Fig. 1.

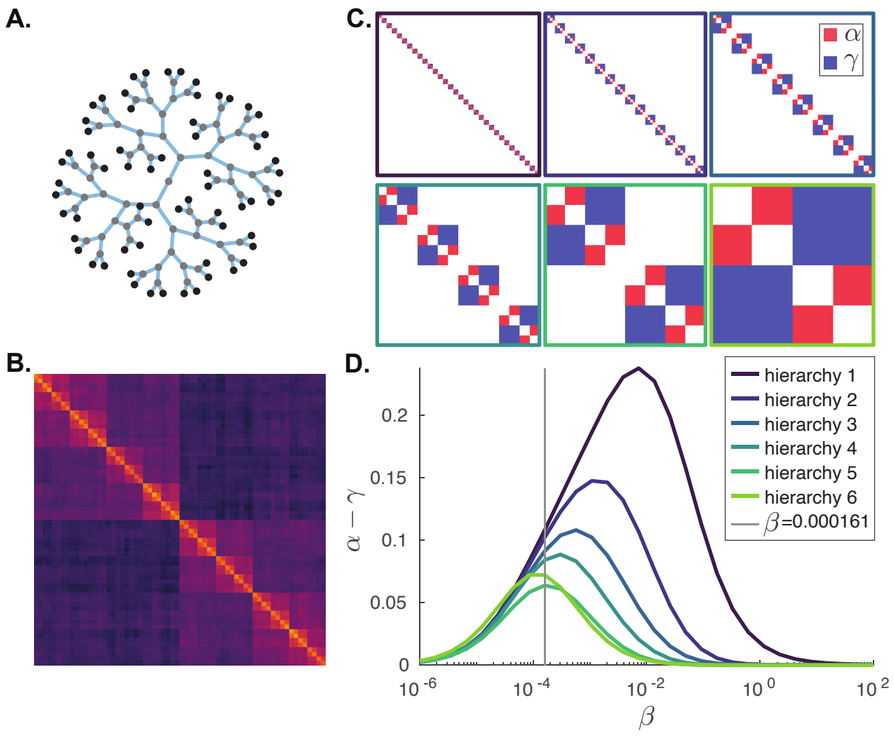


Fig. 1: **Dataset generation process.** (A) hierarchical graph for class nodes (black points), and the latent nodes (gray points). (B) Covariance between entities for an example dataset. (C) Masks that are separating within from cross category values in covariance matrix for each hierarchies, the border color shows hierarchy level. (D) the difference between within and cross category values are computed for a range of beta values that determines the strength between nodes. the covariance in (B) shown by gray vertical line.

Fig. 2.

