A proposal for unifying statistical learning at different scales: Long-Horizon Associative Learning.

Lucas Benjamin[1], Ana Fló[1], Fosca Al Roumi[1], Ghislaine Dehaene-Lambertz[1]

[1]Cognitive Neuroimaging Unit, CNRS ERL 9003, INSERM U992, CEA, Université Paris-Saclay, NeuroSpin center, Gif/Yvette, France.

Corresponding author: Lucas Benjamin

Acknowledgments: This work received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 695710) and GDL from a grant from the French government managed by the Agence Nationale de la Recherche as part of the France 2030 program, under reference ANR-23-IAIIU-0010.

Key-words: learning, infant, language, pattern, EEG

Abstract

Sensory inputs often display complex temporal interdependencies that typically conform to a latent underlying arrangement, whose learning facilitates the efficient interaction with the environment. However temporal dependencies may cover a large range of scales, from consecutive items ***(Saffran et al., 1996)***, to relations in longer sequences ***(Dehaene et al., 2015; Schapiro et al., 2013)***. Whereas numerous models have been proposed to explain statistical learning at different scales ***(Fiser and Lengyel, 2022)***, we argue here that a single and unifying model can account for the different scales of statistical learning including, adjacent and non-adjacent transitions as well as network structures, even in the absence of first order transition probability information. Based on behavioral and MEG results, we previously showed that a long-horizon associative learning process (a biologically plausible implementation of successor representation, or equivalently the Free-Energy Minimization Model – FEMM originally proposed by Lynn and colleagues ***(2020a)***) was able to explain learning based on local statistical properties but also based on high-order network properties ***(Benjamin et al., 2024, 2023a)***. Here, we extend our proposal and describe how this long-term associative learning model accounts for various results from the literature with different scales of dependency between elements. Long-horizon associative learning thus provides a unified framework that spans diverse statistical scales, offering a unique explanation of literature results. We thus aim at replacing the numerous metrics described in the statistical learning literature, each tuned to a specific scale and that occasionally conflict with each other by this unique parsimonious approach.

Introduction

Sensory inputs are rich, diverse, and organized, with complex temporal dependencies. Humans are relatively skilled at recovering the underlying latent arrangement of sequences over a wide range of scales: between consecutive items ***(Saffran et al., 1996)*** and at longer distance ***(Basirat et al., 2014; Bekinschtein et al., 2009; Garvert et al., 2017; Peña et al., 2002; Schapiro et al., 2013)***. For example, using streams of speech syllables, it has been shown that human participants are sensitive to local transition probabilities between consecutive syllables: they use changes in transition probabilities between syllables to detect word boundaries. This ability has been considered a mechanism for word learning in infants ***(Harris, 1955; Saffran et al., 1996)***. They also notice non-adjacent relationships such as "is"…"ing" regardless of the number of syllables between the related elements showing a larger scale sensitivity to statistical relationships. Finally, recent studies have shown that they are also sensitive to network structures and to the higher-order organization of these networks ***(Benjamin et al., 2023a; Garvert et al., 2017; Lynn et al., 2020a, 2020b; Lynn and Bassett, 2020; Mark et al., 2020; Stiso et al., 2022)***. Exposed to streams of visual shapes or auditory sounds which belong to different network clusters, they detect the transition from one cluster to another despite flat adjacent transition probabilities between items.

A few attempts has been made to aggregate some statistical learning results under a common general associative mechanism ***(Endress and Johnson, 2021)***. However, each of these results is mainly explained in the statistical learning literature, by specific models, tailored to each experiment, and supposedly specific cognitive computations, multiplying the number of calculations required for the participant to succeed. This includes local transition probabilities ***(Saffran et al., 1996)***, backward transition probabilities ***(Perruchet and Desaulty, 2008)***, non-adjacent transition probabilities ***(Peña et al., 2002)***, communicability ***(Garvert et al., 2017)***, cosine similarity ***(Schapiro et al., 2013)***, and a compromise between precision and generalizability as a minimization of the free energy function (FEMM) ***(Lynn et al., 2020a)*** (fig 1). These models can be organized according to the scale considered for the temporal dependencies between the elements in the sequence (Fig 1). Broadly, three main scales can be considered: local (i.e. transitions between adjacent elements of the sequence), intermediate (i.e. transitions between non-adjacent elements) and high-order (i.e. at the level of the network structure underlying the sequence). Given that

participants are initially unaware of the underlying structure of a sequence, they should therefore simultaneously compute and store multiple relationships across different temporal scales in parallel. However, the memory requirements to keep track of all these computations is biologically implausible. Additionally, as many of these paradigms have been proposed in the context of language acquisition ***(Benjamin et al., 2023b; Newport and Aslin, 2004; Peña et al., 2002; Saffran et al., 1996)***, the reduced memory span of infants compared to adults must be considered, making a massively parallel multi-computational approach unlikely.

In two previous studies, we provided evidence to support the hypothesis of a common cognitive mechanism underlying statistical learning at different scales. We proposed that over the course of a sequence, the strength of the mental representations of each element slowly decays, allowing the representations of several consecutive elements to overlap in time. This decay mechanism might thus be able to link adjacent items but also more distant items, thus enabling the learning of higher scales of dependencies. To test this idea, we presented adult participants with auditory sequences in which the tone order were constrained by both local and high-order dependencies. Building upon the community network paradigm introduced by Schapiro et al ***(2013)***, the tones were randomly drawn from a network organized in two clusters (called communities) of six elements each. Transitions between tones were thus within the same community or between communities, and transitions probabilities were maintained uniform between each tone of the sequence. Crucially, all possible transitions were not presented during the familiarization phase, a paradigm we called "sparse community" (see Fig1 Local/high order interaction for the network structure). We then tested how familiar these transitions were for participants. We used a 2-by-2 design to compare transitions that were heard or not during habituation and the transitions that were congruent or not with the higher-order community structure (i.e. the two tones belonged to the same community or not, see fig 1 in ***(Benjamin et al., 2023a)***. Using two forced choice ratings we found that participants judged familiar transitions between elements belonging to the same community, even those which were not presented during the training sequence. In other words, they completed the structure by incorporating the missing transitions within clusters. On the other hand, they judged transitions crossing between communities as unfamiliar despite the fact that they have been heard as often as the familiar within-community transitions ***(Benjamin et al., 2023a)***. Subsequently, MEG

recordings in a passive version of this design revealed the same pattern of deviance detection ***(Benjamin et al., 2024)***. We compared a wide variety of models proposed in the statistical and network learning literatures on these measures and showed that the best model to account for participants' learning and generalization of the high-order structure in this sparse design was a long-horizon associative model, implementing a model of tradeoff between precision and computational cost called the free energy minimization model (FEMM) ***(Lynn et al., 2020a)***, mathematically equivalent to the successor representation (SR) model widely used in the reinforcement learning, and spatial navigation literatures ***(Dayan, 1993)***.

Formally, this model is a linear combination of transition probabilities and non-adjacent transition probabilities of all orders with an exponential decreasing weight associated with higher order (see section Formal description of the candidate model and fig 2A). This formal computation can be easily derived from associative learning principle over long horizons, making it biologically plausible (see fig 2B). Thanks to MEG temporal resolution and multivariate pattern analyses (MVPA), we observed an exponentially decaying decoding accuracy of each element like postulated in fig 2B, with a significant overlap of 8 consecutive elements, enabling long-horizon associative learning ***(Benjamin et al., 2024)***. This exponential decay was consistent with previous findings in monkeys' visual cortex ***(Miyashita, 1988)*** or in verbal perseveration behavior in patients ***(Cohen, 1998)***. Recent modelling work tested a comparable hypothesis of successor representation implementation through long lasting representation of elements called traces, and found more evidence for such implementation over other possible candidates ***(Kahn et al., 2024)***. Therefore, the long-horizon associative learning principles might provide a biologically plausible and a parsimonious explanation for statistical learning at different scales. We propose testing our unitary hypothesis by comparing the predictions of our model with data from published experiments spanning different learning scales. While we do not claim that this model improves the individual fits between the data and the models presented in the respective studies, we argue that it provides a unified framework capable of effectively representing a wide range of scenarios.

