Emergence of syntax and word prediction in an artificial neural circuit of the cerebellum

Keiko Ohmae[1] and Shogo Ohmae[1]

1. Neuroscience Department, Baylor College of Medicine, Houston, TX 77030, USA

Correspondence to: shogo.ohmae@bcm.edu (S.O)

ACKNOWLEDGMENTS

We thank Dr. Javier Medina for multiple rounds of brilliant advice and Dr. Mitsuo Kawato, Dr. Tatsuo Sato, Dr. Tatiana Schnur, Dr. Nuo Li, and Dr. Shigeru Kitazawa (listed in the chronological order that the comments were provided) for invaluable insights and comments that improved the manuscript. We also thank Mrs. Zoha Hassan for editing to select subtle nuances of language, Dr. Masashi Kimura (via MENTA: https://menta.work/) for continuous helpful advice on programming, and Dr. Lindsay Bremner (via NeuroEdit: https://www.neuroedit.com/) for excellent editing. This work was supported by grant to SO from the National Institutes of Health (R34 NS118445).

AUTHOR CONTRIBUTIONS

Conceptualization, S.O.; Data curation, S.O.; Formal analysis, S.O. and K.O.; Funding acquisition, S.O.; Investigation, S.O.; Methodology, S.O.; Project administration, S.O.; Resources, S.O.; Software, S.O.; Supervision, S.O.; Validation, S.O.; Visualization, S.O. and K.O.; Writing – original draft, S.O. and K.O.; Writing – review & editing, S.O. and K.O.

CONFLICT OF INTEREST

The authors declare no competing financial interests.

ABSTRACT

The cerebellum, interconnected with the cerebral neocortex, plays a vital role in human-characteristic cognition such as language processing, but knowledge about the underlying circuit computation of the cerebellum remains very limited. Here, to understand the computation underlying the cerebellar language processing, we developed a biologically-constrained cerebellar artificial neural network (cANN) model, which implemented the recently identified cerebello-cerebellar recurrent pathway. We found that while cANN acquires prediction of future words, another function of syntactic recognition emerges in the middle layer of the prediction circuit. The recurrent pathway of the cANN was essential for the two language functions, whereas cANN variants with further biological constraints preserved these functions. Considering the uniform structure of cerebellar circuitry across all functional domains, the single-circuit computation, which is the common basis of the two language functions, can be generalized to fundamental cerebellar functions of prediction and grammar-like rule extraction from sequences, that underpin a wide range of cerebellar motor and cognitive functions. This is a pioneering study to understand the circuit computation of human-characteristic cognition by biologically-constrained ANNs, which can lead new therapeutic interventions for human cognitive dysfunction.

182 words

INTRODUCTION

Language comprehension is a vital cognitive function that supports human communication and knowledge acquisition, but the underlying mechanisms at the level of neuronal activity within specific brain circuits remain largely unexplored. Artificial neural networks (ANNs) provide an invaluable tool to examine how the brain processes language, because there are no established animal models and only limited data from invasive recordings in humans. While recent AI advancements have led to ANN models that attempt to replicate brain functionality, these models do not adequately account for biological constraints. For instance, neocortical models for language processing based on the transformer algorithm[1, 2, 3], which are widely used in language-processing AI like ChatGPT, deviate from actual brain physiology. This is because the transformer-based models lack artificial neurons and process a sentence as a single unit rather than as a sequence of individual words. Thus, creating biologically constrained ANN models is essential to elucidate the computational processes of brain circuits involved in language processing.

Language comprehension after word recognition requires cooperation between the left neocortical language area and the right lateral cerebellum (right Crus I/II)[4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16]. The cerebellum has a large capacity of plasticity, and interestingly, according to recent observations that childhood damage to the right cerebellum causes more permanent and severe language deficits than adult damage[17, 18, 19, 20] and that the right cerebellum is activated during learning of new language tasks[16, 21, 22], cerebellar language acquisition is considered to support subsequent acquisition in the neocortex[15, 17, 18]. To understand the origins of language functions in the neocortex, it is crucial to understand the language processing mechanisms in the cerebellum at the circuit computation level. Here, we examine how language functions emerge in an ANN model that is biologically-constrained by the known connectivity and inputs/outputs of the cerebellar circuit (cANN).

Sentence comprehension involves recalling the meanings of words and combining them appropriately based on grammar to compose the meaning of the sentence, and during this process, two major language functions have been linked to the right lateral cerebellum: The first is (1) next-word prediction[4, 6, 7, 8, 9, 10], which supports faster and more accurate sentence comprehension by placing input words in a predictive context (particularly in noisy environments)[1, 6, 8, 10, 23, 24, 25, 26]. Although neural activity related to this function has also been reported in neocortical language areas[1, 3, 23], stronger evidence including causality has accumulated for the cerebellum[6, 7, 8, 9]. The second function is (2) grammatical processing, especially syntactic recognition about subject–verb–object (S-V-O) information[5, 11, 12, 13, 27]. These two language functions reflect two more general cerebellar functions that support a variety of cognitive functions. One is prediction of external events[4, 5, 10, 15, 28, 29, 30, 31], and (1) next-word prediction is a language case of this general function[6, 7, 8, 10, 15, 25]. The other is rule extraction from the sequence of events[11, 13, 32], and (2) syntactic recognition is a language case of this function[11, 13, 33]. Although the two general functions underlie various cerebellar cognitive functions[4, 10, 15, 28, 31] (see Discussion), it remains unclear how these different functions are

realized in a cerebellar circuit with a uniform cytoarchitecture. To address this gap in knowledge, elucidation of the underlying network dynamics by ANN is needed[15]. Therefore, in the context of language processing, we tested whether a single circuit architecture can support both general functions, thereby shedding light on the general circuit computation of the cerebellum.

Here, we found that the cANN can acquire prediction in language (i.e., next-word prediction) through training, and simultaneously rule extraction from sequences (i.e., syntactic recognition) can also be acquired spontaneously in the middle layer of the same circuit. This suggests that two general functions of the cerebellum, prediction and rule extraction from sequence, can be captured by a single circuit computation of the cerebellum. This also indicates a new treatment for language dysfunction, in which training in word prediction leads to improvement in syntactic comprehension.

RESULTS

The cerebellar ANN model as a next-word prediction circuit

We started by assessing whether the biologically-constrained cANN could learn next-word prediction. Our cANN contains three layers—an input layer (granule cells), a middle layer (Purkinje cells), and an output layer (nuclei neurons), and the cANN consists of a conventionally reported feedforward pathway and a recently identified recurrent pathway (brown, blue pathways in Fig. 1b, respectively). The recurrent pathway includes output cell-input cell pathway, also called the nucleocortical pathway[34, 35, 36, 37, 38, 39, 40], that has been demonstrated to be essential for prediction in the cerebellum[34, 37, 41]. The cANN circuit also incorporates the climbing fiber pathway (gray, Fig. 1b) which delivers prediction errors to the cerebellum[6, 10, 28, 42, 43, 44]. Based on physiological observations that the prediction signal generated by Purkinje cells after a cue event is persistent until the actual event occurs[34, 45, 46], we assumed that the prediction signal of Purkinje cells in the cANN is persistent until the next word (see Discussion). By this persistent prediction signal, the recurrent signal is maintained, allowing it to be integrated with the next word. Additionally, the circuit output signal is also maintained and can be compared with the correct answer signal of the actual next word, to compute the prediction error signal.

We designed the inputs/outputs of the cANN (Fig. 1b) according to previous proposals for the next-word prediction circuit of the cerebellum as in Figure 1a[6, 7, 8, 10, 25]. Based on the sparse coding theory of the input cells (i.e., granule cells), we assumed that words were represented by sparse coding in the input cells: each of the 3000 cells was randomly assigned to one of 3000 words and hence the input signal contained no semantic or grammatical information at all (confirmed later, Fig. 3d). To facilitate the investigation of information processing in this network, we assumed that the correct answer signal (correct answer = actual next word in Fig 1b, gray) was encoded by the same 3000-dimensional sparse coding as in the input cells, so that the correct answer signal contained no syntactic information at all (Fig. 3d). Because the dimensionality of the output cells is theorized to be the same as that of the correct answer signal[28, 47, 48, 49], we set the number of the output cells as 3000. Through the training of cANN, the activity intensity of each output cell can indicate the likelihood of the corresponding

word being the next word (percentages in Fig. 1b). Solely for performance evaluation purposes, we selected the five output cells with the strongest activity and the corresponding five words as the prediction candidates for the next word. Additionally, to replicate the anatomical convergence from Purkinje cells to the output cells (i.e., the number of Purkinje cells > that of the output cells), we created and examined another cANN with this biological convergence (Fig. 5).

For learning, the correct information for the next word was delivered to the circuit word by word and the synaptic weights were updated to reduce the prediction error, in line with the learning rule of the cerebellum (Fig. 1b, gray)[10, 50, 51].

Training the cANN with sentences (e.g., from classic novels) produced significant improvements in both the prediction error and the correct prediction rate (Fig. 1c,d). At the very early stages of training, the correct prediction rate jumped from 0% to about 17% (Fig 1d, gray arrowhead), and at this point the prediction was based on the words that appeared most frequently in the training sentences, regardless of the identity of the immediately preceding word (Extended Data Fig. 1a,b). Then, the correct rate slowly increased to about 40%, where it plateaued; here, the cANN achieved correct prediction rates higher than chance regardless of the frequency of occurrence of the target word (Extended Data Fig. 1f), and the cANN made different next-word predictions after each of the two instances of “the” in a sentence (Fig. 1e, gray boxes), indicating that the circuit utilized words prior to the immediately preceding word.

