

Neuron, Volume 114

Supplemental information

**Representational similarity modulates neural
and behavioral signatures of novelty**

Sophia Becker, Alireza Modirshanechi, and Wulfram Gerstner

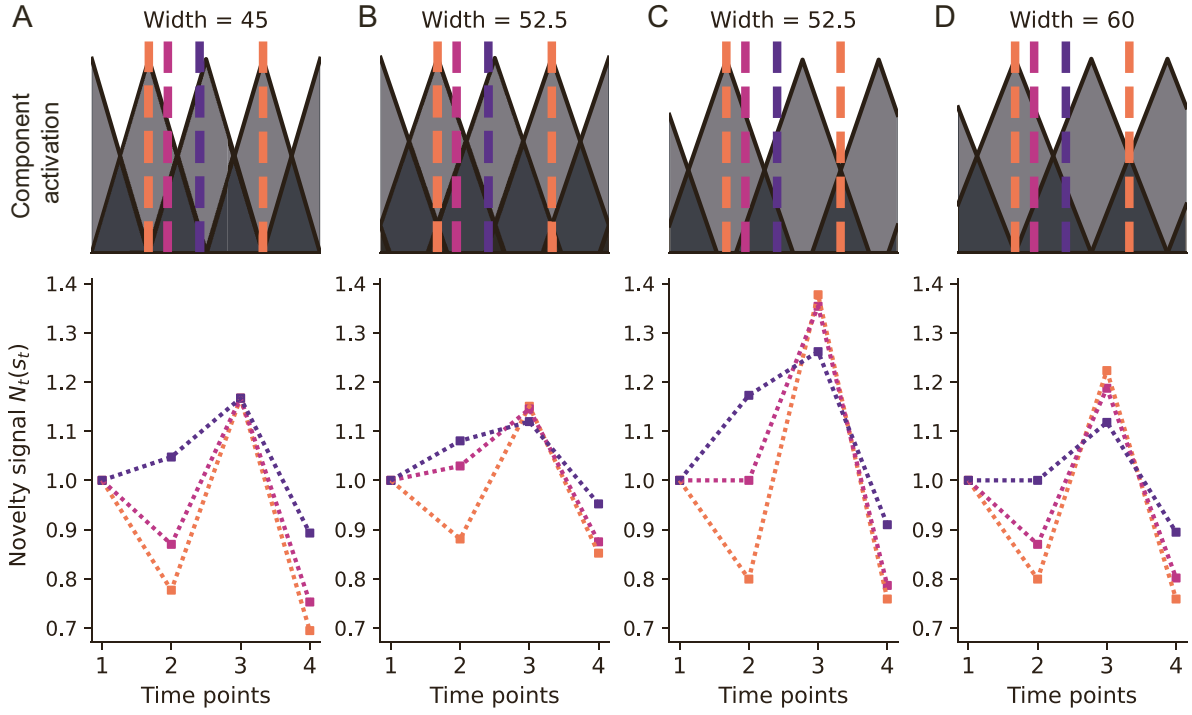


Figure S1: Similarity-based novelty predictions for different component widths (Gabor toy example), related to Figure 1. (A-B) As we increase the width σ_j of each of the four triangular components from $\sigma_j = 45^\circ$ (A) to $\sigma_j = 52.5^\circ$ (B), the third stimulus in the Seq. 3 (purple) also starts to be affected by generalization; it has a lower novelty because it shares similarities with the second stimulus. In Seq. 1-2, the similarity between previous stimuli and the third stimulus is lower, such that its effect on the novelty of the third stimulus is very small (Seq. 2, magenta) or absent (Seq. 1, orange), even for components of width $\sigma_j = 52.5^\circ$. **(C-D)** Using even wider components ($\sigma_j = 60$) enhances the effect of similarities on the novelty of the third stimulus. In panel D, the number of components ($N = 3$) is adjusted such that each stimulus in the interval $[0^\circ, 180^\circ]$ is covered (approximately) equally by components.

