# Dynamic Sparse Training

## Structure and Robustness

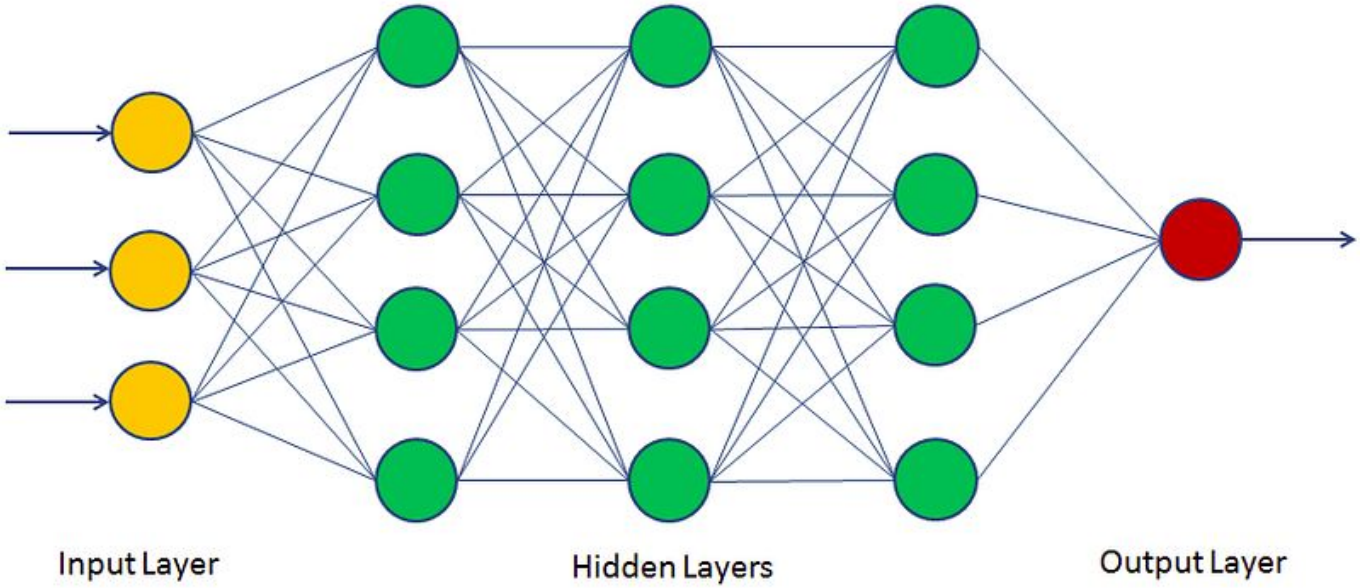
Bendegúz Sulyok
Semmelweis University, Healthcare Management Instruction Centre

