Extracting representational content in deep learning models through second-order isomorphism-based tools

Rudramani Gyawali Singha, **Robert Kim**, and **Nuttida Rungratsameetaweemana**
Department of Biomedical Engineering, Columbia University, New York City, NY

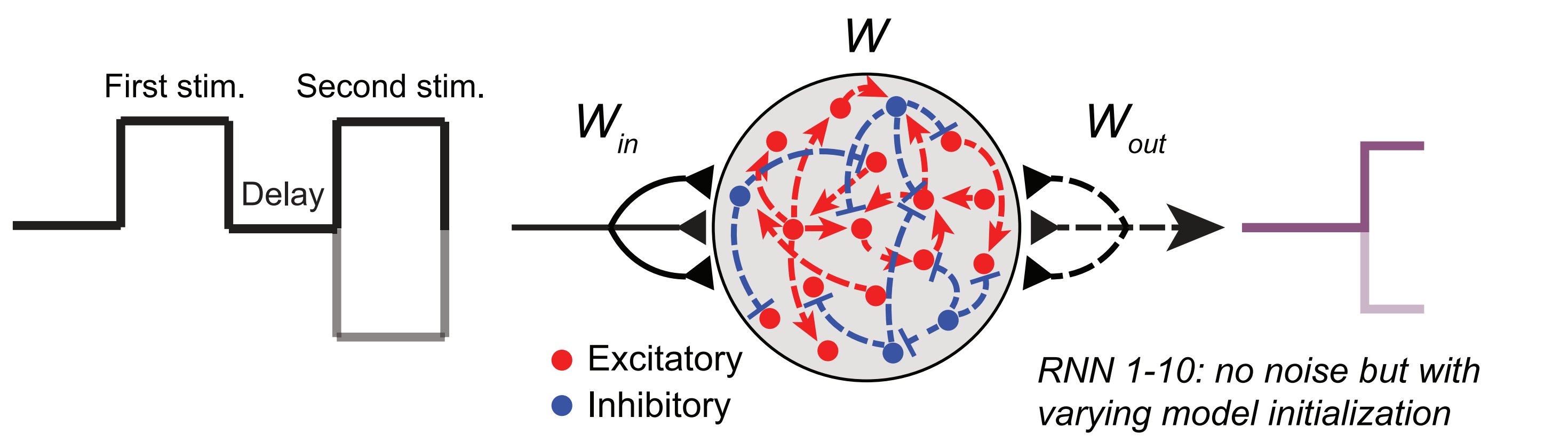
Columbia | Engineering
The Fu Foundation School of Engineering and Applied Science

