BUILDING CONTROL USING DEEP REINFORCEMENT LEARNING

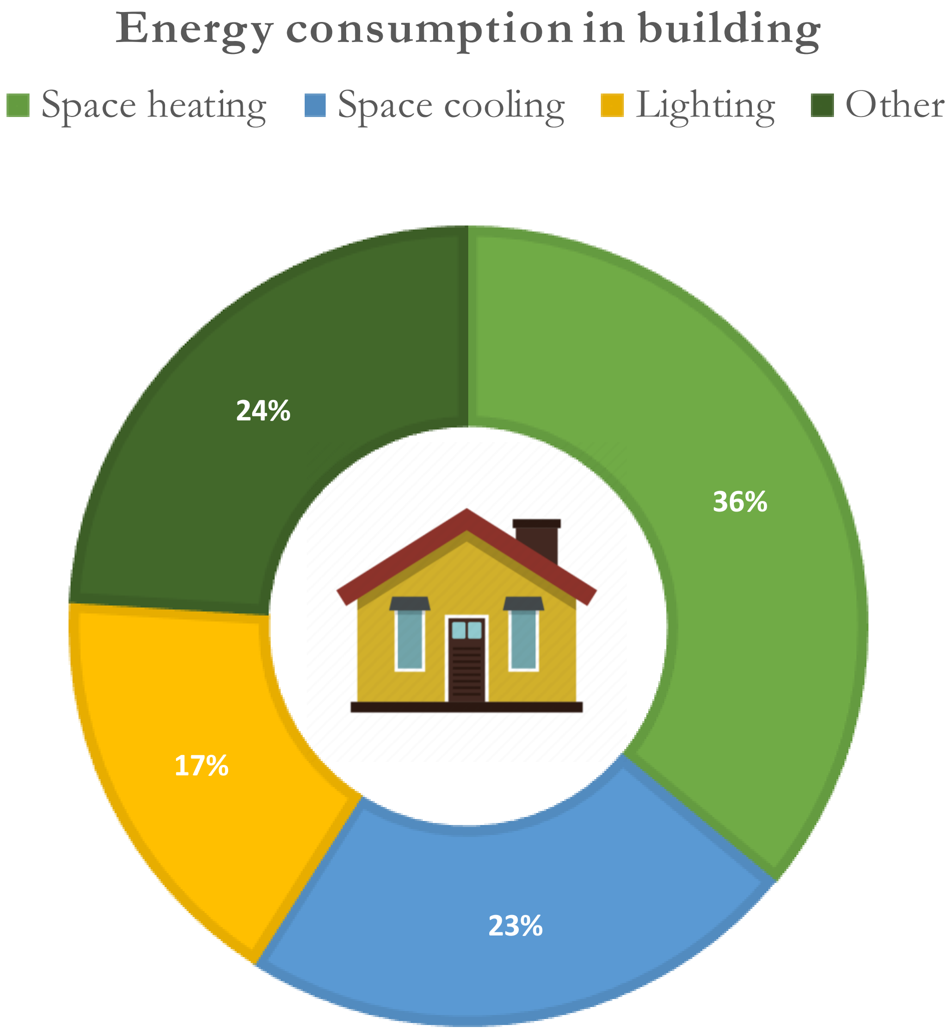
Tanmay Ambadkar

Rosina Adhikari

Background

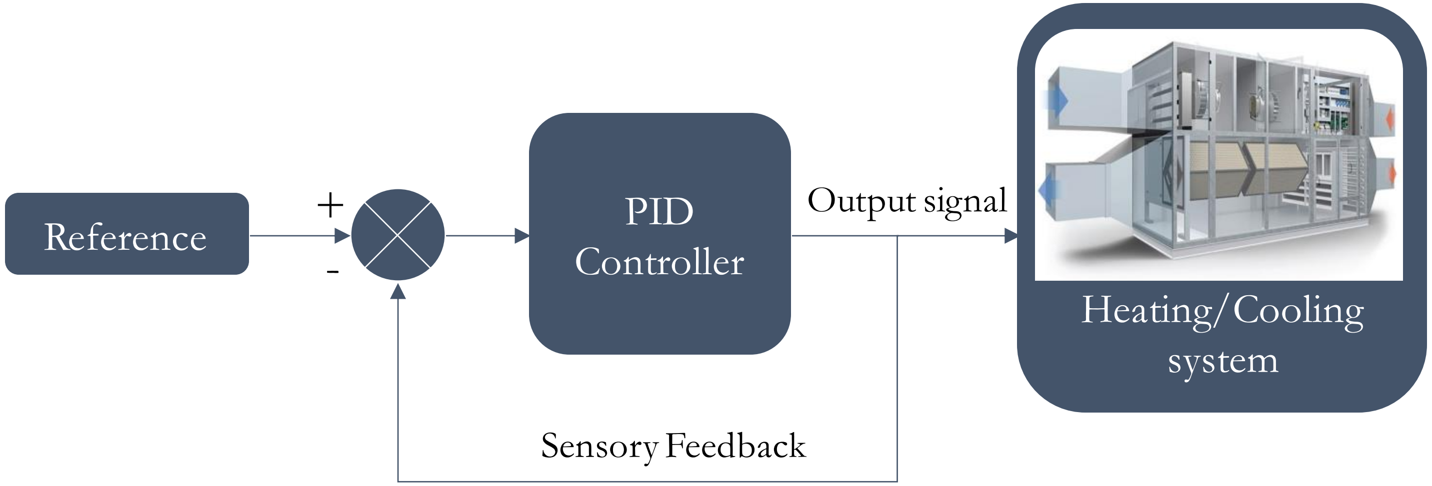
According to the U.S. Energy Information Administration (EIA), in 2020, the residential and commercial **building sectors** represented about **35%** **of the total energy** consumed in the United States.

Space **heating and space cooling** account for around **60%** of total energy consumed in a building.



How are building systems controlled today?

A rule-based control system that operates on a fixed set of rules.