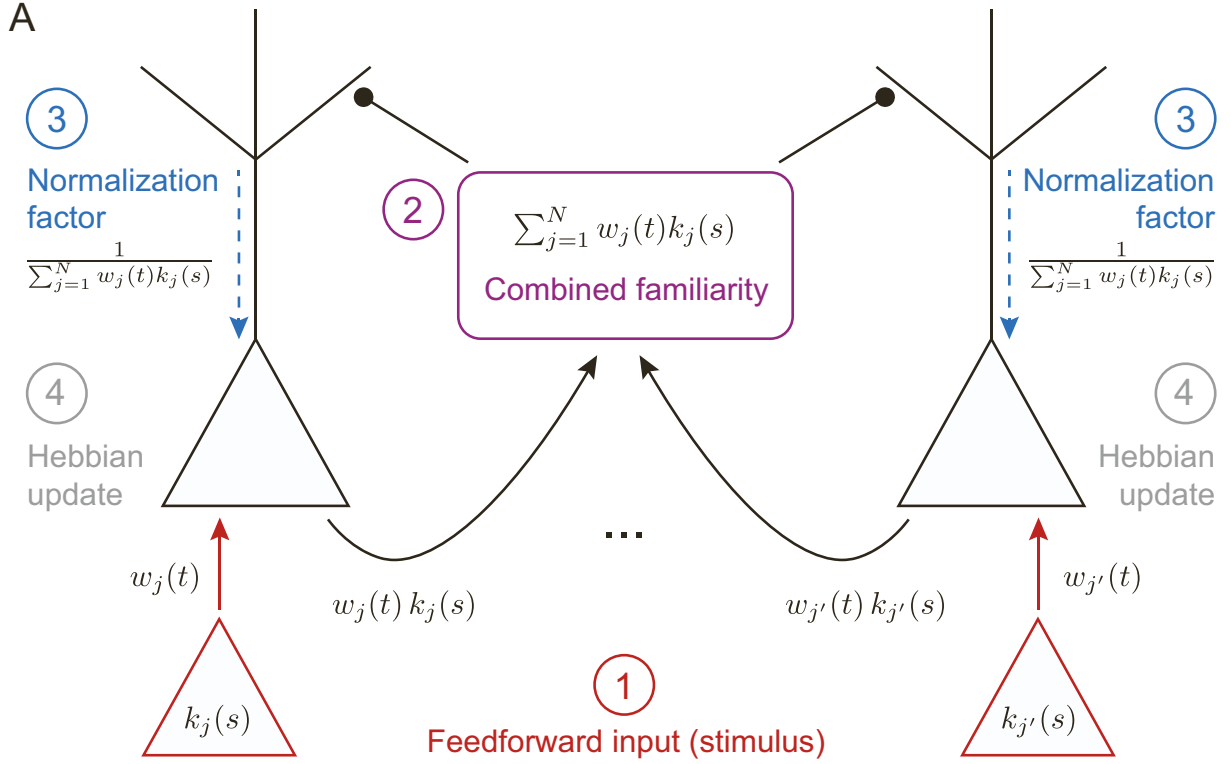


Figure S2: Candidate circuit to implement similarity-based novelty based on shunting inhibition, related to STAR Methods. Component activations $k_j(s)$ are transmitted as presynaptic activations through plastic synapses with weight $w_j^{(t)}$ to postsynaptic pyramidal cells. The postsynaptic activations $w_j^{(t)}k_j(s)$ are transmitted to a shared inhibitory population, from which the combined signal $\sum_j w_j^{(t)}k_j(s)$ is sent back to the pyramidal neurons in the form of dendritic disinhibition. In the presence of shunting inhibition, this dendritic input acts as the normalization factor $1 / \left[\sum_j w_j^{(t)}k_j(s) \right]$ in the Hebbian update of the pre-to-postsynaptic weights $w_j^{(t)}$.