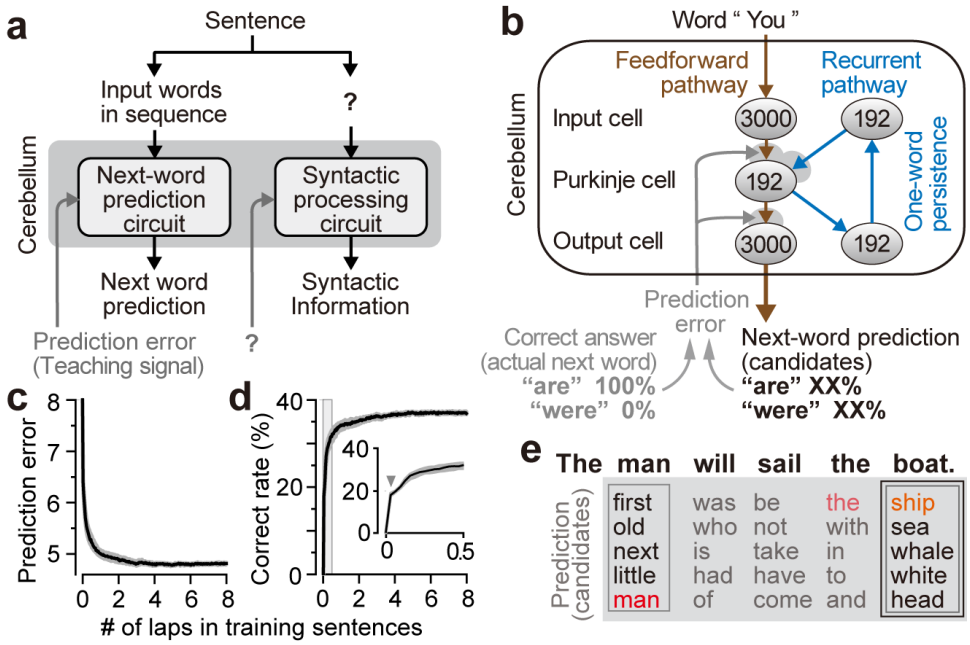

Figure 1. Local circuit model of a cANN to reconstruct cerebellar language functions. **a**, Previously proposed block diagram with Inputs/outputs for next-word prediction. **b**, The three-layer cerebellar circuit, consisting of a feedforward pathway (brown) and a recurrent pathway (blue). The climbing fiber pathway (gray arrows) delivers the prediction error signal to update the

synaptic weights on the Purkinje and output cells (gray shading). **c,d**, Learning curves for the prediction error (**c**; cross entropy; median and IQR) and percentage correct predictions (**d**) for 20 cANNs trained on next-word prediction. Inset, magnified view of the early learning period (gray interval). **e**, Predictions (five candidate words) for the two instances of "the" in a sentence by a well-trained cANN. Correct predictions are highlighted in red, and synonyms are in orange.

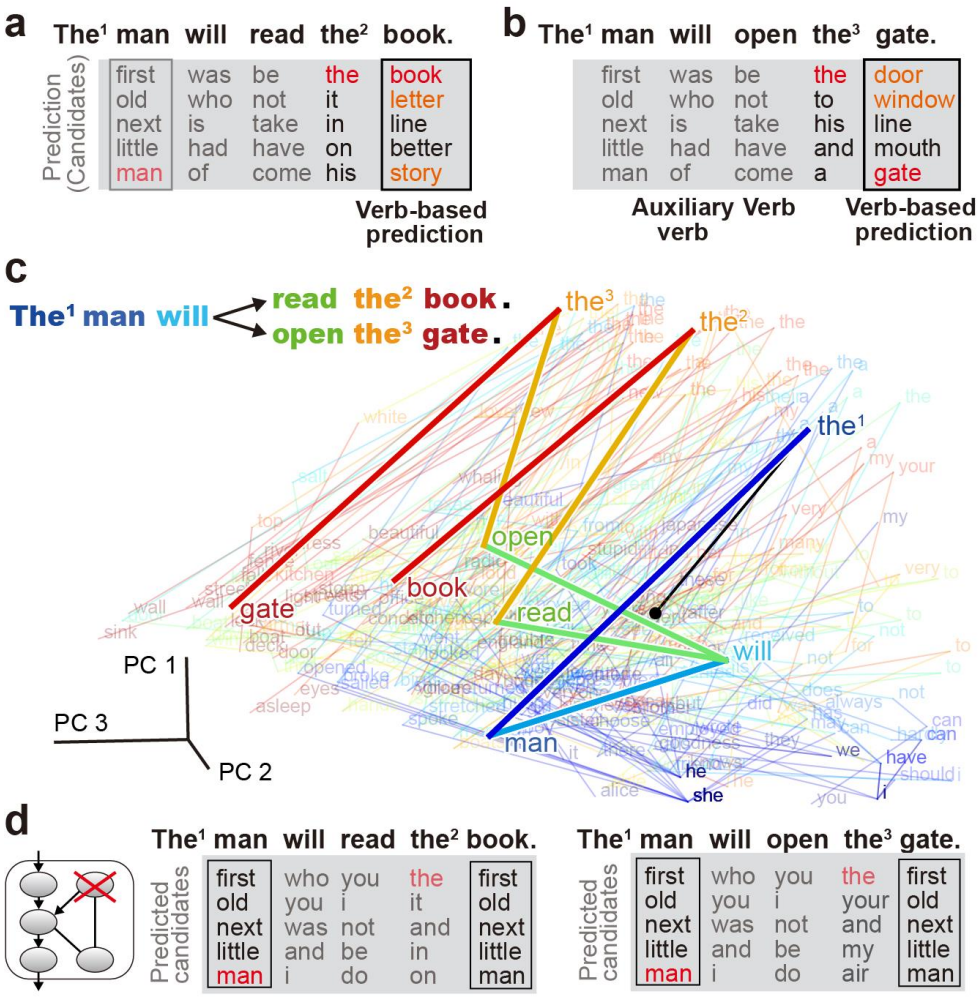

Figure 2. Next-word prediction and network dynamics of a representative cANN. **a,b**, Predictions after three instances of "the" (indexed) in two sentences. **c**, Visualization of the neural activity of Purkinje cells after input of each word by PCA. Words are color-coded in rainbow from the beginning of the sentence to the end, and spontaneous activity is indicated by the black dot. **d**, Predictions when the recurrent signal was blocked.

Next, to replicate the finding from human cerebellar research[7, 8], we examined whether the cANN could predict the noun after a verb. For the sentence “The man will sail [verb] the boat [noun]” used in the previous research, together with two similar sentences, the cANN predicted correct next words and synonyms (Figs. 1e, 2a,b, black boxes). Predicting the noun after a verb had a lower correct rate (median, 22.9%; interquartile range (IQR), 21.0%–23.4%) than the overall prediction rate (38.3%; 37.8%–38.6%), possibly because verbs often have a large number of potential noun partners (e.g. partners of “eat”). By contrast, predictions were more accurate for other word types after a verb (for prepositions: 60.0%, 56.7–63.3; object pronouns: 56.5%, 52.2–60.9). The cANN also made proper predictions for other word types, such as nouns after prepositions (43.0%, 42.1-45.6), and adjectives after “be” verbs (27.5%, 25.0-30.0). From these results, we conclude that the cANN is able to function as a next-word prediction circuit, although the correct rate varies with the word type.

To investigate the mechanism of information processing, we examined the role of the Purkinje cells, where the feedforward signal of a newly arriving word and the recurrent signal containing the previous word information are integrated (Fig. 1b). When we visualized the word-by-word transition of Purkinje cells activity in the sentences using principal component analysis (PCA), three different instances of "the" generated different activity patterns in the Purkinje cells (Fig. 2c, the[1], the[2], the[3]). Since the Purkinje cells are the only element of the circuit upstream of the output cells responsible for predictions, these differences are essential for generating different predictions (Fig. 2a,b, gray and black boxes). When the recurrent input was blocked, the activity in the Purkinje cells became identical for all three instances of “the” (Extended Data Fig. 1c), and the impaired cANN no longer produced different predictions (Fig. 2d), indicating that only the immediately preceding word is encoded. More generally, prediction of the noun after a verb was dramatically degraded in the impaired cANN (correct rate, 2.1%; IQR, 1.1%–2.1%).

The cANN model as a syntax-processing circuit

There is little-to-no evidence as to how the inputs/outputs should be designed for another cANN dedicated to syntactic processing[15] (Fig. 1a). For the cerebellum to assist the neocortex in language acquisition, the cerebellum must acquire language processing on its own without information from the neocortex. Inspired by the potential of the middle layer of recurrent ANNs[52, 53, 54], we investigated whether subject–verb–object (S-V-O) syntactic information could emerge within the next-word prediction circuit of the cANN, whose input signals contain no syntactic information. Indeed, the cANN reliably predicted objective case pronouns after a verb (Extended Data Fig. 1d,e), suggesting that the predictions reflect S-V-O information. Since the Purkinje cells may integrate new and past words according to S-V-O structure, we visualized the activity of the Purkinje cells by PCA (Extended Data Fig. 2a) and by linear support vector machine (Fig. 3a) and found that S words were clustered and well separated from V-O words (blue circle), regardless of the number of words forming the subject (black trace). The separation of S words from V-O words (96.1%; 95.7–96.6; nonlinear support vector machine), V words (96.0%; IQR,

95.7–96.5), and O words (95.8%; 95.4–96.1) was highly accurate (visualized in Extended Data Fig. 2b). Since syntactic information is not included in the correct answer of the next word (confirmed in Fig. 3d, right), these results indicate that the Purkinje cells in the word prediction circuit are capable of extracting highly accurate SVO syntactic information independent of external information. Importantly, even for the same word, the cANN can distinguish when the word appears as a subject (e.g., “the” as a part of a subject, blue arrowhead in Fig. 3a) and when it appears as an object (“the” as a part of object, yellow arrowheads), ensuring that it has S-V-O recognition capability. Thus, to our surprise, syntactic processing does not require a dedicated circuit (Fig. 3b). This may be why the conventional approach of assuming a dedicated circuit as in Figure 1a worked well for next-word prediction, but not for syntactic processing[15].