SEMMELWEIS
UNIVERSITY 1769

# The Efficiency Crisis: over-parameterization

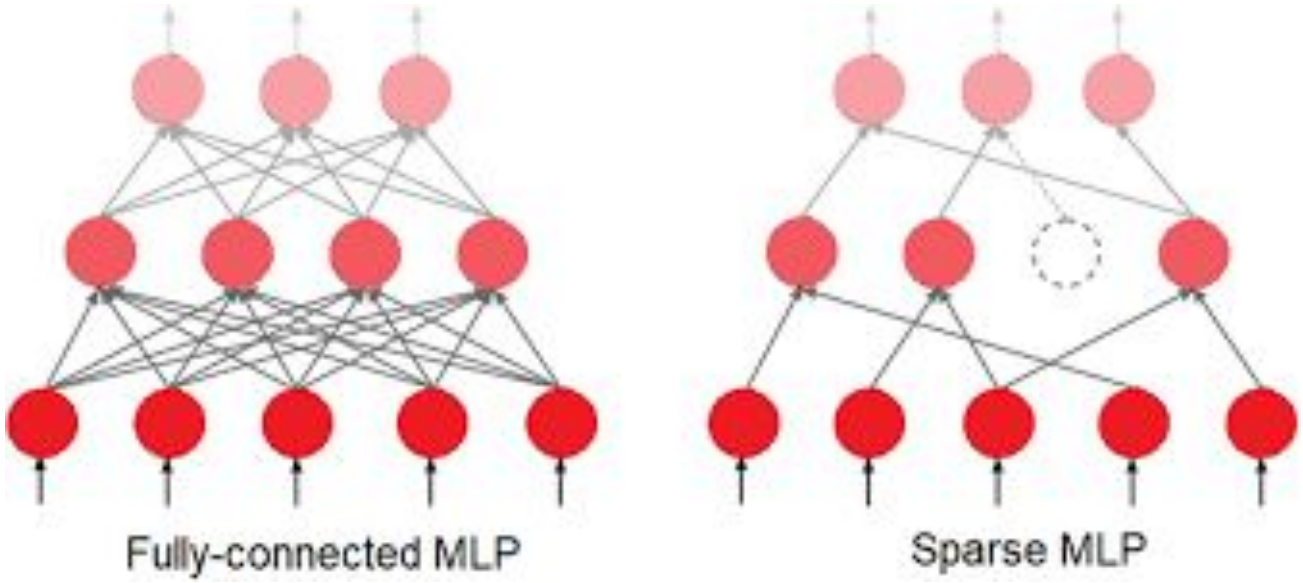
- Multilayer Perceptrons are wasteful
  - Elementary building block
  - Widely used (classification head)
  - High computational overhead

- Post-training compression
  - Model Distillation
  - Quantization
  - Pruning
  - Only helps at inference



# Solution: Dynamic Sparse Training

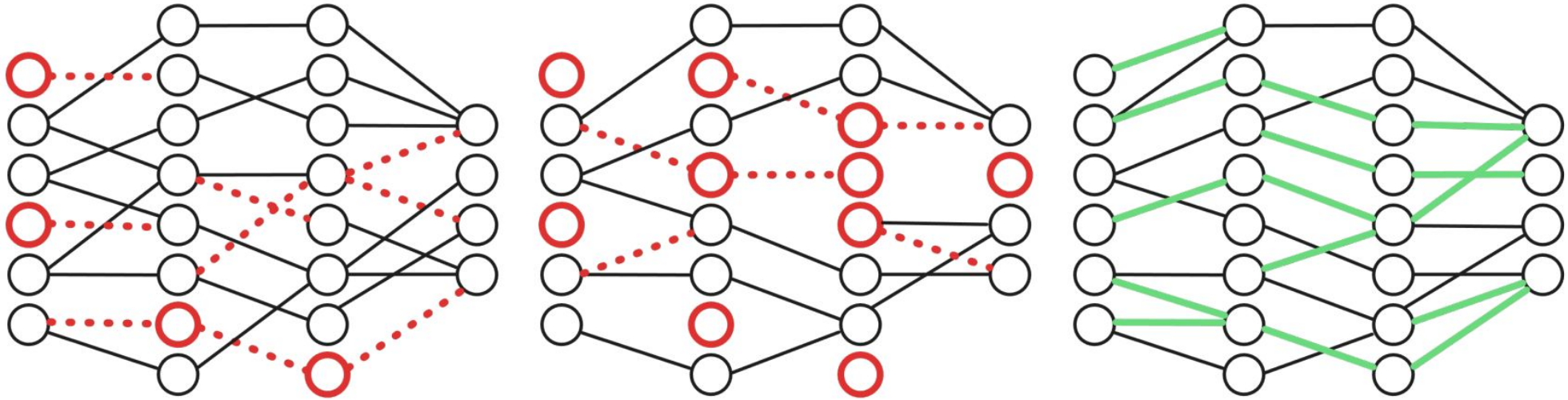
- Intuition: fewer weights ⟶ less resource use
  - Current hardware does not support this
- Use Sparse Linear Layers at training time ⟶ But which links?
- Learn the topology too!
- Keep the sparsity at 99%



