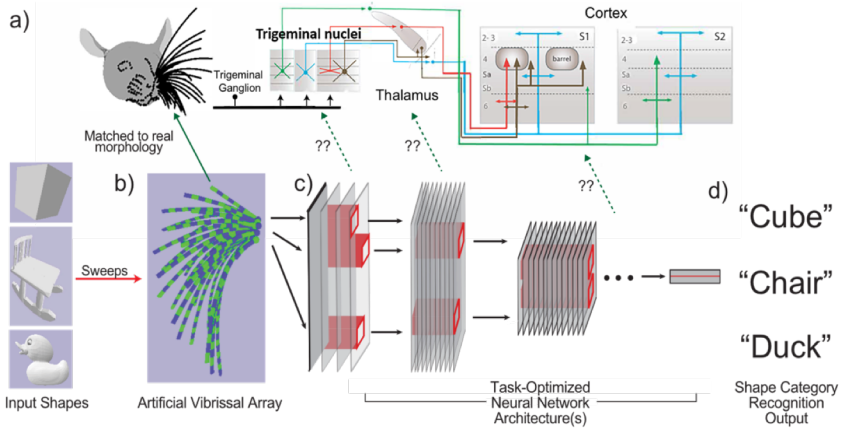
The virtual rat: Building a computational model of the rodent whisker trigeminal system

Chengxu Zhuang, Nadina Zweifel, Jonas Kubilius, Mitra J. Z. Hartmann, Daniel L. K. Yamins

Summary: Rodents "see" the environment mostly through their whiskers. The exquisitely sensitive, actively-controllable whisker array receives raw sensory data in the form of mechanical signals. These signals are carried along the trigeminal pathway through a sequence of increasingly complex processing stages, from brainstem to somatosensory ("barrel") cortex. Although many aspects of these processing stages have been characterized by a long history of experimental studies, the computational operations of the whisker-trigeminal system are poorly understood. In the present work, these computations are modelled through a goal-driven deep neural network (DNN) approach. Using a biophysically-realistic whisker array model (see companion poster, Zweifel et al., 2018), we sweep the array across a wide variety of 3D objects in highly-varying poses, angles, and speeds to generate a large dataset. Next, DNNs from several distinct architectural families are trained on this dataset to solve a shape recognition task. Each architectural family is based on a structurally-distinct hypothesis for processing in the whisker-trigeminal system. These hypotheses correspond to different ways in which spatial and temporal information can be integrated. After training, we find that reasonable performance levels on the challenging shape recognition task are only achieved by specific architectures from several families, while most networks perform poorly. Finally, we show that Representational Dissimilarity Matrices (RDMs), a tool for comparing population codes between neural systems, can separate these higher performing networks. And the data for computing RDMs is in a type that could plausibly be collected in an imaging or neurophysiological experiment. Our results are a proof-of-concept that DNN models of the whisker-trigeminal system are potentially within reach.

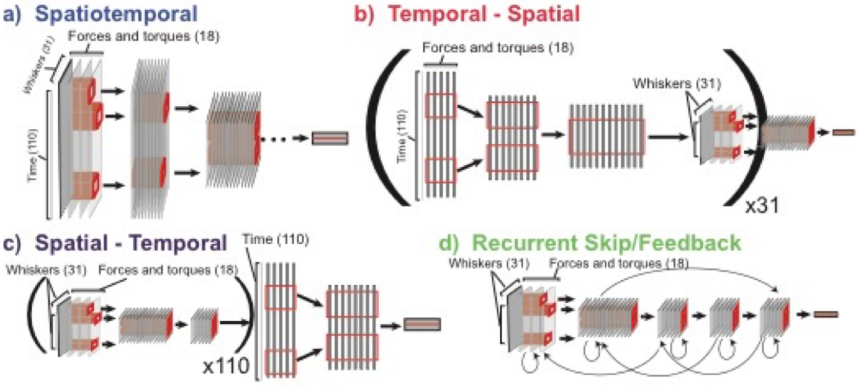
Overview: As shown in **Figure 1**, mechanical signals from the vibrissae are relayed by primary sensory neurons of the trigeminal ganglion to the trigeminal nuclei, the original of multiple parallel pathways to S1 and S2 (**Fig 1ab**). This system is a prime target for DNN modeling because it is likely to be richly representational, but its computational underpinnings are largely unknown. We seek to optimize neural networks of various architectures (**Fig 1c**) to solve shape recognition tasks (**(Fig 1d)**, and then measure the extent to which these networks predict fine-grained response patterns in real neural recordings.

Figure 1.



