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, and Dr. Nuo Li (listed in the chronological order that the comments were provided) for invaluable insights and comments that improved the manuscript. We also thank 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 cooperates with the neocortex in language processing, particularly during language acquisition, but little is known about the circuit computations underlying cerebellar language functions. Here, to simulate language acquisition, we created a biologically constrained cerebellar artificial neural network (cANN) model. We found that as the cANN acquired prediction of future words, a second function—syntactic recognition, the extraction of rules from word sequences—emerged spontaneously in the middle layer of the circuit. This single-circuit computation as the common basis for two language functions can be generalized to prediction and rule-extraction functions underlying a wide range of cerebellar cognitive functions, and also suggests a concrete novel rehabilitation method for patients with language dysfunction.

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. Language comprehension after word recognition requires cooperation between the left neocortical language area and the right lateral cerebellum (right Crus I/II)[1-19]. The cerebellum has a large capacity for plasticity and acquires language functions relatively early in development compared with the rest of the brain[8, 20, 21]. A theory that cerebellar language acquisition supports subsequent language acquisition in the neocortex[8, 20, 21] has been proposed based on recent observations that childhood damage to the right cerebellum causes more permanent and severe language deficits than adult damage[20-23], and that the right cerebellum is activated during learning of new language tasks[15, 24, 25]. Understanding the circuit computations of language functions in the cerebellum is crucial to unravelling the mechanisms of human language processing. In turn, this leads to elucidation of both the origin of language functions in the neocortex and the replacement of neocortical functions by the cerebellum during neocortical dysfunction, such as in Alzheimer's disease[26]. Artificial neural networks (ANNs) are an invaluable tool for examining how the brain processes language because there are no established animal models of language and only limited data from invasive recordings in humans. However, few ANNs for language functions follow the biological constraints of the brain, and those few align only partially with the known circuit connectivity[27, 28]. 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 according to the grammar of the language to compose the meaning of the sentence. During this process, two major language functions have been linked to the right lateral cerebellum. The first of these is next-word prediction[2, 29-33], which supports faster and more accurate sentence comprehension by placing input words in a predictive context (particularly in noisy environments)[29, 30, 33-38]. Although neural activity related to this function has also been reported in neocortical language areas[34, 35], stronger, causal evidence has accumulated for the cerebellum[29-32]. The second function is grammatical processing, especially syntactic recognition concerning subject–verb–object (S-V-O) information[3, 39-42]. These two language functions reflect two general cerebellar functions that support a variety of cognitive functions. One is prediction of external events[2, 3, 6, 8, 11, 33, 43, 44], and next-word prediction is a language case of this general function[8, 29-31, 33, 37]. The second is rule extraction from a sequence of events[39, 42, 45], and syntactic recognition is a language case of this function[39, 42]. Although these two general functions underlie various cognitive functions associated with the cerebellum[2, 3, 6, 8, 9, 11, 33, 43, 44, 46, 47] (see Discussion), it remains unclear how they are realized within the uniform cytoarchitecture of the cerebellum. To elucidate how these two general functions are implemented within a single circuit architecture, it is necessary to understand the circuit computation of each function. ANNs are ideal for this because they allow the visualization of

computations as network dynamics. Therefore, we used our cANN to test whether next-word prediction and syntactic recognition—language-specific counterparts of the two general functions—can be implemented within a single circuit architecture.

We found that, through training, the cANN acquired next-word prediction and it also, simultaneously and spontaneously, acquired syntactic recognition in the middle layer of the same circuit. This suggests that two general functions of the cerebellum, prediction and rule extraction from a sequence, can be captured by a single circuit computation of the cerebellum. This also indicates a concrete new rehabilitation method for language dysfunction: training in word prediction may lead to improvements 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 consists of a conventionally reported feedforward pathway and a recently identified recurrent pathway that starts from the Purkinje cells (Fig. 1b)[48-54]. 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, 29-31, 33, 36, 37]. We assumed that words were represented by sparse coding in the input and output layers: 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). In the output layer, the five words corresponding to the five cells with the strongest activity were denoted, in order, as the prediction candidates for the next word. 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)[33, 55, 56].

