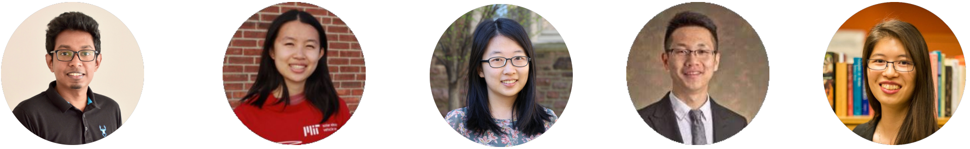
NEURAL INFORMATION PROCESSING SYSTEMS

The Impact of Task Underspecification in Evaluating Deep Reinforcement Learning

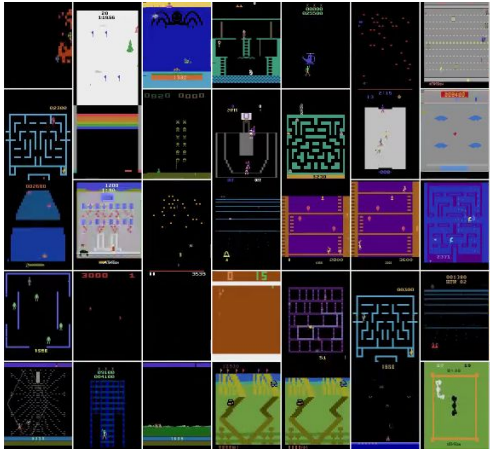
NeurIPS 2022

Vindula Jayawardana, Catherine Tang, Sirui Li, Dajiang Suo, Cathy Wu

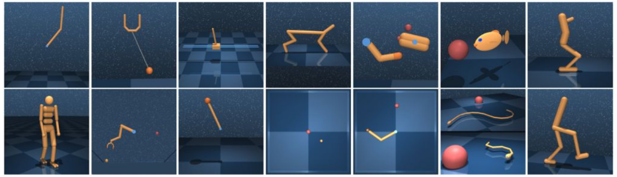


Emerging Case of Task Specific RL

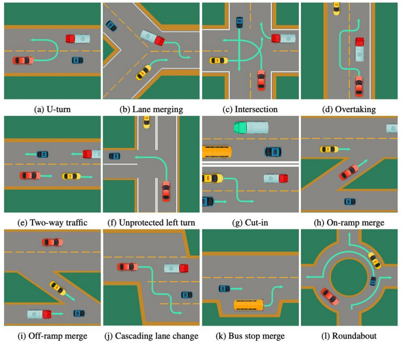
Atari 2600



DM Control Suite



Autonomous Driving

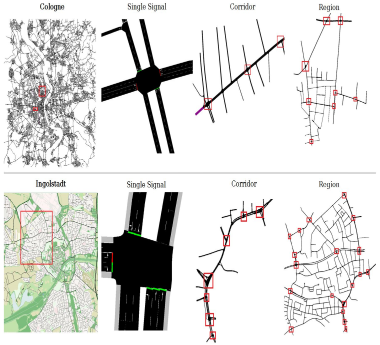


(a) U-turn (b) Lane merging (c) Intersection (d) Overtaking

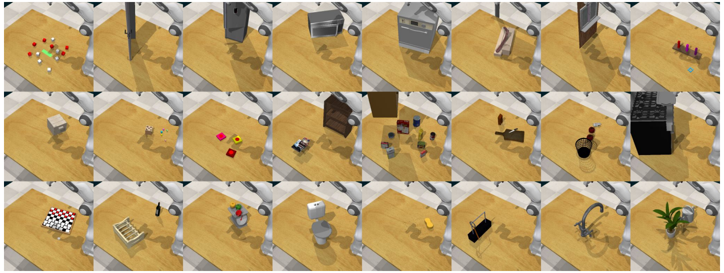
(e) Two-way traffic (f) Unprotected left turn (g) Cut-in (h) On-ramp merge

