connected module receiving CNN input and sparsity matching the first model. (b) Mean correlations of each canonical variable for the models described in (a). Error bars represent standard deviation across recordings. (c) Canonical correlation was also performed between each module and each brain area and all pairwise canonical correlations were correlated with the inter-area canonical correlation in the neural data, quantifying the areawise match between neural and simulated data. (d) Average canonical correlation between each module and each brain region for the regularized model of each model architecture. Top inset shows canonical correlation between each brain region. (e) Summary of the results of (a) and (c) over all network architectures and recording sessions.

Looking at the initial, untrained networks, all architectures matched neural data to a similar degree (~0.49 mean CC), with the exception of the Labeled-line network (~0.38 mean CC), which lacked structure in the visual input space. After training to produce muscle kinematics a very similar pattern is observed, with all architectures obtaining a mean CC around 0.6, except for the

Labeled-line network (~0.56 mean CC). After regularization the pattern is observed once again, with all architectures achieving a mean CC ~0.68, except for the Labeled-line network (~0.62 mean CC), suggesting that neither training on kinematics nor regularization could recover the visual code provided by CNN input. Very similar results were obtained for monkey Z (Fig. S5). Overall, regularized networks performed best, similar to previous work on reach control[25], and all regularized architectures performed equivalently except for the Labeled-line network (monkey M - 2-way ANOVA interaction, $F = 14.9$, $p < 0.001$, Tukey's HSD $p < 0.05$; monkey Z - 2-way ANOVA interaction, $F = 20.3$, $p < 0.001$, Tukey's HSD $p < 0.05$). The strength of the correlation in the first two canonical variables was unaffected by training stage, but fell off quickly depending on training stage (Fig. 5b). Together, these results suggest that the combination of realistic visual inputs and regularization lead to the best fits to neural data, but that high fits across the circuit can be achieved by a variety of architectures.

Three module mRNN with visual input best matches inter-area differences observed in neural data

While we've shown that multiple architectures can achieve high levels of fit to the neural data, it is unclear which architectures can reproduce the inter-area differences observed between brain regions. Put another way, we'd like to know in which architectures the modules have internal dynamics that correspond to the brain regions they were hypothesized to be modeling. To test this, we extended our CCA method to compare the similarity between modules in our simulations to the observed similarity between recorded brain regions. Specifically, pairwise CCA was performed between each module and each brain region, and this distribution of mean canonical correlations was compared to pairwise CCA between brain regions in the recorded neural data (Methods). A correlation near 1 would indicate that the inter-area differences in the model replicated the inter-area differences in the brain, while a value near 0 would indicate no relationship. The procedure for the Homogenous and Sparse models was slightly different, since these architectures did not inherently contain modules to compare. For these models, we iteratively searched for the best groupings of neurons that produced high correlations with neural data, and tested these sets on held-out data (Methods).

We performed this analysis across all training stages, architectures, and recording sessions, and show the results in Figure 5c. The results of this analysis showed that the highest matches between simulation and experiment for the regularized models were achieved by both the Full and Feedforward mRNNs with visual input (mean $r = 0.77$, monkey M - 2-way ANOVA interaction, $F = 20.9$, $p < 0.001$, Tukey's HSD $p < 0.001$; monkey Z - 2-way ANOVA interaction, $F = 23.6$, $p < 0.001$, Tukey's HSD $p < 0.02$), but were equivalent with and without regularization (Tukey's HSD $p > 0.05$). The Labeled-line networks did not replicate the inter-area differences as well as mRNN models. Interestingly, for the No-bottleneck, Homogenous, and Sparse models the trained network were equivalent to the Full and Feedforward models (Tukey's HSD $p > 0.05$), suggesting that is was possible to find partition of neurons that matched the inter-area differences observed in the brain, but that overall these were still weaker explanatory models than the Full and Feedforward models, since the trained networks did not fit neural data as well overall. The mean areawise CCs for the regularized version of each network is shown in Figure 5d, emphasizing that for the feedback mRNN model with visual input the best fit for each brain region came from the corresponding module (i.e. the highest correlation for each row was along the diagonal). When taking the results of these two analyses together (Fig. 5e), the regularized Full and Feedforward models match the neural data best, suggesting that they are both equally applicable explanatory models.

To test the sensitivity of our results to the magnitude of regularization, we repeated all analyses in Figure 5 across each architecture and randomization using 3 additional regularization magnitudes (20%, 50%, 200% of reported) and found that the same results held across these regularization magnitudes. The only exception was that across all regularization magnitudes the

Sparse architecture had a significantly lower mean CC fit to neural data than the feedback mRNN model in monkey Z (ANOVA, $F = 47.3$, $p < 0.001$, Tukey's HSD $p = 0.002$). Further visualizing the canonical variables (CVs) shared by each module and its corresponding brain region revealed the major features that are captured by the model (Fig. S6). These dimensions were also useful for visualizing aspects of neural data that were not well represented, as these tend to get allocated to the CVs with the lowest correlation.

Models flexibly generate grasping movements using simple dynamical rules

We've established that our mRNN models are able to accurately reproduce muscle kinematics, and that the internal neural dynamics of these models well matched neural activity in the brain, as well as matching the inter-area differences observed in recorded neural data. How can we use these models to gain insight into how the brain may perform this task? Since we know the exact formulas that dictate the operation of our models, we can look for simple linear strategies that explain the computations necessary for the task using fixed point analysis[25,35]. In fixed point analysis, we perform an optimization to look for equilibrium points in the activity of the network, linearize the dynamics around these points, and interpret the properties of these linear dynamics (Methods). Whenever the input to a system changes, the fixed point structure changes. Therefore, we opted to perform this analysis jointly across all modules of the network during 3 time epochs within which inputs were stable.

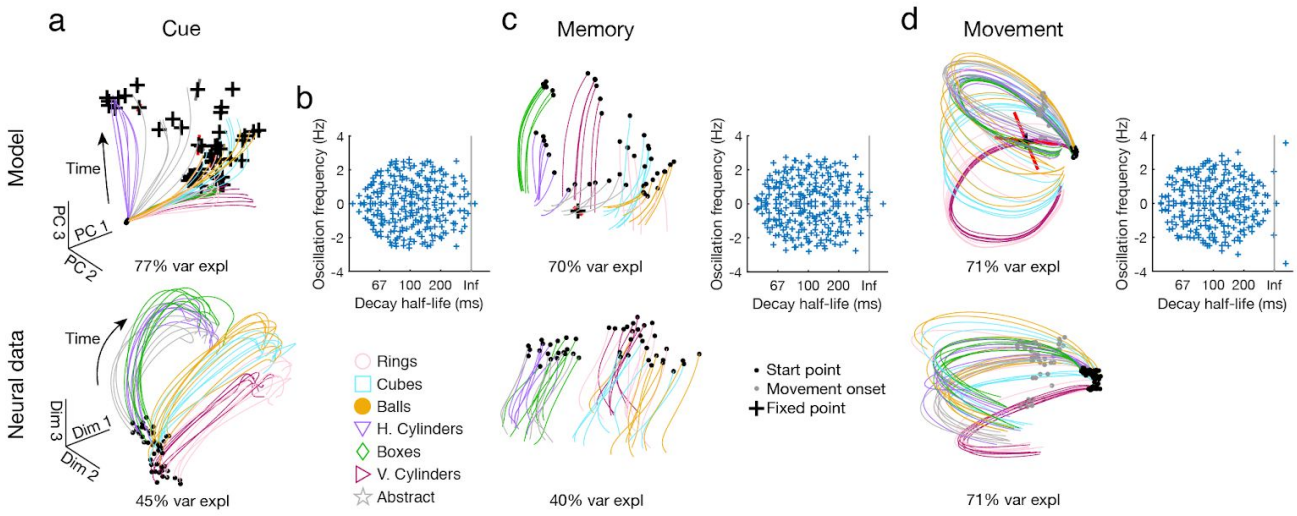

Fig. 6 | Fixed point analysis of regularized full mRNN reveals simple computational strategies. (a - top) Fixed points of an example model (+ symbols) during the cue period (cue onset to cue offset, 700 ms) plotted in the first three PCs alongside trial-averaged activity, showing many fixed points, each corresponding to a different condition. The eigenvectors of the two largest eigenvalues are plotted for each stable (gray) or unstable (red) fixed point, scaled by the magnitude of the eigenvalue. (a - bottom) Neural data from an example session (M5) over the same time period, reduced to 3 PCs and rotated into the PCs of (a - top) using procrustes (Matlab function: *procrustes*). (b) The complex eigenvalue spectrum of the linearized system around a representative fixed point. (c) Same as (a - b) for the memory period (cue offset + 500 ms), showing a single unstable fixed point in the model. (d) Same as (a-b) for the movement period (200 ms before movement onset to 200 ms after movement onset), showing the model with a single unstable fixed point.