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 five words that appeared most frequently in the whole training sentences, regardless of the input, i.e., the identity of the immediately preceding word (Supplementary Fig. 1a,b). Then, the correct rate slowly increased to about 40%, where it plateaued; here, the cANN achieved correct prediction rates significantly higher than chance regardless of the frequency of occurrence of the target word (even in predicting rare words; Supplementary 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.

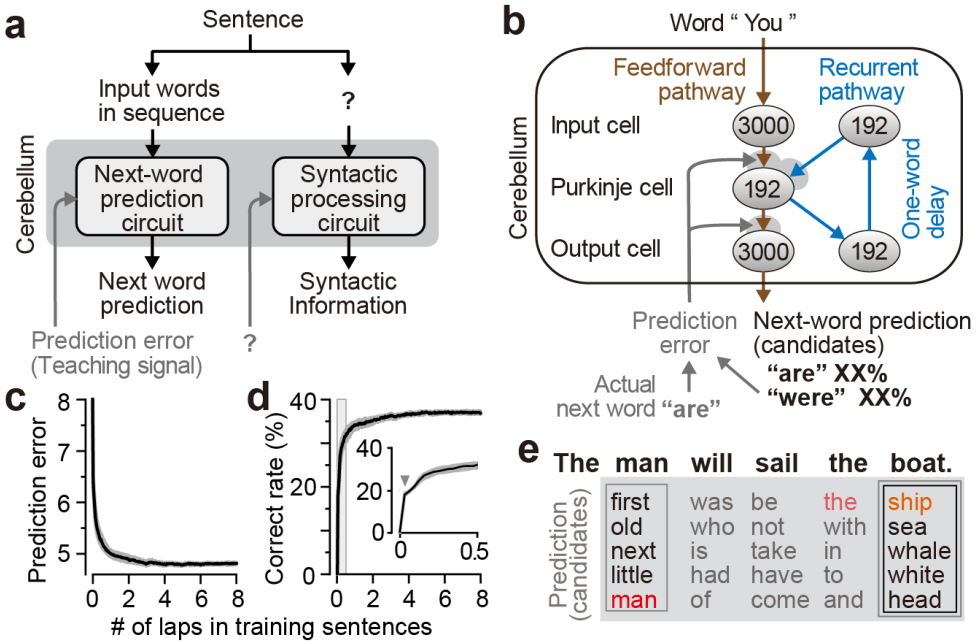

Figure 1. Local circuit model of a cANN to reconstruct cerebellar language functions. **a**, Previously proposed block diagram with Inputs/outputs for next-word prediction. **b**, The three-layer cerebellar circuit, consisting of a feedforward pathway (brown) and a recurrent pathway (blue). The prediction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

Next, to replicate the finding from human cerebellar research[30, 31], 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). 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.

