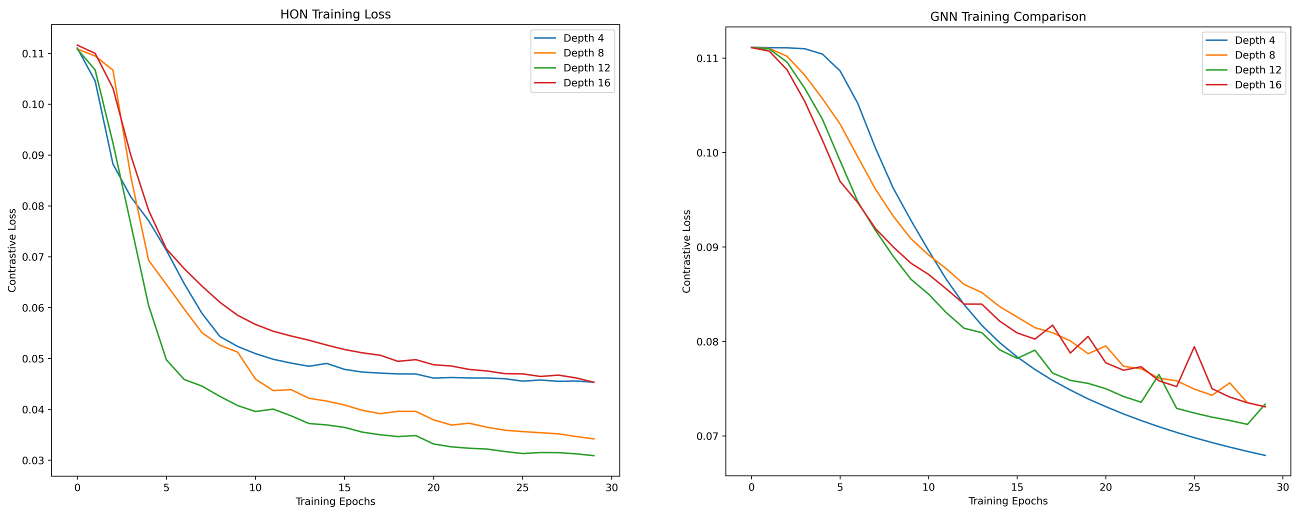


(a) HON training loss curves.

(b) GNN training loss curves.

Figure 4. Training loss curves show the avg. training loss for each training epoch.

A. Message-passing and GNNs

At a high level, message-passing is an iterative algorithm that updates node features by exchanging messages between neighboring nodes. More formally, let G be a graph with vertex set V and edge set E. Each node $i \in V$ is assigned a feature vector h_i. The neighborhood of i, denoted $N(i)$, is the set of nodes that are connected to i with an edge.

For a node $j \in N(i)$, the message i receives from j is defined as:

$$m_{ij} = \Phi_e(h_i, h_j, w_{ij})$$

In this definition, Φ_e is some function which takes both feature vectors and edge weight w_{ij}. The received messages are then aggregated to a single vector m_i:

$$m_i = \sum_{j \in N(i)} m_{ij}$$

h_i is then updated using a function $\Phi_h(h_i, m_i)$.

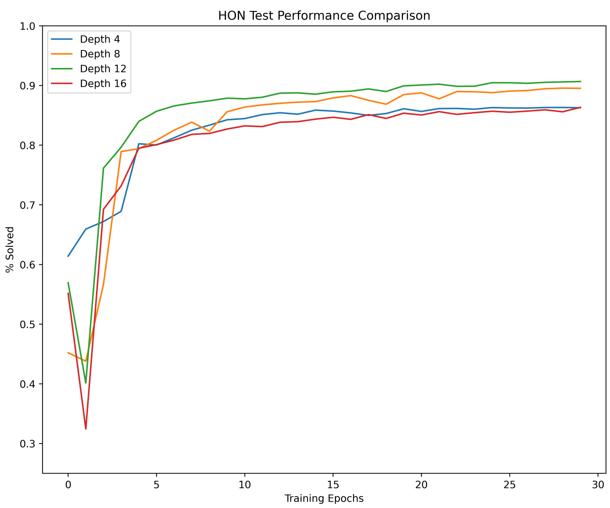
GNNs use the message-passing framework but they learn the functions Φ_e, Φ_h by parameterizing them as multi-layer perceptrons (MLPs). GNNs are multi-layered models where each layer performs an iteration of message-passing. The depth of a GNN is the number of layers it has. The update equations for layer t are formally defined as:

$$m_{ij}^t = \Phi_e^t(h_i^t, h_j^t, w_{ij})$$
$$m_i^t = \sum_{j \in N(i)} m_{ij}^t$$
$$h_i^{t+1} = \Phi_h^t(h_i^t, m_i^t)$$

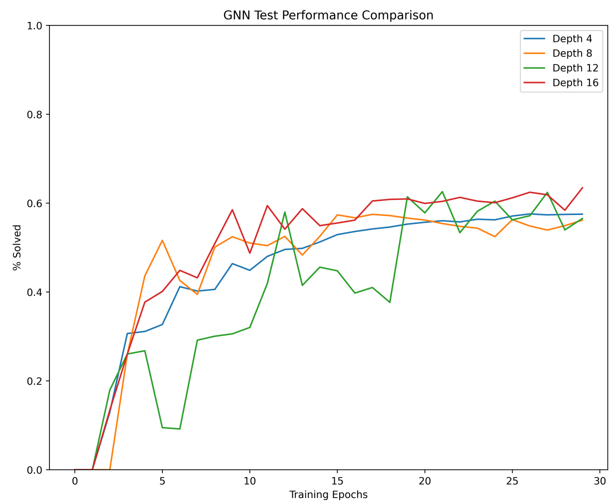
B. Experiments: Model Depth Sweep

Both models saw very little variation in performance from depth. As seen in Figure 4, the depth 4 GNN and depth 12 HON had the lowest training losses. The HON achieves significantly lower training loss and sees much faster and more stable convergence than the GNN.

The test performance shown in Figure 5 shows that despite higher training loss, depth 16 was the highest performing depth for the GNN. The HON once again significantly outperforms the GNN baseline achieving $80 - 90\%$ correct colorings compared to the GNNs $50 - 60\%$ and with quicker, more stable convergence. We chose the depth 16 GNN and depth 12 HON for the full dataset performance comparison.



(a) HON test performance.



(b) GNN test performance.

Figure 5. Test performance curves show the average percentage of correct colorings on test set after each training epoch.