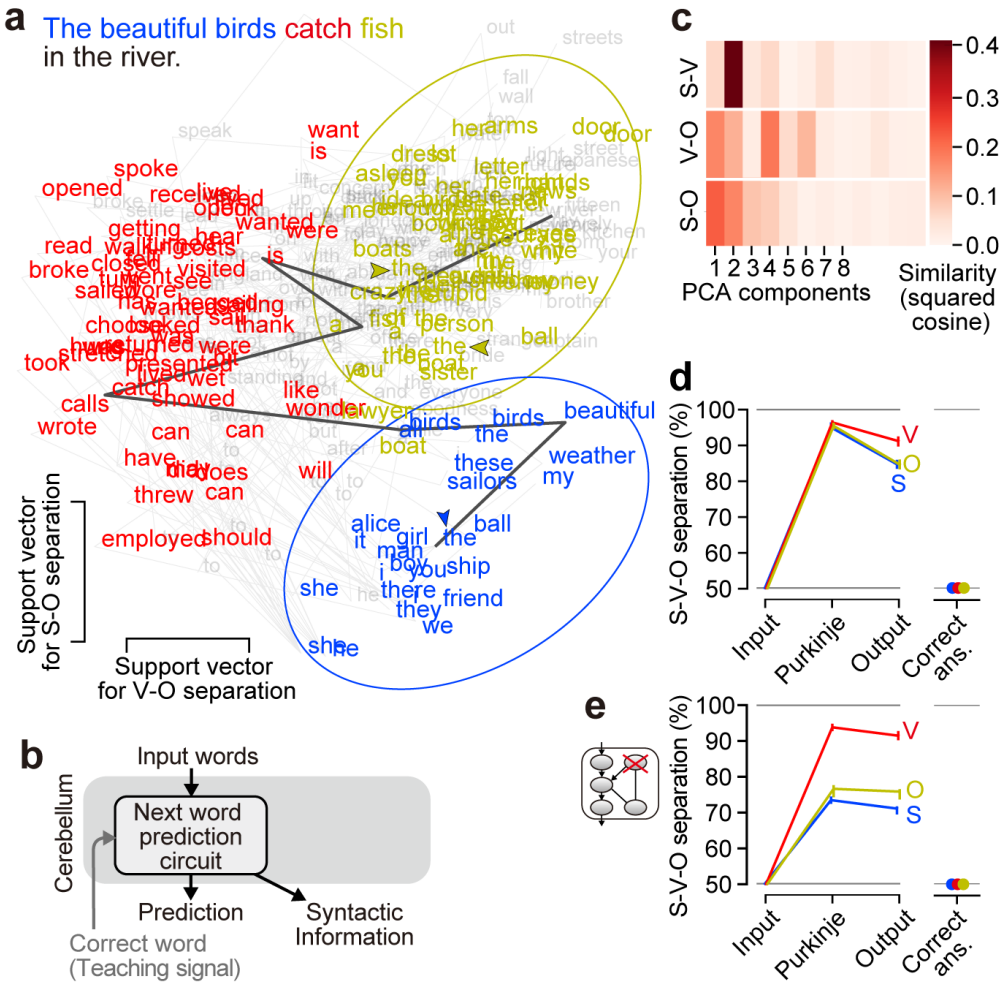

Figure 3. Syntactic processing by the cANN. **a**, Visualization of neural activity of Purkinje cells after inputting a subject (blue), verb (red), and object (yellow), in the dimensions used for classification. The activity transition for a sentence with a longer subject (three words) is highlighted (dark gray trace). **b**, Inputs/outputs for syntactic processing. **c**, Similarity analysis of the

dimensions of S-V, V-O, and S-O separations with PCA dimensions (median). **d**, Syntactic information contained in the feedforward-input, Purkinje, feedforward-output cells, and correct answer signal. The chance level of separation is 50% (gray horizontal line). **e**, Syntactic information when the recurrent signal was blocked.

To investigate how syntactic information emerges in the cANN, we compared the syntactic information along the feedforward pathway (Fig. 3d). We confirmed that the input cells contain no syntactic information at all, and found that syntactic information was predominantly extracted in the Purkinje cells and was degraded in the output cells ($p < 0.001$ for any of S, V, and O, Wilcoxon signed-rank test; visualized by distorted distributions in Extended Data Fig. 2c), probably because their direct task is next-word prediction. Furthermore, when the recurrent signal was blocked, the separation of S and O dropped sharply, revealing the importance of the recurrent signal for processing syntactic information (Fig. 3e).

Examining the extent to which S-V-O syntactic information is a primary piece of information in Purkinje cells, we found that the dimensions used for S-V, V-O, and S-O separation by the support vector machine were all similar to the major dimensions of PCA space in terms of direction (Fig. 3c) and information content (Extended Data Fig. 2d,e). This indicates that S-V-O syntactic information is the primary information represented by the activity of Purkinje cells.

Further biologically constrained cANN variants

We developed three new cANN variants incorporating further biological constraints and assessed their performance. In the first variant, the recurrent pathway was designed according to the cerebellar anatomy, where connections from Purkinje cells to output cells are convergent, and connections from output cells to input cells are divergent. We set the number of output cells to 48 and the number of input cells to 192 in the recurrent pathway, so that Purkinje cell signals must be compressed and decompressed through this pathway. Given that plasticity has been reported for various connections in the cerebellum[55], we assumed plasticity in the recurrent pathway. This cANN variant acquired the equivalent language functions to the original cANN (median of correct prediction rates, 36.5%; median of syntactic separation accuracy, S, 93.4%, V, 95.1%, O, 95.0%). This demonstrates that the robustness of cANN's language functions despite cell number variations in the recurrent pathway.

The second cANN variant, with inhibitory-restricted Purkinje-output connections, also achieved equivalent language functions (median of correct prediction rates, 36.8%; median of syntactic separation accuracy, S, 94.8%, V, 96.2%, O, 95.5%). This demonstrates cANN's robustness regarding sign of connection between the Purkinje cells and the output cells, suggesting that such constraints do not hinder information processing. This aligns with the physiological observation that, although Purkinje cells has only GABAergic connections to the

output cells, they can both inhibit and disinhibit the output cells by increasing and decreasing their firing rates (i.e., the sign of the downstream signal is not limited).

In the original cANN, the distribution of the synaptic weights of input-Purkinje connection resembled a normal distribution (Fig. 4a). If considering the greater-than-zero portion of the weights (w>0; black in Fig. 4a) as the weights of parallel fiber-Purkinje cell synapses (i.e., direct excitatory input cell-Purkinje cell connection), this distribution closely aligns with the physiological distribution of parallel fiber-Purkinje cell synaptic weights, which can be approximated by a half of a normal distribution[56, 57]. However, a sharp peak with zero synaptic weight (w=0; silent synapses) was observed in the physiological data but not in the cANN. Therefore, we created the third cANN variant combining excitatory-limited input-Purkinje projections (imitating direct projection) and inhibitory-limited input-Purkinje projections (imitating indirect projection via molecular layer interneurons). This cANN variant also achieved comparable language functions (median of correct prediction rates, 37.0%; median of syntactic separation accuracy, S, 94.5%, V, 96.6%, O, 95.5%). Interestingly, the distribution of the excitatory-limited synapses exhibited the peak of silent synapses, closer to the physiological observation (Fig. 4b). These findings demonstrate the robustness of the cANN functions to stringent biological constraints, while highlighting the need for such constraints to replicate circuit details such as synaptic weight distributions.

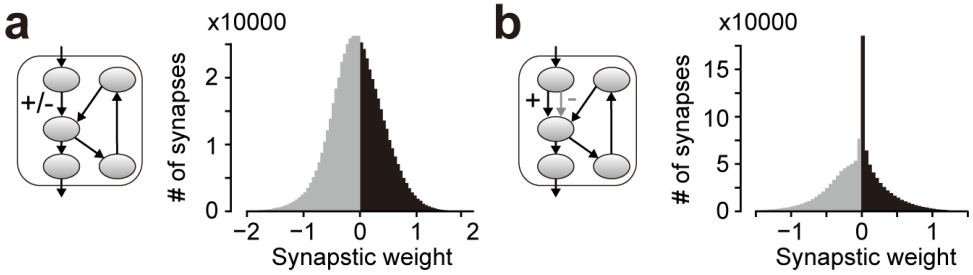

Figure 4. Distribution of synaptic weights at the input cell-Purkinje cell connection. **a**, The original cANN's synaptic weight distribution (right) without sign restriction at input-Purkinje synapses (left). The negative-weight component is in gray, and the positive-weight component is in black. **b**, The synaptic weight distributions of a cANN variant with input-Purkinje synaptic signs restricted to either excitatory (black) or inhibitory (gray).