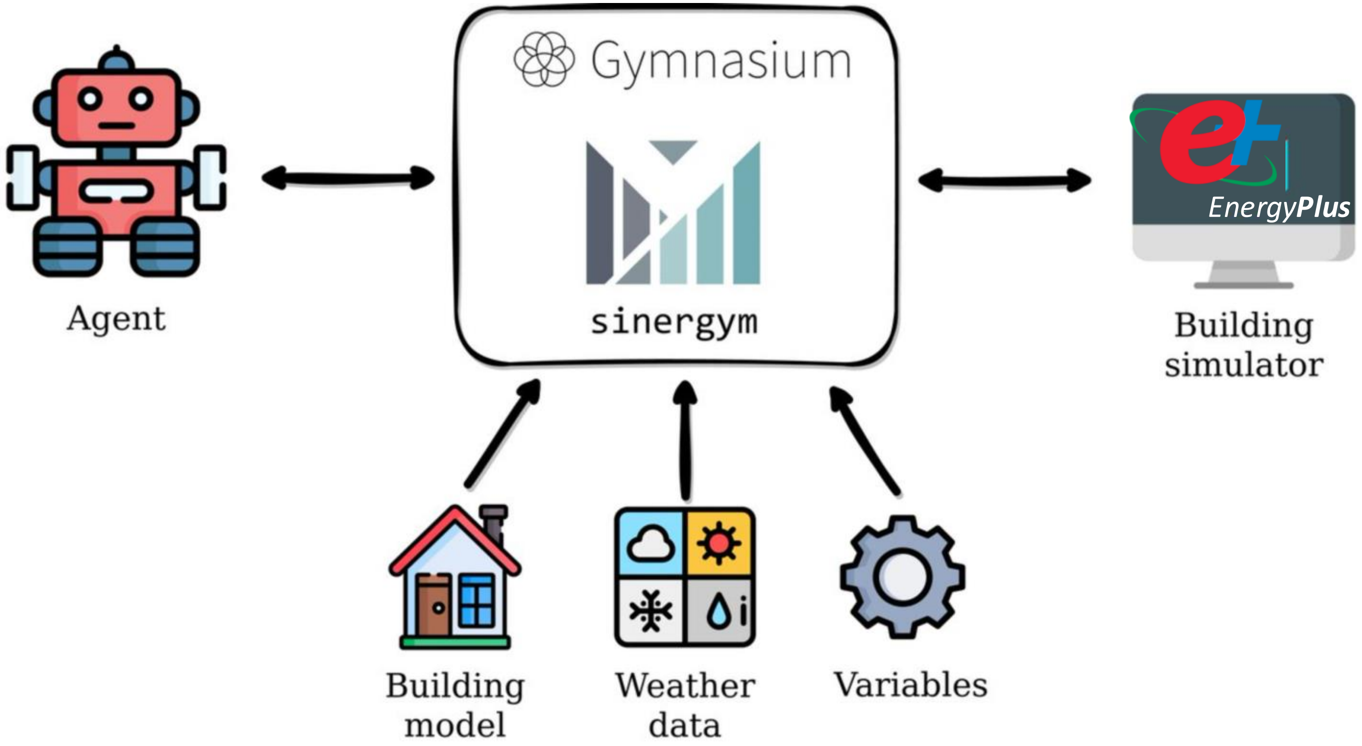


Unable to adjust to complex, dynamic environments inside building, for-example: changing thermal demand, price fluctuation, dynamic occupancy.

Why RL-based Control?

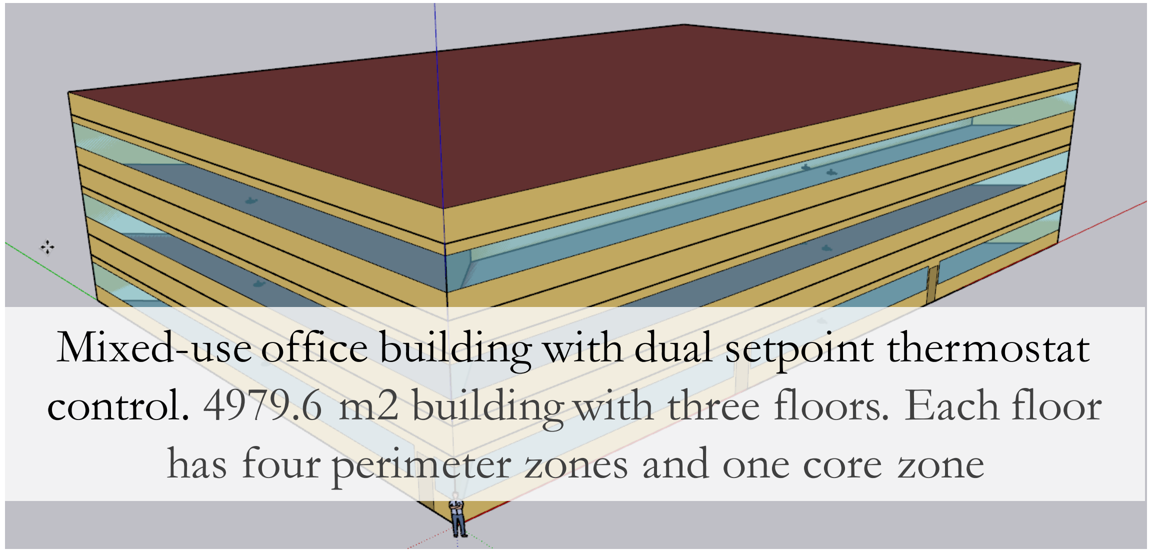
- Adaptive control
- Multi-objective Optimization
- Scalability
- Fault detection

Reinforcement Learning control framework



Jiménez-Raboso, J., Campoy-Nieves, A., Manjavacas-Lucas, A., Gómez-Romero, J., & Molina-Solana, M. (2021). Sinergym: A Building Simulation and Control Framework for Training Reinforcement Learning Agents. In *Proceedings of the 8th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation* (pp. 319–323). Association for Computing Machinery.