Hypotheses: We tested four hypotheses (**Figure 2**) for the integration of spatial and temporal information in the whisker-trigeminal system. The "Spatiotemporal" models (**Fig 2a**) have spatiotemporal integration at all stages. Convolution is performed on both spatial and temporal data dimensions, followed by one or several fully connected layers. In "Temporal-Spatial" networks (**Fig. 2b**), temporal integration is performed separately before spatial integration. Temporal integration consists of one-dimensional convolution over the temporal dimension, separately for each whisker. In spatial integration stages, outputs from

Figure 2.



each whisker are registered to their natural two-dimensional (2D) spatial grid and spatial convolution performed. In "Spatial-Temporal" networks (**Fig. 2c**), spatial convolution is performed first, replicated with shared weights across time points; this is then followed by temporal convolution. Finally, recurrent networks (**Fig. 2d)** do not explicitly contain separate units to handle different discrete timepoints, relying instead on the states of the units to encode memory traces. These networks can have local recurrence as well as long-range skip and feedback connections.

Performance results: Each bar in **Figure 3** represents results from one model. Many specific network choices within all families do a poor job at the task, achieving just-above-chance performance. However, within each family, certain specific choices of parameters lead to much better network performance. Overall, the best performance was obtained for the Temporal-Spatial model, with 15.2% top-1 and 44.8% top-5 accuracy. Training the filters was extremely important for performance; no architecture with random filters performed above chance levels. Architectures with fewer than four layers achieved substantially lower performance than somewhat deeper ones. Number of model parameters was a somewhat important factor in performance within an architectural family, but only to a point, and not between architectural families. The Temporal-Spatial architecture was able to outperform other classes while using significantly fewer parameters. Recurrent networks with long-range feedback were able to perform nearly as well as the Temporal-Spatial model with equivalent numbers of parameters, while using far fewer units. These long-range feedbacks appeared critical to performance, with purely local recurrent architectures (including LSTM and GRU) achieving significantly worse results.

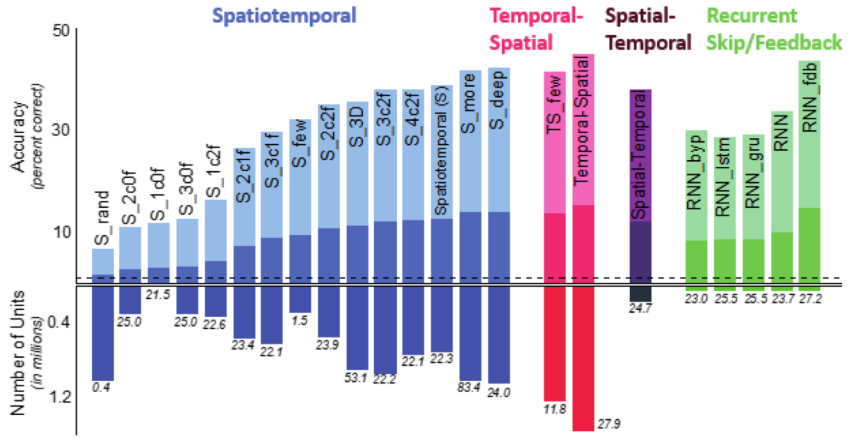


Fig. 3. The positive y-axis is performance measured in percent correct (top1 = dark bar, chance = 0.85%, top5 = light bar, chance = 4.2%). The negative y-axis indicates the number of units in networks, in millions of units. Small italic numbers indicate number of model parameters, in millions. "ncmf" means n convolution and m fully connected layers.

Model Discrimination. As shown in Figure 4, although the top layers of models have convergent RDMs, intermediate layers diverged substantially between models, by amounts larger than either the initial-condition-induced variability within a model layer or the distance between nearby layers of the same model. This observation is important from an experimental design point of view because it shows that different model architectures differ substantially on a well-validated metric that may be experimentally feasible to measure.

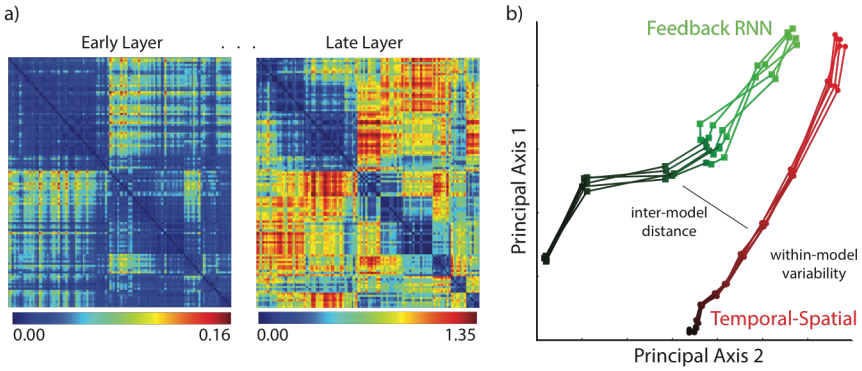


Fig. 4. **(a)** Representational Dissimilarity Matrices (RDMs) for selected layers of a high-performing network from Fig. 3, showing early and late model layers. **(b)** Two-dimensional MDS embedding of RDMs for the feedback RNN (green squares) and Temporal-Spatial (red circles) model. Points correspond to layers, lines are drawn between adjacent layers, with darker color indicating earlier layers.