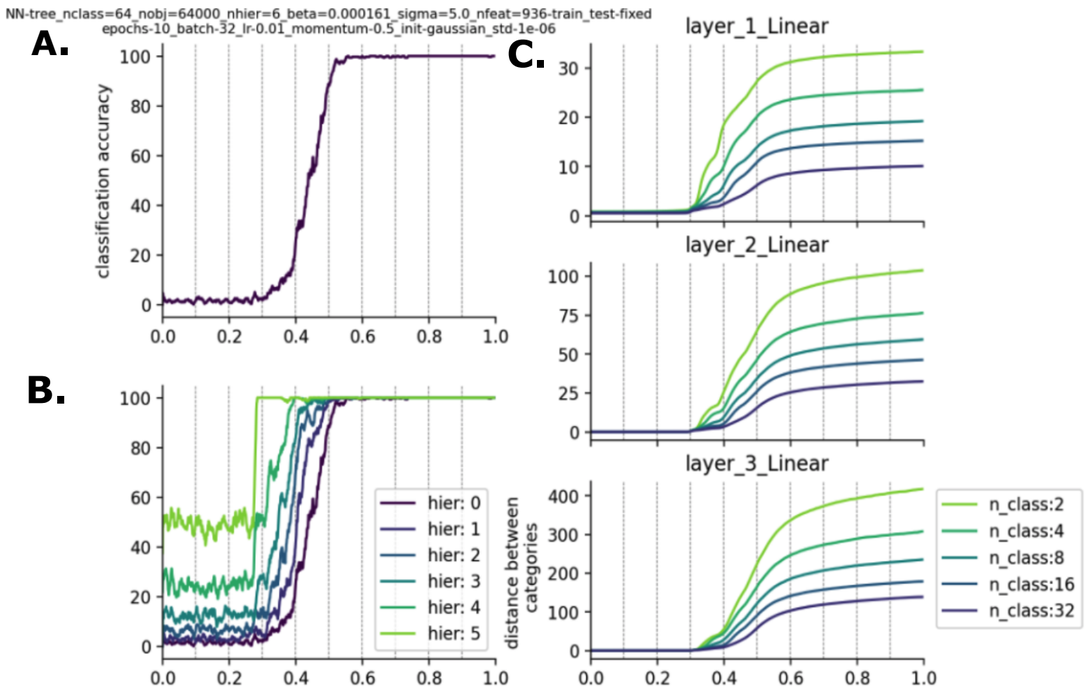


Fig. 2. Replication of prior work for nonlinear networks.
(A) Classification accuracy of the network (based on the lowest hierarchy, hierarchy 0). (B) Accuracies of each hierarchy of the network. (C) Euclidean distances among pairs of classes from each hierarchy as in Rogers & McClelland et al., 2004. In all plots, the x-axis shows relative training time.

Fig. 3.