Abstract

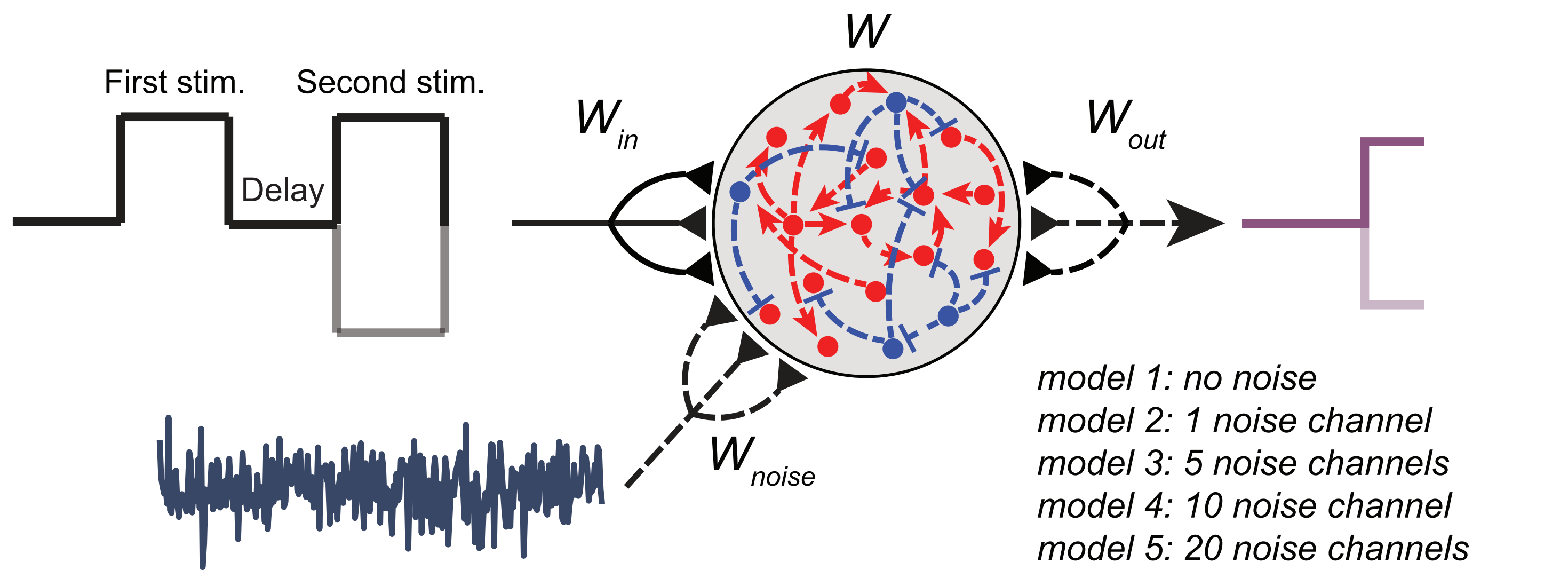
This research investigates the use of second-order isomorphism-based tools to extract representational content in deep learning models. The principle of second-order isomorphism is based on the idea that internal representations and dynamics reflect the structural similarities of external sensory stimuli. We explore the efficacy of representational similarity analysis (RSA) techniques in characterizing the latent semantic features of biophysically realistic deep learning models trained to perform cognitive tasks, such as working memory retention.