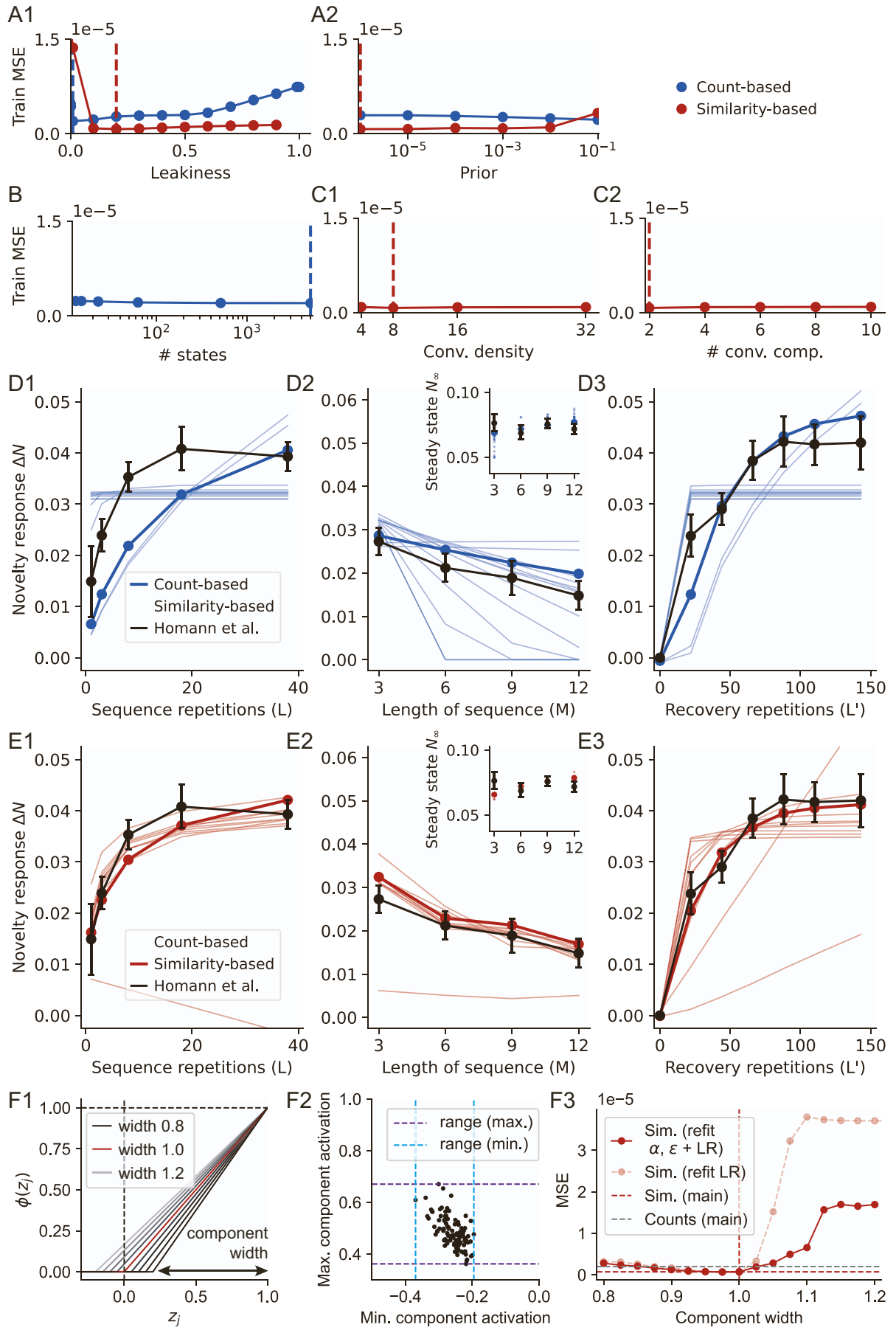


Figure S3: Parameter sensitivity for leaky count-based and leaky similarity-based novelty in the Homann experiment, related to Figure 2. Caption on next page

Figure S3: Parameter sensitivity for leaky count-based and leaky similarity-based novelty in the Homann experiment, related to Figure 2. (A-C) MSE between experimental data¹ and fitted count-based (blue) and similarity-based novelty (red), where one parameter at a time is perturbed from the fitted values: leakiness α (A1), prior ϵ (A2), number of states in count-based novelty (B), convolutional density c (C1) and number N of convolutional components (C2) for similarity-based novelty (STAR Methods). Dashed vertical lines indicate fitted value of the perturbed parameter. **(D-E)** Count-based novelty predictions (D, blue) and similarity-based novelty predictions (E, red) under the perturbations of α as depicted in A1 (thin lines in D-E) and for fitted parameters (bold lines in D-E). Black lines denote trial- and population-averaged novelty responses in mouse V1¹, with error bars showing the SEM across $n = 5$ mice. **(F)** Robustness of similarity-based novelty to changes in the component. F1: Triangular components of different widths that determine component activation $\phi(z_i)$ based on the similarity z_i of a given stimulus with reference Gabor i . F2: Component width are varied within a range such that randomly chosen images from the experiment by Homann et al.¹ have a sufficient probability of activating at least one component. F3: MSE between similarity-based novelty and experimental data, where the component width was perturbed as in F1-F2 (linear regression mapping to the experimental data refit). MSEs are shown under pure weight perturbation (light red) and under weight perturbation with refit of the novelty parameters α (leakiness) and ϵ (prior).