During the cue period, when the object was presented, we found a single fixed point that activity moved towards and was determined by the object presented (Fig. 6a - top), and similar activity

was observed in the neural data (Fig. 6a - bottom). These fixed points tended to have one or two small unstable modes, with the majority of modes attracting activity to a point (Fig. 6b) representing each specific object. Interestingly, during the memory period we found a single fixed point (Fig. 6c - top) that maintained and reorganized activity depending on the condition, similar to the neural data (Fig. 6c - bottom). For example, boxes and horizontal cylinders begin the memory period as separate groups of points, since they are visually distinct, but are pulled much closer together by the end of the memory period, a representation of the fact that they are grasped very similarly. Finally, during the movement we found a single fixed point with multiple modes of oscillation (Fig 6d - top), sometimes unstable, that rotated activity following these oscillatory modes and dependant on the starting position as determined by the memory period. Neural activity showed a very similar pattern (Fig. 6d - bottom).

We repeated this analysis for every network architecture, training stage, and across both animals, but found no difference between the strategies employed by the different models, suggesting that this solution (1) is a parsimonious solution regardless of network architecture, and (2) does not require a particular modular architecture to be implemented. Together, this analysis reveals that networks used simple computational strategies to represent, maintain, reorganize, and unfold activity during movement to generate the required muscle kinematics, and provide a framework for understanding how the brain may complete this task.

Generalizing muscle kinematics and neural population activity for novel objects

Our mRNN model was able to produce simulated neural population dynamics that matched recorded neural activity despite the fact that no neural data was used at any point during the training of the networks. However, all of the grasping objects were presented during training, leaving open the question whether or not our mRNN can generalize to novel objects or predict future neural activity. The results from the fixed point analysis in Figure 6 suggest that networks should be able to generalize to new input, since we found single fixed points with smooth dynamics.

To test this, we trained six additional Full mRNN models with visual input, where 8 objects (one per turntable) were held out during the training of each network (Fig 7a, 6-fold cross-validation, where for each fold every object in a row of Fig. 1c was left out). This is a difficult task, since networks only have information about 40 objects during training. Interestingly, networks were able to generalize to novel objects quite well, attaining normalized kinematics errors similar to those for trained conditions (monkey M - 6.4% regularized, 8.1% trained; monkey Z - 5.4% regularized, 5.1% trained), and were able to produce kinematics unique to the untrained objects (Fig 7b). Projecting the simulated neural data of the novel objects into the CVs determined by the trained objects revealed remarkably good predictions of population level features (Fig. 7c). Crucially, when we projected the simulated and recorded data of the novel objects alone into the CVs determined from the trained objects, and analyzed the canonical correlation across all cross-validation folds and sessions (Fig. 7d-e), the mRNN was able to predict the response to novel objects (monkey M - mean CC 0.64; monkey Z - mean CC 0.65, Fig. S7), especially using the regularized model. Finally, the areawise match of the simulation to the recorded data remained high (Fig 7f-g), similar to the results for trained objects (Fig. 5), suggesting that the structure of the regularized mRNN was able to flexibly generalize to new input and output demands, providing a possible mechanism for how the brain may be able to flexibly handle new objects.

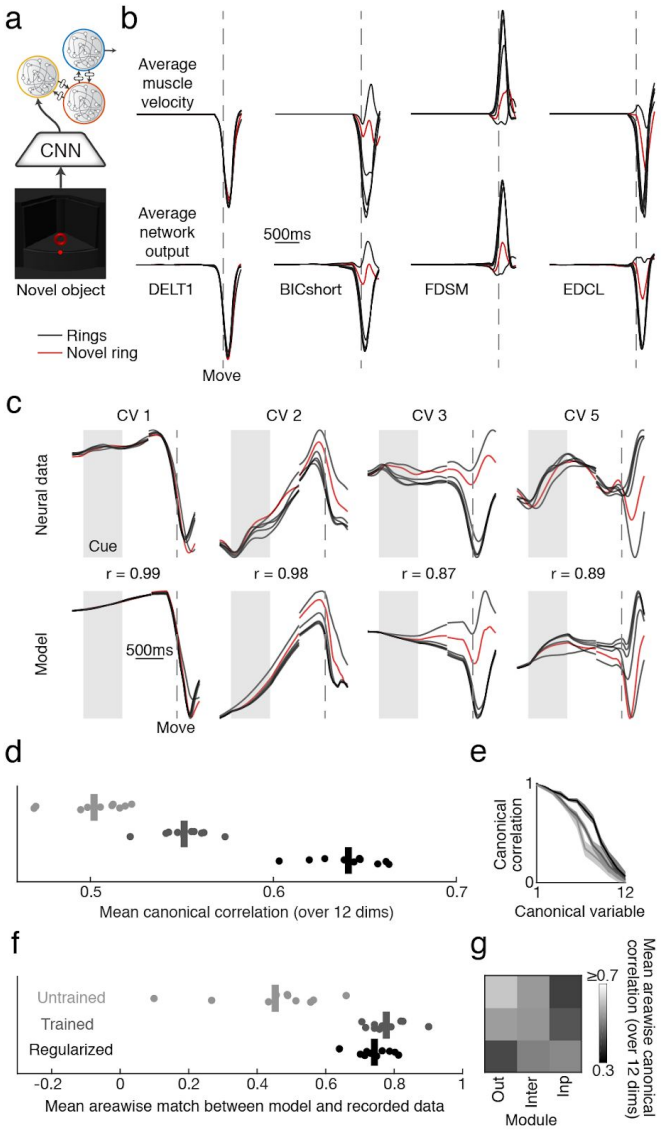

Fig. 7 | **Generalizing muscle kinematics and neural population activity for novel objects in monkey M.** (a) Six additional regularized mRNN networks were trained using a limited set of objects and kinematics (40/48) and subsequently tested on all objects (6-fold cross-validation) to test the ability of the model to generalize to novel objects. Only one turntable (Rings) is shown in order to simplify visual comparison. (b) Example average output kinematics for four muscles (DELT1 - Posterior deltoid, BICshort - Biceps short head, FDSM - flexor digitorum superficialis digit 3, EDCL - extensor digitorum communis digit 5) of an example session (M10), showing a subset of the trained conditions (5 rings) in black, as well as one of the untrained conditions in red. (c) Example canonical variables (CVs) for one recording session (M10) fit to all trained conditions. The novel conditions were projected into the space determined by the trained conditions. Correlation (r-value) between each dimension is shown for novel objects only. (d) Results of canonical correlation analysis across all sessions and cross-validation folds using only the novel objects. Vertical bars represent the mean, and each dot represents a single session. (e) Mean correlations of each canonical variable for the data described in (d). Error bars represent standard deviation across recordings. (f) Canonical correlation was also performed between each module and each brain area and all pairwise canonical correlations were correlated with the inter-area canonical correlation in the neural data, quantifying the areawise match between neural and simulated data. (g) Average canonical correlation between each module and each brain region for the regularized model of each model architecture.

Discussion

Recurrent neural networks are powerful tools for generating complex temporal dynamics. In this work, we demonstrated that modular recurrent neural networks (mRNNs) trained to complete a complex behavioral task can resemble an entire processing pipeline over the course of behavior. These mRNNs took in pixel data and transformed them into the muscle kinematics necessary to grasp various objects. Importantly, no neural data were involved in the model training procedure.

Visual features of objects, as extracted by a convolutional neural network (CNN) trained to identify objects, provided inputs necessary to complete the task, and fit neural data better than a labeled-line code. Our results connect the many works on neural networks for object identity in the ventral stream[15,19,36] to grasp movement generation by showing that the features extracted by such networks are useful for generating grasping movements to learned or novel objects.

Fixed point analysis revealed that models were governed by simple dynamical rules and were able to generalize to novel objects, predicting both the kinematics necessary to grasp them and the corresponding neural population activity. The structure provided by the visual feature space and the

approximately linear dynamics allowed the network to smoothly "interpolate" new objects into the high-dimensional feature space of all known objects and generate the appropriate muscle commands. The single fixed point during the memory period maintains and reorganizes activity, while the single fixed point during movement used those initial conditions to generate the necessary oscillatory patterns for grasping movements, paralleling results in the arm area of motor cortex[24–26,37,38], and suggesting that generation of grasping movements can be understood very similarly to reaching movements under the dynamical systems perspective[23].

Interestingly, all models that included biological regularization, except the Labeled-line networks, equivalently fit neural data overall. Similarly, some unregularized networks were able to match the inter-area differences observed in the recorded data as well as the Full model (in monkey M, but not Z), including the No-Bottleneck model and the Sparse model. However, when taking the results of these two analyses together, the Full and Feedforward models matched neural data best (Fig. 5e). Our interpretation is that (1) biological regularizations are useful for encouraging models to have internal dynamics similar to the brain regardless of model architecture, and (2) modular architectures may not be necessary to complete this task. While modular processing does not appear to be required in this task, we predict that recordings from different cortical layers, or from subcortical regions that have specialized roles, would require modular architectures to properly recapitulate neural data. In cortex, even regions previously thought to be incredibly specialized, such as primary visual cortex, have been shown to contain large amounts of movement related activity[39–41].