# Topology Optimization

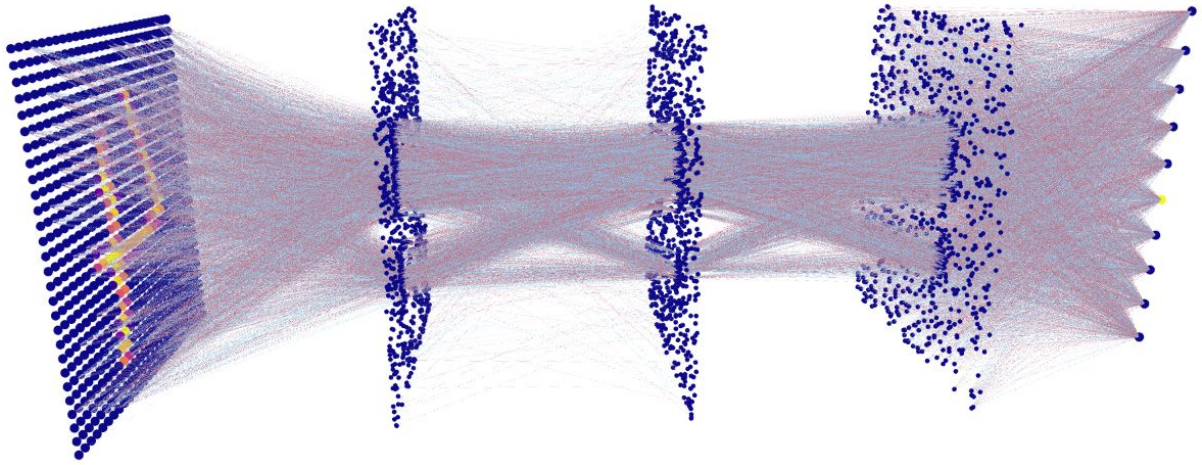
Periodically interrupt weight optimization and update the topology:

a) Remove a % of the weakest links (based on weight magnitude)
b) Optionally prune dead ends too via CR (Chain Removal)
c) Select and activate new links



# Experimental Setup: Data & architecture

- **General Idea:** Take any model and replace DLLs with SLLs
- **Standard visual problem**
  - Small image datasets: MNIST, FashionMNIST, EMNIST (letters only)
  - Flatten + 3 SLLs (1000 neurons, 99% sparsity) + 1 DLL
- **Harder visual problem with CNN head**
  - Larger RGB image dataset: CIFAR10
  - VGG1-BN + 3 SLLs (1000 neurons, 99% sparsity) + 1 DLL