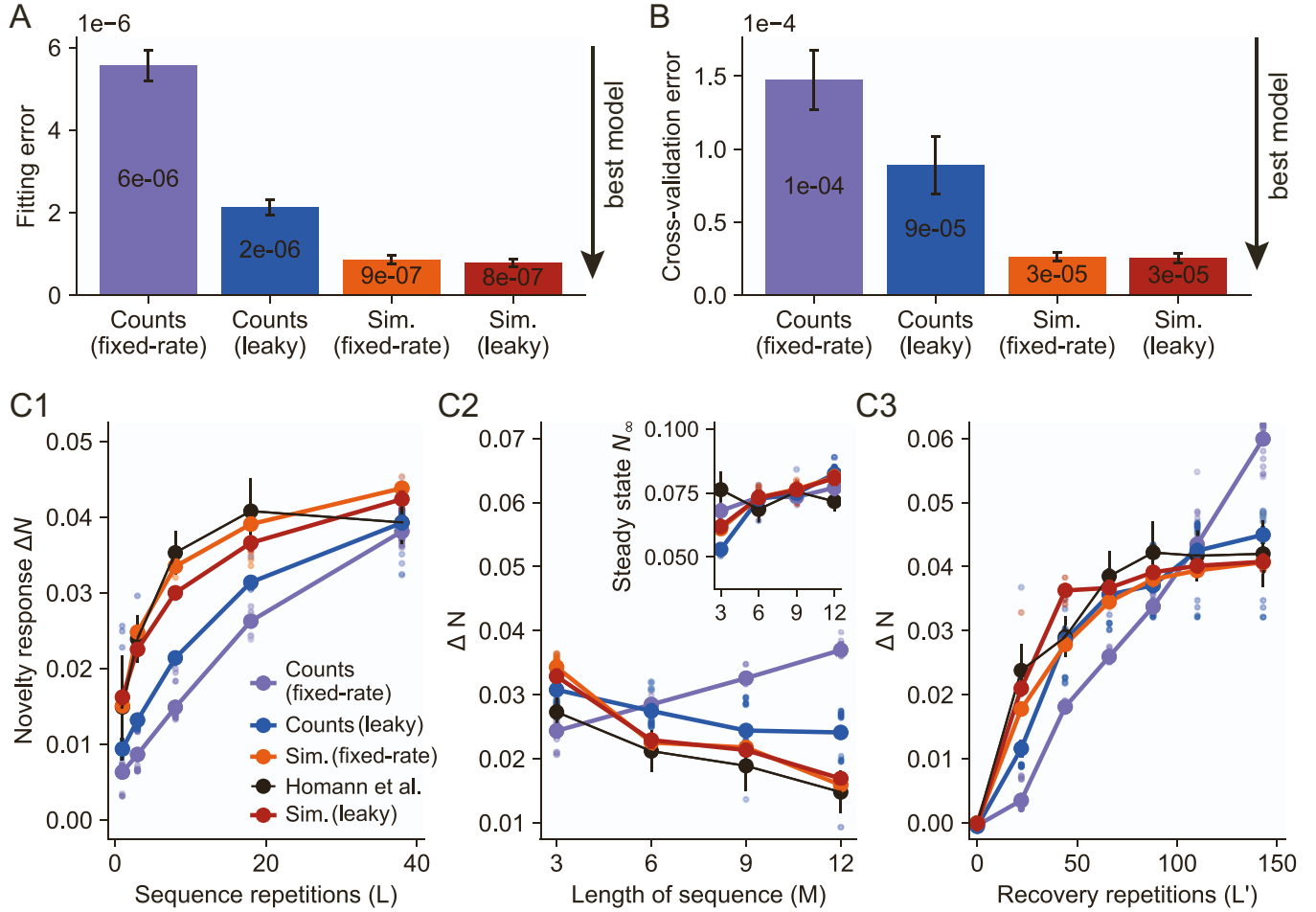


Figure S4: Leaky vs. fixed-rate count-based and similarity-based novelty in the Homann et al. experiment, related to Figure 2. (A-B) Fitted and leave-one-out cross-validated MSE between novelty models and experimental data, for leaky and fixed-rate variants of count-based and similarity-based novelty. Error bars show the SEM, estimated with jackknife resampling of the fitting and cross-validation process, respectively (STAR Methods). (C) Average novelty responses in mouse V1¹ (black line, error bars: SEM across $n = 5$ mice), and predictions of each of the four models (colored lines) on held-out data during the cross-validation. Individual colored data points show model predictions obtained by jackknife resampling of the cross-validation process. Leaky and fixed-rate similarity-based novelty both fit experimental data substantially better than count-based novelty. Fixed-rate count-based novelty captures experimental data worse than leaky count-based novelty, in particular in the M -experiment.

