a Toy model of representation learning via data recovery
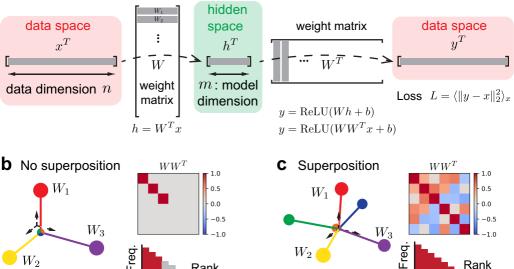

$h = W^T x$

$y = \text{ReLU}(Wh + b)$

$y = \text{ReLU}(WW^T x + b)$

Loss $L = \langle \|y - x\|_2^2 \rangle_x$

b No superposition

c Superposition