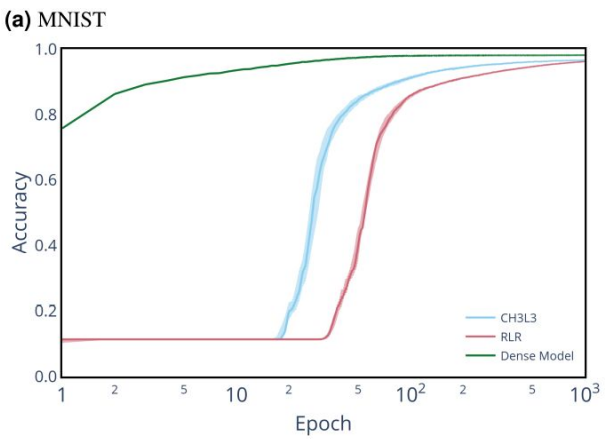


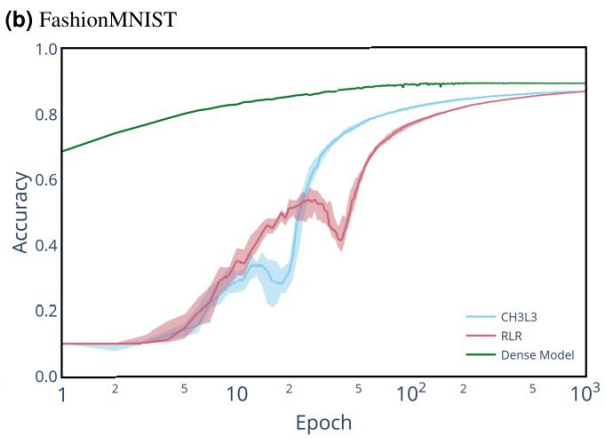
# Experimental Setup: Link regrowth

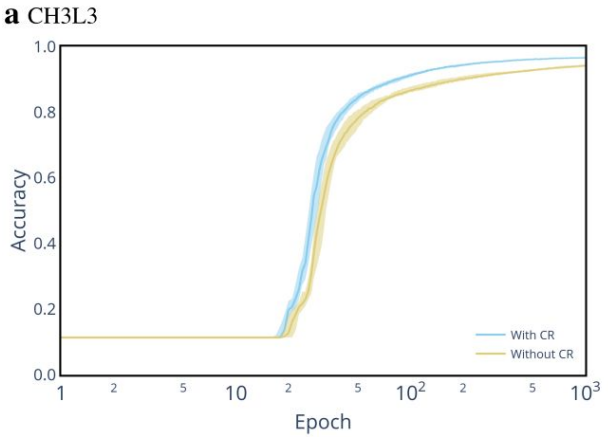
- **RLR (Random Link Regrowth)**
  - Simple yet effective
  - CR not necessary
- **CH3L3[1]**
  - Link prediction method derived from graph theory
  - Promotes dense routing communities
  - Significantly benefits from CR
- **Dense model with no regrowth**
  - Added as a baseline
  - After training pruned to 99% sparsity

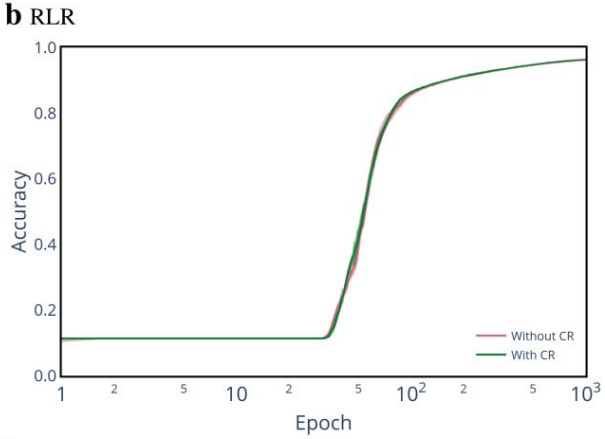
[1] Zhang, Y., Zhao, J., Wu, W., Muscoloni, A., & Cannistraci, C. V. (2024, May). Epitopological learning and cannistraci-hebb network shape intelligence brain-inspired theory for ultra-sparse advantage in deep learning. In The Twelfth International Conference on Learning Representations.

