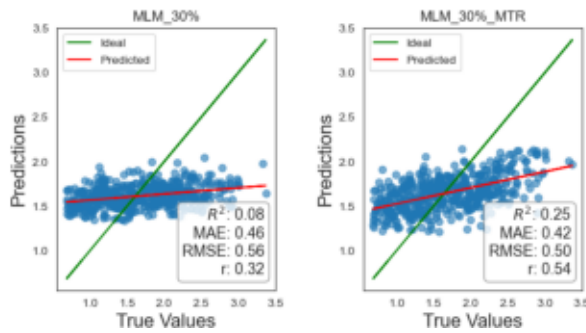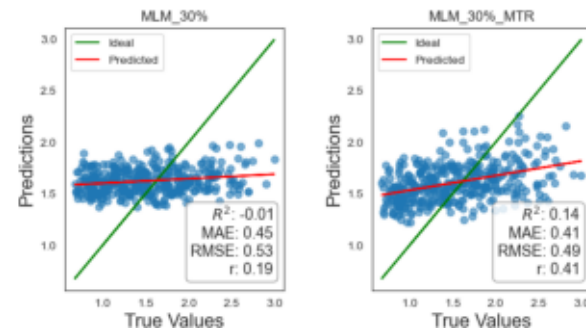Dataset: HLM / Pre-training objective: mlm / split: datasail