The visual inputs to our model were supplied by a CNN that has been generally compared to the ventral stream. This stream has primarily been implicated in object identity processing, while the dorsal stream is largely implicated in spatial localization[42]. Why was this network able to perform so well, despite the fact that AIP lies along the dorsal stream? We propose three reasons. Firstly, AIP is involved in extracting shape information for grasping[43], a process which likely requires the extraction of similar features to those useful for determining object identity. Secondly, AIP is strongly connected to ventral areas in the inferotemporal cortex, including TEa/m[44,45], and areas in the temporal cortex essential for 3D shape perception[46]. These areas are thought to interact during 3D object viewing[47] and are possible routes by which AIP could receive object identity information from the ventral stream. Lastly, in this task objects are only in one place, essentially eliminating the need for a 'where' code that differs between objects, a dominant feature of the dorsal stream.

We found that architectures containing feedback between modules or only feedforward connections produced equivalent matches to neural data, despite the fact that strong feedback connections exist in the anatomical circuit. The likely reason for this is that the task modeled in the current study is very feedforward in nature. Once the monkeys were trained on all objects, the object presented to the monkey on any given trial uniquely determined the grasp plan required to lift the object. These feedback connections would likely come into play in different tasks which have rules that determine how an object should be grasped in a given context. The object identity information that is relayed to AIP from TEa/m is also communicated to ventrolateral prefrontal (VLPF) cortex areas 46v[48] and 12r[49], which relay back to F5 and AIP[14,50,51].These provide an anatomical substrate for context-dependent motor planning in the AIP-F5 circuit, something not explored in the current study. Future experiments should investigate objects in various locations, with rules and context, and try to close the loop by looking at haptic feedback from S2 to AIP[45] and F5[51].

This work builds on many years of work on goal-driven modeling, dynamical systems, and deep neural networks in the visual and motor systems to present a unified view of grasping from pixels to muscles. We believe that the mRNN framework will provide an invaluable setting for hypothesis generation regarding inter-area communication, lesion studies, and computational dynamics in future neuroscience research.

Methods

Animal training and experimental setup

Experimental design has previously been described in detail[2,29]. Briefly, two rhesus monkeys (Macaca mulatta) participated in this study (monkey Z: female, 7.0 kg; monkey M: male, 10.5 kg). Animal housing, care, and all experimental procedures were conducted in accordance with German and European laws governing animal welfare and were in agreement with the guidelines for the care and use of mammals in neuroscience and behavioral research[52].

We developed an experimental setup that allowed us to present a large number of graspable objects to the monkeys while monitoring their behavior, neural activity, and hand kinematics. During each recording session, monkeys grasped a total of 42–48 objects of equal weight that were placed on 8 interchangeable turntables (Figure 1c). Objects were of different shapes and sizes including rings, cubes, spheres, horizontal cylinders, vertical cylinders, and bars. A mixed turntable held objects of different shapes of average size. Additionally, a special turntable held objects of abstract forms, which differed visually, but required almost identical hand configurations for grasping. Monkeys were also trained on a grasping box that cued one of two grasping types, power or precision grip, but for simplicity this data was not included in the current study.

Task paradigm

Monkeys were trained to grasp 48 objects in a delayed grasp, lift, and hold task (Figure 1c,d). While sitting in the dark the monkeys could initiate a trial (self-paced) by placing their grasping hand (left hand in monkey Z, right hand in monkey M) onto a rest sensor that enabled a fixation LED close to the object. Looking at (fixating) this spot for a variable time activated a spot light that illuminated the graspable object. After the light was turned off the monkeys had to withhold movement execution until the fixation LED blinked for 100 ms. After this, the monkeys released the rest sensor, reached for and grasped the object and briefly lifted it up (500 ms). The monkeys had to fixate the LED throughout the task (max. deviation: ~5 deg of visual angle). All correctly executed trials were rewarded with a liquid reward (juice) and monkeys could initiate the next trial after a short delay. Error trials were immediately aborted without reward and excluded from the analysis.

Kinematic recording

Finger, hand, and arm kinematics of the acting hand were tracked with an instrumented glove for small primates. Eight magnetic sensor coils (model WAVE, Northern Digital) were placed onto the fingernails, the hand's dorsum as well as the wrist to compute the centers of 18 individual joints in 3D space, including thumb, digits, wrist, elbow and shoulder. The method and its underlying computational model have been described previously[30]. Recorded joint trajectories were then used to drive a 3D-musculoskeletal model[33,53], which was adjusted to the specific anatomy of each monkey. The model was implemented in OpenSim[31] and allowed extracting a total of 27 DOF in joint angle space, and 50 DOF in muscle tendon length space. All extracted joint angles and muscle lengths were sampled at 100 Hz and low-pass filtered (2nd-order Butterworth filter, 3 Hz low-pass).

Electrophysiological recordings

Single and multiunit activity was recorded simultaneously using floating microelectrode arrays (FMA, Microprobe Inc., Gaithersburg, MD, USA). In each monkey we recorded 192 channels from 6 individual arrays implanted into the cortical areas AIP, F5, and M1 (Figure 1b). In each array, the lengths of the electrodes increased towards the sulcus and ranged from 1.5 (1st row) to 7.1 mm (4th

row). In area F5, one array was placed in the posterior bank of the inferior arcuate sulcus approximately targeting F5a (longer electrodes) and approaching the F5 convexity (F5c; shorter electrodes). The second and more dorsally located array was positioned to target F5p. In AIP, the arrays were implanted into the end of the posterior intraparietal sulcus at the level of area PF and more dorsally at the level of area PFG. In M1, both arrays were placed into the hand area of M1 into the anterior bank of the central sulcus at the level of the spur of the arcuate sulcus[54]. Surgical procedures have been described previously[2]. Neural activity was recorded at full bandwidth with a sampling frequency of 24 kHz and a resolution of 16 bits (model: RZ2 BioAmp Processor; Tucker Davis Technologies, FL, USA). Neural data was synchronously stored to disk together with the behavioural and kinematic data. Raw recordings were filtered offline (bandpass cutoff: 0.3–7 kHz) before spikes were detected (threshold: 3.5x std) and extracted. Spike sorting was processed in two steps: First, we applied super-paramagnetic clustering[55] and then revised the results by visual inspection using Offline Sorter (Plexon, TX, USA) to detect and remove neuronal drift and artefacts. No other pre-selection was applied and single and multiunit activity were analyzed together.

Visual and muscle feature analysis

In order to model the visual features of the objects being presented in the grasping task, we generated simulated images from the monkey's perspective (Fig. S1). We preprocessed and fed these images into a convolutional neural network (CNN), Alexnet[17], that used spatial convolution over pixels to determine the identity of objects in an image. Alexnet was pre-trained on ImageNet[16], a massive set of labeled images. We did not train Alexnet on our images.

To test how well features within the layers of Alexnet could explain neural activity during presentation of the objects (cue period, 700 ms), we first transformed the responses of each layer of Alexnet to the presentation of all objects into its first 20 principal components, which explained 92-99% of the signal variance. Next, we used a support vector machine to regress the features of each layer onto the single trial spike counts during the cue period of each unit separately (Matlab function *fitrlinear*), using 10-fold cross-validation. All regressions had an additional L2 (ridge) penalty of $\lambda = 1/n$, where n was the number of in-fold observations. We then took the median r-value over all units within a recording session, and report the mean of those values across recording sessions in Figure 2b. This method is very similar to regression methods used in previous works of the visual system[32].

In order to make comparisons between regions, we must control for differences in recording quality. Therefore, we calculated the internal consistency of each area, which provides a measure of reliability across trials within a given condition. To calculate internal consistency, trials within each condition were split in half, forming two sets of trials. These sets were correlated with each other for each unit separately, and the resulting r-value was Spearman-Brown corrected to account for the halving of sampling size. We took the median of this distribution and repeated the above analysis 1000 times with different random partitions of trials and took the average over repetitions. Finally, the results of the regression in Figure 2b were normalized by the internal consistency in Figure 2b-inset .

For the analysis of muscle kinematics in Figure 2e-f we performed the same regression analysis using the average muscle velocity of all 50 muscles during movement initiation (200 ms before - 200 ms after movement onset) to predict neural spike count during the same time period.

Modular recurrent neural network

In order to model the planning and execution of a grasping task, we implemented the dynamical system, $\dot{x} = F(x, u)$, using a standard continuous RNN equation of the form

$$\tau \dot{x}_i(t) = -x_i + \sum_{k=1}^{N} J_{ik} r_k(t) + \sum_{k=1}^{I} B_{ik} u_k(t) + b_i^x \qquad (1)$$

where the network has N units and I inputs, x are the activations and r the firing rates in the network, which were related to the activations by the rectified hyperbolic tangent function, such that $r = \{0,\ x < 0;\ tanh(x),\ x \geq 0\}$. The units in the network interact using the synaptic weight matrix, J. The inputs are described by u and enter the system by input weights, B. Each unit has an offset bias, b_i^x. The time integration constant of the network is τ.

