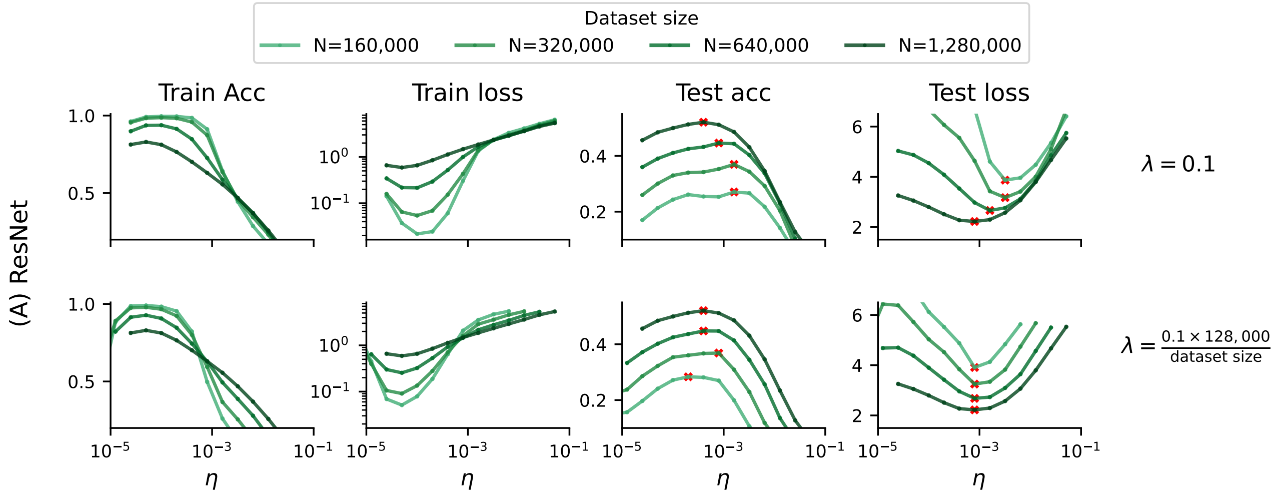

Figure 1: **Under the suggested weight decay scaling, the optimal learning rate is stable across training length.** Similiar to the setting of Fig.1 in the main text, we trained the model for 100 *epochs* with different dataset sizes under a fixed batch size. Using a fixed weight decay (top row), the optimal learning rate (red marks) decreases with the dataset size. Under our suggested weight decay scaling, where $\lambda \propto \frac{1}{\text{dataset size}}$, the optimal learning rate becomes more stable across dataset sizes.