(i) Off-ramp merge (j) Cascading lane change (k) Bus stop merge (l) Roundabout

Traffic Signal Control

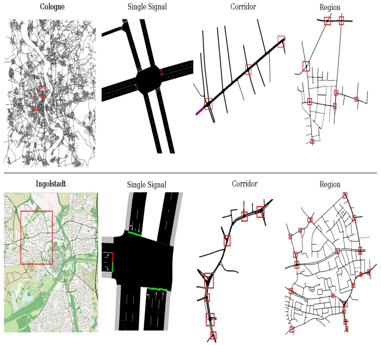


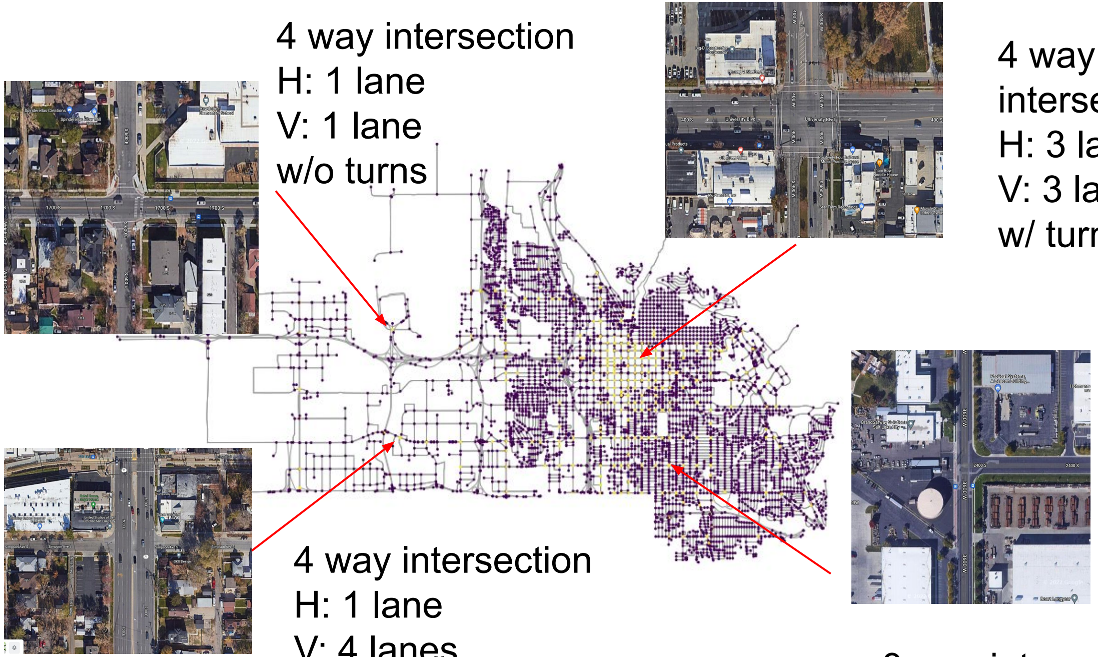
Robotic Manipulation



Curse of Variety in Task Specific RL

Traffic Signal Control





4 way intersection
H: 1 lane
V: 1 lane
w/o turns

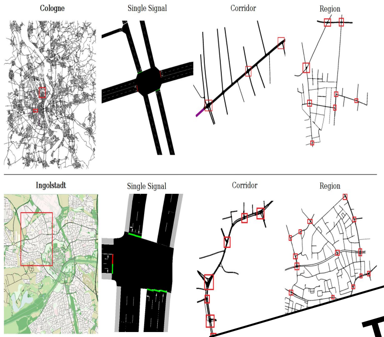
4 way
intersection
H: 3 lanes
V: 3 lanes
w/ turns

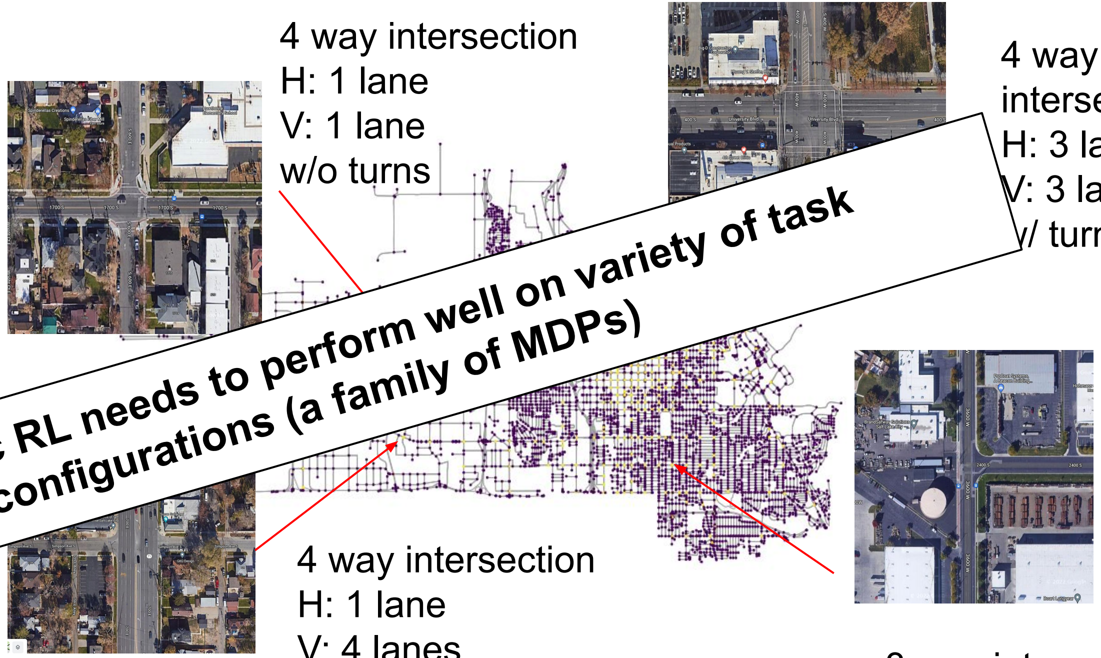
4 way intersection
H: 1 lane
V: 4 lanes
w/ turns