For all simulations N was fixed at 483 (3 x 150 per module, 4 x 8 per inter-module layer), where each module contained 150 units (mN) and each inter-module layer contained 8 units. Note that the inter-module layer used the same rectifying non-linearity as all other units. The inputs were a condition-independent hold signal that was released 200 ms before movement onset and was sent to all modules, and an 8-dimensional signal representing the visual features of the current visual stimulus that was sent only to the input module. The elements of B were initialized to have zero mean (normally distributed values with $SD = \frac{1}{\sqrt{I}}$). The elements of J were initialized to have zero mean (normally distributed values with $SD = \frac{g}{\sqrt{mN}}$) within each module, normally distributed with $SD = \frac{1}{\sqrt{mN}}$ between each module and its corresponding flat layer, and zero for all connections between units in the flat layers. The synaptic scaling factor, g, was set at 1.2 following previous work[56]. We used a fixed time constant of 100 ms for τ, with Euler integration every 10 ms.

The network was required to generate average muscle velocities in 50 dimensions until 400 ms after movement onset, where movement onset was determined by a threshold crossing in elbow position that approximately corresponded to the hand lifting off the handrest. The output of the network was defined as a linear readout of the output module

$$z_i(t) = \sum_{k=1}^{N} W_{ik} r_k(t) + b_i^z \qquad (2)$$

where z represents the 50-dimensional muscle velocity signal and is a linear combination of the internal firing rates using weight matrix W, which was initialized with near zero entries, and b_i^z, which is a bias term for each output dimension.

All non-zero values of the input weights, B, internal connectivity, J, output weights, W, and all biases, were trained using Hessian free optimization[57] (code:https://github.com/sussillo/hfopt-matlab) also utilized in previous work[25,26]. The error function used to optimize the network considered the difference between the output of the linear readout and the desired muscle velocity profiles, v,

$$E_i(t) = z_i(t) - v_i(t) \qquad (3)$$

at each time point, t, and each output dimensions, i, across all trials. We report normalized error, which is the sum of the squared error from Eq. 3 over all times, dimensions, and trials, divided by the total variance of the target signal. In addition to the above error signal, we also implemented two regularizations designed to encourage the network to produce biologically-plausible activity. The two penalties were a cost on the mean firing rate and the sum of the squared input and output weights. The hyper-parameter values for the results in the main text were 3e-2 for the firing rate penalty and 1e-4 for the input/output penalty, and 3 other sets of hyper-parameters values were also tested. However, for the homogeneous and labeled-line architectures, the input/output weight penalty was

normalized by the ratio of trainable parameters in these architectures compared to the full model in order to equalize training pressure between all models tested.

Similar to previous work, we opted not to model any feedback, since the goal of the study was to illustrate the main points parsimoniously and without relying on confronting the issue of what kind of feedback is most biologically plausible in such a network.

All networks were trained until the change in objective from one iteration to the next fell below 2e-6 and the normalized kinematic error was below 5%.

Canonical Correlation Analysis

To compare neural population dynamics between simulations and recorded neural data we performed canonical correlation analysis (CCA), which was carried out on trial-averaged data aligned to both cue onset and movement onset that was concatenated to form a single trajectory. Before CCA, all units in both the neural data and the simulated data were reduced to 12 principal components (PCA) to avoid inflating dimensions of low variance. Data was of the form $ct \times n$, where c is the number of conditions, t is the amount of time per trial, and n is the number of units. CCA produces new dimensions that are linear combinations of the principal components of each data set (neural or simulated) that are highly correlated between data sets and orthogonal to all other canonical variables.

For the areawise analysis in Figure 5c, both PCA (12 dimensions) and CCA was performed independently for each pair of modules and brain regions. To determine the similarity between all pairs of brain regions (Fig. 5d), trials within each condition were split in half and trial-averaged to form two sets of data, which were then reduced using PCA and compared using CCA. We repeated this procedure 10 times with random trial partitions and averaged the resulting canonical correlations.

Assigning modules in Homogeneous and Sparse architectures

Since these architectures did no inherently contain modules, we iteratively searched for partitions of neurons in these models that had a high area-wise correlation with neural data. For every recording session, we split trials into training and testing sets, randomly assigned neurons in the simulated model to one of the three modules, and trial-averaged the data. Next, we iteratively moved individual neurons from one module to another, and accepted this change only if it improved the mean canonical correlation between each module and its corresponding brain region. We repeated this process 1000 times for each recording session, then carried out the analysis from Fig. 5d on the held-out data. In general, this procedure drastically improved how well the Homogeneous and Sparse models could explain the inter-area differences observed in the brain, as correlations for randomly assigned neurons (at the beginning of the iterative process) typically had correlations between -0.2 and 0.2.

Fixed points

To extract simple rules behind the computations of our simulations, we searched for fixed points using standard nonlinear dynamical systems methods combined with linear stability analysis, as has been described in detail previously[25,35]. We searched for a set of points in the high-dimensional state space, $\{x^{1*}, x^{2*}, \ldots\}$, where the dynamics described in Eq. 1 are at an equilibrium, $\dot{x}^* = F(x^*, u^{const}) = 0$, for a given constant input. For some volume around these points, Eq. 1 can be replaced by a linear dynamical system, $\dot{\delta x} = M\delta x$, with $\delta x = x - x^*$ and $M = F'(x^*)$, by definition. These points are considered fixed points if their speed is very slow relative to the speed of the network during normal operation (>1000 times slower for most of our results).

For each time epoch of interest, we repeated the optimization many times and randomly sampled the optimization starting point from activity the networks visited during normal operation, yielding a single fixed point for each condition during the cue period, a single fixed point during the memory period, and a single fixed point during the movement period. In some cases, tight clusters of fixed points (2-3) with similar properties were considered a single fixed point.

Acknowledgements

We would like to thank N Bobb, L Burchardt, M Dörge, R Lbik, K Menz, and M Sartori for assistance, and B Dann for constructive discussions and helpful comments on an earlier version of this manuscript.

Author Contributions

SS and HS carried out experiments; AA, JAM, and SS analyzed data; JAM performed all simulations; JAM wrote the manuscript; All authors edited the manuscript.

Conflict of Interest

Authors report no conflict of interest.

References

1. Luppino, G., Murata, A., Govoni, P. & Matelli, M. Largely segregated parietofrontal connections linking rostral intraparietal cortex (areas AIP and VIP) and the ventral premotor cortex (areas F5 and F4). *Exp. Brain Res.* **128**, 181–187 (1999).
2. Schaffelhofer, S., Agudelo-Toro, A. & Scherberger, H. Decoding a wide range of hand configurations from macaque motor, premotor, and parietal cortices. *J. Neurosci.* **35**, 1068–1081 (2015).
3. Fluet, M.-C., Baumann, M. A. & Scherberger, H. Context-specific grasp movement representation in macaque ventral premotor cortex. *J. Neurosci.* **30**, 15175–15184 (2010).
4. Baumann, M. A., Fluet, M.-C. & Scherberger, H. Context-specific grasp movement representation in the macaque anterior intraparietal area. *J. Neurosci.* **29**, 6436–6448 (2009).
5. Murata, A. *et al.* Object representation in the ventral premotor cortex (area F5) of the monkey. *J. Neurophysiol.* **78**, 2226–2230 (1997).
6. Murata, A., Gallese, V., Luppino, G., Kaseda, M. & Sakata, H. Selectivity for the shape, size, and orientation of objects for grasping in neurons of monkey parietal area AIP. *J. Neurophysiol.* **83**, 2580–2601 (2000).
7. Carpaneto, J. *et al.* Decoding the activity of grasping neurons recorded from the ventral premotor area F5 of the macaque monkey. *Neuroscience* **188**, 80–94 (2011).
8. Gallese, V., Murata, A., Kaseda, M., Niki, N. & Sakata, H. Deficit of hand preshaping after muscimol injection in monkey parietal cortex. *Neuroreport* **5**, 1525–1529 (1994).
9. Tunik, E., Frey, S. H. & Grafton, S. T. Virtual lesions of the anterior intraparietal area disrupt goal-dependent on-line adjustments of grasp. *Nat. Neurosci.* **8**, 505–511 (2005).
10. Fogassi, L. *et al.* Cortical mechanism for the visual guidance of hand grasping movements in the monkey: A reversible inactivation study. *Brain* **124**, 571–586 (2001).

11. Hoffman, D. S. & Strick, P. L. Effects of a primary motor cortex lesion on step-tracking movements of the wrist. *J. Neurophysiol.* **73**, 891–895 (1995).

12. Murata, Y. *et al.* Effects of motor training on the recovery of manual dexterity after primary motor cortex lesion in macaque monkeys. *J. Neurophysiol.* **99**, 773–786 (2008).

13. Passingham, R. E., Perry, V. H. & Wilkinson, F. The long-term effects of removal of sensorimotor cortex in infant and adult rhesus monkeys. *Brain* **106 (Pt 3)**, 675–705 (1983).

14. Fagg, A. H. & Arbib, M. A. Modeling parietal-premotor interactions in primate control of grasping. *Neural Netw.* **11**, 1277–1303 (1998).

15. Yamins, D. L. K. & DiCarlo, J. J. Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* **19**, 356–365 (2016).

16. Deng, J. *et al.* ImageNet: A large-scale hierarchical image database. in *2009 IEEE Conference on Computer Vision and Pattern Recognition* 248–255 (2009).

17. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. in *Advances in Neural Information Processing Systems 25* (eds. Pereira, F., Burges, C. J. C., Bottou, L. & Weinberger, K. Q.) 1097–1105 (Curran Associates, Inc., 2012).

18. Yamins, D. L. K. *et al.* Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 8619–8624 (2014).

19. Cadieu, C. F. *et al.* Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS Comput. Biol.* **10**, e1003963 (2014).

20. Maheswaranathan, N. *et al.* Deep learning models reveal internal structure and diverse computations in the retina under natural scenes. *bioRxiv* 340943 (2018). doi:10.1101/340943

21. Nayebi, A. *et al.* Task-Driven Convolutional Recurrent Models of the Visual System. *arXiv [q-bio.NC]* (2018).

22. 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. *Nat. Neurosci.* (2019). doi:10.1038/s41593-019-0392-5

23. Shenoy, K. V., Sahani, M. & Churchland, M. M. Cortical control of arm movements: a dynamical systems perspective. *Annu. Rev. Neurosci.* **36**, 337–359 (2013).

24. Churchland, M. M. *et al.* Neural population dynamics during reaching. *Nature* **487**, 51–56 (2012).

25. Sussillo, D., Churchland, M. M., Kaufman, M. T. & Shenoy, K. V. A neural network that finds a naturalistic solution for the production of muscle activity. *Nat. Neurosci.* **18**, 1025–1033 (2015).

26. Michaels, J. A., Dann, B. & Scherberger, H. Neural Population Dynamics during Reaching Are Better Explained by a Dynamical System than Representational Tuning. *PLoS Comput. Biol.* **12**, e1005175 (2016).

27. Hennequin, G., Vogels, T. P. & Gerstner, W. Optimal control of transient dynamics in balanced networks supports generation of complex movements. *Neuron* **82**, 1394–1406 (2014).

28. Stroud, J. P., Porter, M. A., Hennequin, G. & Vogels, T. P. Motor primitives in space and time via targeted gain modulation in cortical networks. *Nat. Neurosci.* **21**, 1774–1783 (2018).

29. Schaffelhofer, S. & Scherberger, H. Object vision to hand action in macaque parietal, premotor, and motor cortices. *Elife* **5**, e15278 (2016).

30. Schaffelhofer, S. & Scherberger, H. A new method of accurate hand- and arm-tracking for small primates. *J. Neural Eng.* **9**, 026025 (2012).

31. Delp, S. L. *et al.* OpenSim: open-source software to create and analyze dynamic simulations of movement. *IEEE Trans. Biomed. Eng.* **54**, 1940–1950 (2007).

32. Schrimpf, M. *et al.* Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like? *bioRxiv* 407007 (2018). doi:10.1101/407007

33. Schaffelhofer, S., Sartori, M., Scherberger, H. & Farina, D. Musculoskeletal representation of a large repertoire of hand grasping actions in primates. *IEEE Trans. Neural Syst. Rehabil. Eng.* **23**, 210–220 (2015).

34. Kaufman, M. T. *et al.* The Largest Response Component in the Motor Cortex Reflects Movement Timing but Not Movement Type. *eNeuro* **3**, ENEURO.0085–16.2016 (2016).

35. Sussillo, D. & Barak, O. Opening the black box: low-dimensional dynamics in high-dimensional recurrent neural networks. *Neural Comput.* **25**, 626–649 (2013).

36. Bashivan, P., Kar, K. & DiCarlo, J. J. Neural population control via deep image synthesis. *Science* **364**, eaav9436 (2019).

37. Churchland, M. M., Cunningham, J. P., Kaufman, M. T., Ryu, S. I. & Shenoy, K. V. Cortical preparatory activity: representation of movement or first cog in a dynamical machine? *Neuron* **68**, 387–400 (2010).

38. Russo, A. A. *et al.* Motor Cortex Embeds Muscle-like Commands in an Untangled Population Response. *Neuron* **97**, 953–966.e8 (2018).

39. Saleem, A. B., Diamanti, E. M., Fournier, J., Harris, K. D. & Carandini, M. Coherent encoding of subjective spatial position in visual cortex and hippocampus. *Nature* **562**, 124–127 (2018).

40. Meyer, A. F., Poort, J., O'Keefe, J., Sahani, M. & Linden, J. F. A Head-Mounted Camera System Integrates Detailed Behavioral Monitoring with Multichannel Electrophysiology in Freely Moving Mice. *Neuron* **100**, 46–60.e7 (2018).

41. Stringer, C. *et al.* Spontaneous behaviors drive multidimensional, brainwide activity. *Science* **364**, 255 (2019).

42. Maunsell, J. H. & Newsome, W. T. Visual processing in monkey extrastriate cortex. *Annu. Rev. Neurosci.* **10**, 363–401 (1987).

43. Theys, T., Romero, M. C., van Loon, J. & Janssen, P. Shape representations in the primate dorsal visual stream. *Front. Comput. Neurosci.* **9**, 43 (2015).

44. Webster, M. J., Bachevalier, J. & Ungerleider, L. G. Connections of inferior temporal areas TEO and TE with parietal and frontal cortex in macaque monkeys. *Cereb. Cortex* **4**, 470–483 (1994).

45. Borra, E. *et al.* Cortical connections of the macaque anterior intraparietal (AIP) area. *Cereb. Cortex* **18**, 1094–1111 (2008).

46. Verhoef, B.-E., Vogels, R. & Janssen, P. Inferotemporal cortex subserves three-dimensional structure categorization. *Neuron* **73**, 171–182 (2012).

47. Janssen, P., Verhoef, B.-E. & Premereur, E. Functional interactions between the macaque dorsal and ventral visual pathways during three-dimensional object vision. *Cortex* **98**, 218–227 (2018).

48. Gerbella, M., Borra, E., Tonelli, S., Rozzi, S. & Luppino, G. Connectional heterogeneity of the ventral part of the macaque area 46. *Cereb. Cortex* **23**, 967–987 (2013).

49. Borra, E., Gerbella, M., Rozzi, S. & Luppino, G. Anatomical evidence for the involvement of the macaque ventrolateral prefrontal area 12r in controlling goal-directed actions. *J. Neurosci.* **31**, 12351–12363 (2011).

50. Grafton, S. T. The cognitive neuroscience of prehension: recent developments. *Exp. Brain Res.* **204**, 475–491 (2010).

51. Gerbella, M., Belmalih, A., Borra, E., Rozzi, S. & Luppino, G. Cortical connections of the anterior (F5a) subdivision of the macaque ventral premotor area F5. *Brain Struct. Funct.* **216**, 43–65 (2011).

52. National Research Council, Division on Earth and Life Studies, Institute for Laboratory Animal Research & Committee on Guidelines for the Use of Animals in Neuroscience and Behavioral Research. *Guidelines for the Care and Use of Mammals in Neuroscience and Behavioral Research*. (National Academies Press, 2003).

53. Holzbaur, K. R. S., Murray, W. M. & Delp, S. L. A model of the upper extremity for simulating musculoskeletal surgery and analyzing neuromuscular control. *Ann. Biomed. Eng.* **33**, 829–840 (2005).

54. Rathelot, J.-A. & Strick, P. L. Subdivisions of primary motor cortex based on cortico-motoneuronal cells. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 918–923 (2009).

55. Quiroga, R. Q., Nadasdy, Z. & Ben-Shaul, Y. Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering. *Neural Comput.* **16**, 1661–1687 (2004).

56. Sussillo, D. & Abbott, L. F. Generating coherent patterns of activity from chaotic neural networks. *Neuron* **63**, 544–557 (2009).

57. Martens, J. & Sutskever, I. Learning recurrent neural networks with hessian-free optimization. in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)* 1033–1040 (Citeseer, 2011).