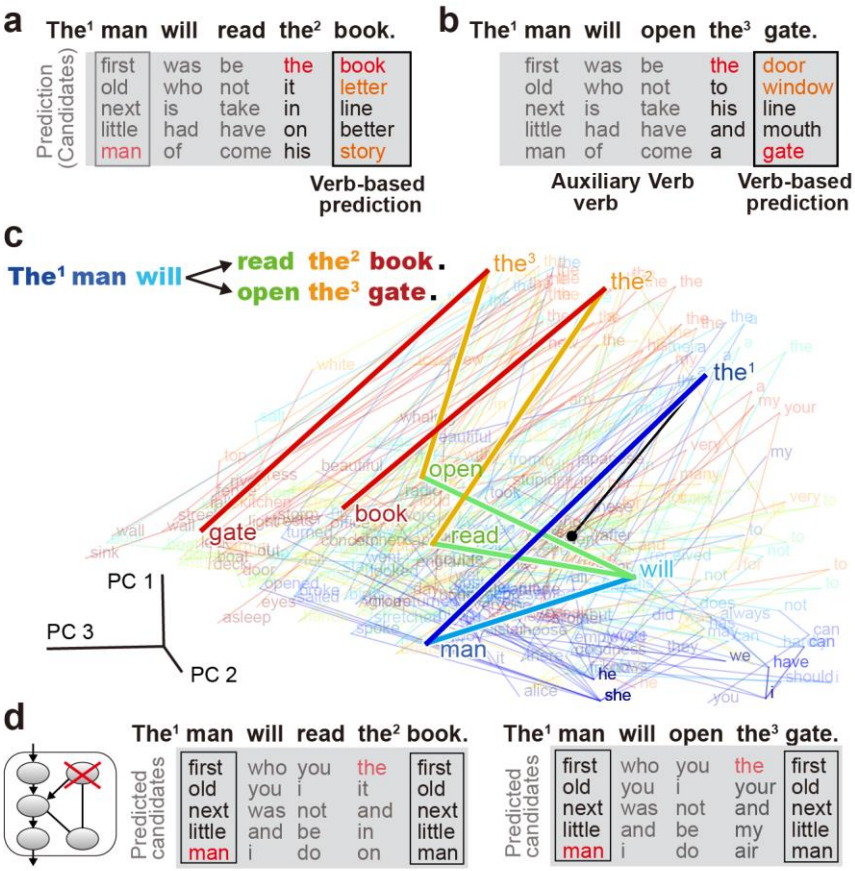

Figure 2. Next-word prediction and network dynamics of a representative cANN. **a,b**, Predictions after three instances of “the” (indexed) in two sentences. **c**, Visualization of the neural activity of Purkinje cells after input of each word by PCA. Words are color-coded in rainbow from the beginning 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 investigate the mechanism of information processing, we examined the role of the Purkinje cells, where the feedforward signal and the recurrent signal delayed by one word are integrated (Fig. 1b). When we visualized the word-by-word transition of Purkinje cells activity in the sentence 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” (Supplementary 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[8] (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[57-59], 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 (Supplementary 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 (Supplementary Fig. 2a) and by linear support vector machine (Fig. 3a) and found that V words were clustered and well separated from S-O words (red circles), regardless of the number of words forming the subject (black trace). The separation of V words from S-O words (96.0%; IQR, 95.7–96.5; nonlinear support vector machine), S words (96.1%; 95.7–96.6), and O words (95.8%; 95.4–96.1) was highly accurate (visualized in Supplementary Fig. 2b). Since syntactic information is not included in the word-by-word teaching signal (see METHODS), 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. 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[8].