3 way intersection
H: 1 lane
V: 1 lane
w/ turns

Curse of Variety in Task Specific RL

Traffic Signal Control





4 way intersection
H: 1 lane
V: 1 lane
w/o turns

4 way
intersection
H: 3 lanes
V: 3 lanes
w/ turns

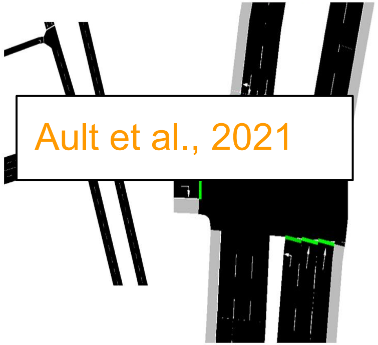
4 way intersection
H: 1 lane
V: 4 lanes
w/ turns

3 way intersection
H: 1 lane
V: 1 lane
w/ turns

Task specific RL needs to perform well on variety of task configurations (a family of MDPs)

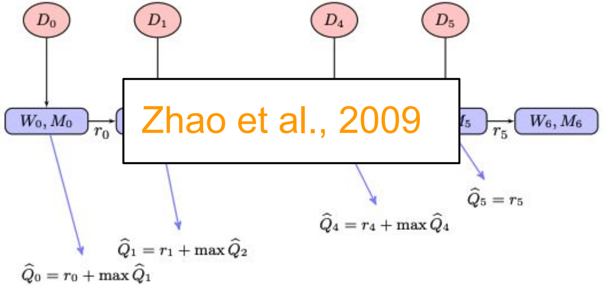
Point MDPs for Evaluations in Task Specific RL

Traffic signal control
(NeurIPS 2021)



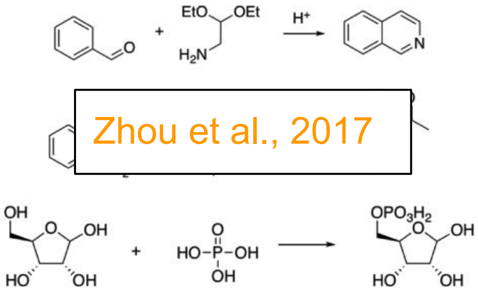
Ault et al., 2021

Chemotherapy designing
(Statistics in medicine 2009)


D_0
D_1
D_4
D_5
W_0, M_0
r_0
r_5
W_6, M_6
$\hat{Q}_5 = r_5$
$\hat{Q}_4 = r_4 + \max \hat{Q}_4$
$\hat{Q}_1 = r_1 + \max \hat{Q}_2$
$\hat{Q}_0 = r_0 + \max \hat{Q}_1$

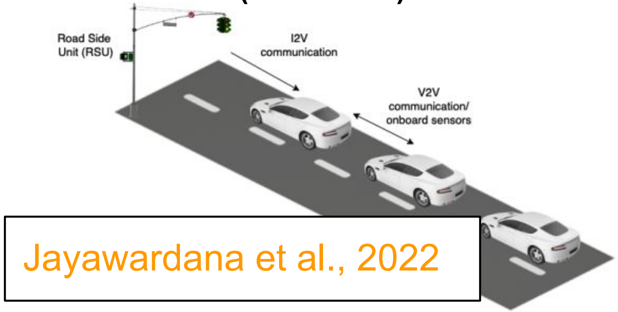
Zhao et al., 2009

Chemical reaction optimization
(*ACS Central Science* 2017)



Zhou et al., 2017

Eco-driving
(ECC 2022)



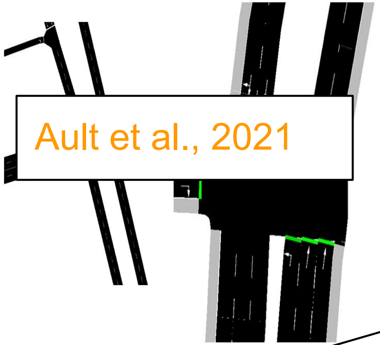
Jayawardana et al., 2022

Bottleneck decongestion
(ITSC 2018)

Vinitsky et al., 2018

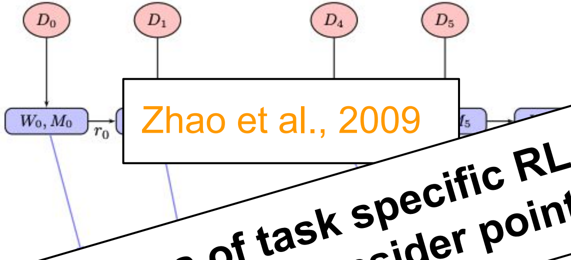
Point MDPs for Evaluations in Task Specific RL

Traffic signal control
(NeurIPS 2021)