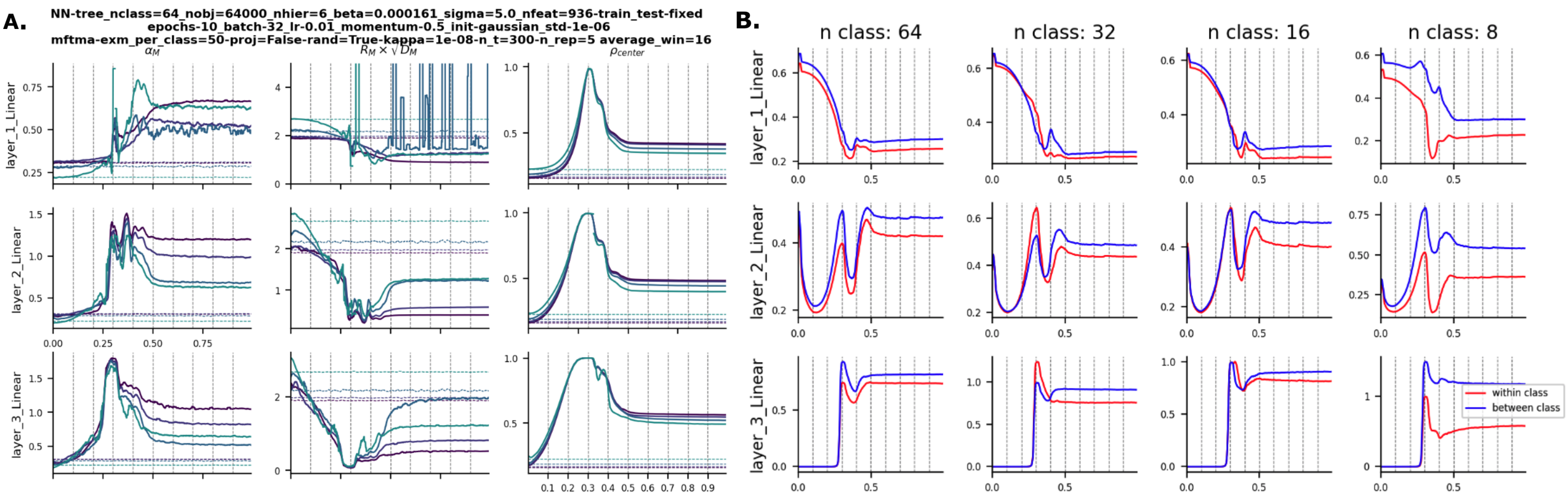


Fig. 3: **Geometrical representation of object manifolds: MFTMA for demonstrating the changes in the measure related to dimensionality, radius and center correlation, during the learning of hierarchies.**
(A) MFTMA analyses for each network layer. Columns denote the manifold capacity, the mean manifold radius multiplied by the square root of the mean manifold dimension, and the center correlation. (B) Average cosine distance (1-uv/|u||v|) for centers belonging to the same branch compared to different branches across hierarchies. Rhe center correlations were computed for different categories conditioned on them being part of the same superclass or not (similar to Fig. 1). The within class correlation shows the average of values for categories if they belong to the same super class. Between class correlation shows values for categories that belong to different superclasses.

Fig. 4.

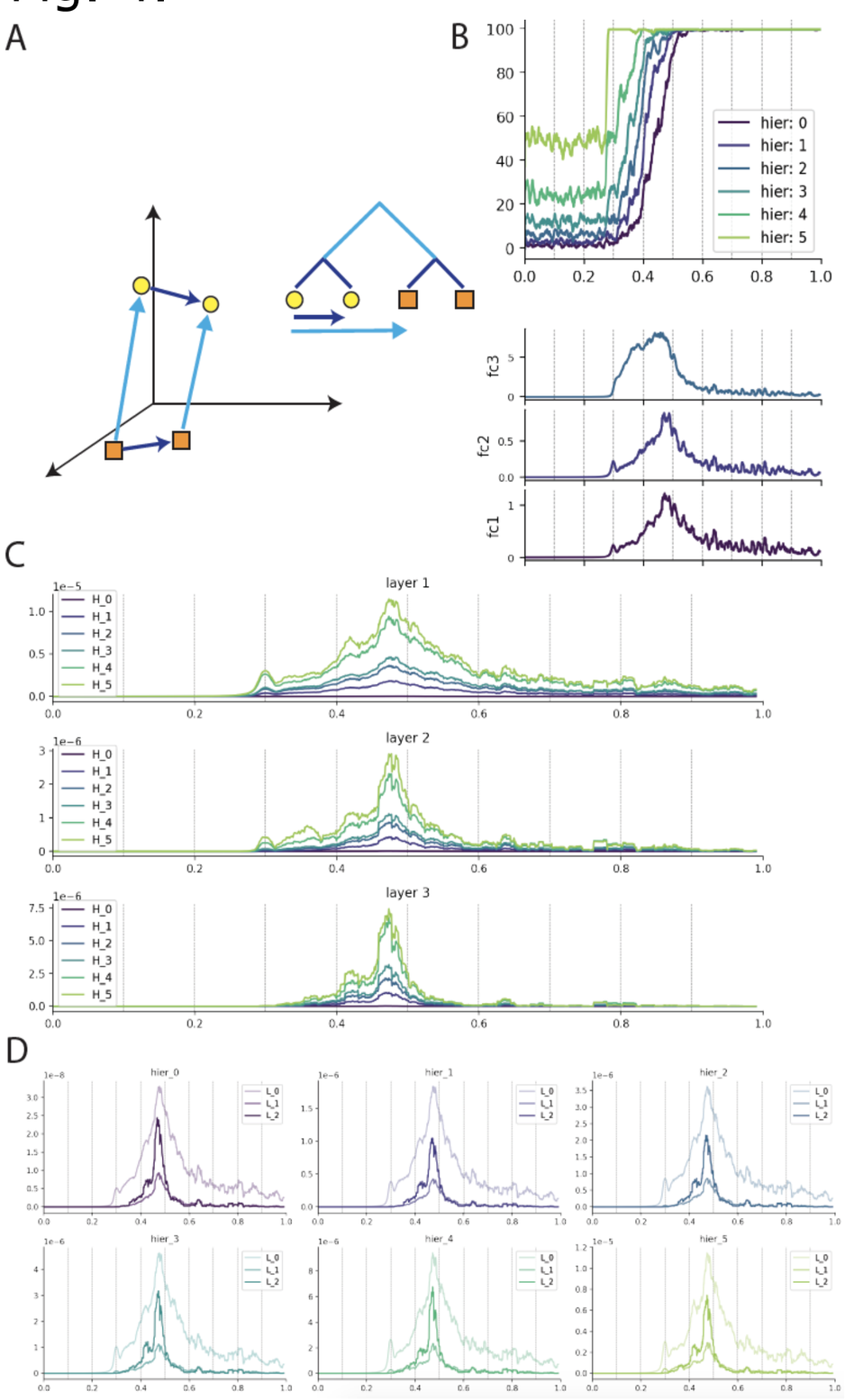
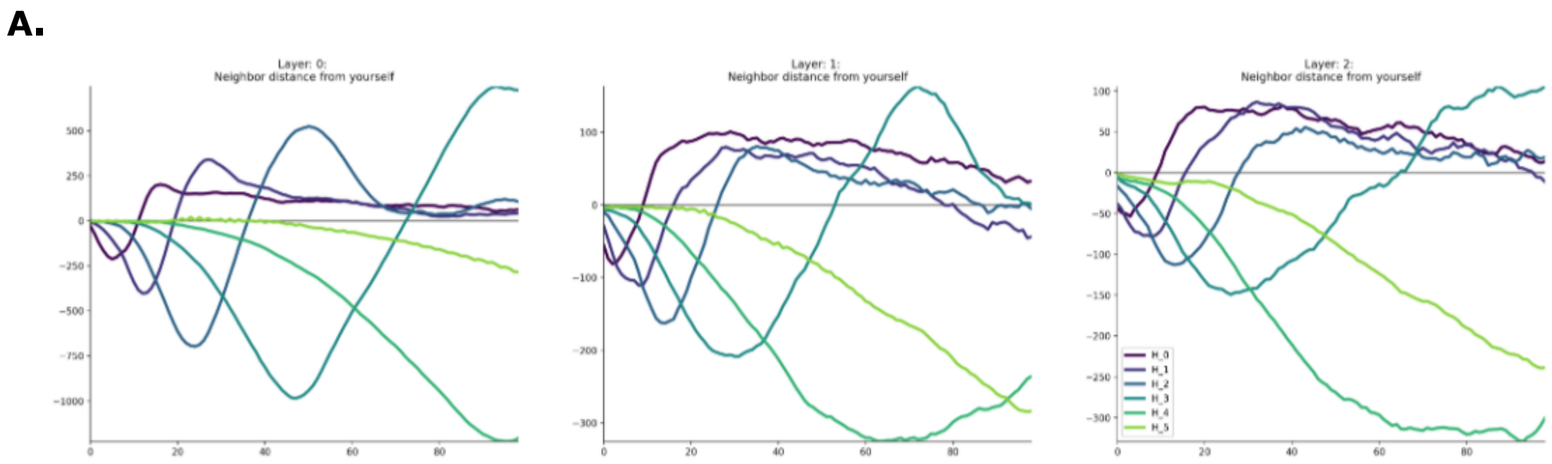


Fig. 4: **Gradient analysis for showing how learning is manifested in the weights.**
(A) demonstrates the approach (as described in overleaf). (B) is the accuracy in the upper panel, and the overall gradient L2 norm (not conditioned on hierarchies) in the lower panel (same as the plots sent before, just different coloring). (C) is conditioned on hierarchies and layers. (D) shows the same data as in (C), but simply plotted according to the hierarchies, so comparison across layers is easier.

Fig. 5.



B.

TBD

Fig. 5: **KNN analysis highlighting changes in the structure of representations.**
(A) demonstrates the mean K-nearest neighbors (K=100, y-axis) distance for each point in training. A positive value on the x-axis means that the points are "looking back in time", i.e. that the representation of a given point is most similar to representations formed earlier in training. Conversely, if the value is negative, the neighbors are more "forward-looking". The higher the hierarchical level, the more forward-looking the representations are. This could suggest early emergence of top hierarchies as they become stable first.