Convergent cANN with the biological Purkinje-output convergence

We examined the performance of another cerebellar cANN model with an anatomical constraint of convergence of the Purkinje-output connection[36, 58] and a physiological constraint of population coding of the output cells[51]. To this end, we modified the original cANN circuit (Fig. 1b) to have only 16 output cells in the feedforward pathway so that the activity pattern of the 16 output cells represents a single word by population coding. Then, the modified circuit predicts only a single word at a time, unlike the original cANN which could probabilistically predict

multiple candidate words. To enable the output of multiple word candidates, based on the physiological observation that the cerebellum has a modular structure which allows for simultaneous generation of multiple outputs, we treated this modified circuit as one module and created a new cANN model with 10 modules in a row (Fig. 5a). Since parallel fibers, projections from the input cells to Purkinje cells, are known to pass across many modules, we assumed the input cells are shared between the modules. This new cANN is referred to as a "convergent cANN", while the original cANN is referred to as a "non-convergent cANN".

For learning, based on the cerebellar theory that enables multiple modules to cooperate and achieve a complex function (MOSAIC model)[31, 48, 58, 59], after each prediction, the prediction error signal was given solely to the module with the closest prediction to the correct answer (= the actual next word). Based on the cerebellar theory that the output cells and the correct answer signal should have the same dimensionality[28, 47, 48, 49], the correct answer was set to be represented in 16 dimensions. To represent the correct answer of 3000 words in 16 dimensions, we assumed a compressed word representation in which similar words (e.g., ship and boat) are represented by similar signal patterns, as observed in the neocortex[60], which is the source of the correct answer signal[10, 25, 28]. To facilitate comparison with the non-convergent cANN model, the following analysis focused on the top five modules in terms of prediction accuracy at the final training stage (see Methods).

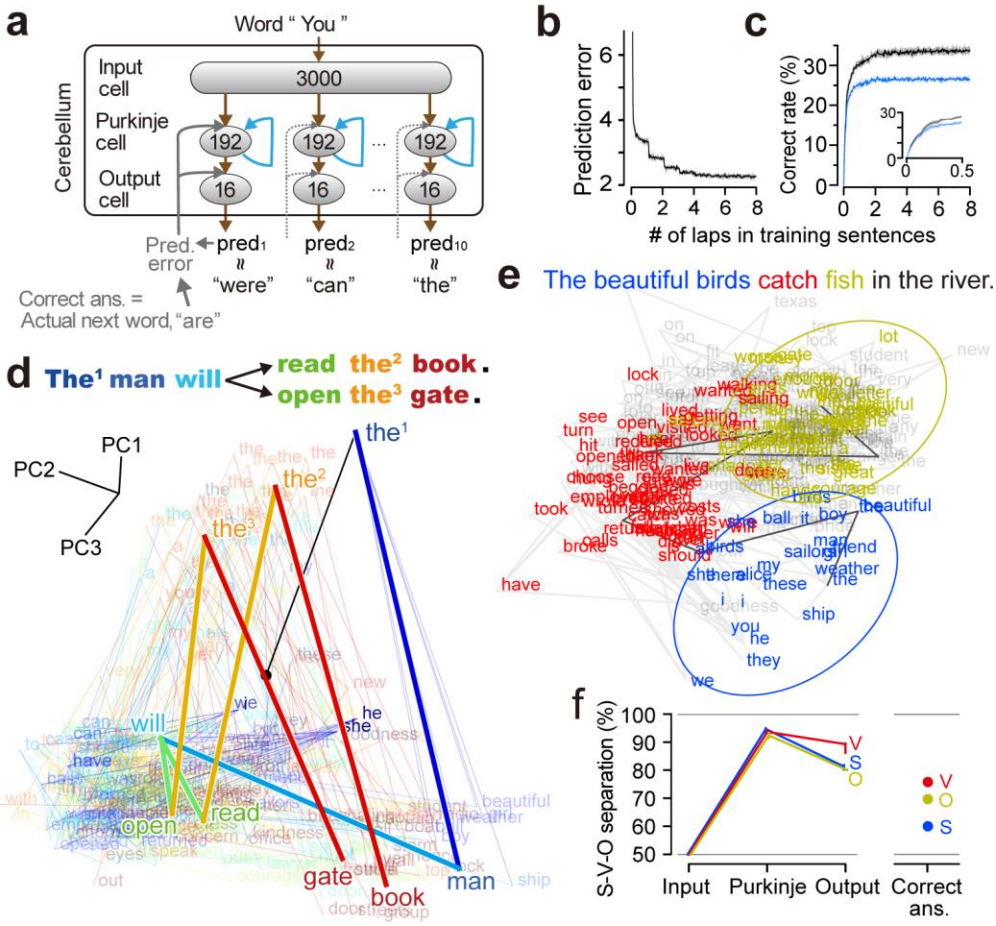

Figure 5. Performance of the convergent cANN.

a, Circuit design of the convergent cANN. The convergent cANN consists of 10 modules, each containing the same circuit as the original cANN in Figure 1b, except for having 16 output cells. After each prediction, the prediction error is given solely to the module that produced the closest output to the correct answer signal of the actual next word (gray solid arrows). **b**,**c**, Learning curves of 10 convergent cANNs for the prediction error (**b**; mean squared error) and percentage of correct predictions (**c**) for all the 10 modules (black) and for the top 5 modules (blue). **d**, Visualization of the neural activity of Purkinje cells of a representative convergent cANN after input of each word, by PCA. Words are color-coded in rainbow from the beginning to the end of the sentence, and spontaneous activity is indicated by the black dot. **e**, Syntactic processing by the convergent cANN. Neuronal activity of Purkinje cells in the convergent cANN after input of a subject (blue), verb (red), and object (yellow) is visualized by the dimensions of the V-O and S-O classifications of the support vector machine. The solid black trace indicates the activity dynamics for the sentence, "The beautiful birds catch fish in the river". **f**, Syntactic information contained in the feedforward-input, Purkinje, feedforward-output cells, and the correct answer signal. The chance level of separation is 50% (gray horizontal line).

Training the convergent cANNs with the training sentences (e.g., from classic novels), again, significantly improved next-word prediction (Fig. 5b,c). The correct prediction rate using top 5 modules was 26.4% (median; interquartile range (IQR), 25.7%–27.1%; Fig 5c, blue; 33.6% by 10 modules), indicating that the convergent cANNs can also acquire next-word prediction. Because learning of the modules occurs independently, one after another, each convergent cANN module had only 1/10 of the learning opportunities compared to the non-convergent cANN, which may have resulted in lower prediction accuracy. In the example sentences (Fig. 5d, top), three different instances of "the" generated different activity patterns of Purkinje cell signals (Fig. 5d, the^1, the^2, the^3), which led different proper predictions. Specifically, after “the^1” for “man”, a representative convergent cANN predicted “general”, “prince”, “roll”, “same”, and “kind”; after “open the^2” for “gate”, it predicted “post” and “door”; and after “read the^3” for “book”, it successfully predicted “book” and “press”. In total, the correct prediction rate of the noun after a verb was 16.0% (median; IQR, 14.9%–17.0%), which was slightly lower than that of the non-convergent cANNs. In contrast, the correct prediction rate of prepositions after verbs was 65.0% (median; IQR, 60.8%–66.7%), slightly higher than that of the non-convergent cANNs. These results indicate that the convergent cANN has capability to make different proper predictions according to context, although the prediction accuracy is slightly lower depending on the word type than the original non-convergent cANN.

We then examined the capability of syntactic processing in the intermediate layer of the convergent cANN and found that signals for subject, verb, and object words were highly clustered and well separated (Fig. 5e). Comparing syntactic information along the feedforward pathway (Fig. 5f), we found that syntactic information was not above the chance level in the input layer, peaked at the Purkinje layer (S, 94.9%, V, 93.8%, O, 92.6%; almost equivalent to the non-convergent cANN), and was significantly degraded in the output layer ($p < 0.01$ for any of S, V, and O, Wilcoxon signed-rank test). Although the correct answer signal contained a small amount of syntactic information (Fig. 5f, right), the Purkinje cells in the convergent-cANN acquired a higher level of syntactic information, indicating that the convergent cANN have the capability to extract syntactic information beyond all external inputs. The important conclusion here is that our cerebellar models can acquire two cerebellar language functions, regardless of the dimensionality or the cording format of the output cells.

DISCUSSION

Anatomical and physiological support for the cANN

Our cANN is highly consistent with a wealth of anatomical and physiological knowledge. This new model effectively accounts for the cerebellar language functions by incorporating the recently identified recurrent pathway into the feedforward circuit, which is the fundamental structure of the conventional cerebellar circuit models. Regarding the recurrent pathway, numerous reports have demonstrated the existence of the key part of the pathway, the output cell-input cell projection[35, 36, 37, 38, 39], and its functional significance on the cerebellar prediction[34, 37, 41]. The recurrent pathway has been implemented in cerebellar modeling[61]. Consistent with this growing evidence for the significance of the recurrent pathway, we found that the pathway is essential for predicting the next word in a sentence and processing syntactic information in the cANN. We assumed that the Purkinje cell signal which is the source signal of the recurrent pathway is persistent until the next word arrives, based on the physiological observation that the predictive signal in Purkinje cells persists from the cue event until the actual event targeted for prediction occurs[34, 45, 46]. This persistent predictive signal enables the recurrent signal to last until the next word, preparing for integration with the next word information of the feedforward pathway. The time interval between words in natural speech and reading is approximately 300 milliseconds (200 words per minute), which aligns well with the reported time scale of predictive signal persistence in the Purkinje cells (over 500 ms)[34, 45, 46].

