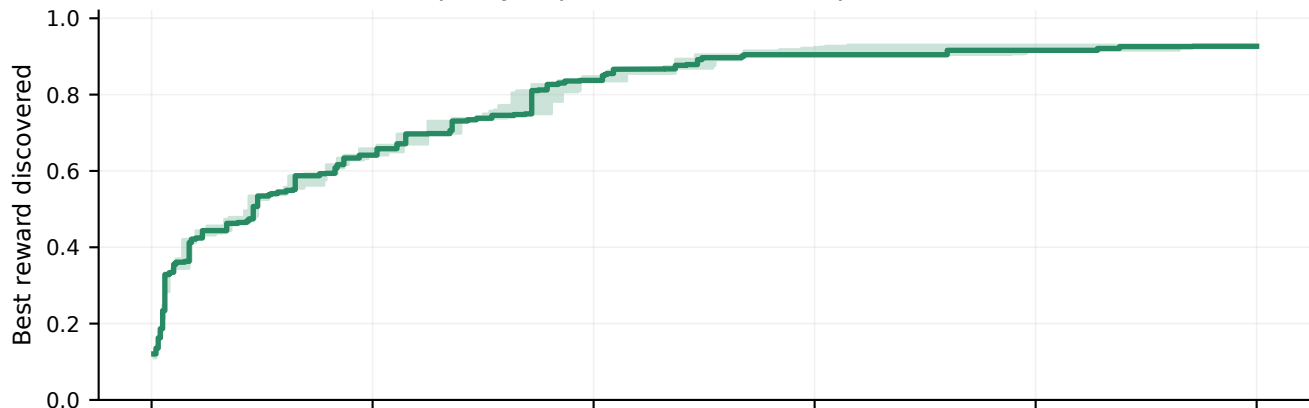


Policy discovery and exploratory volatility are distinct training outcomes

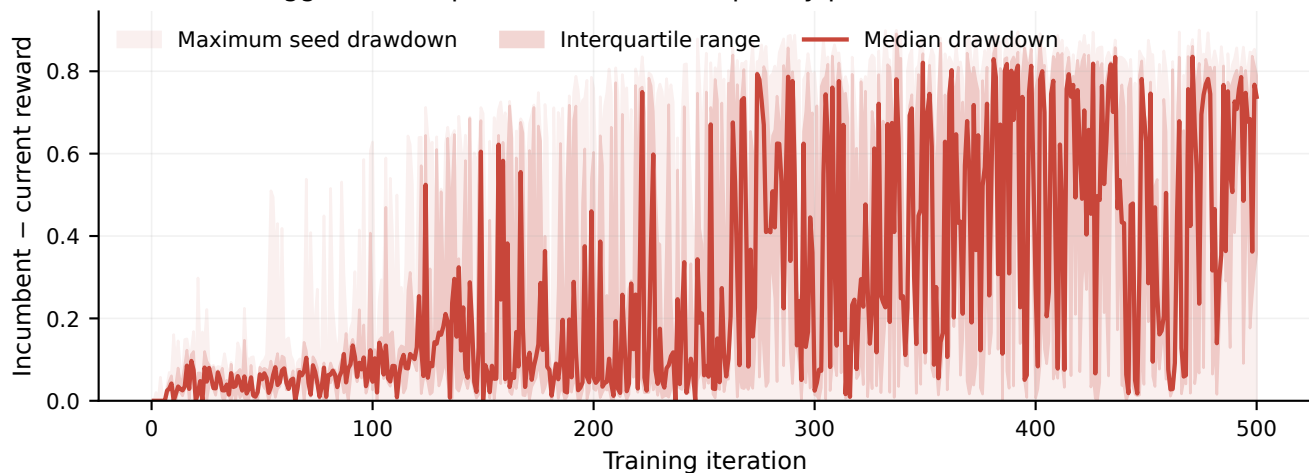
a

Incumbent policy improves as new checkpoints are discovered



b

Aggressive exploration causes temporary performance drawdown



Drawdown measures the loss of the current exploratory policy relative to the best checkpoint already retained.