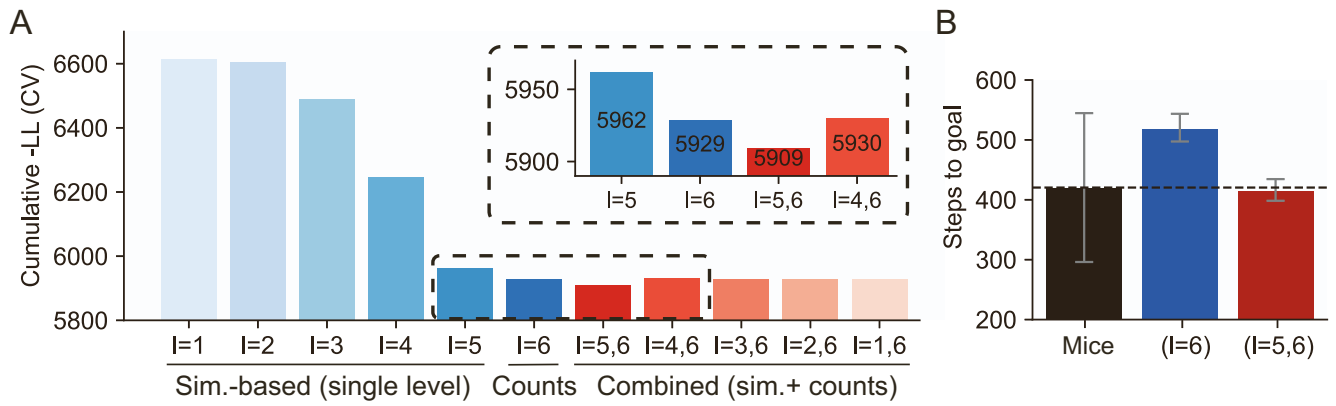


Figure S5: Cumulative cross-validated log-likelihoods and behavioral statistics for all models fitted to the Rosenberg data², related to Figure 3. (A) Cumulative negative LL across cross-validation folds ('CV-LL') in 5-fold cross-validation for (i) count-based novelty ($l = 6$ components), (ii) similarity-based novelty ($l = 1, \dots, 5$ components) and (iii) combined count-similarity-based novelty ($l = 6$ and $l = 1, \dots, 5$ components). Since the CV-LL can be interpreted as log-Bayes factors, differences greater than $\log(20)=3$ and $\log(150)=5$ are considered as significant and strongly significant^{3,4}, respectively. (B) Average steps to goal for mice (black), count-based novelty agent (blue) and the best combined count-similarity-based agent ($l = 5$ and $l = 6$ components). Average is computed across 20 mice and 400 simulations with different initial seeds for each agent. Error bars show bootstrapped SEM.

