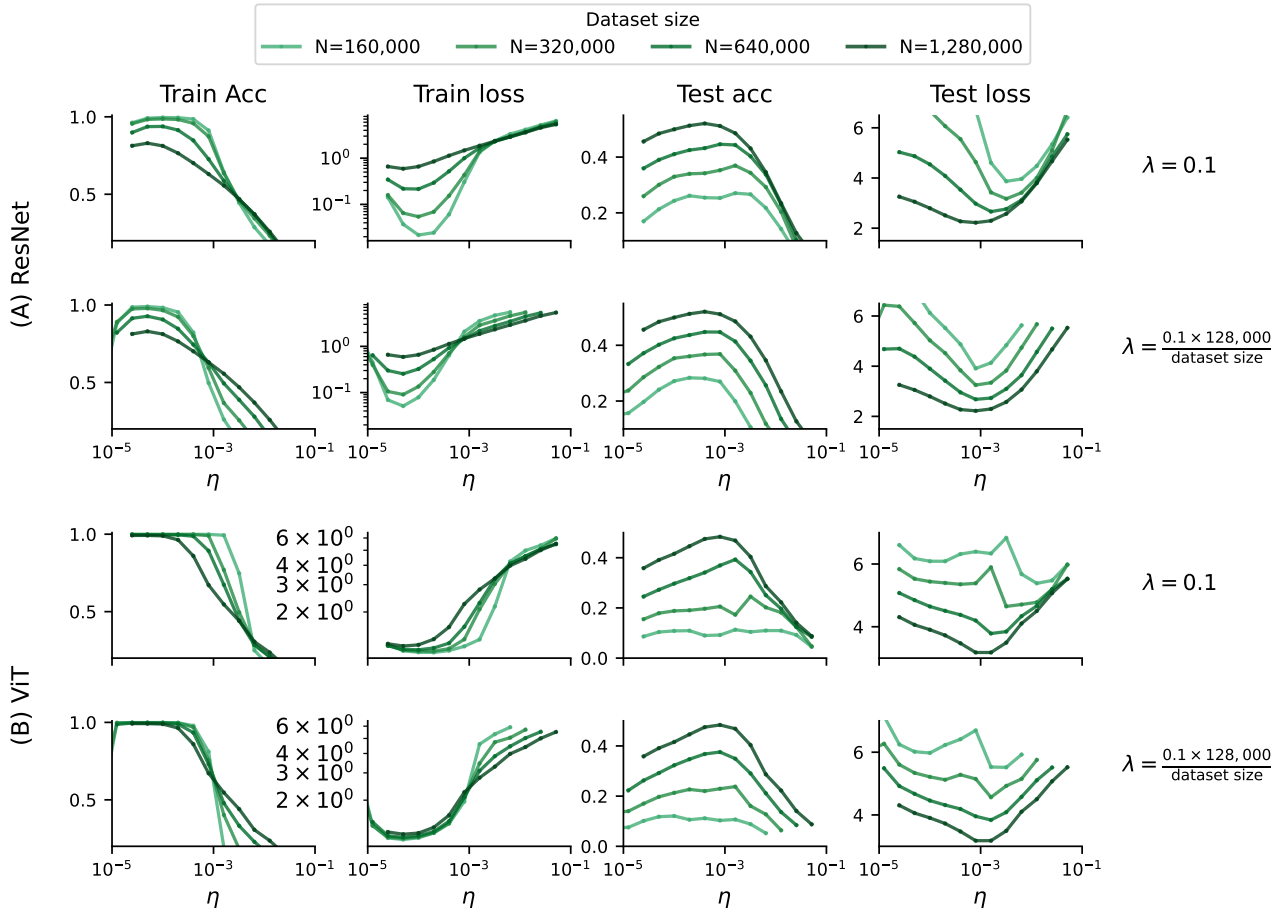

Figure 1: **Under the suggested weight decay scaling, the optimal learning rate is stable across training length.** Mirroring Fig.1 in the main text, we trained the model for 100 *epochs* with different dataset sizes under a fixed batch size. Using a fixed weight decay (top rows in subfig. A, B), the optimal learning rate *decreases* with the dataset size. Under our suggested weight decay scaling (bottom rows in subfig. A, B), where $\lambda \propto \frac{1}{\text{dataset size}}$, the optimal learning rate becomes more stable across dataset sizes. Note that we select the values for λ as 0.1 as they were close-to-optimal for the experiments in Fig.1.