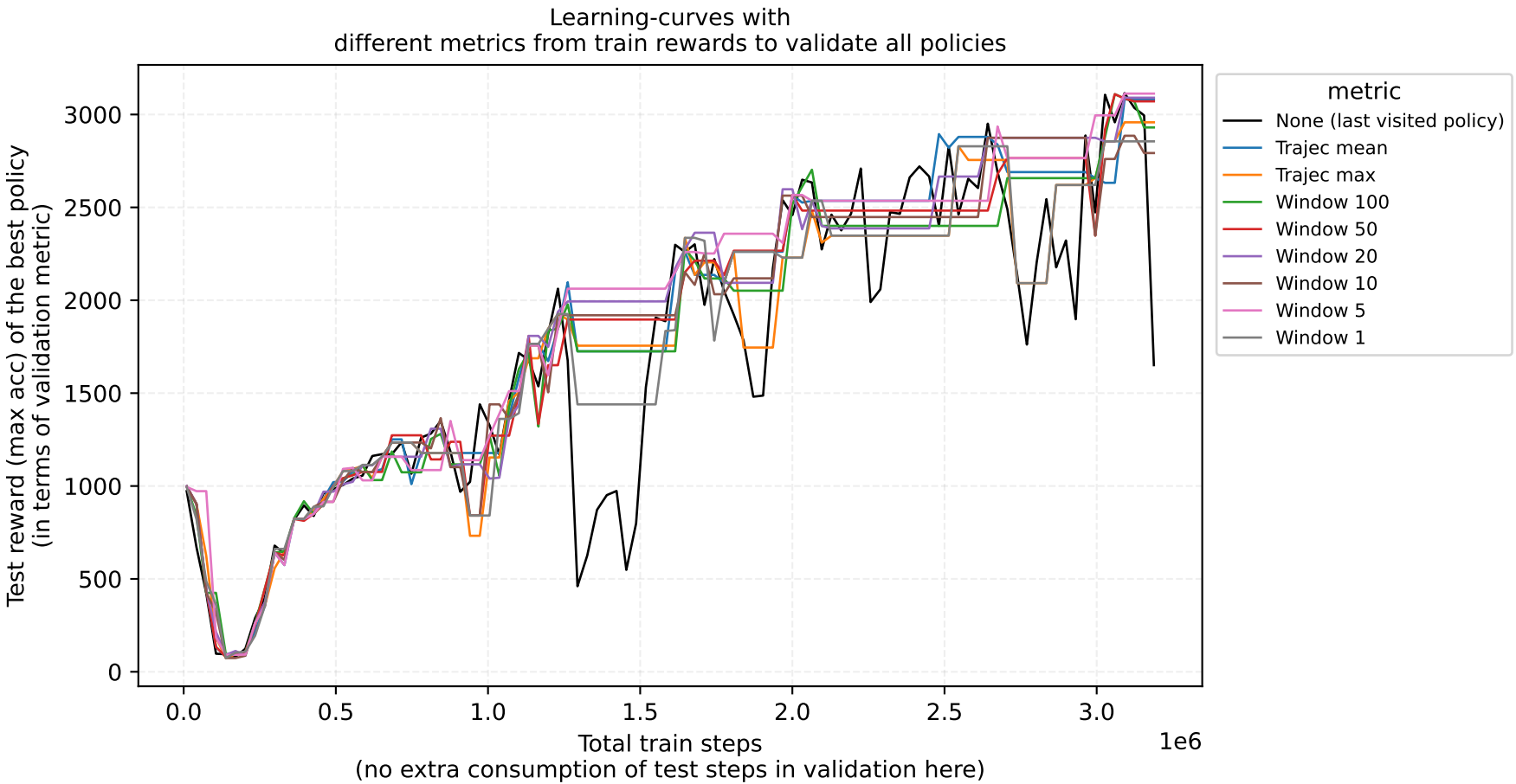
Learning-curves with
different metrics from train rewards to validate all policies
Test reward (max acc) of the best policy
(in terms of validation metric)
3000
2500
2000
1500
1000
500
0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
1e6
Total train steps
(no extra consumption of test steps in validation here)
metric
None (last visited policy)
Trajec mean
Trajec max
Window 100
Window 50
Window 20
Window 10
Window 5
Window 1