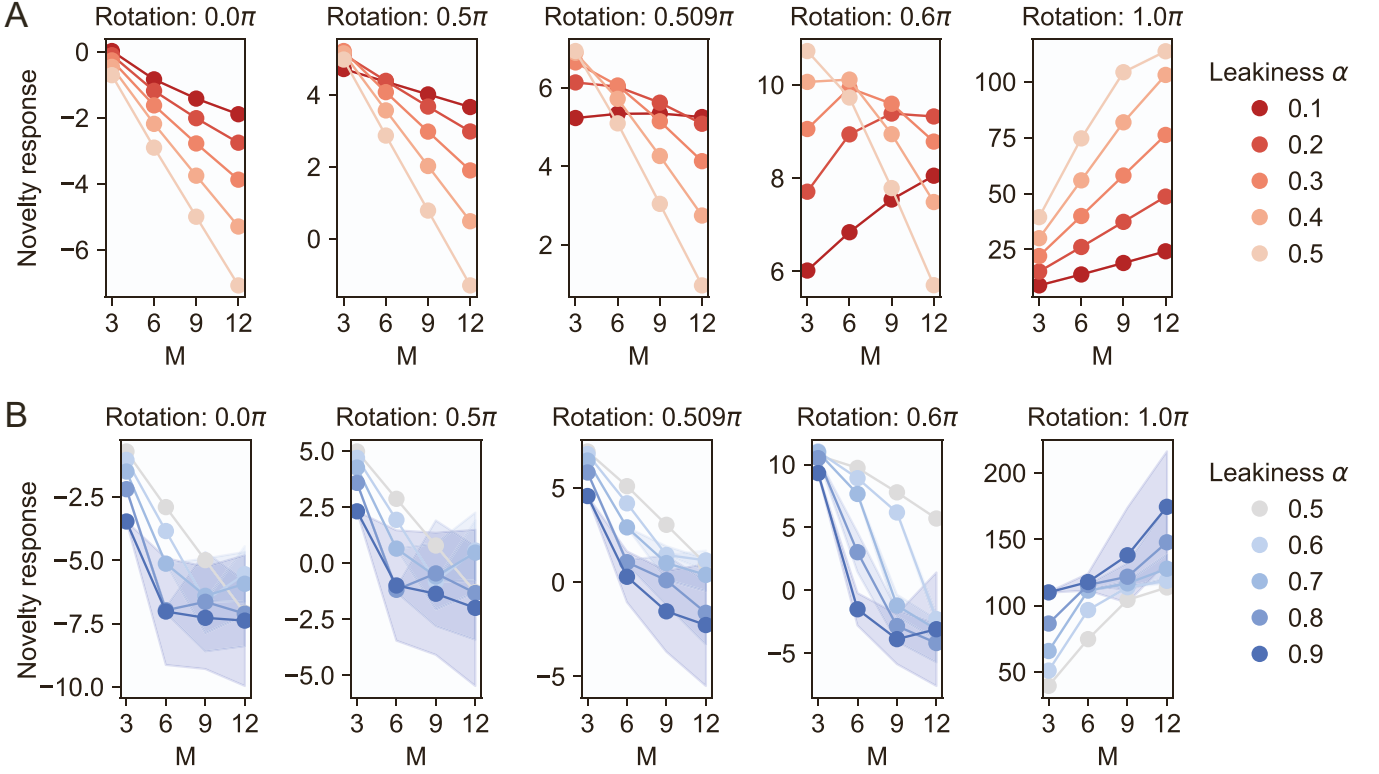


Figure S6: Robustness of similarity-based novelty predictions in the sum-of-parts protocol, related to Figure 4. Sum-of-parts predictions for different values of the leakiness α . The sum-of-parts protocol is simulated for the same five rotation values of the novel image X as in Fig. 4 (left to right panel: 0π , 0.5π , 0.509π , 0.6π , 1π). Each point depicts the predicted novelty response ΔN to the novel image, averaged across 20 simulations with different random permutations of the familiar image sequence (shaded area: SEM across simulations). **(A)** Sum-of-parts predictions are robust for $\alpha \leq 0.5$. The generalization threshold, i.e. the rotation value for which predicted novelty responses switch from the 'generalization' into the 'recency' regime, increase with α . **(B)** Sum-of-parts predictions for $\alpha \geq 0.5$ still robustly show two opposing regimes ('generalization' regime for rotation of 0π , leftmost panel; 'recency' regime for rotation of 1π , rightmost panel), but novelty predictions for large M vary across permutations of the familiar sequence since large α values quickly amplify small variabilities in the novelty responses.