Deep Learning Models

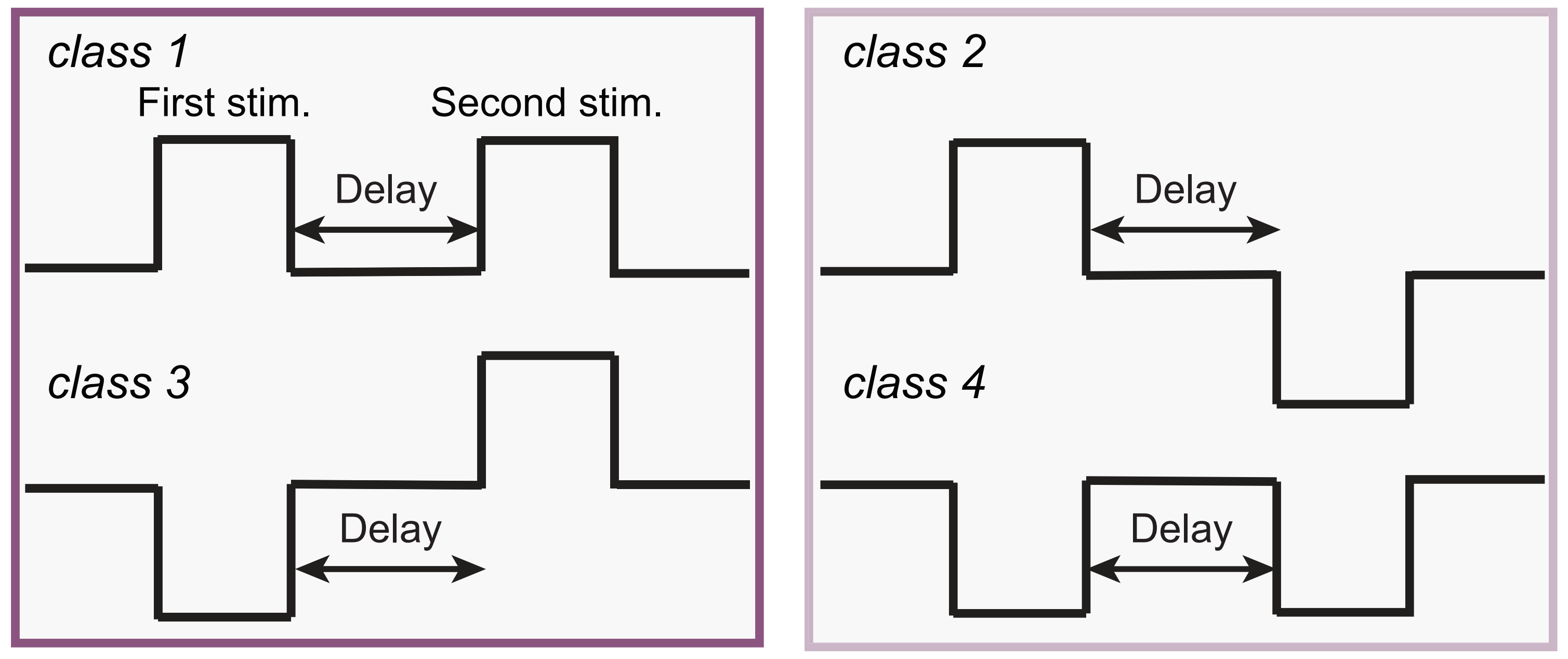
Baseline Models



Experimental Models



Sensory Classes



$$\tau_i \frac{dx_i}{dt} = -x_i(t) + \sum_{j=1}^{N} w_{ij} r_j(t) + \boldsymbol{w}_{noise}\boldsymbol{\psi}(t) + \boldsymbol{w}_{in}\boldsymbol{u}(t) + \boldsymbol{\xi}(t)$$

