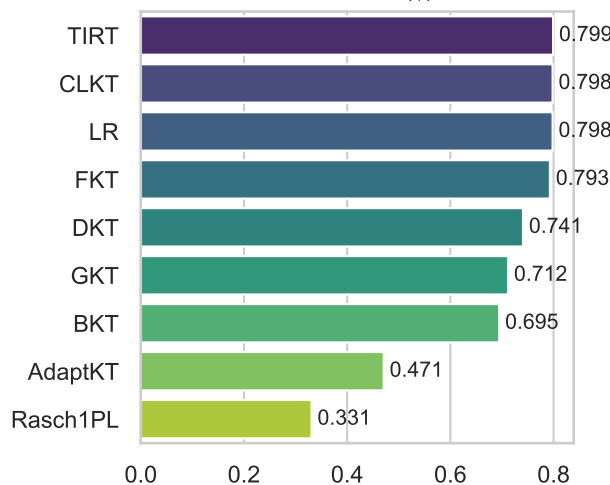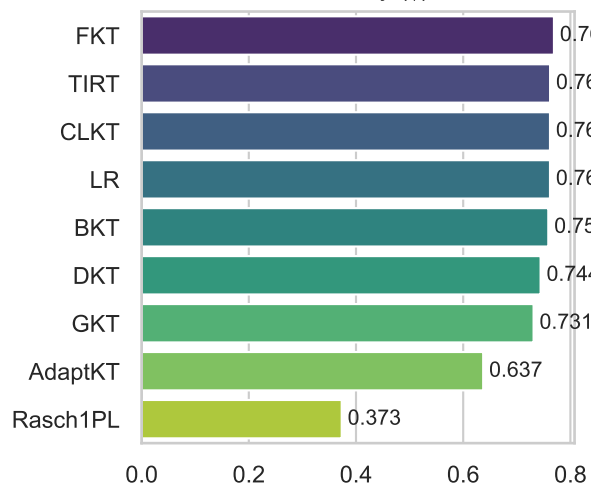Model Comparison on Test Metrics