Environment



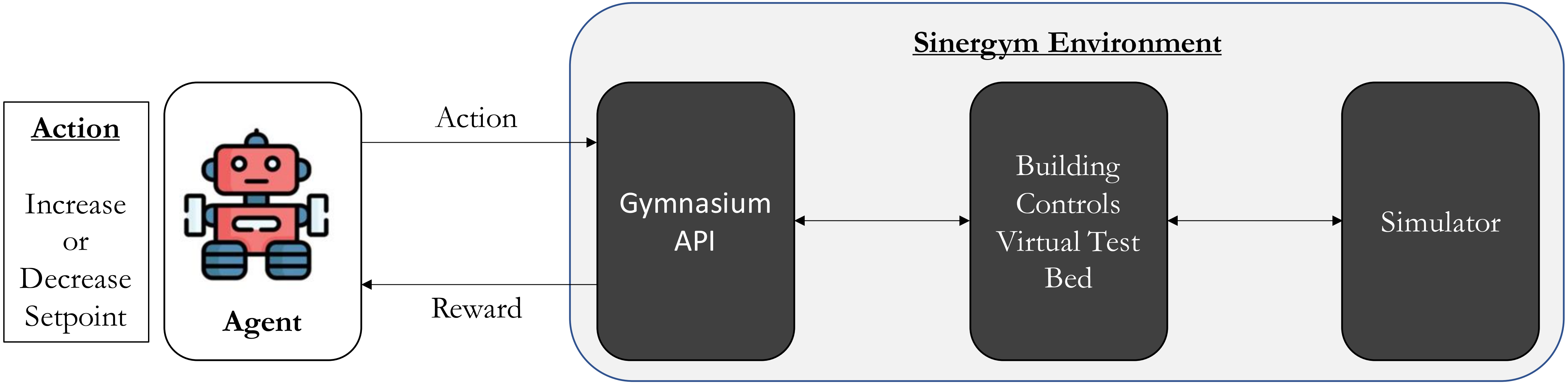
Mixed-use office building with dual setpoint thermostat control. 4979.6 m2 building with three floors. Each floor has four perimeter zones and one core zone

<u>Input Parameters</u>

Thermostat heating setpoint,
Thermostat cooling setpoint
Zone Air Temperature (for each zone)

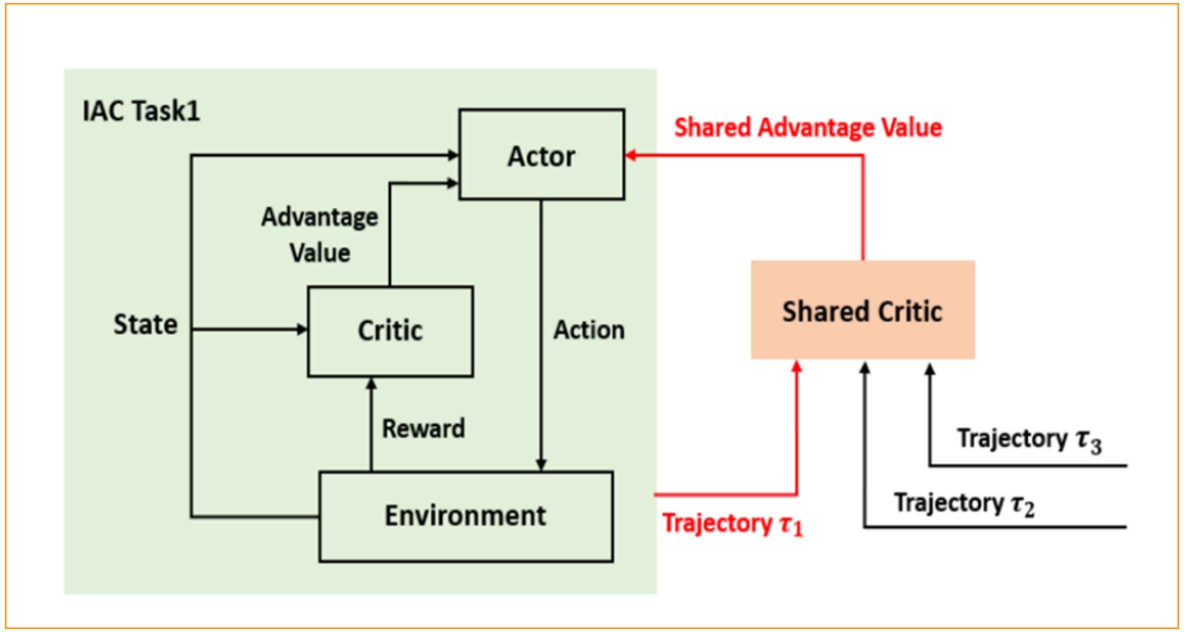
Reward Functions as a trade-off between Energy and Comfort