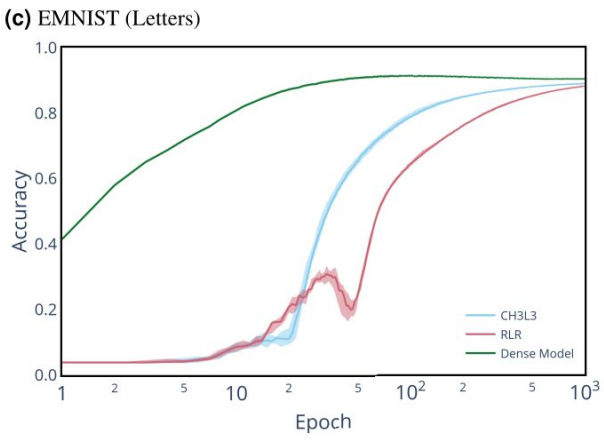
# Training convergence

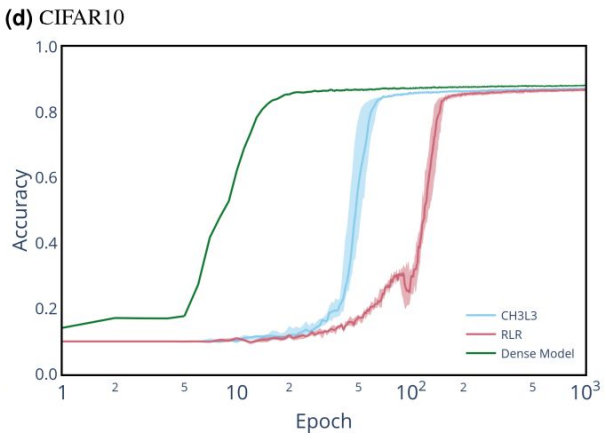


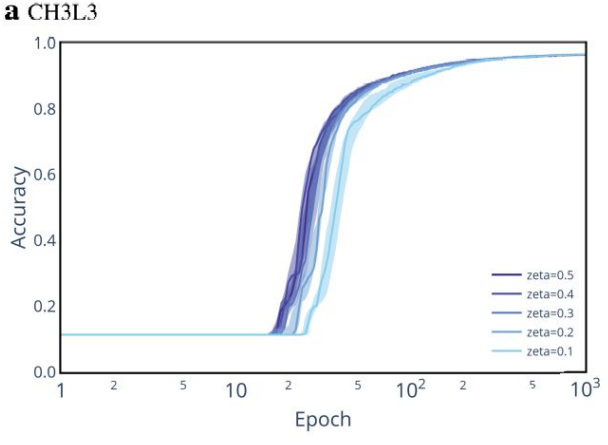



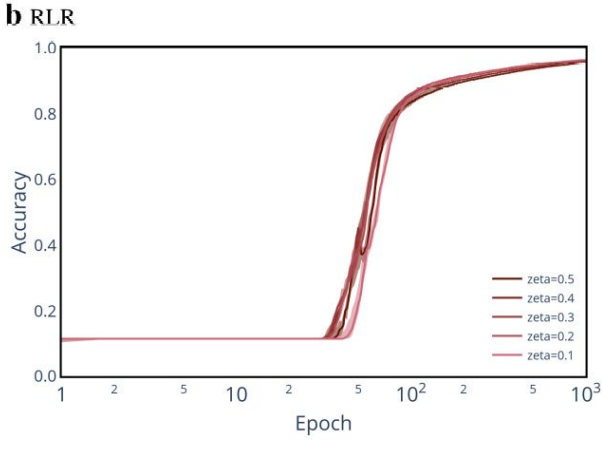



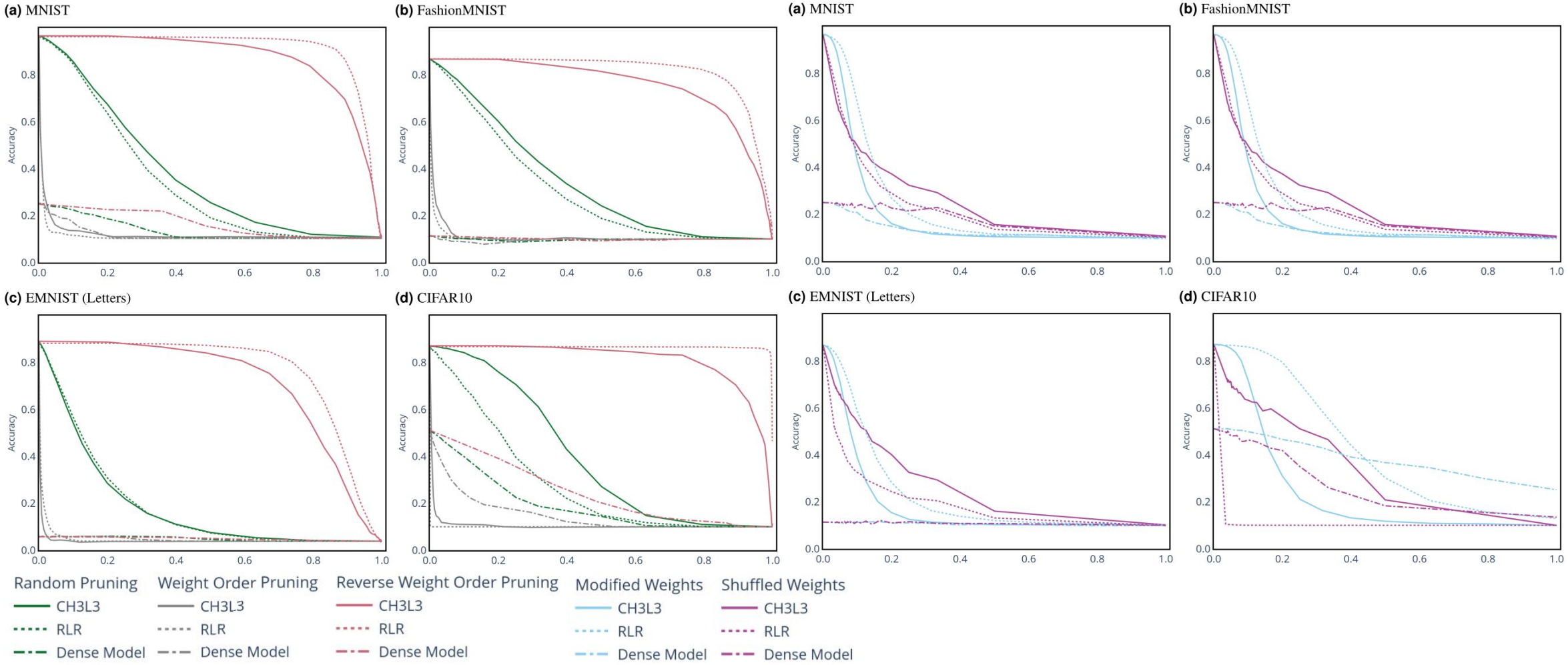



# Robustness analysis

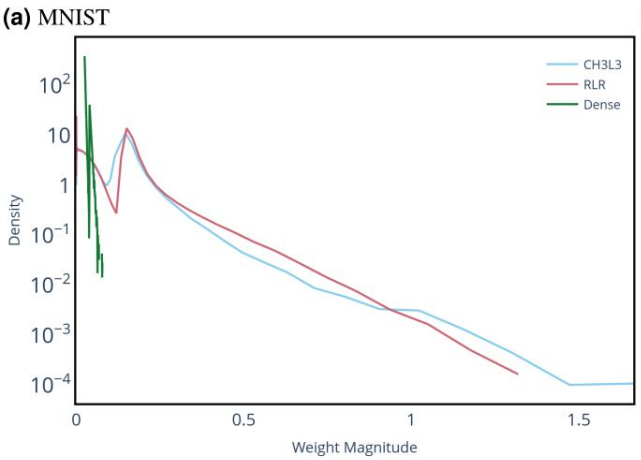


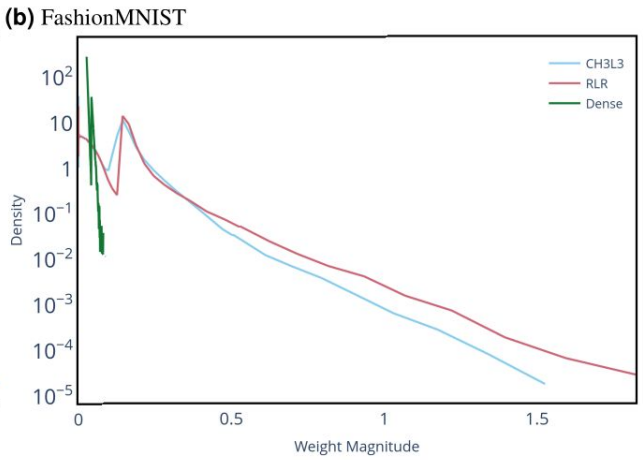