Traditionally the cerebellum has been modeled as feedforward circuits[57, 62, 63, 64, 65, 66, 67], whereas our cANN incorporates the recurrent pathway. Feedforward models, with their limitation of predicting the next word based solely on the immediately preceding word, have significantly reduced capabilities compared to the cANN (see Extended Data Fig. 3). Indeed, when we remove the recurrent pathway from the cANN and trained the cANN variant (emulating a feedforward cerebellar model), both word prediction and syntactic processing became inaccurate (data not shown), resembling the results of blocking the pathway post-training (see Extended Data Figure 1 and Figure 3e). Moreover, conventional feedforward cerebellar models,

based on a two-layer perceptron model with Purkinje cells as the output layer, lack the remarkable middle-layer processing capabilities found in the cANN. In contrast, the cANN can extract syntactic information through middle-layer processing after training the output layer to predict the next word. For the output of the middle-layer processing, we propose candidate output pathways that can be tested in future circuit studies (Extended Data Fig. 4a).

With respect to learning, the cANN utilizes prediction error-based learning, a method frequently employed in ANN models of the brain. Although this type of learning is widely used, the cerebellum is a rare region where its application is physiologically plausible[68]. The cerebellum is clearly considered to learn based on prediction errors in both motor and cognitive processing[10, 28, 36, 42, 69]. The cANN incorporates the inferior olive-climbing fiber pathway, which calculates and conveys prediction errors to the cerebellum (Fig. 1b). In prediction error calculation, there is no need for a retention mechanism for the prediction signal until the next word arrives, since the persistent predictive signal of Purkinje cells ensures that the circuit output persists. Also, a series of studies indicate that the cerebellum can learn more flexibly than classical cerebellar learning rules[40, 42, 43, 55, 70, 71, 72, 73, 74, 75]. In particular, based on experimental findings and proposals that synaptic weights of Purkinje cells and cerebellar nucleus neurons are updated in the direction to reduce prediction error, our cANN updated synaptic weights[10, 50, 51].

Furthermore, we created cANN variants (including the convergent cANN) with further biological constraints, demonstrating the robustness of the two cerebellar language functions of the cANN. Together, our cANN is multilaterally consistent with our biological and physiological knowledge about the cerebellum and is capable of reproducing high-level cognitive functions of the brain.

Significance of a common computational basis for the two functions

Because the cerebellum has a uniform cytoarchitecture[4, 15, 36, 40, 76, 77, 78], much conventional research has been directed toward elucidating the common circuit computation underlying all cerebellar functions. Although many descriptive ideas have been proposed, none have yet successfully explained all the diverse functions of the cerebellum[15]. For example, in the internal-model theory, the cerebellum is a predictive simulator of the external world. This theory can explain next-word prediction by replacing the external world as the prediction target with sentences and viewing the cerebellum as a sentence simulator[6, 7, 8, 10, 15, 25]. However, explaining syntactic processing with the predictive functions of the theory is difficult and requires other ideas such as sequence processing[11, 13, 32].

Recognizing the limited reach of descriptive ideas to the common circuit computation, Diedrichsen, Ivry, and colleagues emphasized the need for a more abstract concept, like the network dynamics of neurons, to explain cerebellar computation[15]. Responding to precisely this need, we here demonstrated that the network dynamics visualized by the cANN can capture the common circuit computation underlying two major language functions. Considering next-word prediction as a language case of cerebellar prediction function (i.e., internal-model function of

the cerebellum) and syntactic processing as a language case of sequence-processing function, we can conclude that a single computational concept—the computation of our cANN—can unify both types of cerebellar general function, at least in language processing. The following general cerebellar computational mechanisms are indicated: (1) prediction-type (i.e., internal-model-type) circuits have inputs/outputs and error-based learning consistent with previous proposals (i.e., the circuit receives a series of events, produces predictions for one-step future events, and is trained by prediction errors) [6, 7, 8, 10, 24, 25, 28], (2) the recurrent signal is essential for long time-step dependent predictions, (3) sequence-processing-type circuits, depending on all the elements of the prediction-type circuits, spontaneously emerge, with utilizing the middle-layer processing as the output. Predictions by internal models are essential for various functions from motor control to cognition[4, 5, 10, 15, 28, 29, 30, 31], and sequence processing is also of broad importance; for example, the motor sequences for tool use in humans have a hierarchical grammar-like structure (called "action grammar")[79, 80]. Therefore, we propose that the circuit computation of our cANN is the generalized basis of the cerebellar computation that underlies a wide range of motor and cognitive functions.

Differing roles of the neocortex and cerebellum in language function

The neocortex and cerebellum are tightly interconnected and cooperate to achieve sophisticated functions such as language processing[5, 10, 14, 15, 28, 69, 81, 82, 83, 84, 85, 86], but it remains unclear how the functions of the two regions differ and how the roles are shared. When combined with clinical findings about language disorders, our cANN provides insights into the differing roles of the neocortex and cerebellum. Although the cerebellar role is limited in adults[15, 87] and so the language functions of the cerebellum are considered subordinate to those of the neocortex, in children, cerebellar language disorders cause permanent and severe deficits[17, 18, 19, 20], including agrammatism, suggesting that the development of the neocortex requires cerebellar support[15, 17, 18]. The ability of the cANN to extract syntactic information independently from the neocortex suggests that syntactic information is sent from the cerebellum during development. If the neocortex can associate S-V-O information with the words themselves to produce information such as word class, this process gradually becomes self-sustaining, consistent with the clinical observations. Segmenting sentences according to S-V-O information may allow the neocortex to process sentences more efficiently[88] (hierarchical processing, Extended Data Fig. 4b,c). Considering that the cerebellum supports the development and maturation of other cognitive functions of the neocortex[10, 15, 89, 90], the developmental importance of sequence processing by the cerebellum may generalize to other cognitive functions in addition to language processing.

Human-characteristic Cognition Revealed by Biologically Constrained ANN Models

ANN models can be categorized according to whether they comply with biological constraints, and if the functions they realize are human-characteristic cognitive functions, for which neuronal data are very limited (e.g., language processing, logical thinking, or social processing). Until

now, no brain circuit model has been developed to fulfill both criteria: adhering to biological constraints and performing sophisticated human-characteristic cognitive functions. However, many models exist which fulfill one or the other demand.

For non-human-characteristic functions (including less complex cognitive functions), numerous ANN models have been proposed, with some adhering to biological constraints and others not. Biologically constrained ANNs have been developed for regions like the cerebellum and basal ganglia, realizing various basic sensorimotor functions including decision-making[62, 63, 64, 65, 66, 67, 91, 92, 93]. ANNs that do not account for essential biological constraints, such as circuit connectivity between cell types, have been created for the cerebral neocortex using simple recurrent neural networks (e.g., three-layer recurrent neural network)[94, 95, 96, 97] and for the visual cortex utilizing deep-learning circuits[98, 99].

For higher-level cognitive processing such as language, far fewer ANNs have been proposed, and none of them are biologically constrained. Namely, models of neocortical language processing are based on language-processing AI circuits, which deviate from actual brain anatomy and physiology (e.g., transformer, LSTM)[1, 2, 3, 100]. In contrast, the cANN is designed based on the cerebellar local circuit and successfully realizes language functions of the brain area. This is the first proposal of a biologically constrained brain model capable of human-characteristic cognitive functions.

Future Directions

We anticipate the creation of biologically constrained ANNs to investigate cognitive functions that are characteristic to humans and difficult to study in animal experiments will be a prominent trend in neuroscience. This study using cANN is a pioneer in such research.

Biologically constrained ANNs also have potential for clinical applications. The following insights gleaned from the findings of the cANN serve as an example. According to previous studies, when two functions share a circuit, training one function can improve the other, even in language functions[79, 80]. Our findings predict that training in predicting the next word will improve the ability to process sentences with complex syntax, leading to the development of training in language processing and rehabilitation of language dysfunction.

Much remains to be elucidated about the grammatical processing performed by the cerebellum. Exploring the cANN's capacity beyond S-V-O syntactic processing, including its potential to word class recognition (suggested by Extended Data Figure 2a) will offer valuable insight into cerebellar grammar processing.

Interestingly, the convergent cANN, which incorporates the further constraint of convergence in Purkinje-output connection into the original cANN, differs substantially from typical circuit design of language-processing AI. Typical AI designs use compressed word representations containing semantic information for input and the sparse one-hot word representation for output. This input format aims to boost output accuracy by providing additional information at the input stage, while the output format is designed to facilitate listing multiple next-word candidates and accommodate sentences with various grammatically correct

branches. Conversely, the convergent cANN employs the sparse coding for input and a compressed word representation for output. This design can be interpreted as aiming to learn information-rich output from information-poor input, proposing a novel brain-inspired AI circuit. Future research is anticipated to uncover the potential of this circuit design.

METHODS

Simulation environment

Code was written in Python and executed with Google Colab (Pro). The source code is available on github (https://github.com/cANN-NLP/NLP_codes).