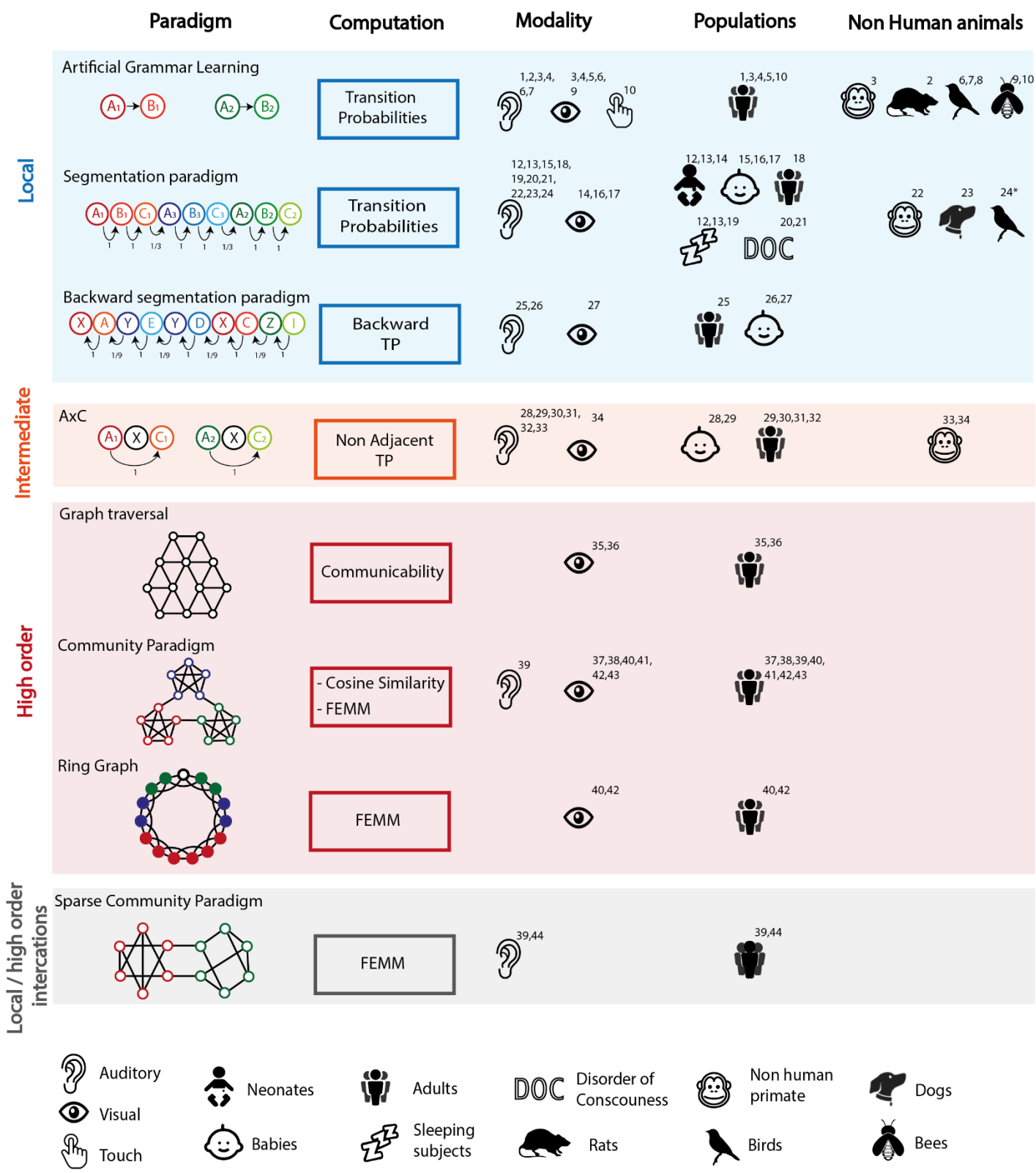
Paradigm
Computation
Modality
Populations
Non Human animals
Local
Intermediate
High order
Local / high order intercations
Artificial Grammar Learning
Transition Probabilities
1,2,3,4, 6,7
3,4,5,6, 9
10
1,3,4,5,10
3
2
6,7,8
9,10
Segmentation paradigm
Transition Probabilities
12,13,15,18, 19,20,21, 22,23,24
14,16,17
12,13,14
15,16,17
18
12,13,19
20,21
DOC
22
23
24*
Backward segmentation paradigm
Backward TP
25,26
27
25
26,27
AxC
Non Adjacent TP
28,29,30,31, 32,33
34
28,29
29,30,31,32
33,34
Graph traversal
Communicability
35,36
35,36
Community Paradigm
- Cosine Similarity
- FEMM
39
37,38,40,41, 42,43
37,38,39,40, 41,42,43
Ring Graph
FEMM
40,42
40,42
Sparse Community Paradigm
FEMM
39,44
39,44
Auditory
Visual
Touch
Neonates
Babies
Adults
Sleeping subjects
DOC
Disorder of Conscouness
Rats
Non human primate
Birds
Dogs
Bees

Figure 1 : Overview of statistical learning paradigms according to local, intermediate and high order temporal dependencies. *We report here on various studies showing learning at different scales of statistical dependence as a function of stimulus modality and in the different populations tested*

1. Harris (1955) 2. Toro & Trobalon (2005) 3. Milne et al (2018) 4. Conway & Christiansen (2006) 5. Reber(1967) 6. Santolin et al (2016) 7. James et al (2020) 8. Menhart et al (2015) 9. Avarguès-Weber et al (2020) 10. Comway & Christiansen (2005) 11. Nicholls et al (2014) 12. Fló et al (Fló et al., 2025, 2022, 2019) 13. Benjamin et al (2023b) 14. Bulf et al (2011) 15. Saffran et al (1996) 16. Fiser & Aslin (2002) 17. Kirkham et al (2002) 18. Saffran et al (1999, 1996) 19. Batterink & Zhang (2022) 20. Xu et al (2022) 21. Benjamin et al (2025). 22. Hauser et al (2001) 23. Boros et al (2021) 24. Takahasi (2010) 25. Perruchet and Desaulty (2008) 26. Pelucchi et al (2009) 27. Tummelsthammer et al (2017) 28. Buiatti et al (2009) 29. Gomez et al (2002) 30. Peña et al (2002) 31. Endress (2010) 32. Newport & Aslin (2004) 33. Newport et al (2004) 34. Sonnweber et al (2015) 35. Garvert et al (2017) 36. Mark et al (2020) 37. Schapiro et al (2013; 2016) 38. Karuza et al (2019, 2017) 39. Benjamin et al (2023a) 40. Lynn et al (2020a) 41. Ren et al (2022) 42. Stiso et al (2022) 43. Kakaei et al (2021) 44. Benjamin et al (2024)

** Although this study does not strictly use the same paradigm, it shows segmentation and learning of tutors' continuous song in juvenile Bengalese Finches.*

Methods

Our analyses are based on data accessible after publication. These studies comprised behavioral measurement of participants' familiarity with different test sequences (familiarity ranking and reaction times in healthy adults, looking preferences in human infants), and neurophysiological data such as ERP differences between test sequences in adults and neonates, and fMRI activations in the entorhinal cortex. They covered the full range of temporal scales shown in Figure 1, allowing us to test our hypothesis of a unified model for learning across these temporal dependencies. We thus qualitatively compared the obtained results with the FEMM prediction but we also, quantitatively estimated the goodness of fit for experiments with sufficient amount of data.

Formal description of the candidate model

The FEMM has been proposed by Lynn and colleagues ***(2020)*** to account for high order statistical learning in network-structured visual sequences, where adjacent transition probabilities alone were deemed insufficient to account for human performance. It was then described as a trade-off between computational accuracy and complexity. Intuitively, performances are explained as memory errors in the process of computing adjacent transition probabilities in sequences. Indeed, participants exposed to a sequence of inputs reinforce the association between element i and i-1 (local transition probabilities). However, errors in this process may lead participants to sometimes bind current element i with element i-2, i-3, i-4…. with a decreasing probability. This intuitive idea can be analytically formalized, and the optimal decreasing probability distribution can be computed by minimizing the free energy function. It results in a probability distribution following the Boltzmann function (exponential decay). Therefore, the estimated mental model of subjects is not limited to transition probability but enables participants to encode high-order structure.

This estimated mental model can be computed as a linear combination of the transition probability matrix (A_1) and non-adjacent transition probabilities of every order ($A_{\Delta t}$) weighted by an exponential decay with Δt (the order of non-adjacency, or size of the memory error). The estimated model can then be written as:

$$\hat{A} = \sum_{\Delta t=0}^{+\infty} Q(\Delta t) A_{\Delta t+1} \ With \ Q(\Delta t) = \frac{1}{Z} e^{-\beta \Delta t}$$

$$(eq. 1)$$

where $A_{\Delta t}$ is the transition probabilities matrix of order Δt between the element of the sequence. Note that for sequences following a Markovian rule (the current element is only dependent of the previous one), the non-adjacent transition probabilities matrices ($A_{\Delta t}$) can be directly computed from the transition probability matrix (A) with the relation: $A_{\Delta t} = A^{\Delta t}$. Then the previous equation can be re-written for markovian sequences:

$$\hat{A} = \sum_{\Delta t=0}^{+\infty} Q(\Delta t) A^{\Delta t+1} \ With \ Q(\Delta t) = \frac{1}{Z} e^{-\beta \Delta t}$$

$$(eq. 2)$$

Which can be simplified as:

$$\hat{A} = \left(1 - e^{-\beta}\right) A \left(I - e^{-\beta} A\right)^{-1}$$

$$(eq. 3)$$

I being the identity matrix

The β parameter was previously estimated to 0.06 in network learning tasks with human adults ***(Benjamin et al., 2023a; Lynn et al., 2020a).*** β plays the role of a generalization parameter: a low value leads to very high generalization (the long-distance dependencies would be strongly taken into account in the computation). On the contrary, a high β value means a weaker generalization, as only short scale dependencies would really influence the FEMM value. To the limits, $\beta = \infty$ corresponds to a total absence of generalization, meaning that only adjacent transition probabilities are learned and $\beta = 0$ allows no learning, as all elements are associated with each other in the same way.