Supplemental Materials

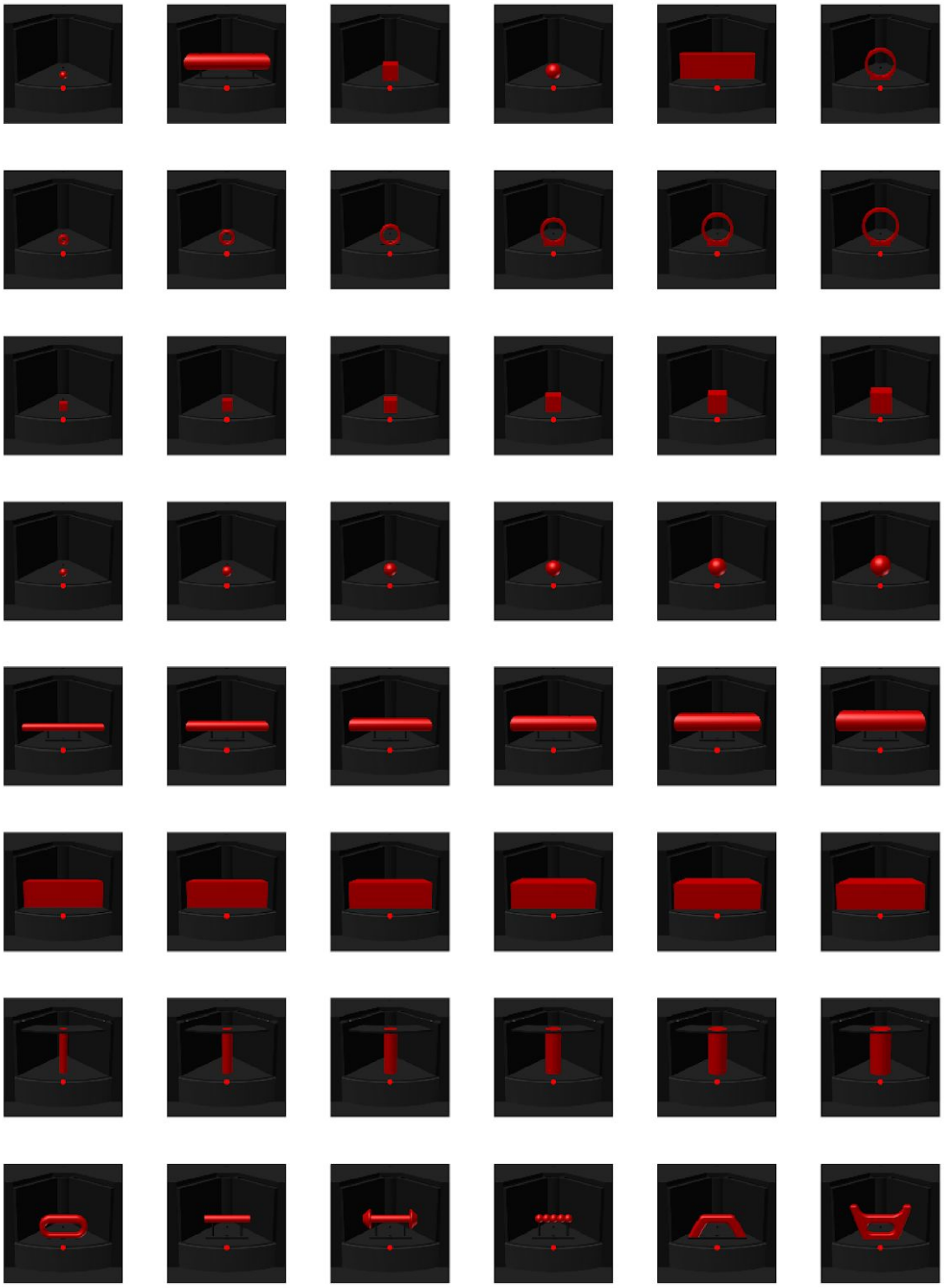

Fig. S1 | Simulated monkey view of objects used as input for Alexnet. Physical objects were CNC manufactured based on mesh models. Red fixation point was added in the approximate location that it was presented to the animals. All input images were RGB and 227x227 pixels in size.

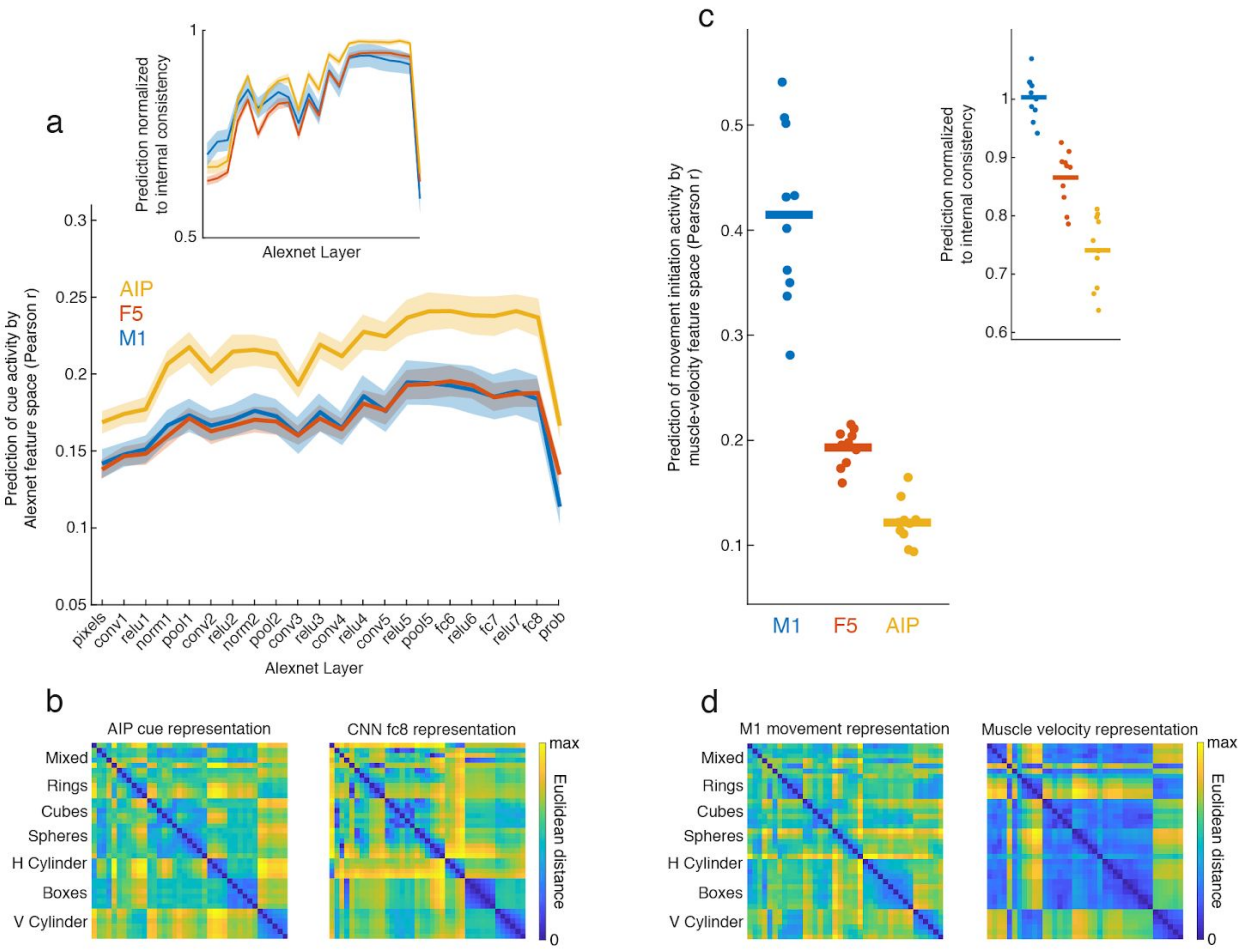

Fig. S2 | Graded shift from visual to kinematic features in the fronto-parietal grasping circuit of monkey Z. (a) The representation of all objects in each layer of the CNN (first 20 principal components) was regressed against the single-trial neural activity of each unit during the cue period, when the object was visible, and the median fit was taken over all units within one recording session. Solid line and error surfaces represent the mean and s.e.m. over all recording sessions of monkey Z. (a - inset) To ensure that results were not due to varying signal quality or firing rate between areas, insets shows regression results normalized to the median internal consistency of each area (i.e. half of trials correlated with the other half condition-wise). (b) Example Euclidean distance between neural representations of each object in AIP during the cue period and in the fc8 layer of the CNN (session Z6). (c) The mean muscle velocity of all grasping conditions during movement initiation (200 ms before to 200 ms after movement onset) was regressed against the single-trial neural activity of each unit during the same time period. Each point represents one recording session of monkey Z. (c - inset) Same normalization procedure as in (a - inset). (d) As in (b), but comparing the movement initiation representation in M1 to the muscle velocity representation in the same time window (session Z6).