Circuit design of cANN

The cANN was created with Google's deep learning libraries, TensorFlow and Keras. To replicate the local circuitry of the cerebellum, the cANN comprised three layers—an input layer (granule cells), a middle layer (Purkinje cells), and an output layer (cerebellar nuclei neurons)—with a recurrent pathway from the output layer to the input layer (Fig. 1b; Apps and Garwicz, 2005; Houck and Person, 2014; Ankri et al., 2015; Houck and Person, 2015; Gao et al., 2016; Raymond and Medina, 2018; Ohmae et al., 2021). In addition, the cANN implemented the climbing fiber pathway which is known to deliver prediction errors to the cerebellum in both motor and cognitive functions (gray, Fig. 1b) (Ito 2008, *Nat Rev*; Sokolov et al. 2017, *Trends Cogn*; Moberget et al. 2014, *JNS*). According to cerebellar learning theory that the cerebellum learns to align its output (prediction) very faithfully with the correct answer signal (Doya, 2000, *Curr Opin*; Ito 2008, *Nat Rev*), and that the representation of cerebellar output and correct answer signal have the same dimensions (Kawato 1999, *Curr Opin*; Kawato et al. 2011, *Curr Opin*), the output layer and the correct answer signal were set to have the same dimensionality. Then, the prediction error is calculated by subtracting the prediction from the correct answer signal.

The next-word prediction circuit was designed according to previous proposals, where words are fed sequentially into the circuit then, at each step, the circuit outputs a prediction of the next word, and information about the correct word information is given used as a teaching signal to improve future predictions. In addition to the circuit connectivity, this biological constraint on the inputs given to the circuit is also critical for simulating the acquisition of function, because, in general, the difficulty of learning is determined by the information supplied in the input signals, including error signals.

Integration of information between the feedforward and recurrent pathways occur as follows: When a sentence begins, the Purkinje cells receive information about the first word from the feedforward input cells and send their activity to the recurrent pathway (and to the feedforward output pathway). Because the Purkinje cell signal is persistent, the recurrent-input cell signal also persists until the feedforward-input cells receive the second word. The Purkinje cells then integrate both signals (i.e., the second word from the feedforward-input cells and the first-word information from the recurrent input cells) and transmit their new signal to the recurrent pathway. In this way, the Purkinje cells sequentially integrate a newly entered word with the previous words. We assumed that the feedforward and recurrent input cells are distinct cell populations because the cerebellum has an enormous number of granule cells in the input layer and individual cells receive only a small number of inputs (on average, only four).

To facilitate the investigation of information processing in this network, we assumed that word representation is coded sparsely in the input layer and the correct answer signal (each of 3000 words is represented by one of 3000 cells; e.g., one cell encodes 'I' and another cell codes 'the'). Then, the immediately preceding word, which is the input to the circuit, and the correct next word (for prediction error) are represented by an activity state in which only one cell is 1 and the rest are 0 (known as a one-hot representation). Theoretically, this representation is known to contain no information associated with the word, such as the syntax or meaning of the word (confirmed in Fig. 3d; cosine similarity is 0 for all word pairs); this allowed us to study the learning and information processing of cANNs in an environment where there is no information supplied externally other than word information.

The 3000 words represented by the 3000 cells were the 3000 most-frequent words in the training sentences. All the other words that appeared less frequently were grouped together as "unknown words" and assigned to the 3001st cell. For the prediction of the next word, the word encoded by the most active output cell denoted the first prediction candidate. Note that in the analysis, even if a prediction of "unknown" was correct, it was not considered a correct prediction of the next specific word so was excluded from the calculation of the correct prediction rate. The correct prediction rate is the percentage of matches between the top five prediction candidates and the correct next word (i.e., this selection of the five candidates was done for analysis purpose only).

Circuit connections denoted by arrows in Figure 1b implement F(Wx+b). For connections from N presynaptic cells to M postsynaptic cells, x is an (N x 1) vector representing the activity of the presynaptic neurons, W is an (M x N) matrix representing the synaptic weights from the presynaptic neurons to the postsynaptic neurons, b is an (M x 1) vector representing the spontaneous activity of the postsynaptic neurons, and F is the function that converts the neuron's input to its output (firing frequency) using the leaky Rectified Linear Unit (leaky ReLU; $y = x$, if $x >= 0$, and $y = \alpha * x$, if $x < 0$). For F from the input cells to the Purkinje cells and the output cells to the input cells, $\alpha = 0.14$, and for F from the Purkinje cells to the output cells, $\alpha = 1.0$. If alpha is 0, the input/output of the neuron is highly nonlinear; if alpha is 1, the input/output is perfectly linear ($y = x$). Although alpha could be modified quite flexibly to train the circuit, the cerebellum was assumed to be quite linear with Purkinje cells and output cells having high spontaneous firing rates (i.e., linear computing; Raymond and Medina, 2018). Since no functionally significant learning has been reported at this time for synaptic weights from the Purkinje cells to the recurrent output cells and from the recurrent output cells to the recurrent input cells, synaptic weights were assumed to be for relaying information (i.e., W = the identity matrix).

The prediction output was compared with the desired output of the true next word (i.e., the ground truth) to calculate the prediction error (E), using the following formula:

$$Prediction\ error\ = cross\ entropy(T, softmax(O))$$

where T is the ground truth of the next word in the one-hot representation and O is the activity of the output cells. The softmax function normalizes the outputs so that the sum of the outputs is one and converts them to probabilities (P), and the cross entropy is the information of the difference between T and P:

$$P_i = softmax(O_i) = \exp(O_i) \,/\, \sum \exp\,(O_i) \quad \text{, for i = 1,…,3001 (index of cells)}$$

$$E = cross\ entropy(T, P) = \sum T_i \log(P_i) = \log\left(P_j\right)$$

for $T_i = 1$ (i = j) and $T_i = 0$ (i ≠ j), where j is the index corresponding to the next word. The prediction error was used to update synaptic weights on the input (i.e. dendritic) side of the Purkinje cells and the output cells (Fig. 1b, gray; Aizenman et al. 1998; Sokolov et al 2017). Encouraged by recent discoveries and proposals suggesting stochastic gradient descent of synaptic updates (i.e., in the direction to reduce prediction error) in the cerebellum (Bouvier et al. 2018; Shadmehr 2020), we assumed a gradient-descending update:

$$W_{ij} = W_{ij} - \varepsilon \frac{\partial E}{\partial W_{ij}}$$

where W represents the synaptic weights and ε is the learning rate.

To make the recurrent pathway more powerful, the number of recurrent input cells could be increased, but increasing the number to more than 192 did not improve the accuracy of word prediction or classification of the syntactic information. This suggests that more sentences are needed to train a greater number of synapses.

Circuit design of convergent cANN

Following the anatomical convergence of Purkinje-output connection, we designed the convergent cANN, in which we set the number of the feed-forward output cells to 16. Based on the cerebellar learning theory mentioned earlier that the representation of cerebellar output and correct answer have the same dimensions, we assumed that the cANN received 16-dimensional correct answer signal. To represent the correct answer of 3000 words in 16 dimensions, we employed a compressed word representation, akin to that in the neocortex (Huth et al. 2016, *Nature*). Concretely, we adopted the 100-dimension word representation, made by Stanford University group's GloVe algorithm and Wikipedia text dataset (glove-wiki-gigaword-100), and reduced the 100 dimensions to 16 using PCA. Consequently, each word was represented as a unique 16-cell activity pattern (a point in 16-dimensional space, fixed throughout the training). The candidate predicted by the 16 output cells was defined as the word whose activity pattern is closest to that of the output cells (≈, in Fig 5a). Since the 16 output cells of one module can output only one word as a prediction, 10 modules were arranged in parallel with a shared input cell layer to replicate the modular structure of the cerebellum. For learning, based on the cerebellar learning theory of how multiple modules cooperate to achieve a complex function (Jacobs et al. 1991; Wolpert & Kawato, 1998; Wolpert et al. 1998, *Trends Cogn Sci*; Haruno et al., 2001; Kawato, Ohmae, et al. 2021), we set only the module with the closest prediction to the correct answer receives the prediction error signal (gray solid line, Fig 5a). Although only one module learns after each prediction, with training on a large number of predictions, all modules should have the chance to make the best prediction and receive the prediction error (due to this,

each convergent cANN module had only 1/10 of learning opportunities compared with the original non-convergent cANN, which may had resulted in slightly lower performance of the convergent cANN).

In the analysis, we examined the five candidate words of the convergent cANN to compare with the original cANN. To obtain the five candidates, we opted for training the 10-module cANN and selecting the five best-performing modules, as this design resulted in higher prediction accuracy than training the 5-module cANN and using all five modules. This difference in performance is attributed to the numerous grammatically correct branches in sentences (e.g., various adverbs and verbs can follow a subject) that five modules were inadequate to encode. The top five modules were selected based on prediction accuracy during training (in the last 1/8 lap). The differences in prediction accuracy between modules were highly consistent between training data and validation data.

Circuit training

Twenty cANNs were trained independently. To train the cANNs, we used 88711 sentences from the Gutenberg Corpus (novels such as *Moby-Dick*, *Alice in Wonderland*, etc.) and the Brown Corpus of the Natural Language Toolkit (NLTK). Sentences that were too short (fewer than three words) or contained too many unknowns (four or more) were excluded. Next-word prediction was terminated at the 16th word if the sentence was too long. 90% (79444) of the sentences were used for training and the remaining 10% were used to evaluate the prediction performance during training. Training was terminated when the prediction performance plateaued (at the end of the eighth lap of the 79444 sentences).

