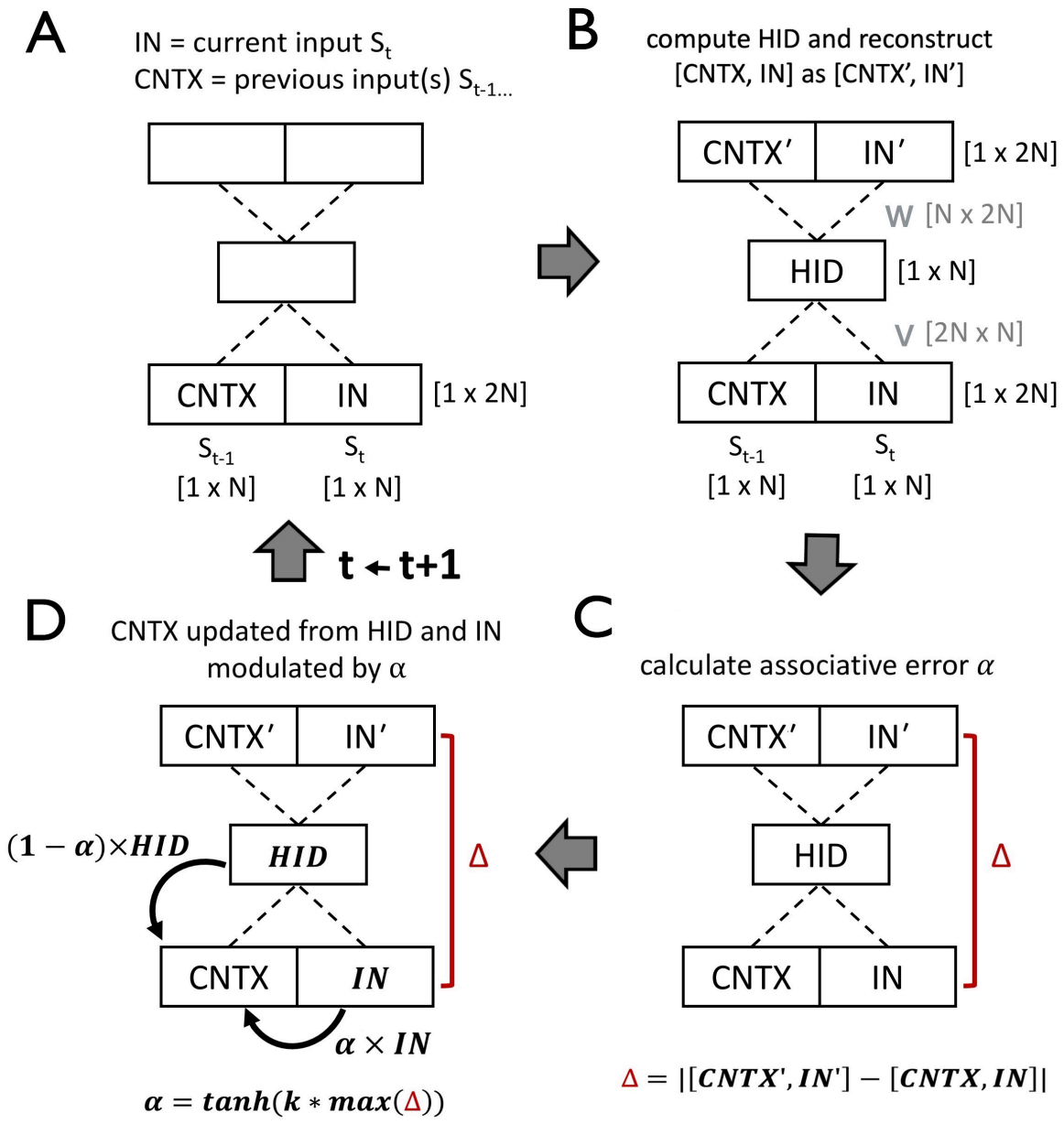


Figure S1. Architecture and updating of a local autoencoder-in-time (AT) unit. Related to STAR Methods and Figure 5. (A) At time t, the input layer of the AT unit is a 1-by-2N vector which contains both the present information S_t in the IN bank and the past information S_{t-1} in the CNTX bank. **(B)** The concatenated vector [CNTX, IN] is multiplied by weight matrix V to form a low-dimensional HID representation (a 1-by-N vector). This HID vector is left-multiplied by a weight matrix W to generate an output layer [CNTX', IN'] which is the reconstruction of input [CNTX, IN]. **(C)** The reconstruction error, Δ, or "surprise", is calculated as the absolute value of [CNTX', IN']- [CNTX, IN]. **(D)** The gating parameter, α, is calculated as tanh(*k**max(*Δ*)). Here, the parameter *k* scales how much the contribution of IN to CNTX is increased by surprise. The CNTX vector is updated as a linear mixture of the IN vector and HID vector, with the linear proportions modulated by α and a level-specific time constant τ. After CNTX is updated, the cycle is complete, and the unit is ready to receive input at time (t+1). IN = input unit, CNTX = context unit, HID = hidden state unit.