Importantly, this mathematical model is biologically plausible through long-horizon associative learning mechanism anchored in Hebb's rule. If the representation of each element of the sequence is hold for enough time in the brain, a significant overlap between several successive element representations will enable simultaneous short and long-distance dependency learning. To match the predictions of the FEMM, the sustained representation in the brain of each element of the sequence should follow the same

exponential decay (see fig 2B), an implementation hypothesis supported by recent data ***(Benjamin et al., 2024; Kahn et al., 2024)***.

A formally equivalent model has already been developed in the reinforcement learning and spatial navigation communities. Indeed, the hippocampal place cells have been proposed to represent maps of probabilistic future states and reward by encoding successor representation instead of positional cognitive maps ***(Dayan, 1993; Momennejad, 2020; Stachenfeld et al., 2017).*** As possible future states are not limited to the next one, but can encompass several steps forward, the theoretical question is similar to the one of integrating different orders of transition probability. Indeed, successor representation has then been formally defined as a pondered sum of probabilistic future states, and can then be written SR = $\sum_{\Delta t} \gamma^{\Delta t} A^{\Delta t}$, with A the transition probability matrix between successive states. This approach is very similar to FEMM with an infinite sum of all powers of the transition matrix, pondered by an exponentially decreasing factor. Here the factor is $\gamma^{\Delta t}$ with $0 < \gamma < 1$. It is important to note that even if those models have been proposed from different backgrounds (statistical learning vs. cognitive map formation), they are mathematically equivalent up to the following relation $\gamma = e^{-\beta}$. Recent work using intracranial recordings confirmed that hippocampus and entorhinal cortex encode of network structure following a successor representation geometry ***(Tacikowski et al., 2024)*** suggesting a possible implication of the hippocampus in the Long-Horizon Associative Learning computation in statistical sequence learning.

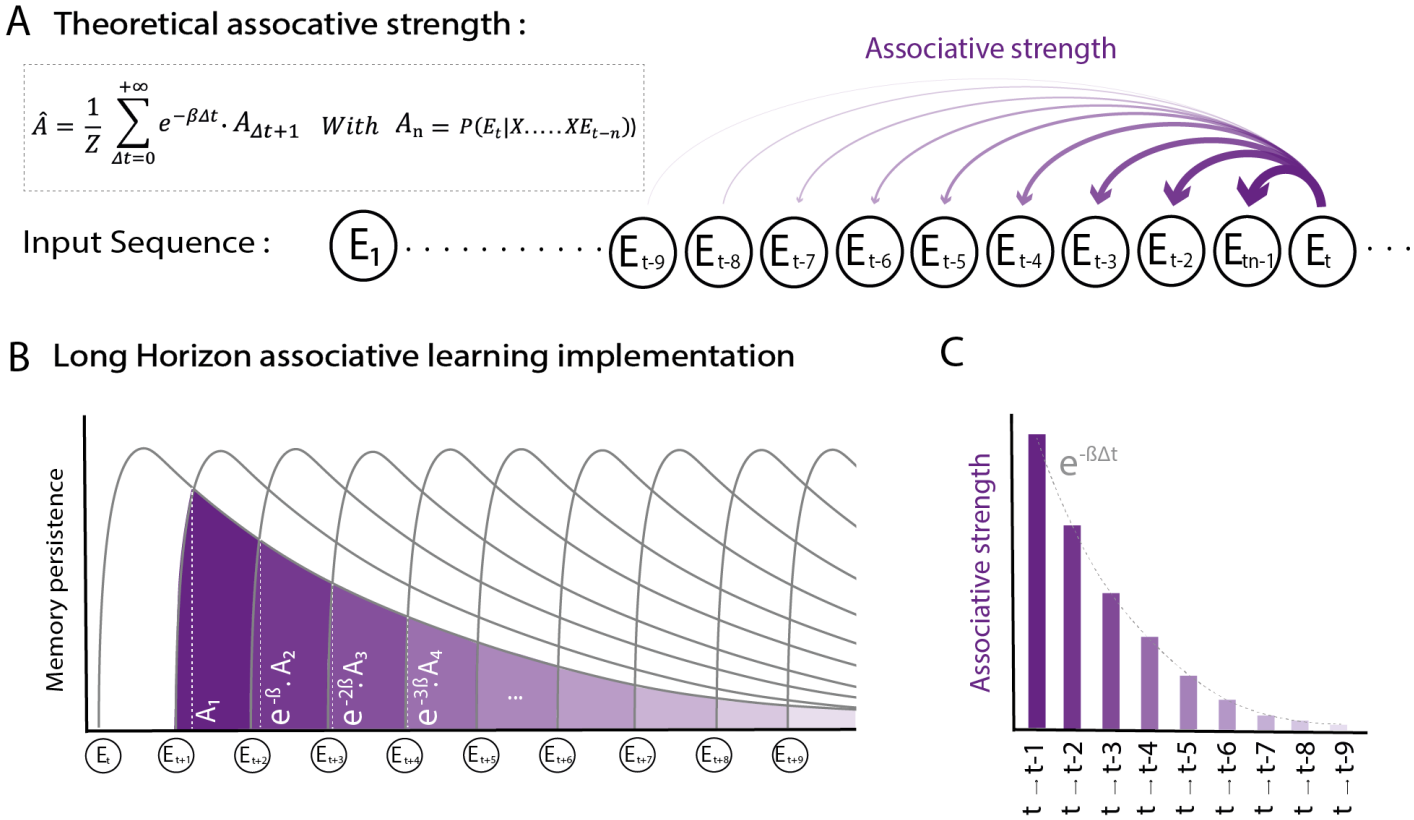

Figure 2 *: Description of the FEMM and its Long-Horizon Associative learning implementation.* ***A.*** *Formal equation of the FEMM enabling to compute the matrix of association between each pair of elements (Â). Â is the results of the association of adjacent and non-adjacent transition probabilities at all order with an exponential decreasing weight. The schema on the right represents the amount of association between a given element of the sequence* E_t *and all its predecessors. It shows that it is strongly associated to* E_{t-1} *(Adjacent TP) but also to* E_{t-2}, E_{t-3}*…. (non-adjacent TP) with a decreasing weight.* ***B****. This graph represents the implementation of FEMM through Long-Horizon associative learning strategy. If the mental representation of each element of the sequence follows an exponential decay, the amount of overlap between successive elements enables Hebbian associations with consecutive and non-consecutive element following a penalty computed with FEMM.* ***C****. Schema of the exponentially decaying weights of association between consecutive (t -> t-1) and nonconsecutive (t –> t-2; t -> t-3; …) elements of the sequence.*

Computation of the model for each paradigm

To examine the applicability of FEMM across various experimental contexts, we carefully curated landmark studies that encompass different scales of statistical learning. Generally, these studies comprise a training period during which participants are exposed to sequences of stimuli, followed by a test period in which items congruent or incongruent with the training are compared. Learning is assumed when

there is a difference in familiarity/deviance rating of the different types of items. Therefore, we conducted simulations to estimate the strength of associations between elements given a specific training, computed an estimation of the familiarity associated with each tested condition and finally qualitatively fit our estimation with the measured data from the literature. When possible, we also made a quantitative comparison between our estimations and the participants' reported performances.

More specifically, we estimated the mental model predicted by FEMM in each paradigm, but to avoid overfitting of each paradigm and keep the most parsimonious model possible, we fixed the β parameter for all paradigms $\beta = 0.06$, i.e. the value proposed by Lynn et al ***(2020)*** and replicated with in our study ***(Benjamin et al., 2023a)***. Then, for each of these paradigms, we used the following approach:

- When the TP matrix was available and the training sequence was a random walk into the associated network (Markovian condition), we used the analytical formula of the FEMM to compute the exact FEMM matrix from the TP matrix (see eq 3.)
- In the case of AXC or artificial grammar paradigms, the local TP matrix was not sufficient to describe the training material as the input followed non-Markovian relationships. Therefore, we estimated the FEMM from the corpus of training sequences, by computing and summing (with the associated decreasing exponential factor) all TP orders within each training sequence (eq. 1). The final $\hat{A}$ matrix was obtained by averaging each matrix of $\hat{A}$ estimation for each training sequence that a participant experienced.