Data analysis

For the 20 cANNs, the median was used as the representative statistical feature and the interquartile range (25th and 75th percentiles) was used for the error bars, in all cases except where noted. To analyze the performance of the cANNs at next-word prediction and syntactic processing after training, we used relatively simple sentences from online sites (https://sentencedict.com/; https://lengusa.com/; https://www.rong-chang.com/nse/) because the novels in the NLTK corpus contain many sentences that are not syntactically common (e.g., conversational calls). We selected 190 sentences containing verbs that the cANNs could handle (i.e., verbs included in the 3000 encoded words). We added the sentence "The man will sail the boat," which was used by Lesage and colleagues to examine language processing in the cerebellum (Lesage et al. 2012), and the similar sentences "The man will read the book" and "The man will open the gate." All test sentences are available in a text file, named as test_sentences.txt, on github (https://github.com/cANN-NLP/NLP_codes). In the training sentences, “read” and “open” appeared more frequently than “sail” (82 occurrences of “sail”, 214 of “read”, and 317 of “open”).

The correct rate for next-word prediction was defined as the probability that one of the first to fifth candidate words would match the correct next word (ground truth). We selected up

to the fifth candidate because that is the standard for the Google deep learning libraries used in this study (keras.metrics.TopKCategoricalAccuracy) and is also the standard in recent AI language processing. To examine the contribution of the recurrent signal, the recurrent signal was blocked by setting the synaptic weight from the recurrent-input cells to Purkinje cells to 0.

To analyze the dynamics of activity of the 192 Purkinje cells, 1328 words in the 190 sentences were fed sequentially into the cANN, and we measured the corresponding 1328 activity states in 192 dimensions of the Purkinje cells. To visualize the transition of Purkinje cell activity in a sentence, the 192 dimensions were compressed into two or three dimensions that maximally preserved the variance of the 1328 points, using principal component analysis (PCA; Python's online source-code library, scikit-learn). To evaluate the accuracy of classification of S-V-O syntactic information in the neuronal activity, we measured the percentage separation of S vs. V/O, V vs. S/O, and O vs. S/V with a nonlinear support vector machine (radial basis function kernel). For this, 90% of the 190 evaluation sentences were used to determine the support vector and the remaining 10% were used to obtain the classification accuracy (%); this process was repeated 256 times to obtain the average accuracy (%). To visualize S-V-O syntactic information in Figure 3a, we used the dimensions of the linear support vector machine, which are best suited for S-V, V-O, and O-S separations (coordinates of vectors orthogonal to the hyperplane best suited for the separations, obtained by svm.coef_ of scikit-learn). In the comparison of the optimal dimensions of the linear support vector machine with the PCA dimensions (Fig. 3c), the similarity index was calculated as the squared dot product of the unit vectors representing the dimensions (squared was used because the sum of the 192 dimensions is one).

REFERENCES

1. Schrimpf M*, et al.* The neural architecture of language: Integrative modeling converges on predictive processing. *Proc Natl Acad Sci U S A* **118**, (2021).

2. Caucheteux C, King JR. Brains and algorithms partially converge in natural language processing. *Commun Biol* **5**, 134 (2022).

3. Caucheteux C, Gramfort A, King JR. Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nat Hum Behav* **7**, 430-441 (2023).

4. Schmahmann JD, Guell X, Stoodley CJ, Halko MA. The Theory and Neuroscience of Cerebellar Cognition. *Annual review of neuroscience* **42**, 337-364 (2019).

5. Marien P*, et al.* Consensus paper: Language and the cerebellum: an ongoing enigma. *Cerebellum (London, England)* **13**, 386-410 (2014).

6. Moberget T, Gullesen EH, Andersson S, Ivry RB, Endestad T. Generalized role for the cerebellum in encoding internal models: evidence from semantic processing. *J Neurosci* **34**, 2871-2878 (2014).

7. Lesage E, Morgan BE, Olson AC, Meyer AS, Miall RC. Cerebellar rTMS disrupts predictive language processing. *Curr Biol* **22**, R794-795 (2012).

8. Lesage E, Hansen PC, Miall RC. Right Lateral Cerebellum Represents Linguistic Predictability. *J Neurosci* **37**, 6231-6241 (2017).

9. D'Mello AM, Turkeltaub PE, Stoodley CJ. Cerebellar tDCS Modulates Neural Circuits during Semantic Prediction: A Combined tDCS-fMRI Study. *J Neurosci* **37**, 1604-1613 (2017).

10. Sokolov AA, Miall RC, Ivry RB. The Cerebellum: Adaptive Prediction for Movement and Cognition. *Trends in cognitive sciences* **21**, 313-332 (2017).

11. Leggio MG, Chiricozzi FR, Clausi S, Tedesco AM, Molinari M. The neuropsychological profile of cerebellar damage: The sequencing hypothesis. *Cortex; a journal devoted to the study of the nervous system and behavior* **47**, 137-144 (2009).

12. Argyropoulos GP. The Subcortical Foundations of Grammaticalization. In: *The Evolution of Language*) (2008).

13. Adamaszek M, Kirkby KC. Cerebellum and Grammar Processing. In: *The Linguistic Cerebellum*) (2016).

14. Wagner MJ, Luo L. Neocortex-Cerebellum Circuits for Cognitive Processing. *Trends Neurosci* **43**, 42-54 (2020).

15. Diedrichsen J, King M, Hernandez-Castillo C, Sereno M, Ivry RB. Universal Transform or Multiple Functionality? Understanding the Contribution of the Human Cerebellum across Task Domains. *Neuron* **102**, 918-928 (2019).

16. Raichle ME*, et al.* Practice-related changes in human brain functional anatomy during nonmotor learning. *Cereb Cortex* **4**, 8-26 (1994).

17. Steinlin M. The cerebellum in cognitive processes: supporting studies in children. *Cerebellum (London, England)* **6**, 237-241 (2007).

18. Steinlin M. Cerebellar disorders in childhood: cognitive problems. *Cerebellum (London, England)* **7**, 607-610 (2008).

19. Riva D, Giorgi C. The cerebellum contributes to higher functions during development: evidence from a series of children surgically treated for posterior fossa tumours. *Brain* **123 (Pt 5)**, 1051-1061 (2000).

20. Scott RB*, et al.* Lateralized cognitive deficits in children following cerebellar lesions. *Dev Med Child Neurol* **43**, 685-691 (2001).

21. Momo K, Sakai H, Sakai KL. Syntax in a native language still continues to develop in adults: honorification judgment in Japanese. *Brain Lang* **107**, 81-89 (2008).

22. Umejima K, Flynn S, Sakai KL. Enhanced activations in syntax-related regions for multilinguals while acquiring a new language. *Scientific reports* **11**, 7296 (2021).

23. Goldstein A*, et al.* Shared computational principles for language processing in humans and deep language models. *Nature neuroscience* **25**, 369-380 (2022).

24. Pickering MJ, Garrod S. An integrated theory of language production and comprehension. *Behav Brain Sci* **36**, 329-347 (2013).

25. Argyropoulos GP. The cerebellum, internal models and prediction in 'non-motor' aspects of language: A critical review. *Brain Lang* **161**, 4-17 (2016).

26. Kuperberg GR, Jaeger TF. What do we mean by prediction in language comprehension? *Lang Cogn Neurosci* **31**, 32-59 (2016).

27. Balsters JH, Whelan CD, Robertson IH, Ramnani N. Cerebellum and cognition: evidence for the encoding of higher order rules. *Cereb Cortex* **23**, 1433-1443 (2013).

28. Ito M. Control of mental activities by internal models in the cerebellum. *Nature reviews* **9**, 304-313 (2008).

29. Desmond JE, Fiez JA. Neuroimaging studies of the cerebellum: language, learning and memory. *Trends in cognitive sciences* **2**, 355-362 (1998).

30. McDougle SD*, et al.* Continuous manipulation of mental representations is compromised in cerebellar degeneration. *Brain*, (2022).

31. Wolpert DM, Miall RC, Kawato M. Internal models in the cerebellum. *Trends Cogn Sci* **2**, 338-347 (1998).

32. Tedesco AM*, et al.* Does the cerebellum contribute to human navigation by processing sequential information? *Neuropsychology* **31**, 564-574 (2017).

33. Lupo M*, et al.* Does the cerebellar sequential theory explain spoken language impairments? A literature review. *Clin Linguist Phon* **35**, 296-309 (2021).

34. Ohmae S, Ohmae K, Heiney S, Subramanian D, Medina JF. A recurrent circuit links antagonistic cerebellar modules during associative motor learning. *bioRxiv preprint*, (2021).

35. Ankri L*, et al.* A novel inhibitory nucleo-cortical circuit controls cerebellar Golgi cell activity. *eLife* **4**, (2015).

36. Apps R, Garwicz M. Anatomical and physiological foundations of cerebellar information processing. *Nature reviews* **6**, 297-311 (2005).

37. Gao Z*, et al.* Excitatory Cerebellar Nucleocortical Circuit Provides Internal Amplification during Associative Conditioning. *Neuron* **89**, 645-657 (2016).

38. Houck BD, Person AL. Cerebellar loops: a review of the nucleocortical pathway. *Cerebellum (London, England)* **13**, 378-385 (2014).

39. Houck BD, Person AL. Cerebellar Premotor Output Neurons Collateralize to Innervate the Cerebellar Cortex. *J Comp Neurol* **523**, 2254-2271 (2015).

40. Raymond JL, Medina JF. Computational Principles of Supervised Learning in the Cerebellum. *Annual review of neuroscience* **41**, 233-253 (2018).