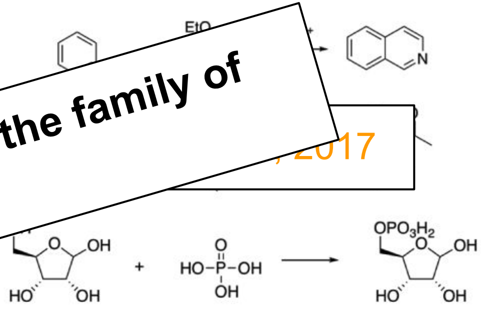
Ault et al., 2021

Chemotherapy designing
(Statistics in medicine 2009)



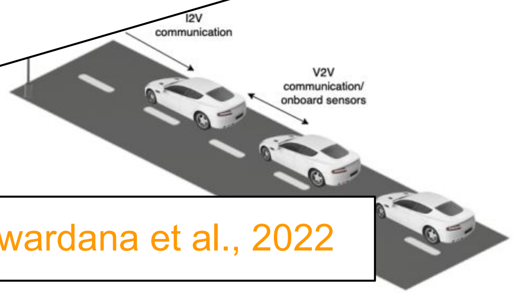
Zhao et al., 2009

Chemical reaction optimization
(*ACS Central Science* 2017)



, 2017

Performance evaluations of task specific RL ignore the family of MDPs and only consider point MDPs



Jayawardana et al., 2022

Bottleneck Decongestion
(ITSC 2018)

Vinitsky et al., 2018

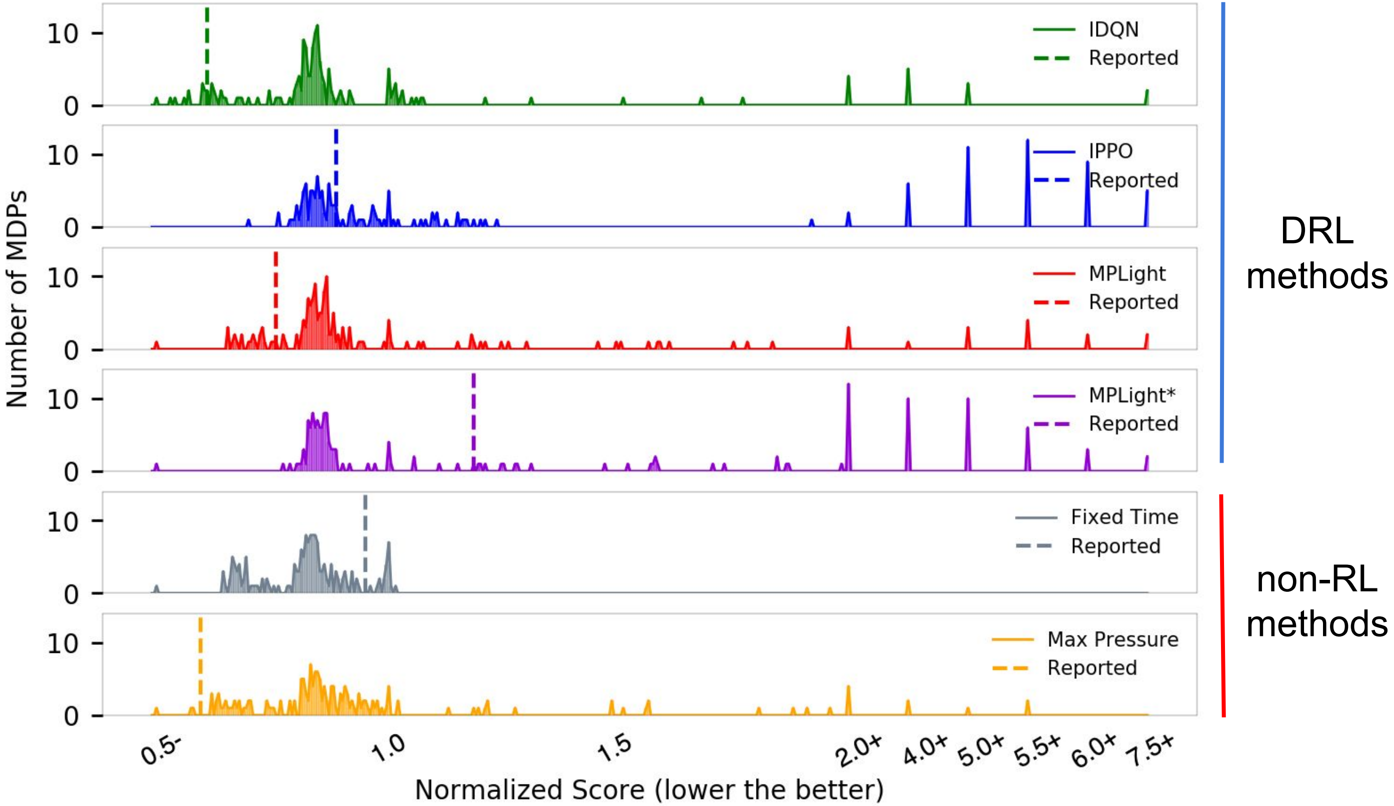
What could go wrong when Point MDPs are used for performance evaluations?