Model	Familiarization mechanism	Generalization across stimuli ('similarity-modulation')	Normalization	Familiarity / novelty phenomenon
DeBaene & Vogels (2009) ⁵	Single-neuron adaptation (input fatigue)	Adaptation scales with similarity of adaptor to neuron's preferred orientation	–	Repetition suppression (monkey IT cortex)
Homann et al. (2022) ¹	Adaptation (gain modulation) in two-layer feedforward network	Lognormal random weights from 1-hot stimulus inputs	–	Novelty responses (mouse V1)
Mill et al. (2011) ⁶	Short-term synaptic plasticity (STP) in two-layer feedforward network	Poisson input population with Gaussian stimulus tuning	–	Stimulus-specific adaptation (rodent A1)
Aitken et al. (2024) ⁷	Familiarity-modulated synapses (STP or Hebbian plasticity)*	Gaussian random weights from 1-hot stimulus inputs	–	Novelty responses in VIP neurons (mouse V1)
Schulz et al. (2021) ⁸	Inhibition of familiar stimulus responses (inhibitory co-tuning + inhibitory spike-time dependent plasticity)	Broader stimulus tuning of inhibitory population	Implicit (activity-dependent shape of learning rule)	Novelty responses (mouse V1)
Bogacz & Brown (2003) ⁹	Anti-Hebbian plasticity, inhibitory Hebbian plasticity	Non-sparse, correlated input patterns	Implicit (anti-Hebbian learning)	Familiarity detection capacity
Tyulmankov et al. (2022) ¹⁰	Anti-Hebbian plasticity (meta-learned)	Correlated input patterns	Implicit (anti-Hebbian learning)	Continual familiarity detection capacity
Mohan et al. (2024) ¹¹	Depression, anti-Hebbian plasticity (inferred from firing rate distributions in RNNs)	Gaussian distributed input currents (iid for different stimuli)	Implicit (anti-Hebbian learning)	Familiarity modulation of stimulus responses (monkey IT cortex)

Table S1: Conceptual links between similarity-based novelty and existing mechanistic novelty models, related to STAR Methods. We focus on three key aspects of similarity-based novelty: (i) familiarization with stimuli during repeated presentation, (ii) generalization across similar stimuli ('similarity-modulation'), and (iii) normalization of familiarization effects. While the first two elements of similarity-based novelty have several plausible implementations in existing novelty models, most mechanistic models do not or only implicitly normalize the effects of familiarization.

References

1. Homann, J., Koay, S.A., Chen, K.S., Tank, D.W., and Berry, M.J. (2022). Novel stimuli evoke excess activity in the mouse primary visual cortex. *Proceedings of the National Academy of Sciences* *119*, e2108882119. doi: <https://doi.org/10.1073/pnas.2108882119>.
2. Rosenberg, M., Zhang, T., Perona, P., and Meister, M. (2021). Mice in a labyrinth show rapid learning, sudden insight, and efficient exploration. *eLife* *10*, e66175. doi: <https://doi.org/10.7554/eLife.66175>.
3. Kass, R.E., and Raftery, A.E. (1995). Bayes factors. *Journal of the American Statistical Association* *90*, 773–795.
4. Jeffreys, H. (1998). *The theory of probability*. OuP Oxford.
5. Baene, W.D., and Vogels, R. (2009). Effects of adaptation on the stimulus selectivity of macaque inferior temporal spiking activity and local field potentials. *Cerebral Cortex* *20*, 2145–2165. doi: <https://doi.org/10.1093/cercor/bhp277>.
6. Mill, R., Coath, M., Wennekers, T., and Denham, S.L. (2011). A neurocomputational model of stimulus-specific adaptation to oddball and markov sequences. *PLoS Computational Biology* *7*, e1002117. doi: <https://doi.org/10.1371/journal.pcbi.1002117>.
7. Aitken, K., Campagnola, L., Garrett, M.E., Olsen, S.R., and Mihalas, S. (2024). Simple synaptic modulations implement diverse novelty computations. *Cell Reports* *43*, 114188. doi: <https://doi.org/10.1016/j.celrep.2024.114188>.
8. Schulz, A., Miehl, C., Berry II, M.J., and Gjorgjieva, J. (2021). The generation of cortical novelty responses through inhibitory plasticity. *eLife* *10*, e65309. doi: <https://doi.org/10.7554/eLife.65309>.
9. Bogacz, R., and Brown, M.W. (2003). Comparison of computational models of familiarity discrimination in the perirhinal cortex. *Hippocampus* *13*, 494–524. doi: <https://doi.org/10.1002/hipo.10093>.
10. Tyulmankov, D., Yang, G.R., and Abbott, L.F. (2022). Meta-learning synaptic plasticity and memory addressing for continual familiarity detection. *Neuron* *110*, 544–557.e8. doi: <https://doi.org/10.1016/j.neuron.2021.11.009>.
11. Mohan, K., Pereira-Obilinovic, U., Srednyak, S., Amit, Y., Brunel, N., and Freedman, D. (2024). Visual familiarity learning at multiple timescales in the primate inferotemporal cortex. Preprint at bioRxiv. <https://doi.org/10.1101/2024.01.05.574412>.