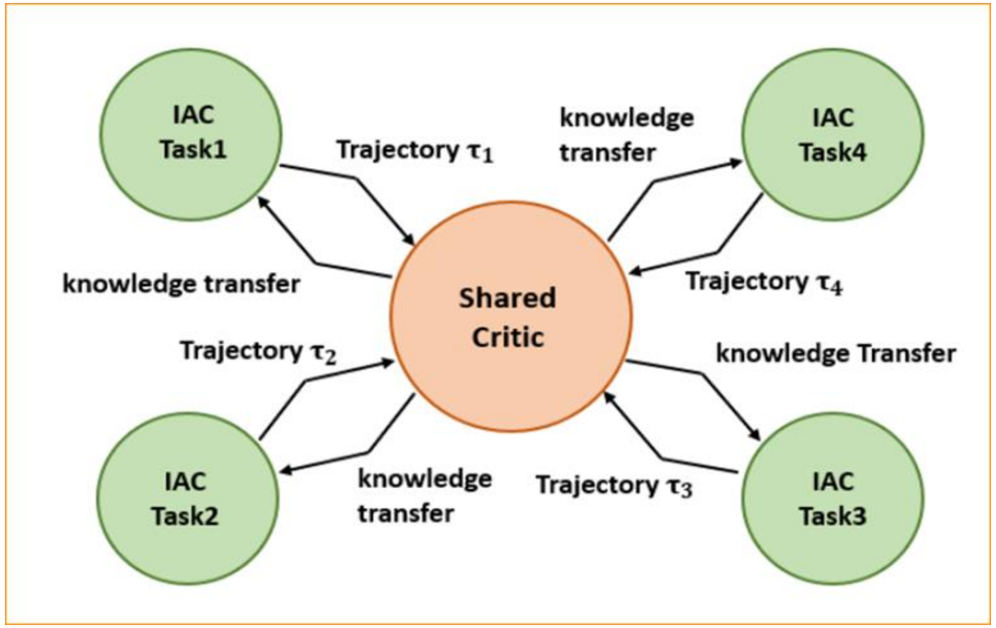
$$R_t = -w\lambda_E E_t - (1-\omega)\lambda_T(|T_t - T_{up}| + |T_t - T_{low}|)$$



Implementation

- Algorithm – PPO (baseline) & MultiTaskPPO





Zhang, G., Feng, L., & Hou, Y. (2021). Multi-task Actor-Critic with Knowledge Transfer via a Shared Critic. In Proceedings of The 13th Asian Conference on Machine Learning (pp. 580–593). PMLR.

Algorithm

Algorithm 1 Multi-task actor-critic with a shared critic

Input: State s: Reward r

Parameter: Task number $i = 1, 2...M$; Number of Episode E; Maximum steps of task per episode T; Transfer weight α;

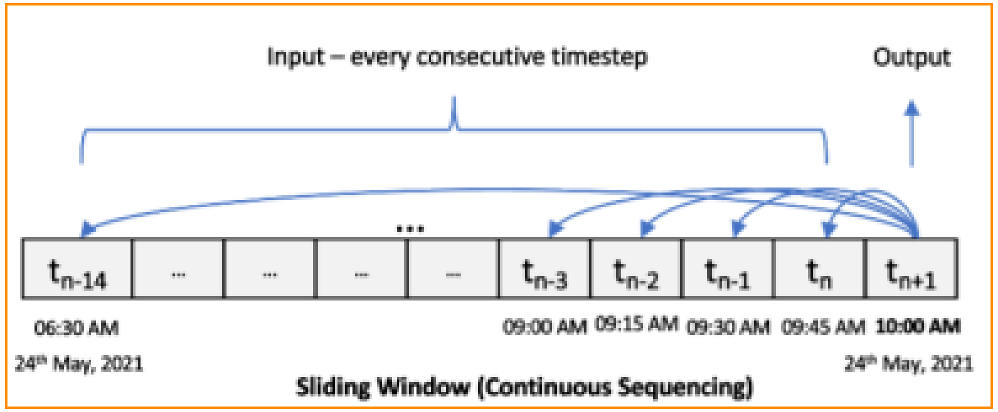
Output: Critic $\eta_{i,E}$ actor $\theta_{i,E}$;