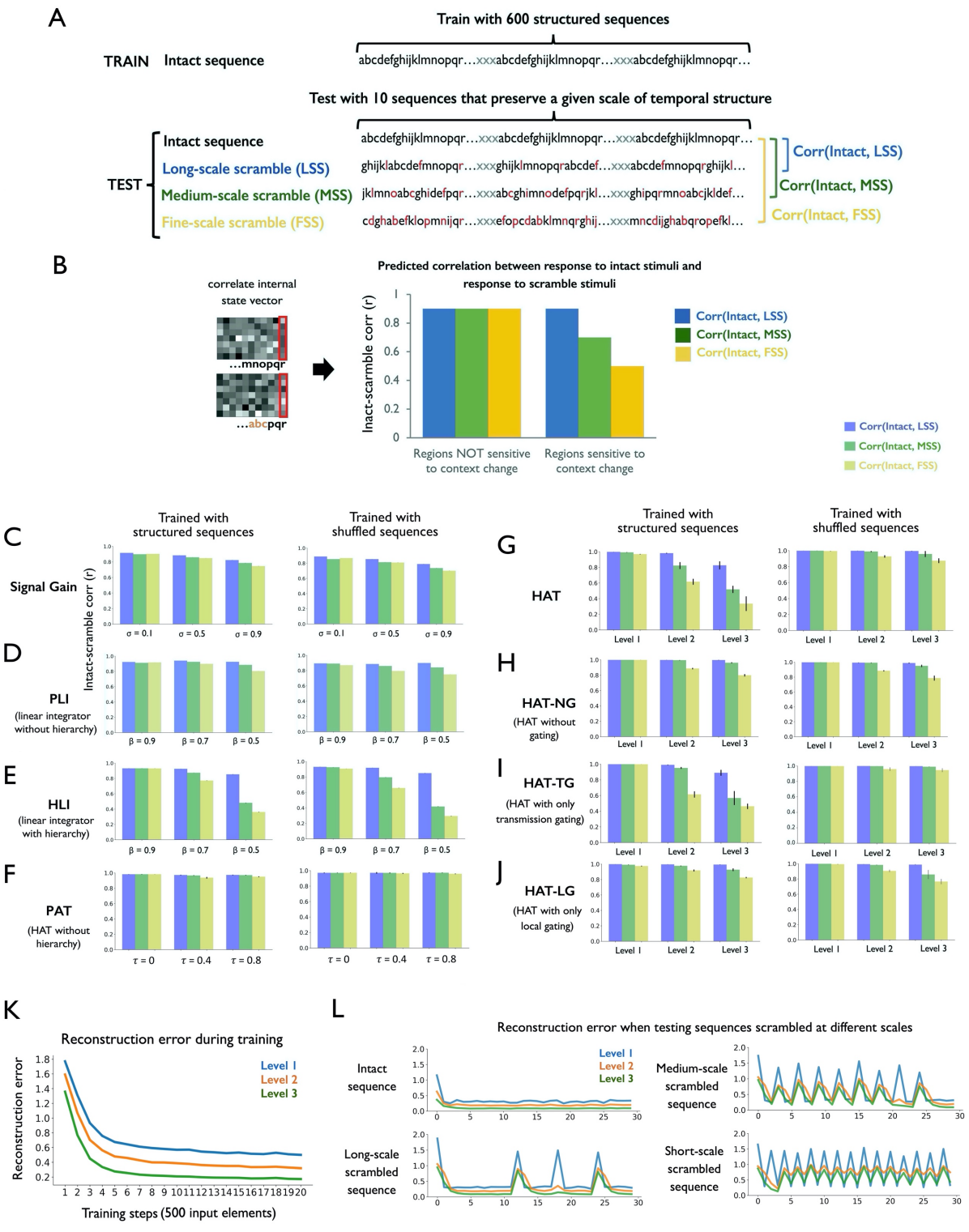


Figure S2. The signal gain model, linear integrator model variants and active integrator model variants account for prior data on hierarchical context dependence. Related to STAR Methods and Figures 1 and 5. (A) Example of training sequences (intact sequences) and testing sequences (long scale, medium scale and fine scale scrambled sequences). Context dependence was measured by correlating the hidden