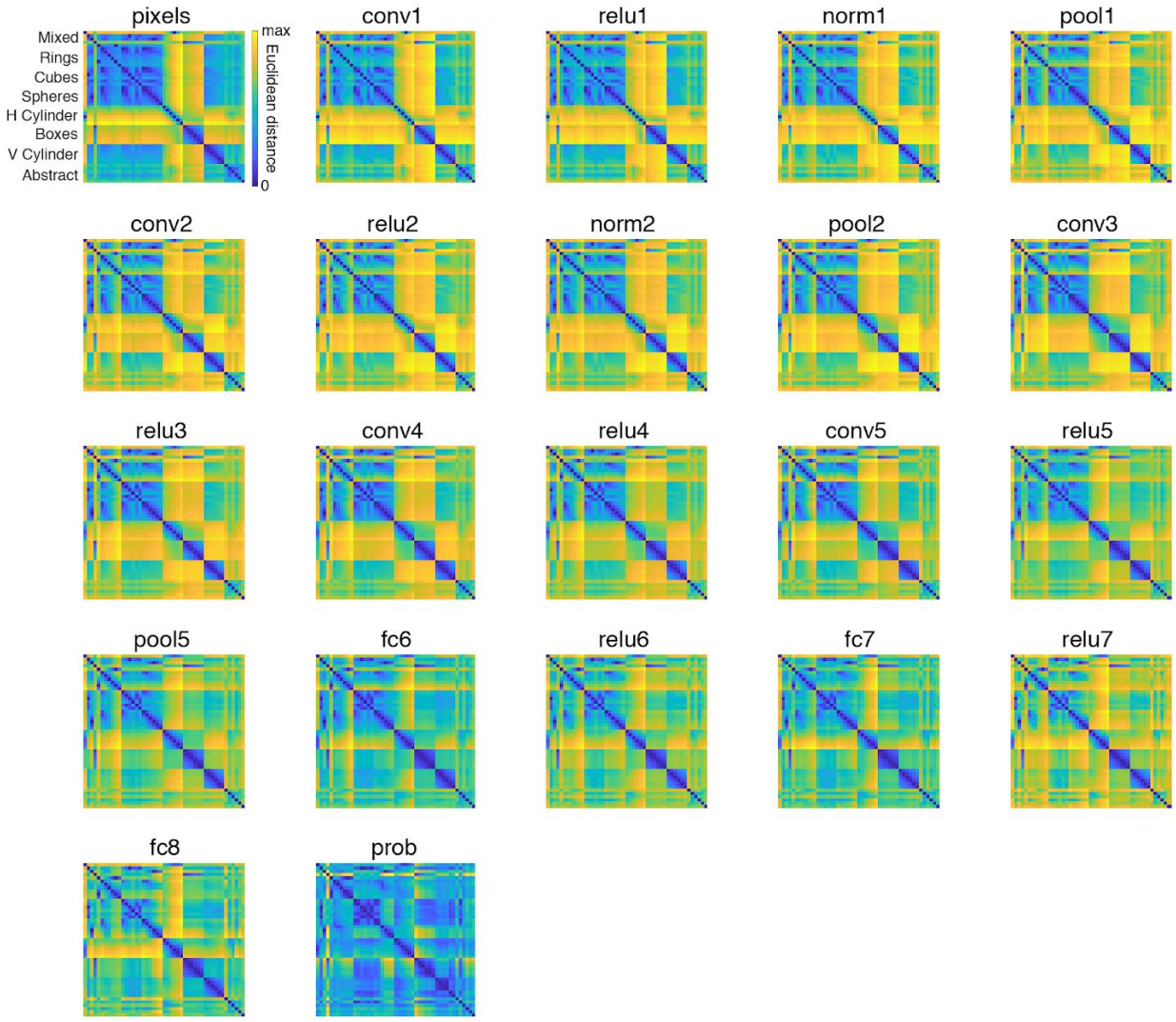

Fig. S3 | Feature representation in Alexnet. Example euclidean distance between representations (first 20 principal components) of each object in each layer of Alexnet.

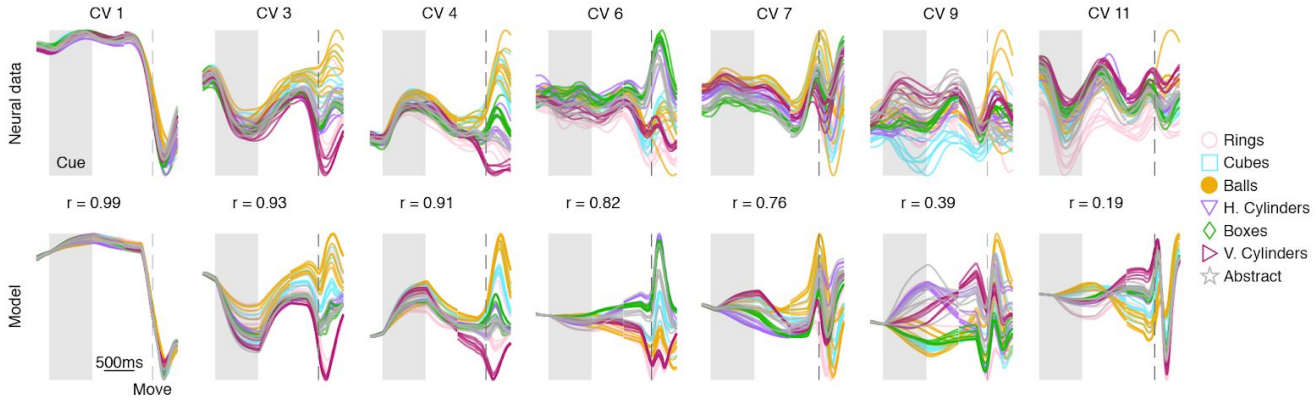

Fig. S4 | Regularized mRNN with visual input matches recorded neural data in monkey Z. (a) Example canonical variables (CVs) from canonical correlation analysis (first 12 principal components) between neural and simulated data across all brain regions and modules (session Z9), showing r-value for each dimension. There are multiple traces for each type of object, representing the different sizes or types within a turntable.

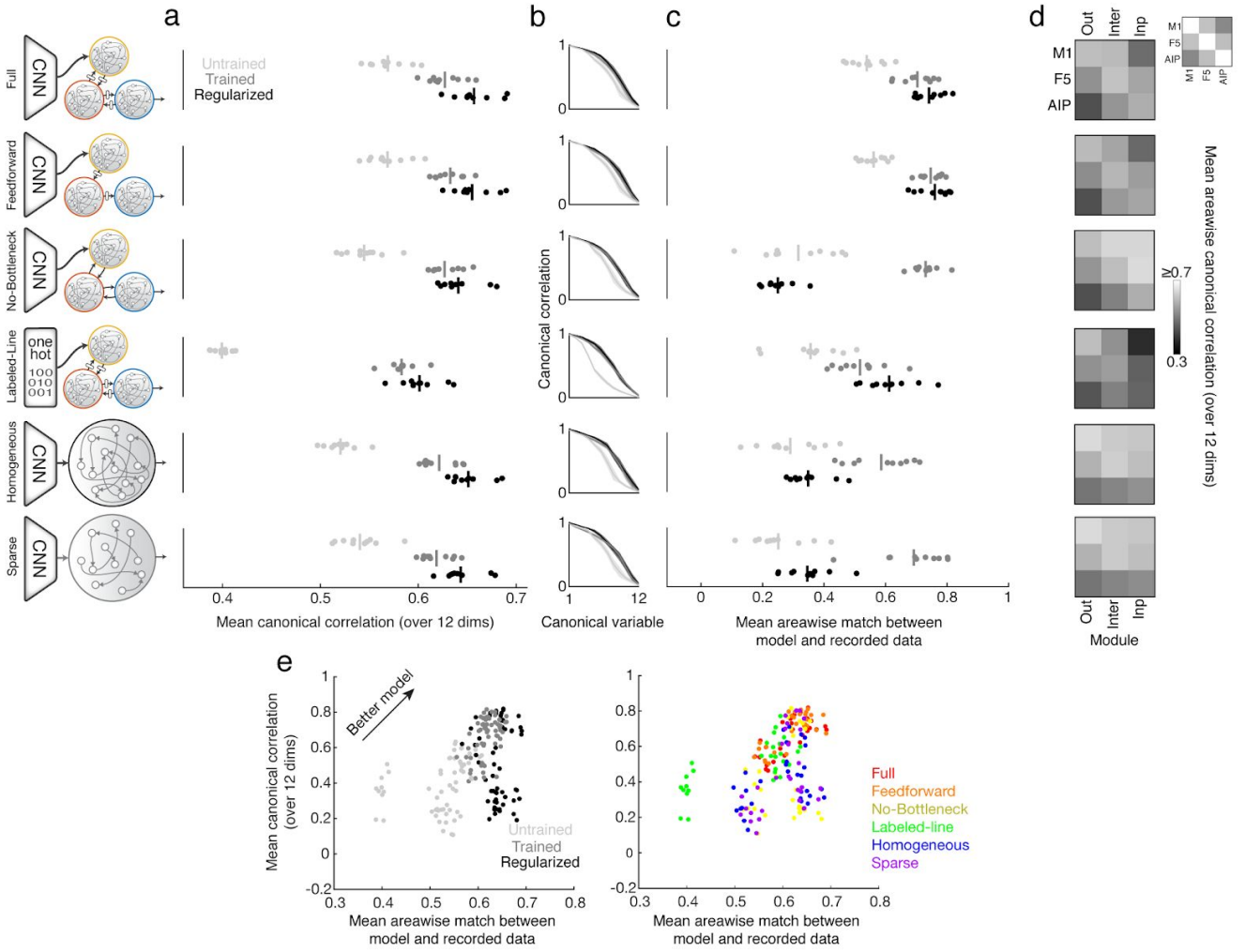

Fig. S5 | Temporal features of regularized neural network model match recorded neural data and align with corresponding brain regions in monkey Z. (a) Results of canonical correlation analysis across all sessions of monkey Z using 6 model architectures for untrained networks, networks trained to produce kinematics, and networks with additional regularizations. Vertical bars represent the mean, and each dot represents a single session. (a - first) Full model with CNN input and three modules. (a - second) same as first model, but with only feedforward connections. (a - third) Same as Full model, but with no flat layer bottleneck between modules. (a - fourth) Three module design receiving a labeled-line input (one-hot), where each condition is represented by a separate input dimension. (a - fifth) A homogeneous, fully-connected module receiving CNN input. (a - sixth) A single, sparsely-connected module receiving CNN input and sparsity matching the first model. (b) Mean correlations of each canonical variable for the models described in (a). Error bars represent standard deviation across recordings. (c) Canonical correlation was also performed between each module and each brain area and all pairwise canonical correlations were correlated with the inter-area canonical correlation in the neural data, quantifying the areawise match between neural and simulated data. (d) Average canonical correlation between each module and each brain region for the regularized model of each model architecture. Top inset shows canonical correlation between each brain region. (e) Summary of the results of (a) and (c) over all network architectures and recording sessions.