# Robustness analysis

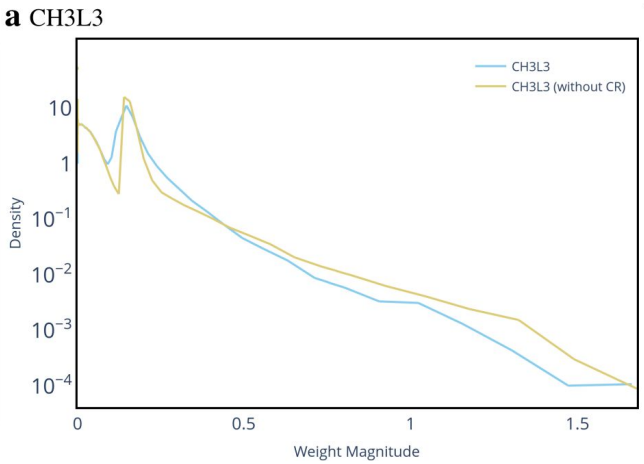
- Random Pruning
- Weight Order Pruning - strong links first
- Reverse Weight Order Pruning - weak links first
- Shuffled Weights - shuffle weights values in bins
- Modified Weights - add Gaussian noise

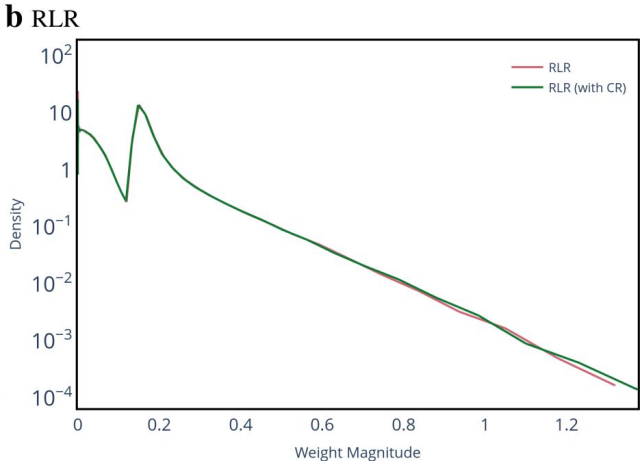
| | Weight Order Pruning | Reverse Weight Order Pruning | Random Failure | Shuffled Weights | Modified Weights | Effective Rank | Stable Rank |
|---|---|---|---|---|---|---|---|
| CH3L3 | 0.122 | 0.869 | 0.374 | 0.255 | 0.189 | 308.049 | 43.137 |
| CH3L3 (without CR) | 0.110 | 0.889 | 0.291 | 0.155 | 0.196 | 282.607 | 42.663 |
| RLR | 0.112 | 0.912 | 0.339 | 0.225 | 0.233 | 2063.546 | 135.037 |
| RLR (with CR) | 0.115 | 0.919 | 0.347 | 0.236 | 0.232 | 2019.569 | 130.005 |