41. Xiao N*, et al.* Positive feedback of efferent copy via pontine nucleus facilitates cerebellum-mediated associative learning. *Cell Rep* **42**, 112072 (2023).

42. Ohmae S, Medina JF. Climbing fibers encode a temporal-difference prediction error during cerebellar learning in mice. *Nature neuroscience* **18**, 1798-1803 (2015).

43. Kitazawa S, Kimura T, Yin PB. Cerebellar complex spikes encode both destinations and errors in arm movements. *Nature* **392**, 494-497 (1998).

44. Ito M. Error detection and representation in the olivo-cerebellar system. *Frontiers in neural circuits* **7**, 1 (2013).

45. Halverson HE, Khilkevich A, Mauk MD. Relating cerebellar purkinje cell activity to the timing and amplitude of conditioned eyelid responses. *J Neurosci* **35**, 7813-7832 (2015).

46. Johansson F, Jirenhed DA, Rasmussen A, Zucca R, Hesslow G. Memory trace and timing mechanism localized to cerebellar Purkinje cells. *Proc Natl Acad Sci U S A* **111**, 14930-14934 (2014).

47. Doya K. Complementary roles of basal ganglia and cerebellum in learning and motor control. *Current opinion in neurobiology* **10**, 732-739 (2000).

48. Kawato M. Internal models for motor control and trajectory planning. *Current opinion in neurobiology* **9**, 718-727 (1999).

49. Kawato M, Kuroda S, Schweighofer N. Cerebellar supervised learning revisited: biophysical modeling and degrees-of-freedom control. *Current opinion in neurobiology* **21**, 791-800 (2011).

50. Bouvier G*, et al.* Cerebellar learning using perturbations. *eLife* **7**, (2018).

51. Shadmehr R. Population coding in the cerebellum: a machine learning perspective. *J Neurophysiol* **124**, 2022-2051 (2020).

52. Elman JL. Finding Structure in Time. *Cognitive Science* **14**, 179-211 (1990).

53. Socher R, Lin CC-Y, Ng AY, Manning CD. Parsing natural scenes and natural language with recursive neural networks. In: *ICML*) (2011).

54. Chang F, Dell GS, Bock K. Becoming syntactic. *Psychological Review* **113**, 234-272 (2006).

55. Gao Z, van Beugen BJ, De Zeeuw CI. Distributed synergistic plasticity and cerebellar learning. *Nature reviews* **13**, 619-635 (2012).

56. Isope P, Barbour B. Properties of unitary granule cell-->Purkinje cell synapses in adult rat cerebellar slices. *J Neurosci* **22**, 9668-9678 (2002).

57. Brunel N, Hakim V, Isope P, Nadal JP, Barbour B. Optimal information storage and the distribution of synaptic weights: perceptron versus Purkinje cell. *Neuron* **43**, 745-757 (2004).

58. Kawato M, Ohmae S, Hoang H, Sanger T. 50 Years Since the Marr, Ito, and Albus Models of the Cerebellum. *Neuroscience* **462**, 151-174 (2021).

59. Haruno M, Wolpert DM, Kawato M. Mosaic model for sensorimotor learning and control. *Neural Comput* **13**, 2201-2220 (2001).

60. Huth AG, de Heer WA, Griffiths TL, Theunissen FE, Gallant JL. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* **532**, 453-458 (2016).

61. Khilkevich A, Canton-Josh J, DeLord E, Mauk MD. A cerebellar adaptation to uncertain inputs. *Sci Adv* **4**, eaap9660 (2018).

62. Albus JS. A theory of cerebellar function. *Mathematical Biosciences* **10**, 25-61 (1971).

63. Marr D. A theory of cerebellar cortex. *J Physiol* **202**, 437-470 (1969).

64. Buonomano DV, Mauk MD. Neural Network Model of the Cerebellum: Temporal Discrimination and the Timing of Motor Responses. *Neural Computation* **6**, 38-55 (1994).

65. Medina JF, Mauk MD. Computer simulation of cerebellar information processing. *Nature neuroscience* **3 Suppl**, 1205-1211 (2000).

66. Yamazaki T, Tanaka S. The cerebellum as a liquid state machine. *Neural Netw* **20**, 290-297 (2007).

67. Narain D, Remington ED, Zeeuw CI, Jazayeri M. A cerebellar mechanism for learning prior distributions of time intervals. *Nat Commun* **9**, 469 (2018).

68. Doya K. What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? *Neural Netw* **12**, 961-974 (1999).

69. Porrill J, Dean P, Stone JV. Recurrent cerebellar architecture solves the motor-error problem. *Proc Biol Sci* **271**, 789-796 (2004).

70. Hull C. Prediction signals in the cerebellum: beyond supervised motor learning. *eLife* **9**, (2020).

71. Boyden ES, Katoh A, Raymond JL. Cerebellum-dependent learning: the role of multiple plasticity mechanisms. *Annual review of neuroscience* **27**, 581-609 (2004).

72. Carey M, Lisberger S. Embarrassed, but not depressed: eye opening lessons for cerebellar learning. *Neuron* **35**, 223-226 (2002).

73. Medina JF. The multiple roles of Purkinje cells in sensori-motor calibration: to predict, teach and command. *Current opinion in neurobiology* **21**, 616-622 (2011).

74. Aizenman CD, Manis PB, Linden DJ. Polarity of long-term synaptic gain change is related to postsynaptic spike firing at a cerebellar inhibitory synapse. *Neuron* **21**, 827-835 (1998).

75. Kitazawa S, Wolpert DM. Rhythmicity, randomness and synchrony in climbing fiber signals. *Trends Neurosci* **28**, 611-619 (2005).

76. Apps R, Hawkes R. Cerebellar cortical organization: a one-map hypothesis. *Nature reviews* **10**, 670-681 (2009).

77. Ramnani N. The primate cortico-cerebellar system: anatomy and function. *Nature reviews* **7**, 511-522 (2006).

78. De Zeeuw CI. Bidirectional learning in upbound and downbound microzones of the cerebellum. *Nat Rev Neurosci* **22**, 92-110 (2021).

79. Thibault S*, et al.* Tool use and language share syntactic processes and neural patterns in the basal ganglia. *Science (New York, NY* **374**, eabe0874 (2021).

80. Higuchi S, Chaminade T, Imamizu H, Kawato M. Shared neural correlates for language and tool use in Broca's area. *Neuroreport* **20**, 1376-1381 (2009).

81. Li N, Mrsic-Flogel TD. Cortico-cerebellar interactions during goal-directed behavior. *Current opinion in neurobiology* **65**, 27-37 (2020).

82. Ohmae S, Uematsu A, Tanaka M. Temporally specific sensory signals for the detection of stimulus omission in the primate deep cerebellar nuclei. *J Neurosci* **33**, 15432-15441 (2013).

83. Bostan AC, Dum RP, Strick PL. Cerebellar networks with the cerebral cortex and basal ganglia. *Trends in cognitive sciences* **17**, 241-254 (2013).

84. Kinno R, Ohta S, Muragaki Y, Maruyama T, Sakai KL. Differential reorganization of three syntax-related networks induced by a left frontal glioma. *Brain* **137**, 1193-1212 (2014).

85. Tanaka M*, et al.* Roles of the Cerebellum in Motor Preparation and Prediction of Timing. *Neuroscience* **462**, 220-234 (2021).

86. Proville RD*, et al.* Cerebellum involvement in cortical sensorimotor circuits for the control of voluntary movements. *Nature neuroscience* **17**, 1233-1239 (2014).

87. Fiez JA, Raichle ME. Linguistic processing. *Int Rev Neurobiol* **41**, 233-254 (1997).

88. Ding N, Melloni L, Zhang H, Tian X, Poeppel D. Cortical tracking of hierarchical linguistic structures in connected speech. *Nature neuroscience* **19**, 158-164 (2016).

89. Wang SS, Kloth AD, Badura A. The cerebellum, sensitive periods, and autism. *Neuron* **83**, 518-532 (2014).

90. Badura A*, et al.* Normal cognitive and social development require posterior cerebellar activity. *eLife* **7**, (2018).

91. Lo CC, Wang XJ. Cortico-basal ganglia circuit mechanism for a decision threshold in reaction time tasks. *Nature neuroscience* **9**, 956-963 (2006).

92. O'Reilly RC, Frank MJ. Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. *Neural Comput* **18**, 283-328 (2006).

93. Chen G, Scherr F, Maass W. A data-based large-scale model for primary visual cortex enables brain-like robust and versatile visual processing. *Sci Adv* **8**, eabq7592 (2022).

94. Mante V, Sussillo D, Shenoy KV, Newsome WT. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* **503**, 78-84 (2013).

95. Chen G, Kang B, Lindsey J, Druckmann S, Li N. Modularity and robustness of frontal cortical networks. *Cell* **184**, 3717-3730 e3724 (2021).

96. Sussillo D, Abbott LF. Generating coherent patterns of activity from chaotic neural networks. *Neuron* **63**, 544-557 (2009).

97. Wang XJ. Decision making in recurrent neuronal circuits. *Neuron* **60**, 215-234 (2008).

98. Yamins DL, DiCarlo JJ. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience* **19**, 356-365 (2016).

99. Yamins DL, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc Natl Acad Sci U S A* **111**, 8619-8624 (2014).

100. Costa R, Assael IA, Shillingford B, de Freitas N, Vogels T. Cortical microcircuits as gated-recurrent neural networks. *Advances in neural information processing systems* **30**, (2017).