Transformer test loss vs. learning rate and embedding dimension, at n = 10242.00 embed dim 32 embed dim 128 1.75 embed dim 512 1.50 -1.25 -Test loss 1.00 -0.75 0.50 0.25 $0.00 \pm$ 10-2 10^{-3} 10^{-4} learning rate