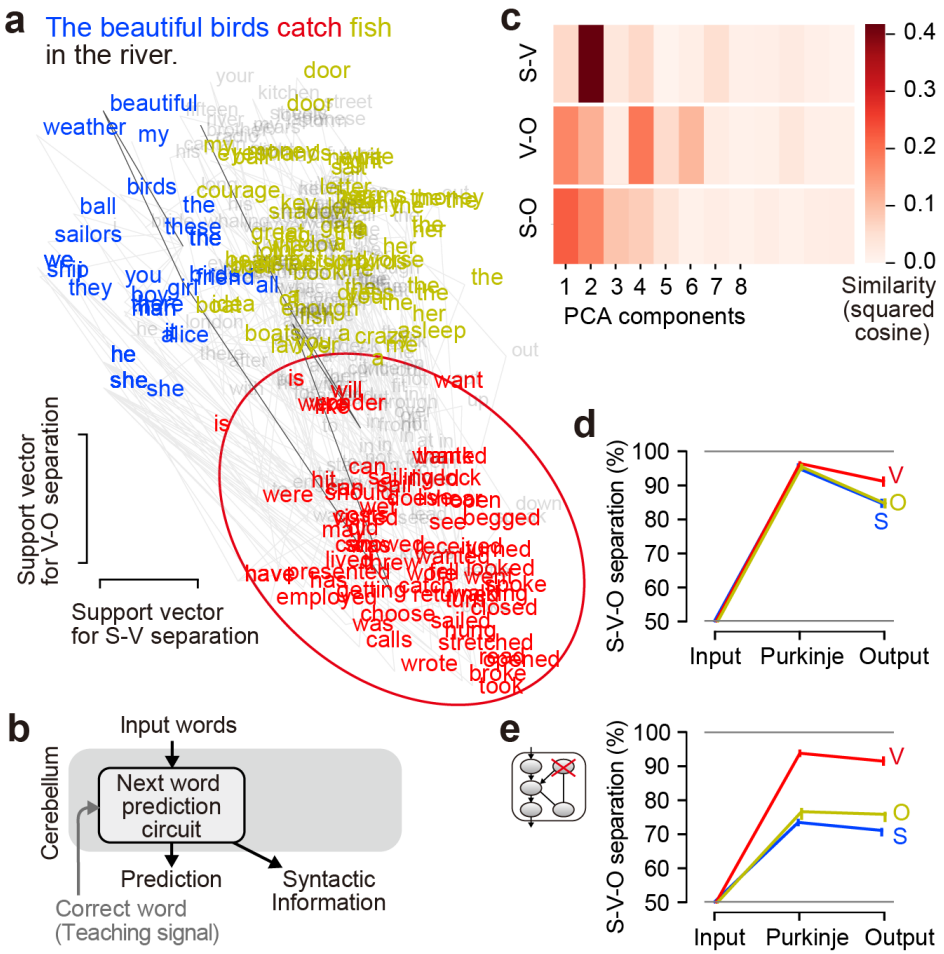

Figure 3. Syntactic processing by the cANN. **a**, Visualization of neural activity of Purkinje cells after inputting a subject (blue), verb (red), and object (yellow), in the dimensions used for classification. The activity transition for a sentence with a longer subject (three words) is highlighted (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, and feedforward-output cells. 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 Supplementary 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 (Supplementary Fig. 2d,e). This indicates that S-V-O syntactic information is the primary information represented by the activity of Purkinje cells.

DISCUSSION

Anatomical and physiological support for the cANN

Our cANN is highly consistent with a wealth of anatomical and physiological knowledge. While conventional local circuit models of the cerebellum are feedforward circuits[60-65], recent reports have demonstrated the existence and functional significance of a recurrent connection starting from the Purkinje cells[48-54, 66-68]. Since both next-word prediction and syntactic processing are inaccurate in a cANN without the recurrent pathway (data not shown) and conventional feedforward models (Supplementary Fig. 3), the implementation of the recurrent pathway in our cANN is not only consistent with anatomical findings but is also essential for the unified realization of the two language functions. Furthermore, the cANN proposes candidate output pathways for syntax information that can be tested in future circuit studies (Supplementary Fig. 4a). With respect to learning, a series of studies indicate that the cerebellum can learn more flexibly than classical cerebellar learning rules[54, 69-77]. 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[33, 55, 56]. Together, our cANN is multilaterally consistent with the knowledge about the cerebellum.

Significance of a common computational basis for the two functions

Because the cerebellum has a uniform cytoarchitecture[2, 8, 50, 54, 78-80], much conventional research has been directed toward elucidating the common circuit computation underlying all cerebellar functions. Although many descriptive ideas have been proposed, none has yet successfully explained all the diverse functions of the cerebellum[8]. 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 substituting sentences as the prediction target, in place of the external world, and viewing the cerebellum as a sentence simulator[8, 29-31, 33, 37]. However, explaining syntactic processing with the predictive functions of the internal-model theory is difficult and requires other ideas, such as sequence processing[39, 42, 45].