representation between the intact and different levels of scrambled sequences. The target element (i.e. the last element of each sub-sequence) for correlation is marked with red. **(B)** The predicted correlation of hidden representations in regions that are more / less sensitive to temporal context. **(C)** The signal gain model was able to account for both of the key empirical phenomena of hierarchical context dependence (P1 and P2, above). Because the noise added to the internal representations varied in magnitude across processing stages and across levels of scrambling, the signal gain model generated the pattern of hierarchy of context dependence: the higher "stages" of the signal gain model generated lower correlation between intact and scrambled stimuli (left). We observed a similar pattern when testing the model on temporal structures that it was never trained on (right). One could plausibly amend the signal gain model to generate a learning effect, by positing that the noise level is increased when processing unfamiliar stimuli. Overall, we conclude that the signal gain model could account for the phenomenon of hierarchical context dependence. **(D)** The PLI model showed more sensitivity to context change when the β parameter was decreased (i.e. when the model preserved more temporal context, analogous to the higher-levels of a hierarchical model). However, this context-dependence effect was not specific to sequences that were seen during training – it was also observed when training and testing employed completely different sequences. **(E)** The HLI model showed more sensitivity to context change at higher levels. This hierarchical context dependence effect was stronger than in the PLI model. However, this context dependence was not specific to sequences that were seen during training. **(F)** The PAT model trained with structured sequences showed more sensitivity to context change when the τ parameter was increased (i.e. when the model preserved more hidden representation, analogous to the higher-level circuit). The effect was mild and was not specific to sequences that were seen during training. **(G)** The HAT model, when trained with structured sequences, exhibited a hierarchy of context dependence across different levels of the model. Importantly, this context dependence effect in HAT was much stronger when the model was trained and tested on the same sequences. In other words, the context dependence in HAT depends on the model's learning of temporal structure. **(H)** The HAT variant with no gating mechanism (HAT-NG) showed a similar pattern to the PLI and signal gain results: the higher levels of the model showed more context dependence, but the pattern generated was not specific to the structure of the training sequences. **(I)** The HAT variant with only transmission gating (HAT-TG), when trained with structured sequences, showed a hierarchy of context dependence across different levels of the model. This context effect was even stronger when the model was trained and tested on the same sequences. **(J)** The HAT variant with only local gating (HAT-LG) showed a hierarchy of context dependence across different levels of the model, but the pattern generated was not specific to the structure of the training sequences. **(K)** During training with the structured sequences, all levels of the HAT model exhibited a decrease in reconstruction error Δ with increasing training duration. **(L)** Layer-specific reconstruction error in the HAT model when testing with different levels of scrambled sequences: intact sequence, long-scale scrambled sequences, medium-scale scrambled sequences and fine-scale scrambled sequences. More finely scrambled sequences generated larger reconstruction error signals. LSS = long scale scramble, MSS = medium scale scramble, FSS = fine scale scramble, PLI = parallel linear integrator, HLI = hierarchical linear integrator, PAT = parallel autoencoder-in-time, HAT = hierarchical autoencoders in time.