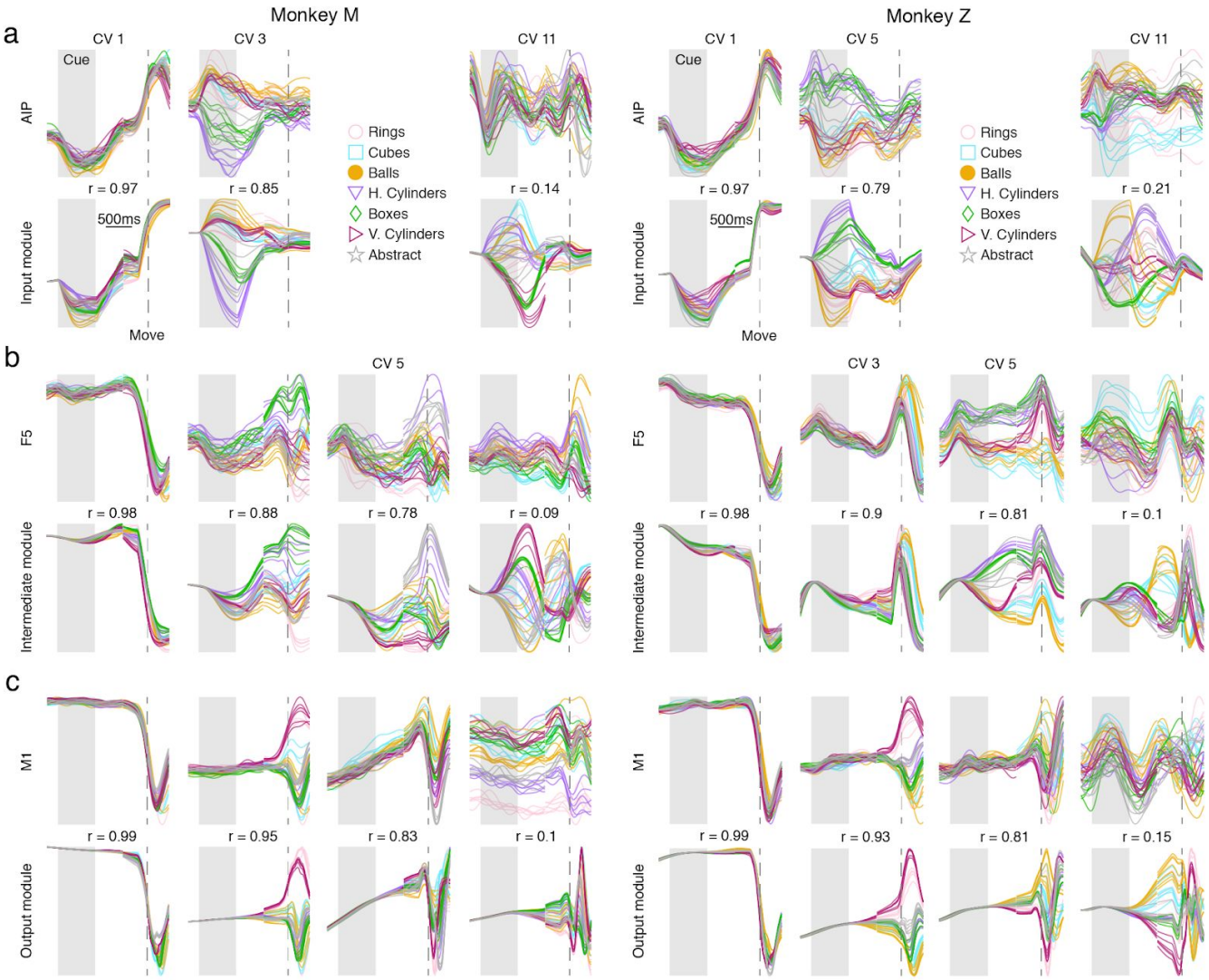

Fig. S6 | Regularized mRNN with visual input reproduces neural population features of each brain region. (a) Example canonical variables (CVs) of representative recording sessions (M10 & Z9) showing features in AIP captured by the input module. r-value for each dimension is shown, and there are multiple traces for each type of object, representing the different sizes or types within a turntable. The main features captured in AIP were a condition-independent signal that modulated during the cue and during movement onset and a strong cue-related signal that maintained some condition specificity throughout the trial. The main feature not captured by the model was a transient response to the cue that lasted less than 200 ms. (b) same as (a) for comparison between F5 and the intermediate module. In F5, a condition-independent signal that tracks movement initiation was most highly correlated, followed by multiple dimensions showing strong condition-dependance throughout the trial and tracking time within memory. (c) Same as (a) for comparison between M1 and the output module. M1 showed the movement initiation signal, and the majority of condition-specific activity was localized to the movement itself. Finally, CV 11 revealed a stable turntable signal that was not present in the model and present from the beginning of each trial. This signal reflected the fact that objects were grouped by turntable, which changed in a block design, an aspect of the task that we chose not to bake into the model, instead modeling each trial independently.

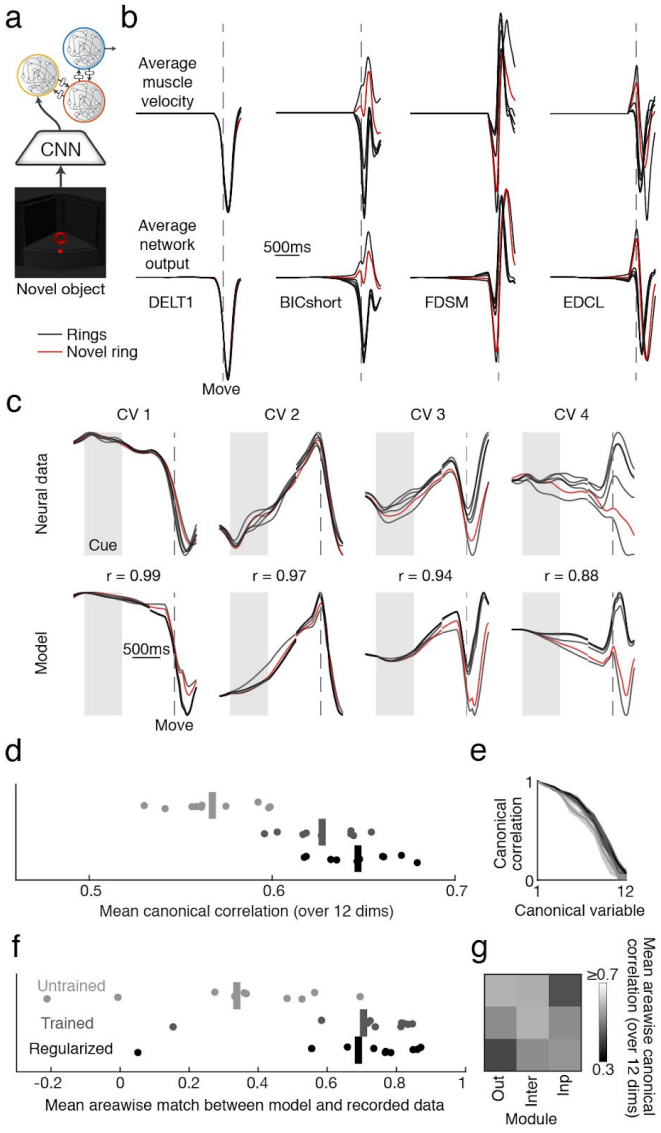

Fig. S7 | Generalizing muscle kinematics and neural population activity for novel objects in monkey Z. (a) Six additional regularized mRNN networks were trained using a limited set of objects and kinematics (40/48) and subsequently tested on all objects (6-fold cross-validation) to test the ability of the model to generalize to novel objects. Only one turntable (Rings) is shown in order to simplify visual comparison. (b) Example average output kinematics for four muscles (DELT1 - Posterior deltoid, BICshort - Biceps short head, FDSM - flexor digitorum superficialis digit 3, EDCL - extensor digitorum communis digit 5) of an example session (Z9), showing a subset of the trained conditions (5 rings) in black, as well as one of the untrained conditions in red. (c) Example canonical variables (CVs) for one recording session (Z9) fit to all trained conditions. The novel conditions were projected into the space determined by the trained conditions. Correlation (r-value) between each dimension is shown for novel objects only. (d) Results of canonical correlation analysis across all sessions and cross-validation folds using only the novel objects. Vertical bars represent the mean, and each dot represents a single session. (e) Mean correlations of each canonical variable for the data described in (d). Error bars represent standard deviation across recordings. (f) Canonical correlation was also performed between each module and each brain area and all pairwise canonical correlations were correlated with the inter-area canonical correlation in the neural data, quantifying the areawise match between neural and simulated data. (g) Average canonical correlation between each module and each brain region for the regularized model of each model architecture.