Recognizing the limited reach of descriptive ideas to encapsulate the common circuit computation, Ivry and colleagues emphasized the need for a more abstract concept, like the network dynamics of neurons, to explain cerebellar computation[8]. 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 an internal-model type of language function and syntactic processing as a sequence-processing type of language function, we can conclude that a single computational concept—the computation of our cANN—can unify both types of function in language processing. The following general cerebellar computational mechanisms are indicated: (1) prediction-type (i.e., internal-model-type) circuits have inputs/outputs and a training style consistent with previous proposals[6, 29-31, 33, 36, 37], (2) a recurrent signal is essential for predictions based on past inputs, (3) sequence-processing-type circuits rely on all the elements of the prediction-type circuits, particularly the middle layer. Predictions by internal models are essential for various functions from motor control to cognition[2, 3, 6, 8, 11, 33, 43, 44], and sequence processing is also of broad importance; for example, the motor sequences for tool use in humans have a hierarchical syntax-like structure[46, 47]. Therefore, we propose that the circuit computation of our cANN forms the generalized principle of cerebellar computation that underlies a wide range of motor and cognitive functions.

Differing roles of the neocortex and cerebellum in language function

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[81] 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[20-23], including agrammatism, suggesting that the development of the neocortex requires cerebellar support[8, 20, 21]. 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[82] (hierarchical processing, Supplementary Fig. 4b,c). Considering that the cerebellum supports the development and maturation of other cognitive functions of the neocortex[9, 33, 83], the developmental importance of sequence processing by the cerebellum may generalize to other cognitive functions in addition to language processing.

Future directions

When two functions share a circuit, training one function can improve the other, even in language functions[46, 47]. With that knowledge, 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.

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

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). 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 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 of the delay in the recurrent pathway, the recurrent input cells receive this information one word late, at which time 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 integrates 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 and output layers (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 an 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. Note that 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. For the prediction of the next word, the word encoded by the most active output cell denoted the first prediction candidate, and so on in order up to the fifth prediction candidate.

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 = α * x, if x < 0). For F from the input cells to the Purkinje cells and the output cells to the input cells, α = 0.14, and for F from the Purkinje cells to the output cells, α = 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:

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

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

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 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." 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. Tsai, P.T.*, et al.* Autistic-like behaviour and cerebellar dysfunction in Purkinje cell Tsc1 mutant mice. *Nature* **488**, 647-651 (2012).
2. Schmahmann, J.D., Guell, X., Stoodley, C.J. & Halko, M.A. The Theory and Neuroscience of Cerebellar Cognition. *Annu. Rev. Neurosci.* **42**, 337-364 (2019).
3. Marien, P.*, et al.* Consensus paper: Language and the cerebellum: an ongoing enigma. *Cerebellum (London, England)* **13**, 386-410 (2014).
4. 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).
5. Tanaka, M.*, et al.* Roles of the Cerebellum in Motor Preparation and Prediction of Timing. *Neuroscience* **462**, 220-234 (2021).
6. Ito, M. Control of mental activities by internal models in the cerebellum. *Nat. Rev. Neurosci.* **9**, 304-313 (2008).
7. Bostan, A.C., Dum, R.P. & Strick, P.L. Cerebellar networks with the cerebral cortex and basal ganglia. *Trends in cognitive sciences* **17**, 241-254 (2013).
8. Diedrichsen, J., King, M., Hernandez-Castillo, C., Sereno, M. & Ivry, R.B. Universal Transform or Multiple Functionality? Understanding the Contribution of the Human Cerebellum across Task Domains. *Neuron* **102**, 918-928 (2019).
9. Wang, S.S., Kloth, A.D. & Badura, A. The cerebellum, sensitive periods, and autism. *Neuron* **83**, 518-532 (2014).
10. Ohmae, S., Kunimatsu, J. & Tanaka, M. Cerebellar Roles in Self-Timing for Sub- and Supra-Second Intervals. *J Neurosci* **37**, 3511-3522 (2017).
11. Desmond, J.E. & Fiez, J.A. Neuroimaging studies of the cerebellum: language, learning and memory. *Trends in cognitive sciences* **2**, 355-362 (1998).
12. Gao, Z.*, et al.* A cortico-cerebellar loop for motor planning. *Nature* **563**, 113-116 (2018).
13. Proville, R.D.*, et al.* Cerebellum involvement in cortical sensorimotor circuits for the control of voluntary movements. *Nat. Neurosci.* **17**, 1233-1239 (2014).
14. Strick, P.L., Dum, R.P. & Fiez, J.A. Cerebellum and nonmotor function. *Annual review of neuroscience* **32**, 413-434 (2009).
15. Raichle, M.E.*, et al.* Practice-related changes in human brain functional anatomy during nonmotor learning. *Cereb Cortex* **4**, 8-26 (1994).
16. Li, N. & Mrsic-Flogel, T.D. Cortico-cerebellar interactions during goal-directed behavior. *Curr. Opin. Neurobiol.* **65**, 27-37 (2020).
17. 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).
18. Kinno, R., Ohta, S., Muragaki, Y., Maruyama, T. & Sakai, K.L. Differential reorganization of three syntax-related networks induced by a left frontal glioma. *Brain* **137**, 1193-1212 (2014).