Case Study: Traffic Signal Control

- Evaluate DRL for traffic signal control using RESCO benchmark.
    - **Four DRL algorithms**: *IDQN, IPPO, MPLight, MPLight**
    - **Two traditional algorithms**: *Fine-tuned Fixed Time, Max Pressure*
- Performance evaluated based on **normalized average delay per vehicle**
    - Score normalization based on untuned-fixed time controller
- 345 signalized intersections in Salt Lake City in Utah are binned to 164 unique signalized intersections (**164 unique point MDPs**)
- **164 performance scores per algorithm**

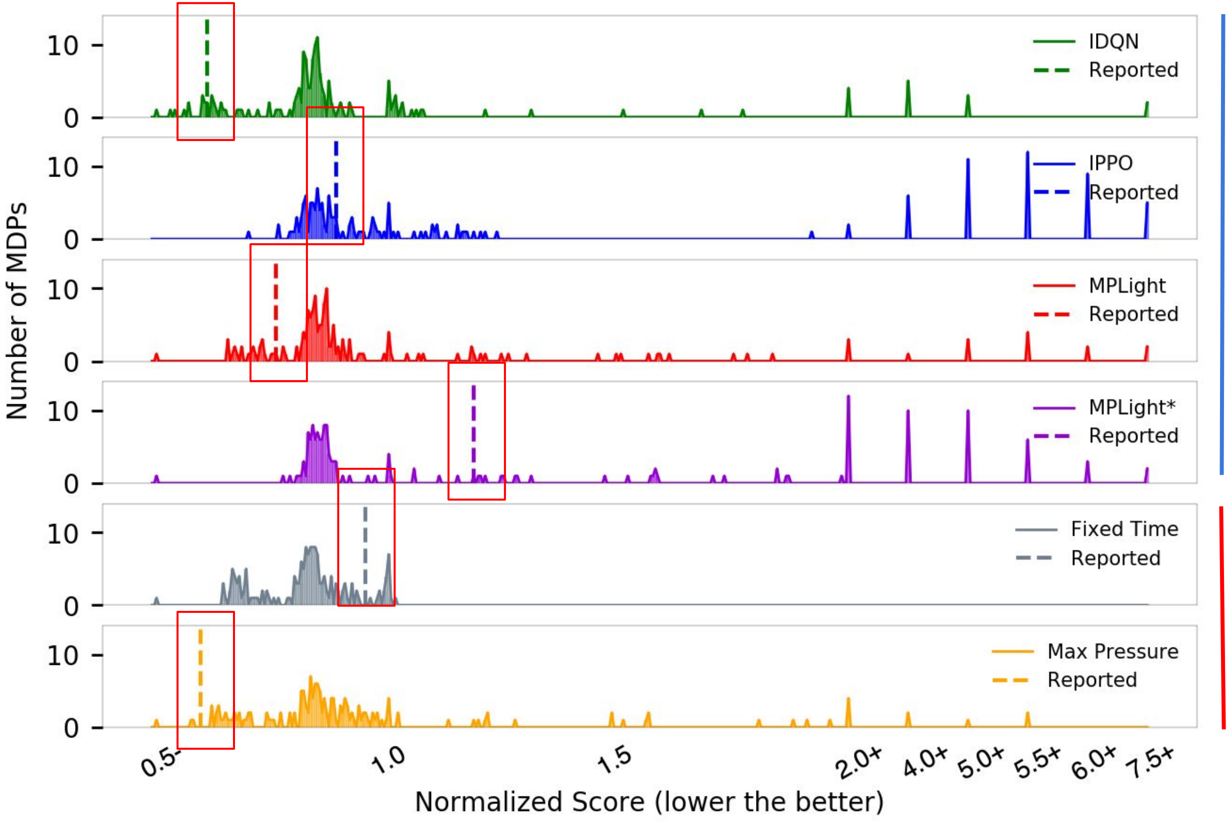
James Ault and Guni Sharon. Reinforcement learning benchmarks for traffic signal control. In Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track, 2021.

Case Study: Score Distribution



*Reported performances are based on re-evaluations of the methods on Ingolstadt single intersection.