Once the participants' mental model was computed based on the training material for each paradigm, we estimated the familiarity of each test condition.

- When experimental conditions corresponded to pairs, the estimated familiarity referred directly to the corresponding association on the FEMM matrix $(\hat{A})$ previously obtained. This is the case for all network paradigms (fig 3 E, F, G & I)
- For longer test conditions, we computed a pondered sum of each pair of elements within the test sequence. For example, the familiarity of a test sequence with three elements *abc* is estimated following the equation : $\hat{A}(ab) + \hat{A}(bc) + e^{-\beta}\hat{A}(ac)$. This formula was used for all paradigms with test sequences of more than 2 elements (fig 3 A, B, C &D).

- In the specific case of Lynn et al ***(2020a)***, we followed their method to compare whether learning was different for two types of networks: the community and the Lattice network. We computed the average FEMM familiarity of all transitions of the network for both networks.

This method enabled us to obtain quantitative predictions of estimated FEMM familiarity for each condition in each paradigm, which we could then compare with experimental measures like behavioral judgment, looking time, reaction time or fMRI activity.

Results

All the results are presented in figure 3.

Local scale

Segmentation paradigms

At the local scale, learning transition probabilities has been proposed to account for several results in the literature, among which the important study on speech segmentation by 8 months-old infant by Saffran and colleagues ***(1996)***. In this study and its many variations ***(Batterink and Zhang, 2022; Benjamin et al., 2025, 2023b; Bulf et al., 2011; Fiser and Aslin, 2002; Fló et al., 2022, 2019; Kirkham et al., 2002; Saffran et al., 1999, 1996; Xu et al., 2022)***, participants heard structured sequences of syllables (or other auditory, visual of haptic stimuli) with embedded trisyllabic pseudo-words (ABC), so that local transitions between consecutive syllables within a word are predictable (TP=1), while between two words the transition probability drops at 1/3 (three other words are possible). After having been exposed to the structure during a few minutes, participants are typically exposed to isolated triplets corresponding either to a word ($A_iB_iC_i$) or to a so-called part-word ($B_iC_iA_k$ for example) straddling a word boundary. A difference between Word and Part-word conditions (in looking time, familiarity rating, EEG responses…) is usually interpreted as a correct learning segmentation of the sequence based on the learning of transition probabilities between syllables. In previous work, we showed that adults and sleeping neonates had similar capacities in computing transition probabilities and segmentation in such sequences ***(Benjamin et al., 2023b; Fló et al., 2025, 2022)***. Moreover, comatose patients showed neural correlates of segmentation in a similar task suggesting a partially preserved computation of local statistics ***(Benjamin et al., 2025; Xu et***

al., 2022) suggesting that computing adjacent transition probabilities in sequences is a low level and automatic process.

Comparison with the model: To account for those results with the FEMM, we computed the FEMM mental model from the TP matrix of the training sequence and estimated the familiarity with both Words and Part-words conditions. Familiarity was up to 19% higher for Words compared to Part-words (see estimated FEMM familiarity in **figure 3A**). As adjacent transition probability is the first order (and most prominent) component of the FEMM, it is not surprising that a TP difference induces a FEMM difference. FEMM can therefore explain the effects reported in the literature for this design.

Grammar learning paradigms

Some studies investigating grammar learning, such as Milne et al's work ***(2018)***, can be seen as testing adjacent and non-adjacent dependencies simultaneously. We chose this specific study because all the training and test sequences were explicitly stated in the article, allowing us to precisely test the model. After being exposed to 8 different training sequences (each sequence repeated 6 times), participants were presented with 5 items sequences and had to judge if those agreed or not with the grammar. The authors showed that participants' judgment for test sequences was correlated with local transition probabilities in both visual and auditory modalities.

Comparison with the model: We extracted the percentage of time participants judged the sequence as a violation of the grammar. To account for those results with our model, we first estimated the FEMM mental model based on the 8 training sequences that participants experienced. Then we computed the predicted FEMM familiarity for each test sequence. Correlations between those predictions and the behavioral responses of human adults showed an overall good agreement in both auditory and visual modalities (Visual $R=-0.65$ $p<0.05$, Auditory $R=-0.8$ $p<0.01$). Thus, FEMM significantly accounts for those results.

Intermediate scale

AXC paradigm

At the intermediate level, sensitivity to non-adjacent dependencies has been described as essential to language learning. Indeed, to account for grammatical rules such as tense agreement in English ("is ... ing", “has … ed”), one needs to learn non-adjacent associations, independently of the central element. It has then been shown that participants could learn non-adjacent transition probabilities of the form P(C|XA), in AXC triplets where the syllable A predicts C with variable X syllables in between. For instance, Peña et al ***(2002)*** showed that participants could learn such dependencies when presented with AXC triplets with subliminal pauses between each triplet. Specifically, the authors showed that after listening to a sequence including AXC triplets with pauses between each word, participants preferred test words that respected the A-C dependency (Rule Word AX’C where X’ has never been presented in AXC triplet) over Part-Words (XC_iA_k) participants have heard.

A surprising property of this paradigm is that by increasing the variety of possible Xs to be inserted in the middle of an A-C relationship, the familiarity of the A-C relationship improves. This has been measured by Gomez ***(2002)***. In this study, she compared the performances of a two forced choice task between pseudo-words respecting or not the A-C relationship (Words - A_iXC_i- vs non-Words - A_iXC_k-). She measured subjects’ performances for training sequences with dictionaries of 2, 6, 12, or 24 possible Xs in the middle of the A-C relationship, and showed that the adults’ performances increased with the number of possible Xs. As the A-C relationship is independent of the number of possible Xs in the middle, learning only the non-adjacent dependency cannot explain by itself this variation in familiarity with the number of X.

Comparison with the model: We first estimated the FEMM familiarity of Words, Part-words and Rule-Words in the classical case where three Xs can be presented in the middle of each pair ***(Buiatti et al., 2009)*** and when training is provided with isolated words. We estimated the model on each triplet of the stream following the authors' interpretation that the pause created a segmentation of the sequence ***(Peña et al, 2002)***. We showed that in test, Word triplets (A_iXC_i) had higher familiarity rating than triplets respecting the rule ($A_iX’C_i$), themselves higher than Part-Words (XC_1A_2) (see fig 1 in ***(Peña et***

al., 2002), Fig 3C), mimicking familiarity measures in humans *(Buiatti et al., 2009; Newport and Aslin, 2004; Pena, 2002).* In agreement with Peña et al *(2002)*, we also found that in the absence of pauses during training (in that case the model is estimated on the continuous stream), Rule-Words were no longer estimated more familiar compared to the Part-Words.

We also compared familiarity for every number of possible X between 2 and 29 elements. We showed that the difference in familiarity between Words (A_iXC_i) vs non-Words (A_iXC_j) also increased with the number of Xs in the model and qualitatively matched the behavioral results in Gomez' study *(2002).* In fig 3D, the left panel represents the FEMM prediction for Words vs non-Words for different number of possible X ranging from 2 to 29. The right-hand panel shows familiarity as measured by Gomez et al *(2002)*. The model predicts a logarithmic increase of the Word vs non-Word dissimilarity with the number of possible X while the empirical results seem more linear. Nonetheless, differentiation between linear and logarithmic trend with only four measured values may be challenging. A more careful exploration of this question, with more conditions, would be interesting to better estimate the effect of the number of X on the participants' performance.

Higher order scale

Garvet et al's network paradigm

At a higher-order scale, description of the structure is often based on a network formalism, and a variety of networks and their properties have been explored. For instance, Garvert and colleagues *(2017)* scanned participants with fMRI while they were traversing an abstract network composed of 12 objects. Among other results, the authors reported a correlation between the brain signal and Communicability, another metric of distance between nodes in networks. The conceptual idea behind Communicability and FEMM is similar, but the penalization of long-distance dependencies is greater in Communicability than FEMM. Using a searchlight method, they found a cluster in the hippocampus and entorhinal cortex correlating with communicability.

Comparison with the model: Based on the network structure, we computed the FEMM and thus estimated a predicted familiarity for each transition of the network. We could then re-analyze the average activation of the cluster that they evidenced to be significantly correlated with communicability, and we showed that its mean activation per transition also significantly correlated with FEMM predictions of the transitions ($p<0.05$) (**Fig 3E**).