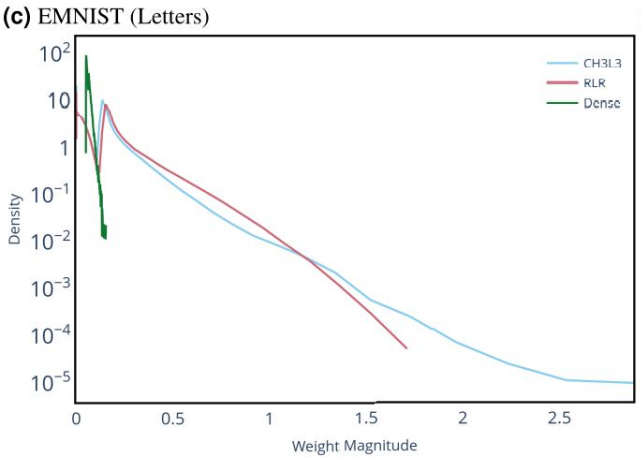
# Weight distribution

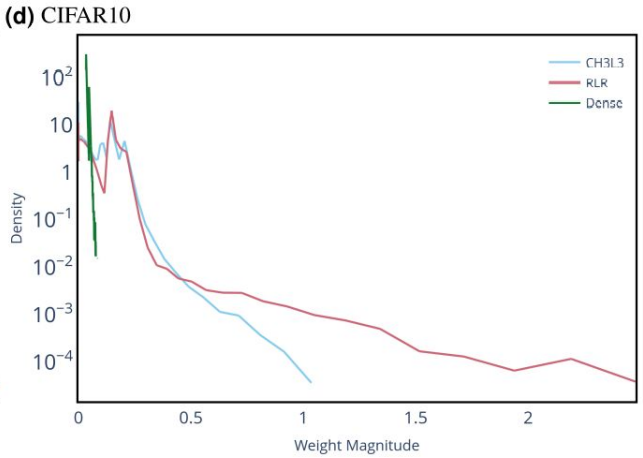


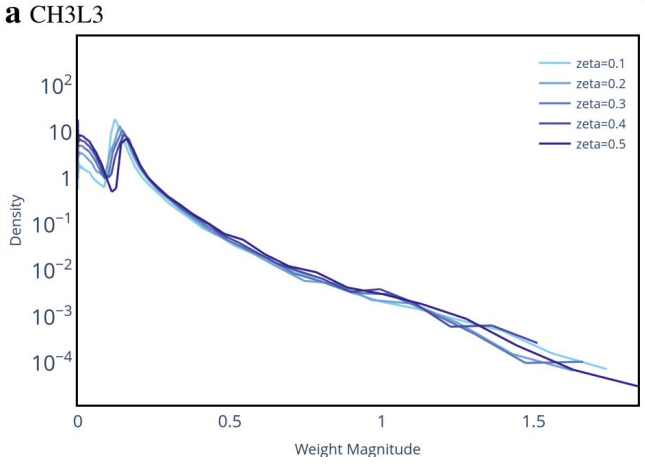



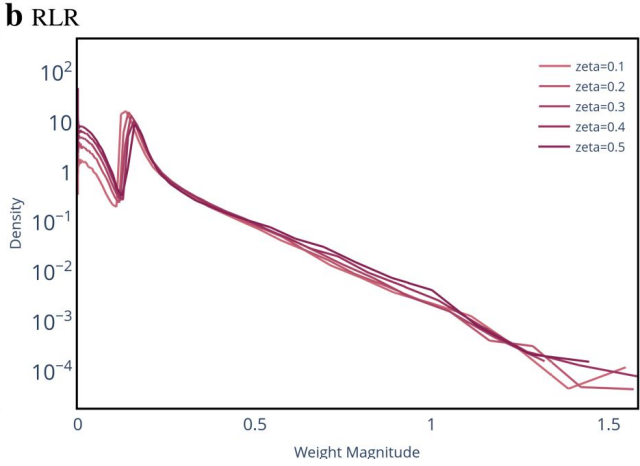



# Key Takeaways

- Optimization is highly effective at the 99% sparsity constraint
- The two methods result in robustness against different perturbation types
- CH3L3:
  - Compresses the network (exploration < exploitation)
  - Relies on CR
  - Accelerates convergence and
- RLR:
  - Spreads out the network (exploitation < exploration)
  - Entirely unaffected by CR
  - Strong edges are structurally important

# Thank you for your attention!

We acknowledge the support of the AccelNet-MultiNet program, a project of the National Science Foundation (award #1927425 and #1927418)

We acknowledge the support of Semmelweis University Health Services Management Training Centre

SEMMELWEIS
UNIVERSITY 1769