19. Schmahmann, J.D. & Pandya, D.N. Disconnection syndromes of basal ganglia, thalamus, and cerebrocerebellar systems. *Cortex; a journal devoted to the study of the nervous system and behavior* **44**, 1037-1066 (2008).
20. Steinlin, M. The cerebellum in cognitive processes: supporting studies in children. *Cerebellum (London, England)* **6**, 237-241 (2007).
21. Steinlin, M. Cerebellar disorders in childhood: cognitive problems. *Cerebellum (London, England)* **7**, 607-610 (2008).
22. 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).
23. Scott, R.B.*, et al.* Lateralized cognitive deficits in children following cerebellar lesions. *Dev Med Child Neurol* **43**, 685-691 (2001).
24. Momo, K., Sakai, H. & Sakai, K.L. Syntax in a native language still continues to develop in adults: honorification judgment in Japanese. *Brain Lang* **107**, 81-89 (2008).
25. Umejima, K., Flynn, S. & Sakai, K.L. Enhanced activations in syntax-related regions for multilinguals while acquiring a new language. *Scientific reports* **11**, 7296 (2021).
26. Jacobs, H.I.L.*, et al.* The cerebellum in Alzheimer's disease: evaluating its role in cognitive decline. *Brain* **141**, 37-47 (2018).
27. Pulvermuller, F., Tomasello, R., Henningsen-Schomers, M.R. & Wennekers, T. Biological constraints on neural network models of cognitive function. *Nature reviews* **22**, 488-502 (2021).
28. Costa, R., Assael, I.A., Shillingford, B., de Freitas, N. & Vogels, T. Cortical microcircuits as gated-recurrent neural networks. *Advances in neural information processing systems* **30** (2017).
29. Moberget, T., Gullesen, E.H., Andersson, S., Ivry, R.B. & Endestad, T. Generalized role for the cerebellum in encoding internal models: evidence from semantic processing. *J Neurosci* **34**, 2871-2878 (2014).
30. Lesage, E., Hansen, P.C. & Miall, R.C. Right Lateral Cerebellum Represents Linguistic Predictability. *J Neurosci* **37**, 6231-6241 (2017).
31. Lesage, E., Morgan, B.E., Olson, A.C., Meyer, A.S. & Miall, R.C. Cerebellar rTMS disrupts predictive language processing. *Curr Biol* **22**, R794-795 (2012).
32. D'Mello, A.M., Turkeltaub, P.E. & Stoodley, C.J. Cerebellar tDCS Modulates Neural Circuits during Semantic Prediction: A Combined tDCS-fMRI Study. *J Neurosci* **37**, 1604-1613 (2017).
33. Sokolov, A.A., Miall, R.C. & Ivry, R.B. The Cerebellum: Adaptive Prediction for Movement and Cognition. *Trends Cogn. Sci.* **21**, 313-332 (2017).
34. Schrimpf, M.*, et al.* The neural architecture of language: Integrative modeling converges on predictive processing. *Proc Natl Acad Sci U S A* **118** (2021).
35. Goldstein, A.*, et al.* Shared computational principles for language processing in humans and deep language models. *Nature neuroscience* **25**, 369-380 (2022).