Community network

Schapiro and colleagues ***(2013)*** proposed a community network paradigm to demonstrate the sensitivity of humans to clusters in network structure, despite the absence of information at the local scale. Indeed, this network is composed of three communities (groups of nodes almost entirely connected to each other but weakly connected to other groups) and all the nodes in the network have exactly four neighbors. The sequence derived from such network is then completely flat in terms of adjacent transition probabilities, with each pair of elements in the resulting sequence having exactly 0.25 transition probability. Despite this lack of local transition probability information, Schapiro and colleagues ***(2013)*** as well as many others later on ***(Benjamin et al., 2024, 2023a; Kahn et al., 2024, 2018; Karuza et al., 2019, 2017; Lynn et al., 2020a)*** showed that humans could discriminate transitions staying within communities from transitions switching between communities. Lynn and colleagues ***(2020a)*** also compared the community network with lattice network structures. Like the community network, it is composed of twelves nodes with four neighbors per nodes. They found that participants trained on a lattice network were more familiar with the structure (faster reaction time) compared to participants trained on a community network.

Comparison with the model: The FEMM was initially proposed in this study ***(Lynn et al., 2020a)*** to specifically account for those results. In particular, the authors showed that the drop in familiarity when switching community was correctly predicted by the model (see **fig 3F**). Moreover, they showed that the learning difference between the community and the lattice network was also correctly accounted by the FEMM. Indeed, the average of the FEMM familiarity on every transition of the lattice network is

higher than the one on the community network, which is congruent with faster reaction time when learning the lattice compared to the community network (see **fig 3H**).

Ring graph

In this important study, another type of network was investigated by Lynn and colleagues: the ring graph. In this 15 nodes network, each node is connected to four neighbors (+1 / +2 / -1 / -2). Then, from a node, four are directedly reachable by a single transition (distance 1), four others are reachable in two steps (distance 2) and the six others can be reached in three transitions (distance 3). Measuring the familiarity of each pair of this network using reaction time revealed a preference for connected pairs (distance 1) over unconnected pairs of distance 2, itself preferred to unconnected pairs of distance 3.

Comparison with the model: This also matches the FEMM predictions as shown by Lynn and colleagues ***(2020a)*** see **Fig 3G.** Indeed, participants' reaction time to Distance 3 pairs was higher than for Distance 2 pairs, itself higher than for connected pairs.

Local and high order interaction

Sparse and high sparse community networks

In previous studies, we introduced the sparse community design ***(Benjamin et al., 2024, 2023a)***, in which one (sparse community) or two (high sparse community) possible transitions per node inside each community in the network were removed. Thus, it creates a 2-by-2 design where a transition between two sounds can be locally congruent (Familiar vs New transitions i.e. the transition between these two sounds was present during the learning or not) or respect higher-order community membership (Within vs Between transitions: these two sounds belong to the same community or not). We first measured familiarity judgments of participants and showed that they were sensitive to the community structure. Specifically, their pattern of errors showed that they hallucinated non existing transitions between two elements belonging to the same community, despite having never heard them (New Within transitions). We called this the completion effect, as their mental model of the sparse community network was not sparse, missing transition being reconstructed. Conversely, transitions between the two clusters were

rated as less familiar than transition within the clusters, despite being heard the same amount of time during the training, also showing that the latent structure was better memorized than the presented transitions. To explain this behavior, we compared different models proposed in the statistical learning and network learning literature, including transition probabilities, non-adjacent transition probabilities, ngrams, communicability… and we showed that it was the FEMM that provided the best fit to the behavior **(see Fig 3I).**

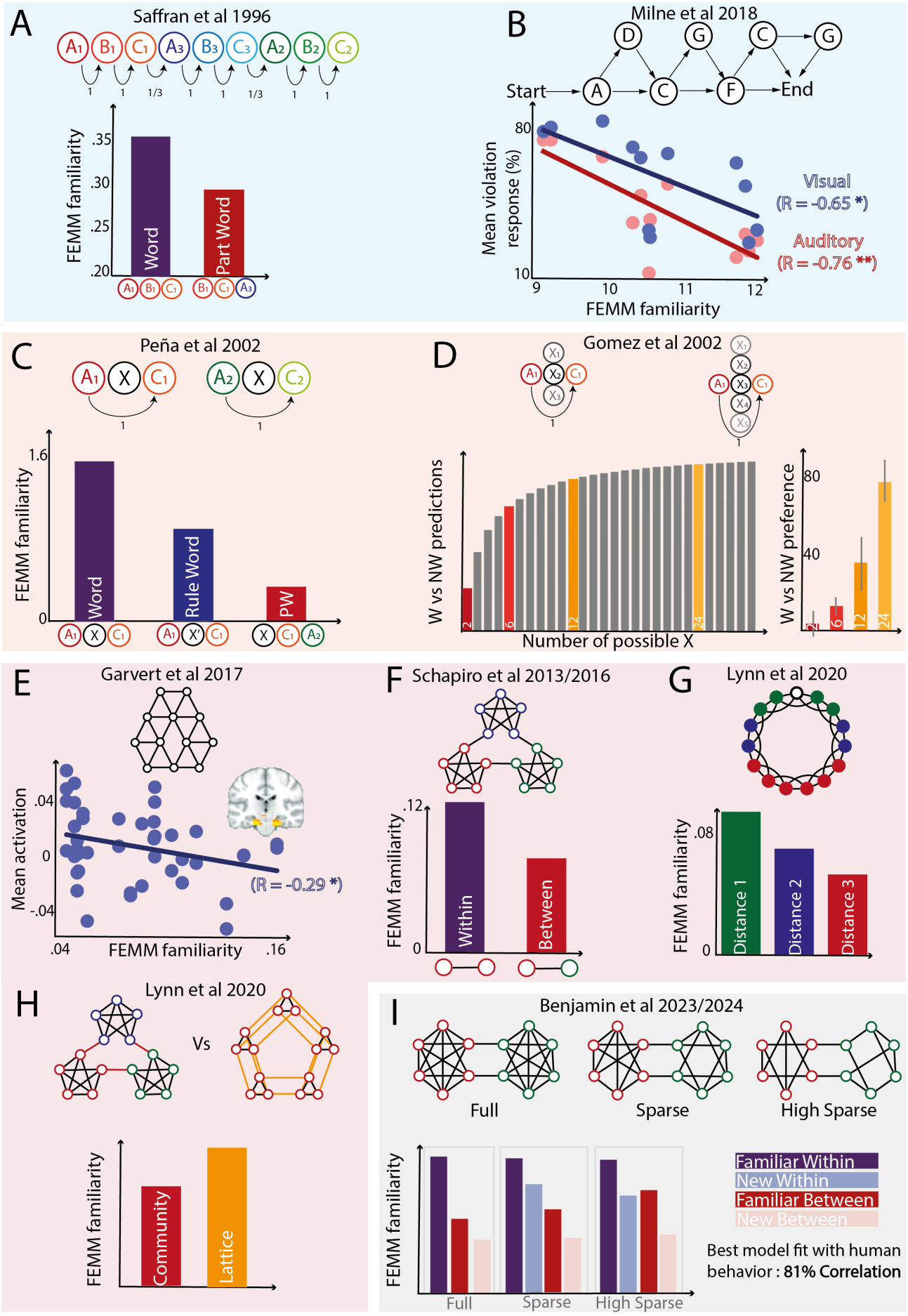
A
Saffran et al 1996
FEMM familiarity
Word
Part Word
B
Milne et al 2018
Start
End
Mean violation response (%)
FEMM familiarity
Visual
(R = -0.65 *)
Auditory
(R = -0.76 **)
C
Peña et al 2002
FEMM familiarity
Word
Rule Word
PW
D
Gomez et al 2002
W vs NW predictions
Number of possible X
W vs NW preference
E
Garvert et al 2017
Mean activation
FEMM familiarity
(R = -0.29 *)
F
Schapiro et al 2013/2016
FEMM familiarity
Within
Between
G
Lynn et al 2020
FEMM familiarity
Distance 1
Distance 2
Distance 3
H
Lynn et al 2020
Vs
FEMM familiarity
Community
Lattice
I
Benjamin et al 2023/2024
Full
Sparse
High Sparse
FEMM familiarity
Familiar Within
New Within
Familiar Between
New Between
Best model fit with human behavior : 81% Correlation

Figure 3 *: Re-analyzing results from different statistical learning scales with FEMM. Comparison between FEMM simulation and results from the literature.*

A) *Segmentation tasks of embedded words in a continuous sequence have been proposed as a demonstration of sensitivity to adjacent transition probabilities. Bar plots display the FEMM estimated familiarity for Word and Part-Word conditions.*

B) *Artificial grammars have also been used to investigate local transition probability learning. The mean violation response measured in Milne et al 2018 significantly correlates with FEMM familiarity in both visual (blue) and auditory (red) modalities .*

