Training P-H-D ($\sigma=0.6$)

REINFORCE
DQN

training returns vs episode