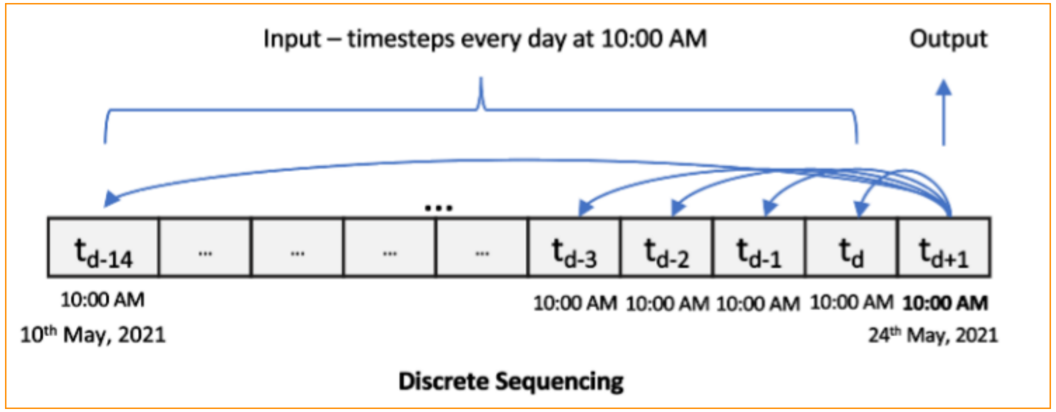
Randomly initialise actor π_{θ_i} and critic network $V_{\eta_{ni}}$ for task i, $i = 1, 2...m$;
Initialize episode counter e, $e = 0$
while $e_i < E_i$ **do**
  for each task i **do**
    Set step counter $t = 0$
    while $t < T_i$ and not terminal state **do**
      Select action $a_{i,t} \sim \pi_{\theta_i}$.
      Execute $a_{i,t}$ and state $r_{i,t+1}$ and $s_{i,t+1}$
      Store tuple $(s_{i,t}, a_{i,t}, r_{i,t+1}, s_{i,t+1})$
      Update step counter: $t \leftarrow t + 1$
    end while
    for each sample **do**
      Compute advantage value using Eq.10
      Compute shared advantage value using Eq.8
    end for
    Compute critic gradient using Eq.12
    Compute actor gradient using Eq.11
    Compute shared critic gradient using Eq.6
  end for
  Update episode counter $e \leftarrow e + 1$
end while
return Critic and actor weights: $\theta_{i,E}, \eta_{i,E}$;

Zhang, G., Feng, L., & Hou, Y. (2021). Multi-task Actor-Critic with Knowledge Transfer via a Shared Critic. In Proceedings of The 13th Asian Conference on Machine Learning (pp. 580–593). PMLR.

Models

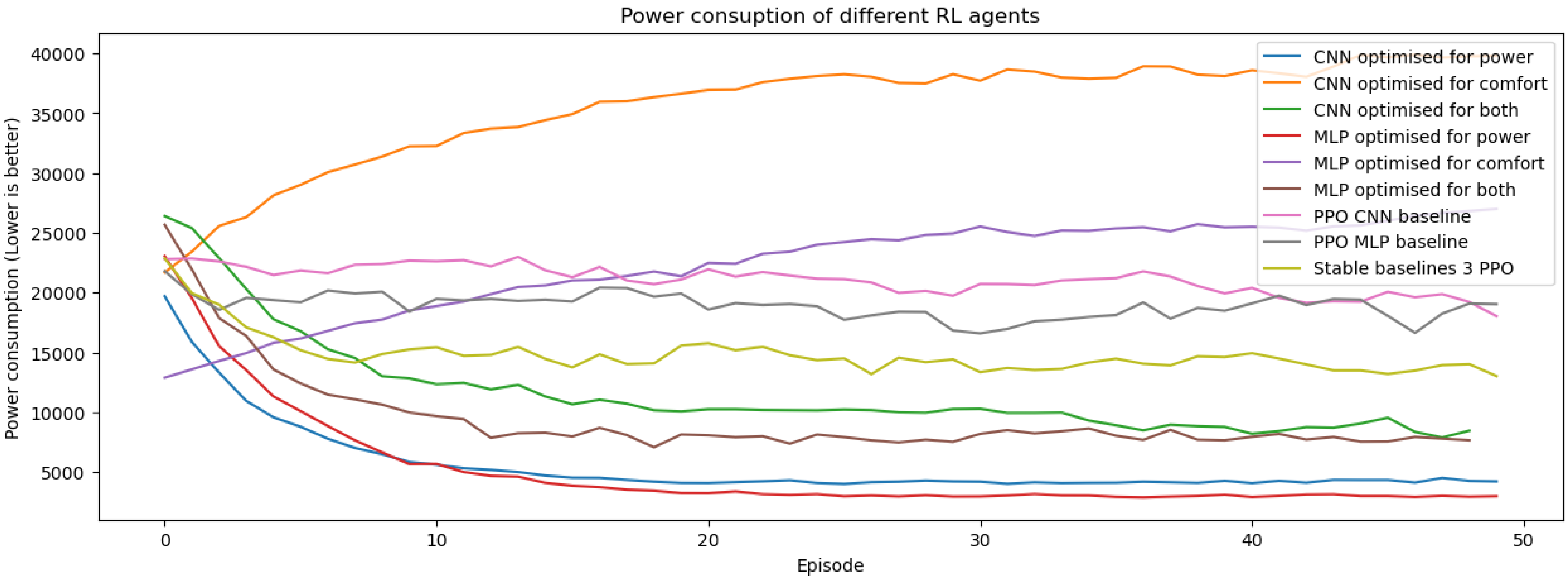
- MLP
- Time-discrete CNN (1D)