36. Pickering, M.J. & Garrod, S. An integrated theory of language production and comprehension. *Behav Brain Sci* **36**, 329-347 (2013).
37. Argyropoulos, G.P. The cerebellum, internal models and prediction in 'non-motor' aspects of language: A critical review. *Brain Lang* **161**, 4-17 (2016).
38. Kuperberg, G.R. & Jaeger, T.F. What do we mean by prediction in language comprehension? *Lang Cogn Neurosci* **31**, 32-59 (2016).
39. Leggio, M.G., Chiricozzi, F.R., Clausi, S., Tedesco, A.M. & 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).
40. Balsters, J.H., Whelan, C.D., Robertson, I.H. & Ramnani, N. Cerebellum and cognition: evidence for the encoding of higher order rules. *Cereb Cortex* **23**, 1433-1443 (2013).
41. Argyropoulos, G.P. The Subcortical Foundations of Grammaticalization. in *The Evolution of Language* 10-17 (2008).
42. Adamaszek, M. & Kirkby, K.C. Cerebellum and Grammar Processing. in *The Linguistic Cerebellum* 81-105 (2016).
43. McDougle, S.D.*, et al.* Continuous manipulation of mental representations is compromised in cerebellar degeneration. *Brain* (2022).
44. Wolpert, D.M., Miall, R.C. & Kawato, M. Internal models in the cerebellum. *Trends Cogn. Sci.* **2**, 338-347 (1998).
45. Tedesco, A.M.*, et al.* Does the cerebellum contribute to human navigation by processing sequential information? *Neuropsychology* **31**, 564-574 (2017).
46. Thibault, S.*, et al.* Tool use and language share syntactic processes and neural patterns in the basal ganglia. *Science (New York, N.Y* **374**, eabe0874 (2021).
47. 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).
48. Ohmae, S., Ohmae, K., Heiney, S., Subramanian, D. & Medina, J.F. A recurrent circuit links antagonistic cerebellar modules during associative motor learning. *bioRxiv preprint* (2021).
49. Ankri, L.*, et al.* A novel inhibitory nucleo-cortical circuit controls cerebellar Golgi cell activity. *Elife* **4** (2015).
50. Apps, R. & Garwicz, M. Anatomical and physiological foundations of cerebellar information processing. *Nat. Rev. Neurosci.* **6**, 297-311 (2005).
51. Gao, Z.*, et al.* Excitatory Cerebellar Nucleocortical Circuit Provides Internal Amplification during Associative Conditioning. *Neuron* **89**, 645-657 (2016).
52. Houck, B.D. & Person, A.L. Cerebellar loops: a review of the nucleocortical pathway. *Cerebellum* **13**, 378-385 (2014).
53. Houck, B.D. & Person, A.L. Cerebellar Premotor Output Neurons Collateralize to Innervate the Cerebellar Cortex. *J. Comp. Neurol.* **523**, 2254-2271 (2015).