C) *At the intermediate scale, sensitivity to non-adjacent dependencies have been demonstrated using AXC paradigms. The reported preference for RuleWord over PartWord is also correctly accounted by the FEMM.*

D) *Varying the number of possible X in the AxC paradigm causes a variation of participant's' performances in Word vs NonWord discrimination despite a similar non-adjacent transition probability P(C|XA) = 1 in all cases. The left barplot represents the FEMM estimation for number of possible Xs ranging from 2 to 29 showed an increased Word vs NonWord discrimination performance, similar to Gomez et al behavioral results (right barplot).*

E) *Using fMRI, Garvert and colleagues* (2017) *showed that when exposed to a sequence deriving from a network structure, a cluster in the entorhinal cortex significantly correlated with Communicability. Here, we show that the average activation within this cluster also significantly correlated with FEMM.*

F) *In the community paradigm, Schapiro and colleague* (2013) *showed that participants could learn the cluster structure and differentiate transitions Within from transitions Between communities despite no local transition probability difference. This difference is well modeled by the FEMM as shown in the barplot comparing within from between transitions' familiarity.*

G) *The ring graph, has also been investigated with this framework. Lynn et al* (2020a) *reported that two connected nodes (white to green) where jugged more familiar than unconnected nodes two transitions apart (white to blue), themselves more familiar than unconnected nodes three transitions apart (white to red). The estimated FEMM familiarity displayed in the barplot correctly predicts this behavior.*

H) *Lynn et al* (2020a) *also showed that, in accordance to the FEMM predictions, learning a community structure was easier (faster reaction time) than a lattice structure. Indeed, the average FEMM familiarity is higher for lattice (orange) than for community (red) network.*

I) *Finally, in a previous experiment we introduced the sparse and high sparse community designs. We modified the community paradigm by removing possible transitions within each community, thus mixing local and high order statistical properties. The FEMM predicted a pattern of familiarity including the generalization of the structure where non-existent transitions within communities would be judged much more familiar than those between communities. The behavioral and MEG results experimentally confirmed this prediction. Critically, we compared many statistical models on this paradigm and showed that FEMM significantly outperformed the other statistical models of the literature.*

Discussion

We have shown here that a single model, can account for a wide range of results in the literature, spanning various scales of statistical learning in sequences. The model accurately fits the experimental data, and its ability to explain such a diverse set of results renders this approach more parsimonious than the numerous specialized models proposed in the literature. We proposed this biologically plausible Long-Horizon associative learning as a unifying approach. This framework emphasizes the similarities between the statistical learning and network learning literatures, suggesting that learning in both contexts, while often being studied independently, might arise from similar associative computations. While network learning is often conceptualized as abstract structure learning ***(Tacikowski et al., 2024),*** requiring explicit representations of maps, our approach suggests that many findings might be explained by low level mechanisms akin to those described in the statistical learning literature.

1) A Single Free Parameter to Capture Integration Across Statistical Scales

FEMM is a particularly interesting model because it relies on a single free parameter, β, to balance accuracy and integration across multiple statistical scales. At the extremes, when $\beta \to \infty$, the model becomes sensitive only to the local transition probabilities. Conversely, when $\beta \to 0$, all scales are weighted equally, leading to uniform familiarity between all stimulus pairs in the sequence. For intermediate values, the β parameter functions as a generalization factor: the lower the β, the more the memorized structure ($\hat{A}$) overgeneralizes the actual structure (A), sacrificing accuracy. In contrast, as β increases, the memorized structure ($\hat{A}$) more closely aligns with the actual transition probabilities of the structure (A). Thus, identifying the real β value in different populations (e.g., human infants, non-human animals) might offer valuable insights into their sequence learning abilities and reveal potential causal mechanisms underlying specific cognitive performances.

In all the results presented above, we set the β parameter to a fixed value ($\beta = 0.06$, as previously determined by Lynn et al ***(2020a)*** and later confirmed in Benjamin et al ***(2023a)***) to avoid over-fitting the model to each individual dataset and to properly evaluate its true predictive power. However, its estimation may vary depending on the measurement method and/or the level of participant involvement in the task. For example, in an active behavioral task using the sparse community paradigm, where

participants judged the familiarity of tone pairs after a learning phase, we estimated the average β for the population to be 0.06. In contrast, using the same paradigm in a passive MEG study, we estimated β at 0.5 from the MEG signal. In other words, the strength of associations between tones decreased more rapidly in the passive MEG study than in the active behavioral task. This discrepancy may be due to the differing sensitivities of the measurement methods or to active engagement in the behavioral task. Memorization effort in the behavioral task might lead to stronger top-down processes to maintain long distance elements in working memory, thereby increasing the weighting of long-range dependencies. Given that the role of attention in statistical learning remains a topic debated in the literature ***(Ambrus et al., 2020; Batterink and Choi, 2021; Batterink and Paller, 2019; Benjamin et al., 2025, 2021; Liu et al., 2023; Toro et al., 2005; Xu et al., 2022)***, further investigation into this potential relationship is needed. At the modeling level, enhanced attention should decrease the β value by actively sustaining neural activity longer for each element in the sequence (see fig2A & B). Another hypothesis would be that different brain areas encode different representations, more or less temporally limited or general (accounted by a gradient of the β values), enabling multiscale representation in different areas ***(Momennejad, 2020)***.

Other factors might also modulate the β value, such as individual differences. Indeed, although estimating generalization at the individual level is challenging, Lynn and colleagues ***(2020a)*** showed that an individually fitted β parameter correlated with participants' memory capacity in a n-back task, providing initial evidence that individuals may differ in their generalization/accuracy trade-off. Specifically designed task with test-retest reliability of this parameter would help investigate this question. Thus, estimating the β value at different ages or in different species might provide valuable insights into sequence learning abilities in this specific group. Regarding human development, we may hypothesize that the β value is initially high due to the limited ability of immature neurons to sustain firing, which would promote a focus on local transitions rather than higher-order integration. Indeed, the computation of local transition probabilities has already been observed in sleeping neonates, with accuracy and learning speed comparable to that of adults. EEG allows to study implicit learning and to recover modulating effects between elements of the sequence before behavioral paradigms based on familiarity and recognition start being sensitive. Furthermore, EEG allows to test very young infants. Therefore, it would be worthwhile to study sequence learning with different levels of integration at the same ages using EEG. If the β value proves to be a measurable and functionally relevant parameter, it

could offer a novel approach to understanding developmental changes and enable meaningful cross-species comparisons in sequence learning.

2) Biological implementation and relation to other models of decay for sequence learning

In our previous MEG experiments, we proposed and provided evidence for an implementation of this theoretical model though Long-Horizon Associative learning. Specifically, we suggested that association between several successive elements could simply arise from overlapping mental representation of each element of the sequence (fig 2B/C). Our MEG data revealed exponentially decaying representations of each tone in the sequence, with significant overlap extending across up to eight consecutive tones. This measured exponential decay provide supports to the proposal of an exponential decay in the weights associated with each order of transition probabilities. The counterpart of associations strength is a mismatch amplitude modulation, which can be similarly explained by an exponential decay ***(Wacongne et al., 2012)***.

3) Model limitations

We wanted here to highlight the significant role that low-level associative mechanisms have in a wide range of statistical learning studies. Although powerful, this model has limitations and is not intended to explain all aspects of human sequence learning. Other types of regularities, such as rule-based patterns in sequences are not captured by the FEMM. Alternating sequences (e.g., ABABABAB…) or more complex rules like repetition alternation (e.g., AABBAABB…) and symmetry (e.g., ABCDDCBA) are more easily detected through pattern recognition rather than by learning the strength of associations between successive elements. Recent studies suggest that such sequences may be stored in a compressed, language-like manner ***(Al Roumi et al., 2023, 2021; Dehaene et al., 2022; Planton et al., 2021; Quilty-Dunn et al., 2022)***. It is still debated whether rule-based and statistical learning share a common cognitive mechanism ***(Fiser & Lengyel, 2022)*** or involve distinct processes ***(Maheu et al., 2020)*** happening in parallel ***(Al Roumi et al., 2023)***. Further studies are needed to clarify how these forms of learning interact and whether rule-based learning might emerge as a strategy to reduce the cognitive cost of maintaining long-range associations.

Conclusion

We have shown here how long-horizon associative learning processes could constitute a unifying framework to bridge different aspects of sequence learning. It unifies under a common cognitive model and a unique brain implementation the learning of the different scales of statistical dependencies in sequences. Moreover, this framework also makes various predictions that have yet to be tested. Indeed, if the learning of transition probabilities and network structures in auditory sequences is based on the same mechanism, the properties demonstrated with transition probabilities paradigms should, at least partially, be transferred to network learning. Network learning thus should inherit the automaticity of statistical learning, enabling sleeping newborns to learn community structures in the same way as they learn transition probabilities. Similarly, successful learning of transition probabilities in non-human animals should predict equally successful learning of simple network properties (although statistical learning mechanisms can partially differ across species ***(Avarguès-Weber et al., 2020)***). Moreover, this model makes of hierarchical structures learning an emergent property of statistical learning, as long-horizon associative learning computed on a network composed of "communities of communities" is theoretically sufficient to encode this hierarchy of organization.