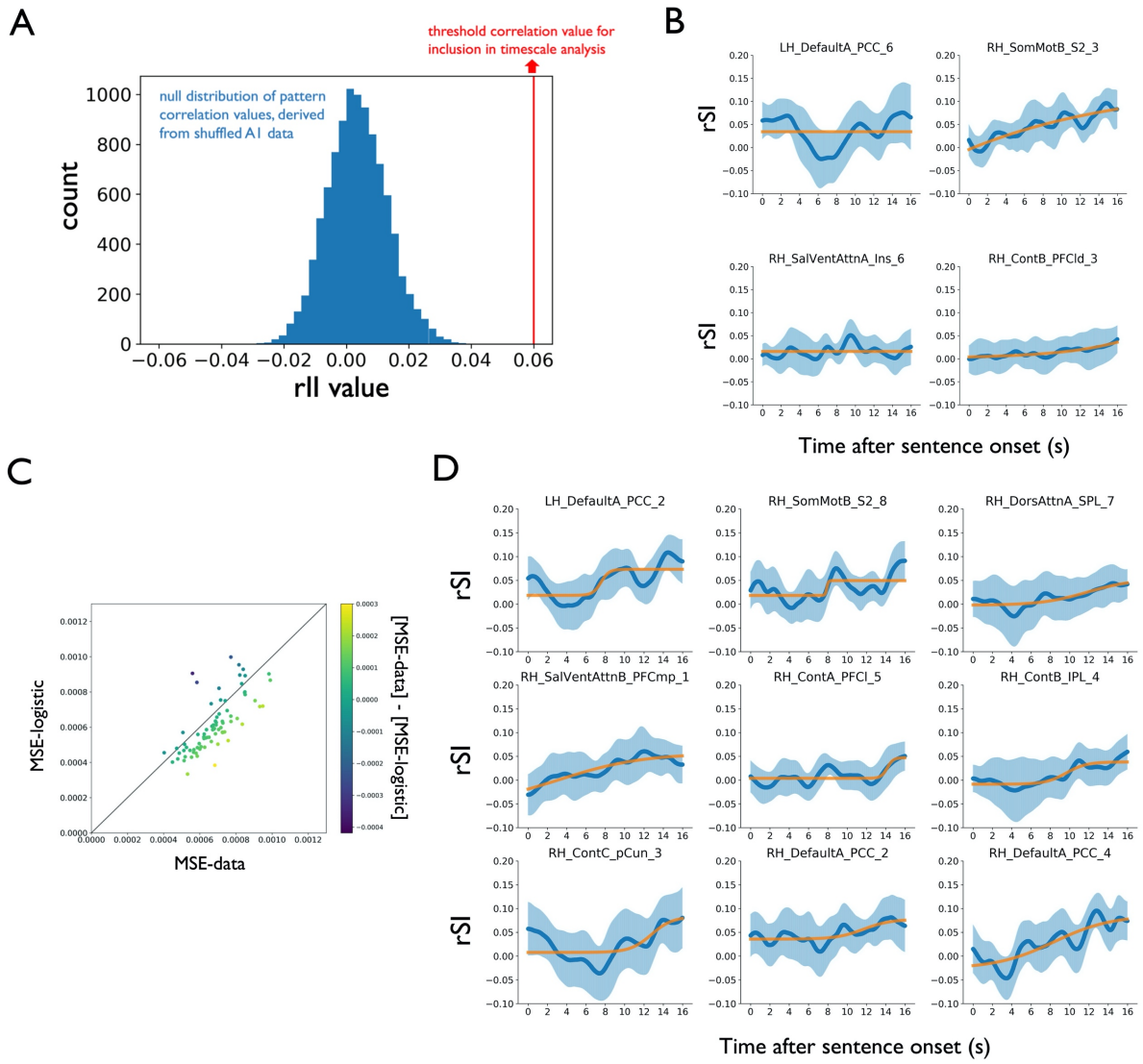


Figure S3. Validation of ROIs for ISPC analysis and logistic model fits - raw rSI and logistic fitting curves of a set of excluded ROIs. Related to STAR Methods. (A) The within-group ISPC (rII) was computed within an auditory cortex "A1+" parcel, which was functionally defined in a separate naturalistic narrative dataset (Simony et al., 2016). The surrogate distribution of rII values was computed by computing ISPC against non-matching sentences (shuffling the sentence order, see Supplemental Methods Section 4). In order to visualize the most meaningful timescale parameters in regions that responding reliably (in Figures 2 and 3), we chose a threshold of rII=0.06. This threshold was not chosen in order to correspond to an arbitrary statistical threshold, but nonetheless it is clear that rII= 0.06 lies far outside the null distribution of rII values. Thus, we used 0.06 as a conservative threshold for ROIs that showed reliable stimulus-locked response. The ROIs included in Figures 2, 3 and 4 (all exhibiting rSS > 0.06) generated a reliable stimulus-locked response to the scrambled stimulus. **(B)** A set of 4 anatomical regions of interest (ROIs) in which the parameters of the logistic function could not be confidently recovered after fitting the rSI curves. We visually identified parcels in which the rSI curve did not appear to follow a logistic curve. These parcels occur near left posterior cingulate cortex, right somatomotor cortex, right insula and the right prefrontal cortex. **(C)** The out-of-sample mean square error (MSE) in predicting rSI curves using either the raw in-sample curve or logistic model fit to the in-sample data. The error was measured using a split- half cross-

