Learning-curves with different metrics from train rewards to validate all policies metric None (last visited policy) Trajec mean Test reward (max acc) of the best policy (in terms of validation metric) 8000 Trajec max Window 100 Window 50 Window 20 Window 10 6000 Window 5 Window 1 4000 2000 0 100000 200000 300000 400000 500000 Total train steps (no extra consumption of test steps in validation here)