References

1. **Al Roumi F, Marti S, Wang L, Amalric M, Dehaene S.** 2021. Mental compression of spatial sequences in human working memory using numerical and geometrical primitives. Neuron **109**:2627-2639.e4. doi:10.1016/j.neuron.2021.06.009

2. **Al Roumi F, Planton S, Wang L, Dehaene S.** 2023. Brain-imaging evidence for compression of binary sound sequences in human memory. eLife **12**:e84376. doi:10.7554/eLife.84376

3. **Ambrus GG, Vékony T, Janacsek K, Trimborn ABC, Kovács G, Nemeth D.** 2020. When less is more: Enhanced statistical learning of non-adjacent dependencies after disruption of bilateral DLPFC. Journal of Memory and Language **114**:104144. doi:10.1016/j.jml.2020.104144

4. **Avarguès-Weber A, Finke V, Nagy M, Szabó T, d'Amaro D, Dyer AG, Fiser J.** 2020. Different mechanisms underlie implicit visual statistical learning in honey bees and humans. Proc Natl Acad Sci USA **117**:25923–25934. doi:10.1073/pnas.1919387117

5. **Basirat A, Dehaene S, Dehaene-Lambertz G.** 2014. A hierarchy of cortical responses to sequence violations in three-month-old infants. Cognition **132**:137–150. doi:10.1016/j.cognition.2014.03.013

6. **Batterink LJ, Choi D.** 2021. Optimizing steady-state responses to index statistical learning: Response to Benjamin and colleagues. Cortex. doi:10.1016/j.cortex.2021.06.008

7. **Batterink LJ, Paller KA.** 2019. Statistical learning of speech regularities can occur outside the focus of attention. Cortex **115**:56–71. doi:10.1016/j.cortex.2019.01.013

8. **Batterink LJ, Zhang S.** 2022. Neuropsychologia Simple statistical regularities presented during sleep are detected but not retained. Neuropsychologia **164**:108106. doi:10.1016/j.neuropsychologia.2021.108106

9. **Bekinschtein T, Dehaene S, Rohaut B, Tadel F, Cohen L, Naccache L.** 2009. Neural signature of the conscious processing of auditory regularities. Proceedings of the National Academy of Sciences of the United States of America **106**:1672–1677. doi:10.1073/pnas.0809667106

10. **Benjamin L, Dehaene-Lambertz G, Fló A.** 2021. Remarks on the analysis of steady-state responses: spurious artifacts introduced by overlapping epochs. Cortex. doi:10.1016/j.cortex.2021.05.023

11. **Benjamin L, Fló A, Al Roumi F, Dehaene-Lambertz G.** 2023a. Humans parsimoniously represent auditory sequences by pruning and completing the underlying network structure. eLife **12**:e86430. doi:10.7554/eLife.86430

12. **Benjamin L, Fló A, Palu M, Naik S, Melloni L, Dehaene-Lambertz G.** 2023b. Tracking transitional probabilities and segmenting auditory sequences are dissociable processes in adults and neonates. Developmental Science **26**:e13300. doi:10.1111/desc.13300

13. **Benjamin L, Sablé-Meyer M, Fló A, Dehaene-Lambertz G, Roumi FA.** 2024. Long-Horizon Associative Learning Explains Human Sensitivity to Statistical and Network Structures in Auditory Sequences. J Neurosci **44**:e1369232024. doi:10.1523/JNEUROSCI.1369-23.2024

14. **Benjamin L, Zang D, Fló A, Qi Z, Su P, Zhou W, Wang L, Wu X, Gui P, Dehaene-Lambertz G.** 2025. The role of conscious attention in auditory statistical learning: Evidence from patients with impaired consciousness. iScience **28**:111591. doi:10.1016/j.isci.2024.111591

15. **Boros M, Magyari L, Török D, Bozsik A, Deme A, Andics A.** 2021. Neural processes underlying statistical learning for speech segmentation in dogs. Current Biology **31**:5512-5521.e5. doi:10.1016/j.cub.2021.10.017

16. **Buiatti M, Peña M, Dehaene-Lambertz G.** 2009. Investigating the neural correlates of continuous speech computation with frequency-tagged neuroelectric responses. NeuroImage **44**:509–519. doi:10.1016/j.neuroimage.2008.09.015

17. **Bulf H, Johnson SP, Valenza E.** 2011. Visual statistical learning in the newborn infant. Cognition **121**:127–132. doi:10.1016/j.cognition.2011.06.010

18. **Cohen L.** 1998. Competition between past and present. Assessment and interpretation of verbal perseverations. Brain **121**:1641–1659. doi:10.1093/brain/121.9.1641

19. **Conway CM, Christiansen MH.** 2005. Modality-Constrained Statistical Learning of Tactile, Visual, and Auditory Sequences. Journal of Experimental Psychology: Learning, Memory, and Cognition **31**:24–39. doi:10.1037/0278-7393.31.1.24

20. **Dayan P.** 1993. Improving Generalisation for Temporal Difference Learning: The Successor Representation 14.

21. **Dehaene S, Al Roumi F, Lakretz Y, Planton S, Sablé-Meyer M.** 2022. Symbols and mental programs: a hypothesis about human singularity. Trends in Cognitive Sciences **26**:751–766. doi:10.1016/j.tics.2022.06.010

22. **Dehaene S, Meyniel F, Wacongne C, Wang L, Pallier C.** 2015. The Neural Representation of Sequences: From Transition Probabilities to Algebraic Patterns and Linguistic Trees. Neuron **88**:2–19. doi:10.1016/j.neuron.2015.09.019

23. **Endress AD.** 2010. Learning melodies from non-adjacent tones. Acta Psychologica **135**:182–190. doi:10.1016/j.actpsy.2010.06.005

24. **Endress AD, Johnson SP.** 2021. When forgetting fosters learning: A neural network model for statistical learning. Cognition 104621. doi:10.1016/j.cognition.2021.104621

25. **Fiser J, Aslin RN**. 2002. Statistical learning of new visual feature combinations by infants. Proceedings of the National Academy of Sciences of the United States of America **99**:15822–15826. doi:10.1073/pnas.232472899

26. **Fiser J, Lengyel G.** 2022. Statistical Learning in Vision. Annu Rev Vis Sci **8**:265–290. doi:10.1146/annurev-vision-100720-103343

27. **Fló A, Benjamin L, Palu M, Dehaene-Lambertz G.** 2025. Statistical learning beyond words in human neonates. eLife **13**:RP101802. doi:10.7554/eLife.101802

28. **Fló A, Benjamin L, Palu M, Dehaene-Lambertz G.** 2022. Sleeping neonates track transitional probabilities in speech but only retain the first syllable of words. Scientific Reports **12**:4391. doi:10.1038/s41598-022-08411-w

29. **Fló A, Brusini P, Macagno F, Nespor M, Mehler J, Ferry AL.** 2019. Newborns are sensitive to multiple cues for word segmentation in continuous speech. Developmental Science **22**. doi:10.1111/desc.12802

30. **Garvert MM, Dolan RJ, Behrens TEJ.** 2017. A map of abstract relational knowledge in the human hippocampal–entorhinal cortex. eLife **6**:1–20. doi:10.7554/eLife.17086

31. **Gomez R.** 2002. Variability and detection of invariant structure. SciFed Journal of AIDS & HIV Research **1**:431–436. doi:10.23959/sfahrj-1000006

32. **Harris ZS.** 1955. From Phoneme to Morpheme. Language **31**:190–222. doi:10.2307/411036

33. **Hauser MD, Newport EL, Aslin RN.** 2001. Segmentation of the speech stream in a non-human primate: Statistical learning in cotton-top tamarins. Cognition **78**:53–64. doi:10.1016/S0010-0277(00)00132-3

34. **James LS, Sun H, Wada K, Sakata JT.** 2020. Statistical learning for vocal sequence acquisition in a songbird. Scientific Reports **10**:1–18. doi:10.1038/s41598-020-58983-8

35. **Kahn AE, Bassett DS, Daw ND.** 2024. Trial-by-trial learning of successor representations in human behavior. doi:10.1101/2024.11.07.622528

36. **Kahn AE, Karuza EA, Vettel JM, Bassett DS.** 2018. Network constraints on learnability of probabilistic motor sequences. Nature Human Behaviour **2**:936–947. doi:10.1038/s41562-018-0463-8