validation method (STAR Methods). For all ROIs, the MSE is similar when predicting the out-of-sample rSI curve with predictions from a logistic model and when using the raw in-sample rSI curve. For most ROIs, the error from the logistic fit is actually lower than from the raw in-sample data. This suggests that the logistic function is a valid model for the rSI curves. **(D)** A set of 9 anatomical ROIs in which the alignment time quantified by logistic fitting was not reliable across subjects. We identified these parcels by bootstrapping the logistic function parameters. The alignment time was computed for each fold of the bootstrap, from which we derived a distribution of alignment values. When the 5^{th}-95^{th} percentile range was more than 6s, we considered the ROI unreliable. The unreliable parcels occurred near bilateral posterior cingulate cortex, right precuneus, right prefrontal cortex, right superior and inferior temporal lobule, and right somatomotor cortex. The parcels are individually labeled with their names from the Schaefer parcellation (Schaefer et al., 2018). A1 = primary auditory cortex, rII = intact-intact inter-subject pattern correlation, rSS = scramble-scramble inter-subject pattern correlation, rSI = scramble-intact inter-subject pattern correlation.

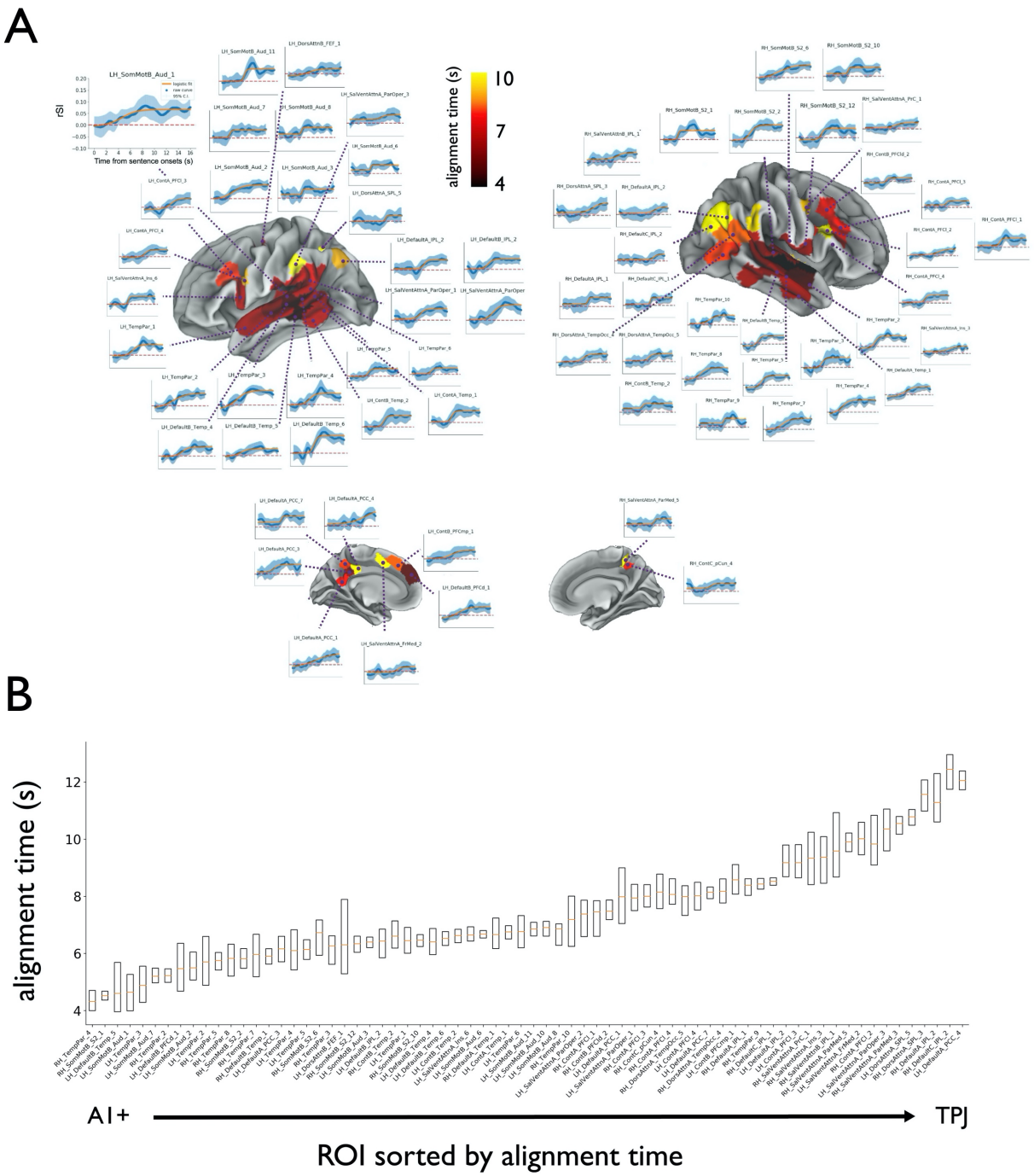


Figure S4. The temporal profiles of context construction mapped for each ROI individually. **Related to Figure 3. (A)** The raw $rSI_{DE:CE}$ curves (blue curves) are overlaid with their corresponding logistic fits (orange lines) for each ROI. The shaded blue area indicates a parametric 95% confidence interval on each rSI measurement at each time point. Colors on the cortical map indicate the alignment time from the logistic fits. **(B)** Bootstrapped alignment time in individual ROIs. For each ROI, the orange line shows the median of 1000 bootstrapped alignment time. The upper line of the box indicates the last data point less than Q3, and the lower line indicates the last data point higher than Q1. rSI = intact-scramble inter-subject pattern correlation. ROI = region of interest, Q3 = third quartile, Q1 = first quartile.

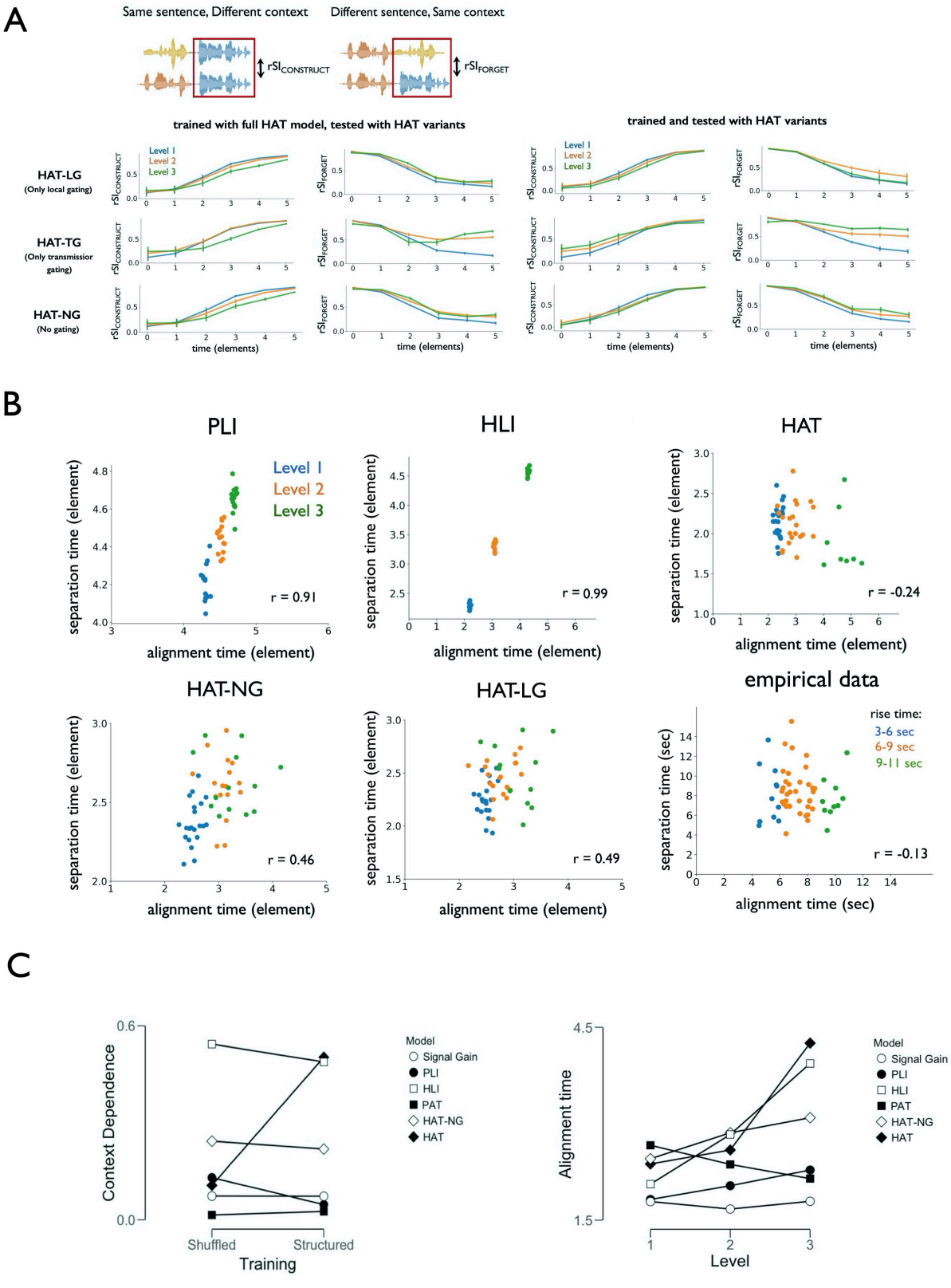
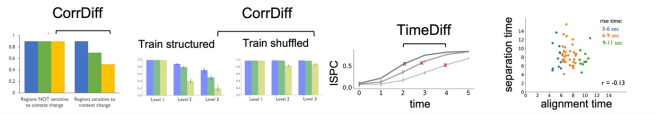