Case Study: Score Distribution

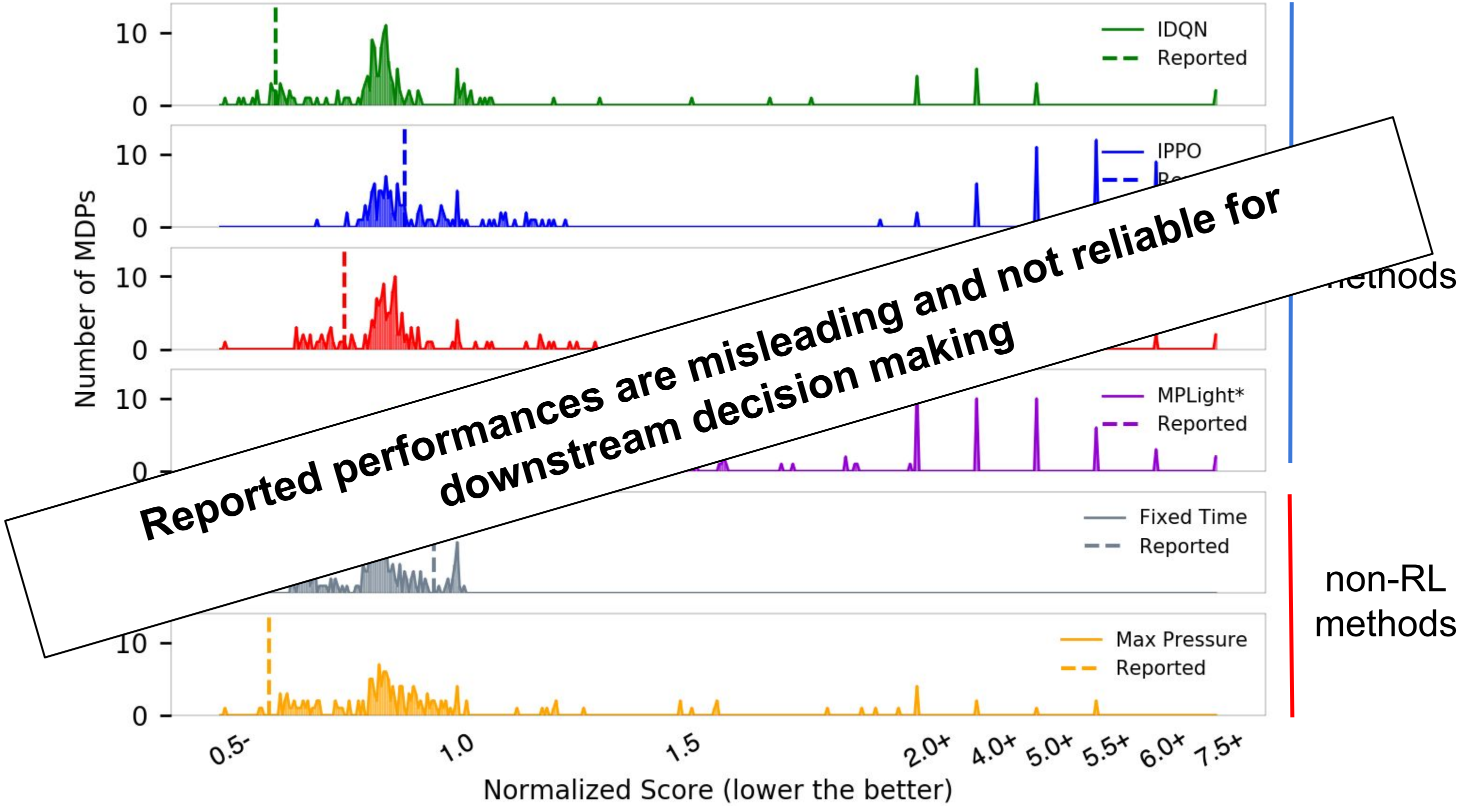


DRL methods

non-RL methods

*Reported performances are based on re-evaluations of the methods on Ingolstadt single intersection.

Case Study: Score Distribution



Reported performances are misleading and not reliable for downstream decision making

*Reported performances are based on re-evaluations of the methods on Ingolstadt single intersection.

How to fix this issue and perform reliable evaluations?

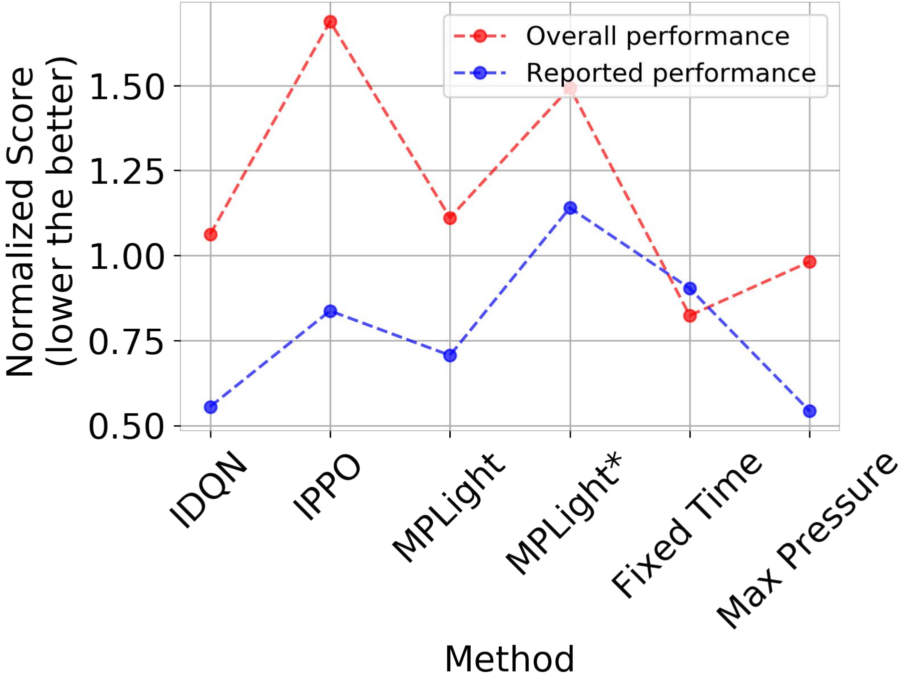
Overall Performance Across an MDP Family

- Overall performance of a method *R* on task *T* given a point MDP family set *U*

$$E_R^T = \sum_{i=1}^{|U|} \underbrace{s_{R,i}}_{\text{normalized performance score}} \times \underbrace{p_{T,i}}_{\text{normalized importance score}}$$

- **Assumption:** Given a task *T*, $p_{T,i}$ is known.