37. **Kakaei E, Aleshin S, Braun J.** 2021. Visual object recognition is facilitated by temporal community structure. Learn Mem **28**:148–152. doi:10.1101/lm.053306.120

38. **Karuza EA, Kahn AE, Bassett DS.** 2019. Human sensitivity to community structure is robust to topological variation. Complexity **2019**. doi:10.1155/2019/8379321

39. **Karuza EA, Kahn AE, Thompson-Schill SL, Bassett DS.** 2017. Process reveals structure: How a network is traversed mediates expectations about its architecture. Scientific Reports **7**:1–9. doi:10.1038/s41598-017-12876-5

40. **Kirkham NZ, Slemmer JA, Johnson SP.** 2002. Visual statistical learning in infancy: evidence for a domain general learning mechanism. Cognition **83**:B35–B42. doi:10.1016/S0010-0277(02)00004-5

41. **Liu H, Forest TA, Duncan K, Finn AS.** 2023. What sticks after statistical learning: The persistence of implicit versus explicit memory traces. Cognition **236**:105439. doi:10.1016/j.cognition.2023.105439

42. **Lynn CW, Bassett DS.** 2020. How humans learn and represent networks. Proceedings of the National Academy of Sciences of the United States of America **117**. doi:10.1073/pnas.1912328117

43. **Lynn CW, Kahn AE, Nyema N, Bassett DS.** 2020a. Abstract representations of events arise from mental errors in learning and memory. Nature Communications **11**. doi:10.1038/s41467-020-15146-7

44. **Lynn CW, Papadopoulos L, Kahn AE, Bassett DS.** 2020b. Human information processing in complex networks. Nature Physics **16**:965–973. doi:10.1038/s41567-020-0924-7

45. **Mark S, Moran R, Parr T, Kennerley SW, Behrens TEJ.** 2020. Transferring structural knowledge across cognitive maps in humans and models. Nature Communications **11**:1–12. doi:10.1038/s41467-020-18254-6

46. **Menyhart O, Kolodny O, Goldstein MH, DeVoogd TJ, Edelman S.** 2015. Juvenile zebra finches learn the underlying structural regularities of their fathers' song. Front Psychol **6**. doi:10.3389/fpsyg.2015.00571

47. **Milne AE, Petkov CI, Wilson B.** 2018. Auditory and Visual Sequence Learning in Humans and Monkeys using an Artificial Grammar Learning Paradigm. Neuroscience **389**:104–117. doi:10.1016/j.neuroscience.2017.06.059

48. **Miyashita Y.** 1988. Neuronal correlate of visual associative long-term memory in the primate temporal cortex. Nature **335**:817–820. doi:10.1038/335817a0

49. **Momennejad I.** 2020. Learning Structures: Predictive Representations, Replay, and Generalization. Current Opinion in Behavioral Sciences **32**:155–166. doi:10.1016/j.cobeha.2020.02.017

50. **Newport EL, Aslin RN.** 2004. Learning at a distance I. Statistical learning of non-adjacent dependencies. Cognitive Psychology **48**:127–162. doi:10.1016/S0010-0285(03)00128-2

51. **Newport EL, Hauser MD, Spaepen G, Aslin RN.** 2004. Learning at a distance II. Statistical learning of non-adjacent dependencies in a non-human primate. Cognitive Psychology **49**:85–117. doi:10.1016/j.cogpsych.2003.12.002

52. **Nicholls EK, Hempel de Ibarra N.** 2014. Bees associate colour cues with differences in pollen rewards. Journal of Experimental Biology jeb.106120. doi:10.1242/jeb.106120

53. **Pelucchi B, Hay JF, Saffran JR.** 2009. Learning in reverse: Eight-month-old infants track backward transitional probabilities. Cognition **113**:244–247. doi:10.1016/j.cognition.2009.07.011

54. **Pena M.** 2002. Rôle du calcul statistique dans l'acquisition du langage.

55. **Peña M, Bonatti LL, Nespor M, Mehler J.** 2002. Signal-driven computations in speech processing. Science **298**:604–607. doi:10.1126/science.1072901

56. **Perruchet P, Desaulty S**. 2008. A role for backward transitional probabilities in word segmentation? Memory & Cognition **36**:1299–1305. doi:10.3758/MC.36.7.1299

57. **Planton S, van Kerkoerle T, Abbih L, Maheu M, Meyniel F, Sigman M, Wang L, Figueira S, Romano S, Dehaene S.** 2021. A theory of memory for binary sequences: Evidence for a mental compression algorithm in humans. PLoS Computational Biology. doi:10.1371/journal.pcbi.1008598

58. **Quilty-Dunn J, Porot N, Mandelbaum E.** 2022. The Best Game in Town: The Re-Emergence of the Language of Thought Hypothesis Across the Cognitive Sciences. Behavioral and Brain Sciences 1–55. doi:10.1017/S0140525X22002849

59. **Ren X, Zhang H, Luo H.** 2022. Dynamic emergence of relational structure network in human brains. Progress in Neurobiology **219**:102373. doi:10.1016/j.pneurobio.2022.102373

60. **Saffran JR, Aslin RN, Newport EL.** 1996. Statistical Learning by 8-Month-Old Infants. Science **274**:1926–1928.

61. **Saffran JR, Johnson EK, Aslin RN, Newport EL.** 1999. Statistical learning of tone sequences by human infants and adults. Cognition **70**:27–52. doi:10.1016/S0010-0277(98)00075-4

62. **Santolin C, Rosa-Salva O, Regolin L, Vallortigara G.** 2016. Generalization of visual regularities in newly hatched chicks (Gallus gallus). Anim Cogn **19**:1007–1017. doi:10.1007/s10071-016-1005-2

63. **Schapiro Anna C., Rogers TT, Cordova NI, Turk- NB, Botvinick MM.** 2013. Neural representations of events arise from temporal community structure **16**:486–492. doi:10.1038/nn.3331.Neural

64. **Schapiro Anna C, Rogers TT, Cordova NI, Turk-Browne NB, Botvinick MM.** 2013. Neural representations of events arise from temporal community structure. Nat Neurosci **16**:486–492. doi:10.1038/nn.3331

65. **Schapiro AC, Turk-Browne NB, Norman KA, Botvinick MM.** 2016. Statistical learning of temporal community structure in the hippocampus. Hippocampus **26**:3–8. doi:10.1002/hipo.22523

66. **Sonnweber R, Ravignani A, Fitch WT.** 2015. Non-adjacent visual dependency learning in chimpanzees. Anim Cogn **18**:733–745. doi:10.1007/s10071-015-0840-x

67. **Stachenfeld KL, Botvinick MM, Gershman SJ**. 2017. The hippocampus as a predictive map. Nat Neurosci **20**:1643–1653. doi:10.1038/nn.4650

68. **Stiso J, Lynn CW, Kahn AE, Rangarajan V, Szymula KP, Archer R, Revell A, Stein JM, Litt B, Davis KA, Lucas TH, Bassett DS.** 2022. Neurophysiological Evidence for Cognitive Map Formation during Sequence Learning. eNeuro **9**:ENEURO.0361-21.2022. doi:10.1523/ENEURO.0361-21.2022

69. **Tacikowski P, Kalender G, Ciliberti D, Fried I.** 2024. Human hippocampal and entorhinal neurons encode the temporal structure of experience. Nature 1–8. doi:10.1038/s41586-024-07973-1

70. **Takahasi M, Yamada H, Okanoya K.** 2010. Statistical and Prosodic Cues for Song Segmentation Learning by Bengalese Finches (Lonchura striata var. domestica): Song segmentation learning in finches. Ethology **116**:481–489. doi:10.1111/j.1439-0310.2010.01772.x

71. **Toro JM, Sinnett S, Soto-Faraco S.** 2005. Speech segmentation by statistical learning depends on attention. Cognition **97**:25–34. doi:10.1016/j.cognition.2005.01.006

72. **Toro JM, Trobalón JB.** 2005. Statistical computations over a speech stream in a rodent. Perception and Psychophysics **67**:867–875. doi:10.3758/BF03193539

73. **Tummeltshammer K, Amso D, French RM, Kirkham NZ.** 2017. Across space and time: infants learn from backward and forward visual statistics. Dev Sci **20**:e12474. doi:10.1111/desc.12474

74. **Wacongne C, Changeux JP, Dehaene S.** 2012. A neuronal model of predictive coding accounting for the mismatch negativity. Journal of Neuroscience **32**:3665–3678. doi:10.1523/JNEUROSCI.5003-11.2012

75. **Xu C, Li H, Gao Jiaxin, Li L, He F, Yu J, Ling Y, Gao Jian, Li J, Melloni L, Luo B, Ding N**. 2022. Statistical learning in patients in the minimally conscious state. Cerebral Cortex bhac222. doi:10.1093/cercor/bhac222