Figure S5. Predictions of temporal integration phenomena by different computational models. Related to Figure 5. (A) Predictions of context construction ($rSI_{CONSTRUCT}$) and context forgetting (rSI_{FORGET}) for

variants of the HAT model with limited gating mechanisms. (left) Each of these HAT variants was trained with the full HAT model (all mechanisms intact) but they were then tested with a limited gating mechanism (Supplemental Methods), to evaluate the effect of gating. $rSI_{CONSTRUCT}$ and rSI_{FORGET} generated by HAT-LG (HAT with only local gating, top), HAT-TG (HAT with only transmission gating, middle) and HAT-NG (HAT with neither local gating nor transmission gating, bottom). (right) Each of these HAT variants was both trained and tested with a limited gating mechanism: $rSI_{CONSTRUCT}$ and rSI_{FORGET} generated by HAT-LG (top), by HAT-TG (middle) and HAT-NG (bottom). For HAT-LG and HAT-NG, the performance patterns resembled that of the linear integrator models, in which the higher levels of the model showed both longer alignment time and longer separation times. For the HAT-TG model, we found that the higher levels of the model generated somewhat nonspecific sequence representations (similar internal representations across diverse inputs). The patterns of the HAT-TG model were difficult to confidently interpret, because the separation curves were no longer logistic. Nonetheless, it appeared that HAT-TG was again similar to linear integrator models, which could reproduce the hierarchically varied alignment time for context construction, but not the distinct separation times for context forgetting. Note that higher levels of the HAT model generated more nonspecific representations when transmission gating was removed: this implies that higher levels of the model can more readily identify the beginning of a distinct new sequence when they receive "surprise" signals generated from the levels below. **(B)** Model predictions and empirical results of correlation between alignment time (time for integrating prior information) and separation time (time for forgetting prior information). PLI predicted that the alignment time is positively correlated with the separation time ($r=0.91$, $p<0.0001$). HLI predicted that the alignment time is positively correlated with the separation time ($r=0.99$, $p<0.0001$). HAT predicted that the alignment time is not correlated with the separation time ($r=-0.24$, $p=0.08$). HAT-NG predicted that the alignment time is positively correlated with the separation time ($r=0.46$, $p=0.0003$). HAT-LG predicted that the alignment time is positively correlated with the separation time ($r=0.49$, $p<0.0001$). The empirical results showed that for each individual ROI, the alignment time for context construction was not correlated with the separation time for context forgetting in that ROI ($r=-0.13$, $p=0.3$). Overall, we found that only the HAT model (which showed no correlation between alignment time and separation time) was compatible with the empirical results. Thus, within the set of models tested, context gating mechanisms are essential for capturing the empirical dissociation between alignment time and separation time. **(C)** Measurement of changes in context dependence with learning (left), and alignment time across levels (right) for a suite of models and model variants. (left) Model performance generating the phenomenon of hierarchical context dependence (Figure S2) when tested on familiar sequences (structured) or unfamiliar sequences (shuffled). (right) Model performance in reproducing the phenomenon of hierarchical context construction, which manifests as a level-by-level increase in alignment times. Only four models successfully generated the hierarchically varied alignment time, as quantified by logistic fitting – the PLI model, the HLI model, the HAT-NG model and the HAT full model. Out of the four models, the alignment time difference was much larger for HLI (Δ Alignment Time = 1.88) and HAT model (Δ Alignment Time = 1.88) than for the PLI model (Δ Alignment Time = 0.46) and HAT-NG model (Δ Alignment Time = 0.64). The signal gain model showed no difference between alignment time at Level 1 and Level 3 of the model. As noted above, the PAT model did not successfully learn distinct internal representations for different sequence items; as a result is showed larger alignment time at Level 1 than Level 3: this occurred because the ISPC values in the PAT model were higher for Level 3 than for other levels (even for random pairs of stimuli), this biased the alignment time downward, because the alignment curve ramped upward from a very high baseline value. HAT = hierarchical autoencoders in time, rSI = scramble-intact inter-subject pattern correlation, PLI = parallel linear integrator, HLI = hierarchical linear integrator, HAT-NG = HAT with no gating mechanism, PAT = parallel autoencoders in time.