Case Study: Re-Evaluation

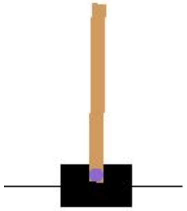


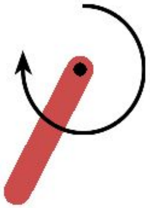
Rank	Reported	Overall
1	Max Pressure	Fixed Time
2	IDQN	Max Pressure
3	MpLight	IDQN
4	IPPO	MPLight
5	Fixed Time	MPLight*
6	MPLight*	IPPO

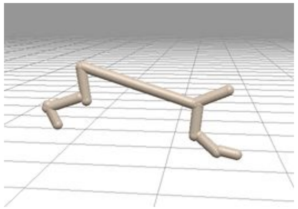
*Reported performances are based on re-evaluations of the methods on Ingolstadt single intersection.

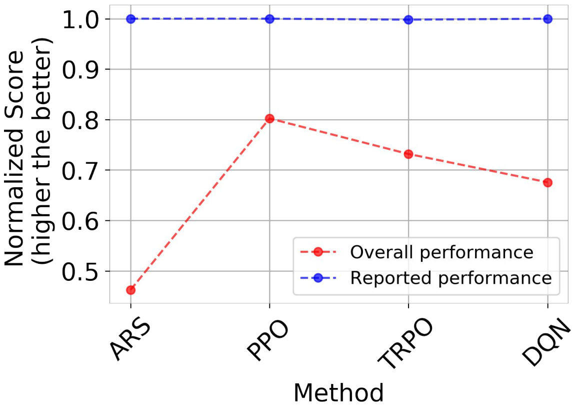
Results reported here should not be illustrated as evidence against using DRL for traffic signal control and should only be used as evidence of shortcomings in point MDP based evaluations. Further studies are encouraged to study the overall benefits of DRL for traffic signal control without the point MDP based assumptions.

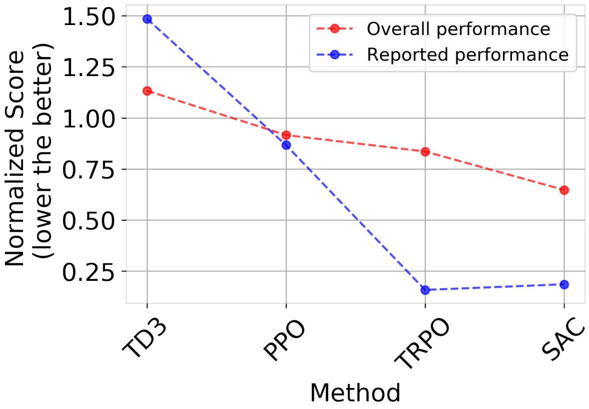
Re-Evaluation of Popular Control Tasks

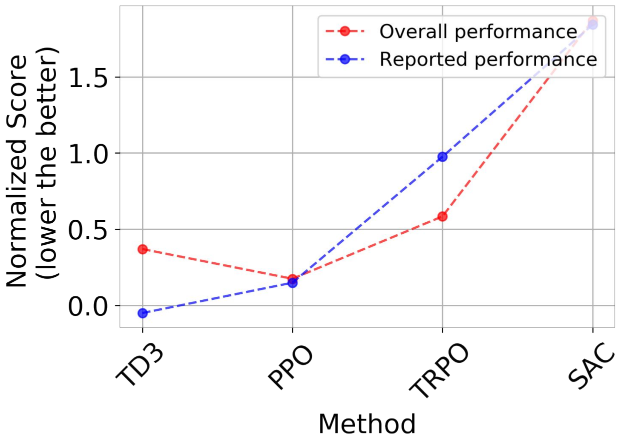












*Reported performance of each task is measured by training and evaluating DRL methods on commonly used single point-MDP given in common benchmark suites.