54. Raymond, J.L. & Medina, J.F. Computational Principles of Supervised Learning in the Cerebellum. *Annu. Rev. Neurosci.* **41**, 233-253 (2018).
55. Bouvier, G.*, et al.* Cerebellar learning using perturbations. *eLife* **7** (2018).
56. Shadmehr, R. Population coding in the cerebellum: a machine learning perspective. *J Neurophysiol* **124**, 2022-2051 (2020).
57. Elman, J.L. Finding Structure in Time. *Cognitive Science* **14**, 179-211 (1990).
58. Socher, R., Lin, C.C.-Y., Ng, A.Y. & Manning, C.D. Parsing natural scenes and natural language with recursive neural networks. in *ICML* (2011).
59. Chang, F., Dell, G.S. & Bock, K. Becoming syntactic. *Psychological Review* **113**, 234-272 (2006).
60. Brunel, N., Hakim, V., Isope, P., Nadal, J.P. & Barbour, B. Optimal information storage and the distribution of synaptic weights: perceptron versus Purkinje cell. *Neuron* **43**, 745-757 (2004).
61. Medina, J.F. & Mauk, M.D. Computer simulation of cerebellar information processing. *Nature neuroscience* **3 Suppl**, 1205-1211 (2000).
62. Yamazaki, T. & Tanaka, S. The cerebellum as a liquid state machine. *Neural Netw.* **20**, 290-297 (2007).
63. Marr, D. A theory of cerebellar cortex. *J Physiol* **202**, 437-470 (1969).
64. Albus, J.S. A theory of cerebellar function. *Math. Biosci.* **10**, 25-61 (1971).
65. Buonomano, D.V. & Mauk, M.D. Neural Network Model of the Cerebellum: Temporal Discrimination and the Timing of Motor Responses. *Neural Computation* **6**, 38-55 (1994).
66. Giovannucci, A.*, et al.* Cerebellar granule cells acquire a widespread predictive feedback signal during motor learning. *Nat. Neurosci.* **20**, 727-734 (2017).
67. Lisberger, S.G. & Fuchs, A.F. Role of primate flocculus during rapid behavioral modification of vestibuloocular reflex. II. Mossy fiber firing patterns during horizontal head rotation and eye movement. *J Neurophysiol* **41**, 764-777 (1978).
68. Lisberger, S.G. & Sejnowski, T.J. Motor learning in a recurrent network model based on the vestibulo-ocular reflex. *Nature* **360**, 159-161 (1992).
69. Hull, C. Prediction signals in the cerebellum: beyond supervised motor learning. *eLife* **9** (2020).
70. Gao, Z., van Beugen, B.J. & De Zeeuw, C.I. Distributed synergistic plasticity and cerebellar learning. *Nat. Rev. Neurosci.* **13**, 619-635 (2012).
71. Ohmae, S. & Medina, J.F. Climbing fibers encode a temporal-difference prediction error during cerebellar learning in mice. *Nat. Neurosci.* **18**, 1798-1803 (2015).
72. Boyden, E.S., Katoh, A. & Raymond, J.L. Cerebellum-dependent learning: the role of multiple plasticity mechanisms. *Annual review of neuroscience* **27**, 581-609 (2004).
73. Carey, M. & Lisberger, S. Embarrassed, but not depressed: eye opening lessons for cerebellar learning. *Neuron* **35**, 223-226 (2002).
74. Medina, J.F. The multiple roles of Purkinje cells in sensori-motor calibration: to predict, teach and command. *Curr. Opin. Neurobiol.* **21**, 616-622 (2011).

75. Aizenman, C.D., Manis, P.B. & Linden, D.J. Polarity of long-term synaptic gain change is related to postsynaptic spike firing at a cerebellar inhibitory synapse. *Neuron* **21**, 827-835 (1998).
76. Kitazawa, S., Kimura, T. & Yin, P.B. Cerebellar complex spikes encode both destinations and errors in arm movements. *Nature* **392**, 494-497 (1998).
77. Kitazawa, S. & Wolpert, D.M. Rhythmicity, randomness and synchrony in climbing fiber signals. *Trends Neurosci* **28**, 611-619 (2005).
78. Apps, R. & Hawkes, R. Cerebellar cortical organization: a one-map hypothesis. *Nat. Rev. Neurosci.* **10**, 670-681 (2009).
79. Ramnani, N. The primate cortico-cerebellar system: anatomy and function. *Nature reviews* **7**, 511-522 (2006).
80. De Zeeuw, C.I. Bidirectional learning in upbound and downbound microzones of the cerebellum. *Nat. Rev. Neurosci.* **22**, 92-110 (2021).
81. Fiez, J.A. & Raichle, M.E. Linguistic processing. *Int Rev Neurobiol* **41**, 233-254 (1997).
82. 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).
83. Badura, A.*, et al.* Normal cognitive and social development require posterior cerebellar activity. *Elife* **7** (2018).