Models	hierarchical architecture	nonlinear temporal integration	context gating mechanism	Hierarchical context dependence Δ Corr(Level 1, Level 3)	Learning effect Δ Corr(Structured, Shuffled)	Hierarchically varied alignment time Δ Alignment Time (Level 1, Level 3)	Alignment Time ≠ Separation Time Corr (Alignment Time, Separation Time)
Signal Gain Model	✗	✗	✗	0.07	0.00	0.00	N/A
Parallel Linear Integrator (PLI)	✗	✗	✗	0.05	-0.08	0.46	0.88
Hierarchical Linear Integrator (HLI)	✓	✗	✗	0.49	0.05	1.88	1.00
Parallel Autoencoders in time (PAT)	✗	✓	✗	0.03	0.01	-0.51	N/A
HAT-no gating (HAT-NG)	✓	✓	✗	0.22	0.02	0.64	0.71
Hierarchical Autoencoders in time (HAT)	✓	✓	✓	0.50	0.40	1.88	-0.24

Table S1. Comparison of model performance across a set of four empirical phenomena: hierarchical context dependence; learning-dependent integration; hierarchical alignment times; and the decoupling of alignment time and separation time. Related to STAR Methods and Figure 5. Summary of model architectures, along with performance across four metrics. Hierarchical context dependence (first column of model results) is the basic hierarchical context phenomenon reported in prior studies and modeled in Figure S2. The learning effect (second column of model results) is a measure of whether hierarchical context dependence is selectively observed for familiar (structured) vs unfamiliar (shuffled) sequences, as tested in Figure S2B and Figure S5C. The hierarchical variation in alignment time (third column of model results) is a measure of the difference in alignment time between lower and upper levels of the model, intended to capture the empirical phenomenon of hierarchical context construction (Figure 3) modeled in Figure S5C. Finally, the absence of a correlation between alignment time and separation time (fourth column of model results) is the phenomenon reported here (Figure 5) and modeled in Figure S5B. *Comparison across models*: By examining the performance across each class of model, we can begin to infer the essential computational roles of hierarchical architecture, context gating, and nonlinear integration. *Hierarchical architecture:* In order to test the importance of hierarchical architecture, we can compare the HLI and HAT models against variants which operate with "parallel" levels. In these models (i.e. PLI and PAT), each unit receives input directly from the environment, but integrates the information with distinct time constants. These "parallel" integrator models produced smaller context effects than the equivalent models with a stage-by-stage integration process (Figure S5C). Thus, in light of the anatomical evidence for hierarchical organization, hierarchical processing appears to be an important feature for temporal processing. *Context gating*: Is a variable time constant (i.e. gated integration with prior context) necessary to account for the data? When we tested HAT variants with reduced or absent gating mechanisms, they generated predictions similar to the linear integrator models: they exhibited hierarchical context effects, but alignment time and forgetting time were robustly correlated ($r = 0.46$, $p = 0.0003$, Figure S5B and Figure S5A, right). In addition, HAT variants that lacked gating learned less distinct representations of sequence elements (Figure S5A) and their integration processes were less affected by sequence familiarity (Figure S5C). Thus, a context gating mechanism appears essential for HAT-like models to capture the empirical data. *Nonlinear integration*: Is nonlinear integration necessary to account for the overall pattern of empirical data? By nonlinear integration, we mean models in which the state at time t+1 is a nonlinear function of the state at t and the input at t. A nonlinear integration mechanism is not sufficient to account for the

empirical data on its own, because HAT variants without gating (but with nonlinear integration) exhibited small learning effects and their alignment times correlated with their separation times (Figure S5B, C). However, none of the linear models could account for the full pattern of data (i.e. the four rightmost columns in this table), and the full HAT model, which incorporated nonlinearity, was the most successful. Thus, within the panel of models tested, nonlinear integration is not sufficient to account for the data, but it improves model performance when combined with a context gating mechanism