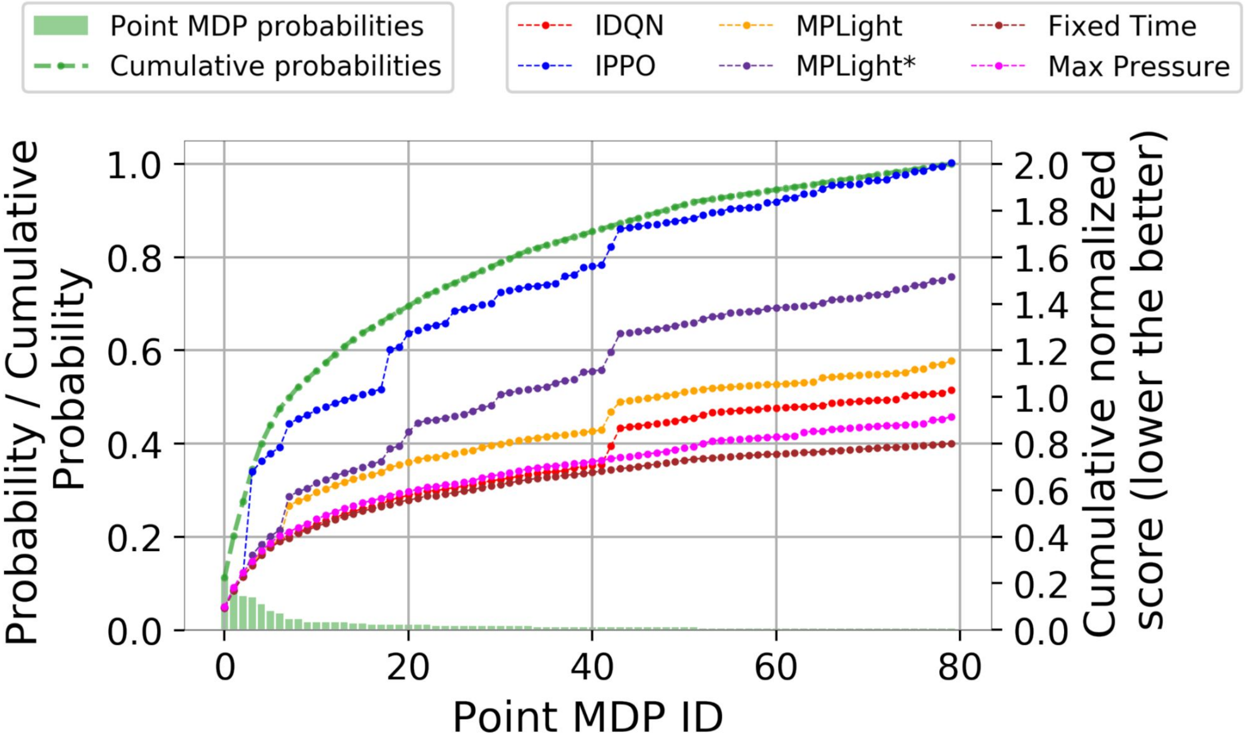
Challenges in MDP Family-based Evaluations

MDP Families for Benchmarking

- Create benchmark control tasks that depict MDP families.
- Publish datasets of MDP families of control tasks including point-MDP distributions.
- Incentivize publication of such datasets and control task at leading conferences.

Performance Profiles

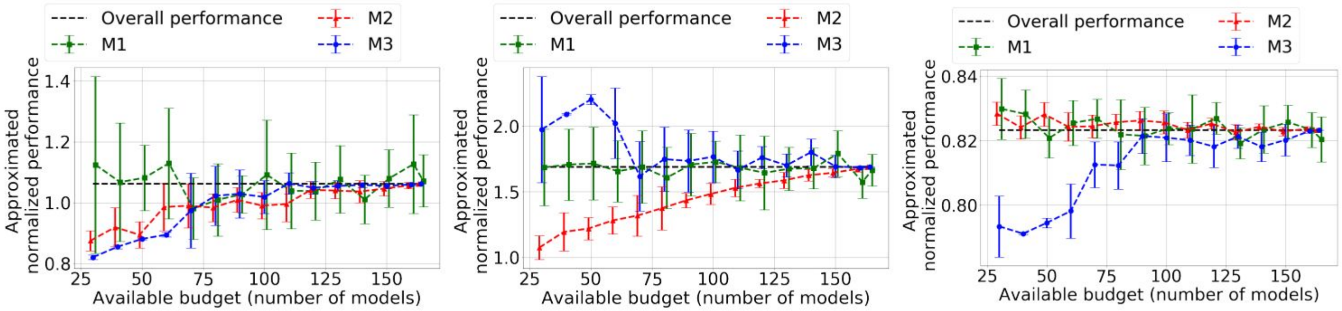
Performance Profile of Traffic Signal Control



Performance Approximations

- Adopt performance approximations using clustering and random sampling under a computational budget.
- Standardize the evaluations by making the selected point MDPs public.

M1: random sampling with replacements from the point MDP distribution
M2: random sampling without replacements
M3: clustering point MDPs using k-means and assigning probability mass of all point MDPs that belong to same cluster to its centroid



(a) IDQN (b) IPPO (c) Fixed Time

M3: only 50% of point-MDPs are needed to predict the overall performance

Takeaways

- **Point-MDP based performance evaluations can be misleading**
- Use MDP families instead of point-MDP based evaluations
- Use performance profiles for better reporting
- Do performance approximations when the budget is limited

More details: https://vindulamj.github.io/rl-evaluations/