$$r_i(t) = \sigma(x_i(t)) = \frac{1}{1+\exp(-x_i(t))}$$

$$o(t) = \boldsymbol{w}_{out}\boldsymbol{r}(t) + b$$

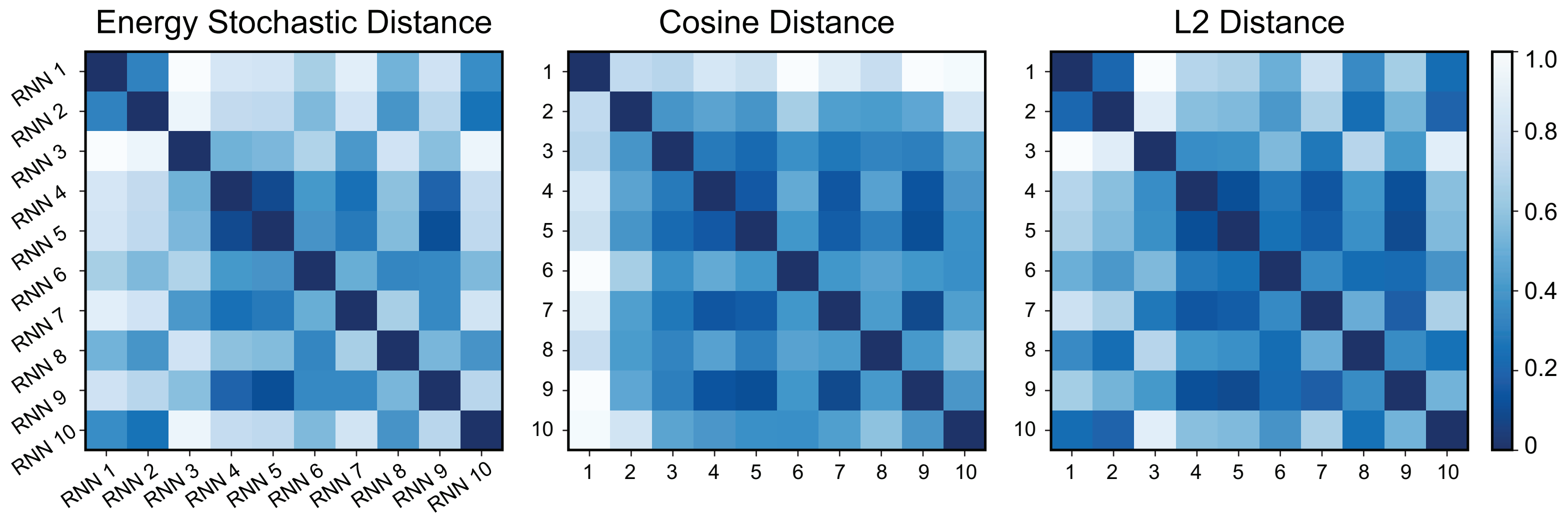
Methods

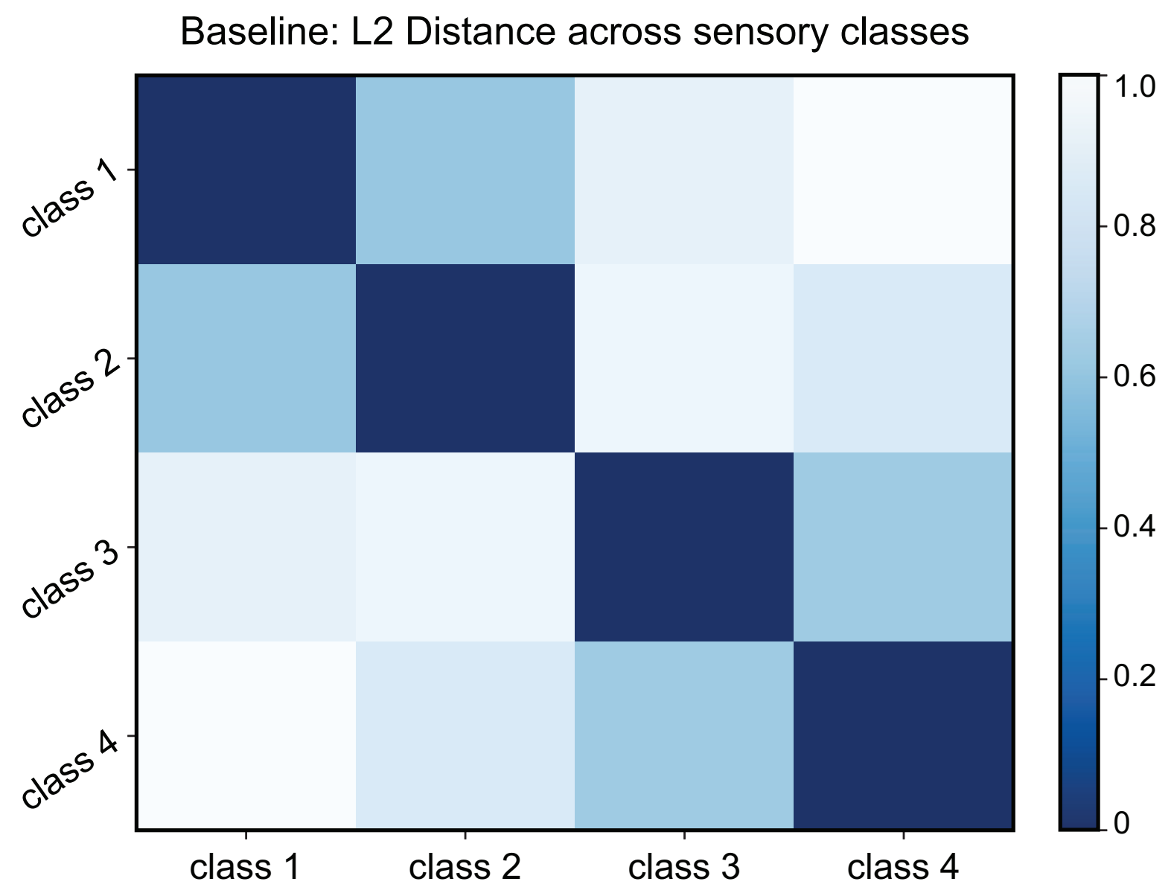
We utilized biologically plausible recurrent neural networks (RNNs) trained at varying noise levels to execute the delayed-match-to-sample task. For the control sensitivity tests, we trained 10 RNNs with no noise and analyzed the synaptic current activities (x) to capture the internal representations of each network. For the experimental comparisons, we trained RNN models with varying noise levels (i.e., model 1-5). We employed multiple RSA methods to characterize and compare neural representation similarity across networks and noise levels in these two conditions:

1. Gaussian Stochastic: 2-Wasserstein distance
2. Energy Stochastic: Non-parametric distance for higher-order
3. Correlation: Linear relationship between two variables
4. Cosine: Cosine of the angle between two vectors
5. L1: Absolute difference measure
6. L2 (NanEud): Euclidean distance measure (with NaNs)
7. Minkowski: Generalization of Euclidean and Manhattan
8. Chebyshev: Maximum absolute difference measure

Results: Control Sensitivity Tests

To establish a baseline, we applied the above RSA methods to the 10 RNNs with varying initialization. These models were trained to perform the task without any noise parameter. The RSA distance matrices from three methods (energy stochastic, cosine, and L2) are shown below:

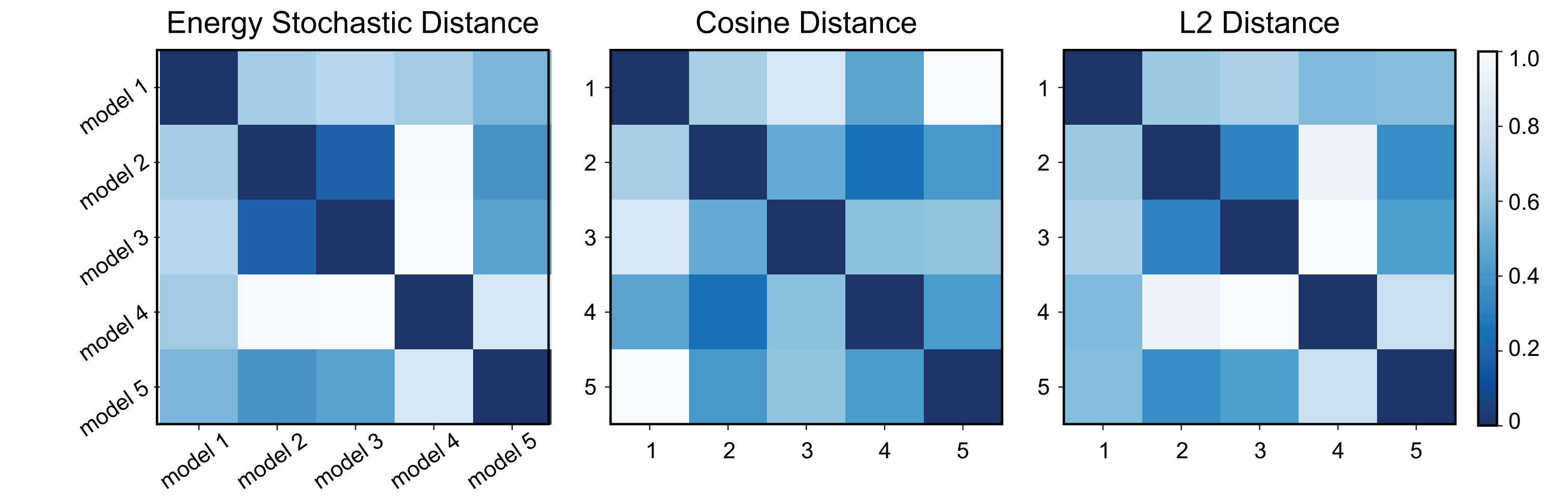


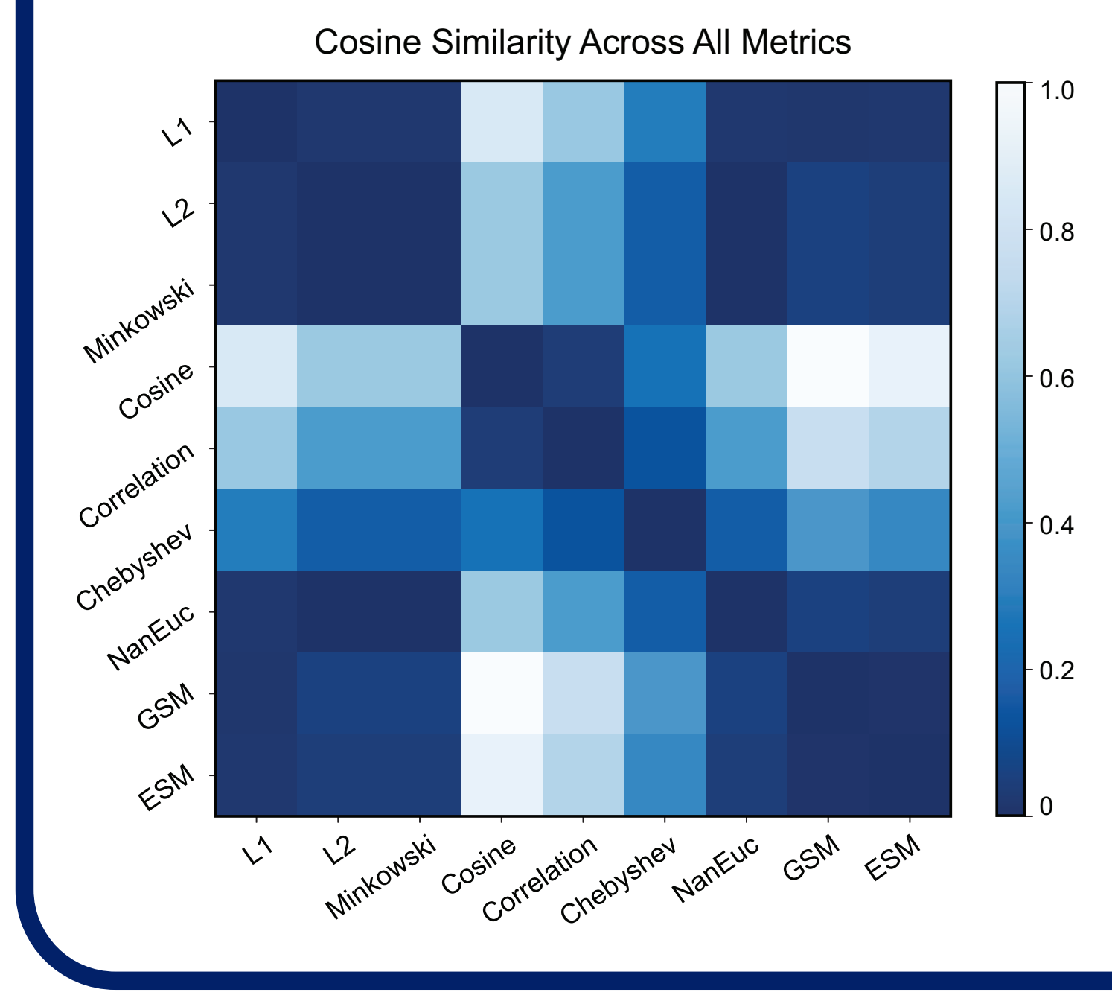


To evaluate how the four classes of sensory presentations are represented in the RNNs, we also computed the L2 distance across these sensory classes and across the 10 RNNs.

Results: Experiment Comparisons

To systematically assess the impact of noise on neural representations, we applied the same set of RSA methods to RNNs trained with different degrees of noise (models 1–5). The results from three methods (energy stochastic, cosine, and L2) are shown below:





Finally, we compare the effectiveness of different RSA metrics by computing the cosine distance measure.

Reference & acknowledgement

1. Rungratsameetaweemana, N., Kim, R., & Sejnowski, T. *bioRxiv*, 2022.
2. Duong, L., Zhou, J., Nassar, J., Berman, J., Olieslagers, J., & Alex H. Williams. *In International Conference on Learning Representations*, 2023.
3. Kriegeskorte, N., Mur, M., & Bandettini, P. A. *Frontiers in systems neuroscience*, 2008.

contact: rgs2151@columbia.edu and nr2869@columbia.edu

Conclusions

Our research provides a comprehensive evaluation of representational similarity analysis techniques in quantifying representational content in deep learning models trained to perform cognitive tasks. By examining the impact of varying levels of intrinsic noise, we identify the most effective methods for assessing the semantic features of these models. Our findings contribute to the development of robust analytical tools for studying the dynamics of deep learning models in relation to external inputs.