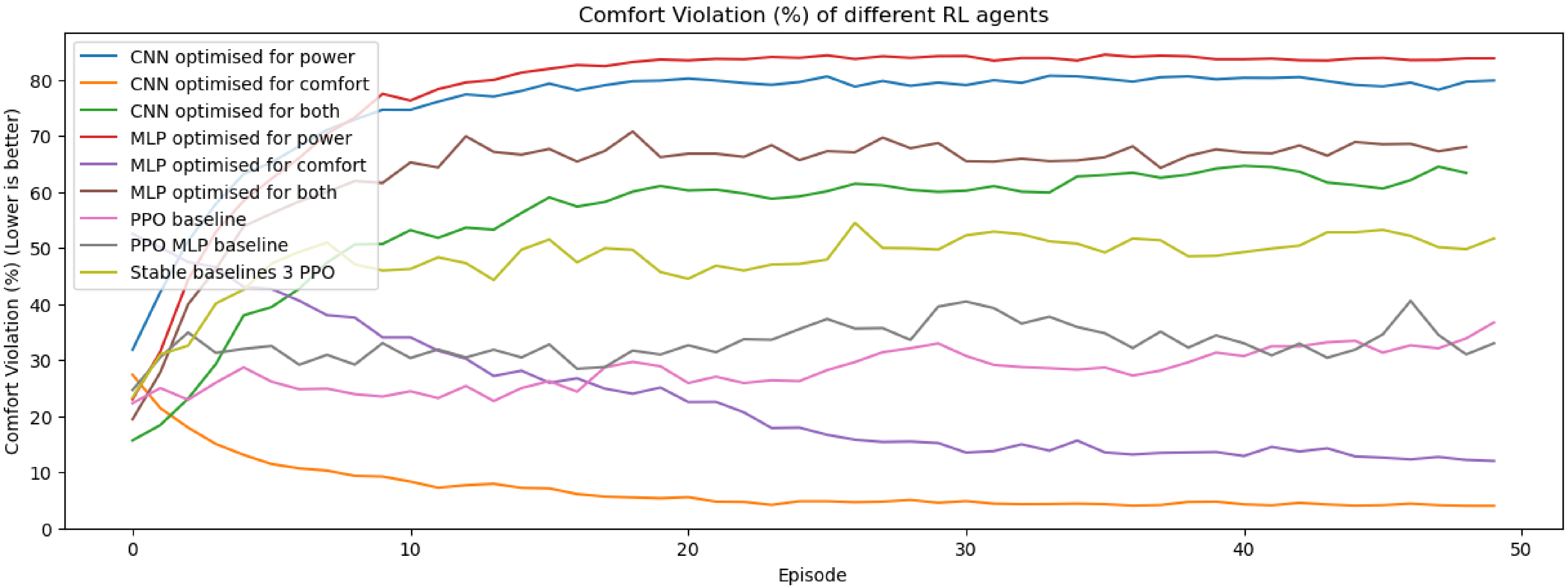


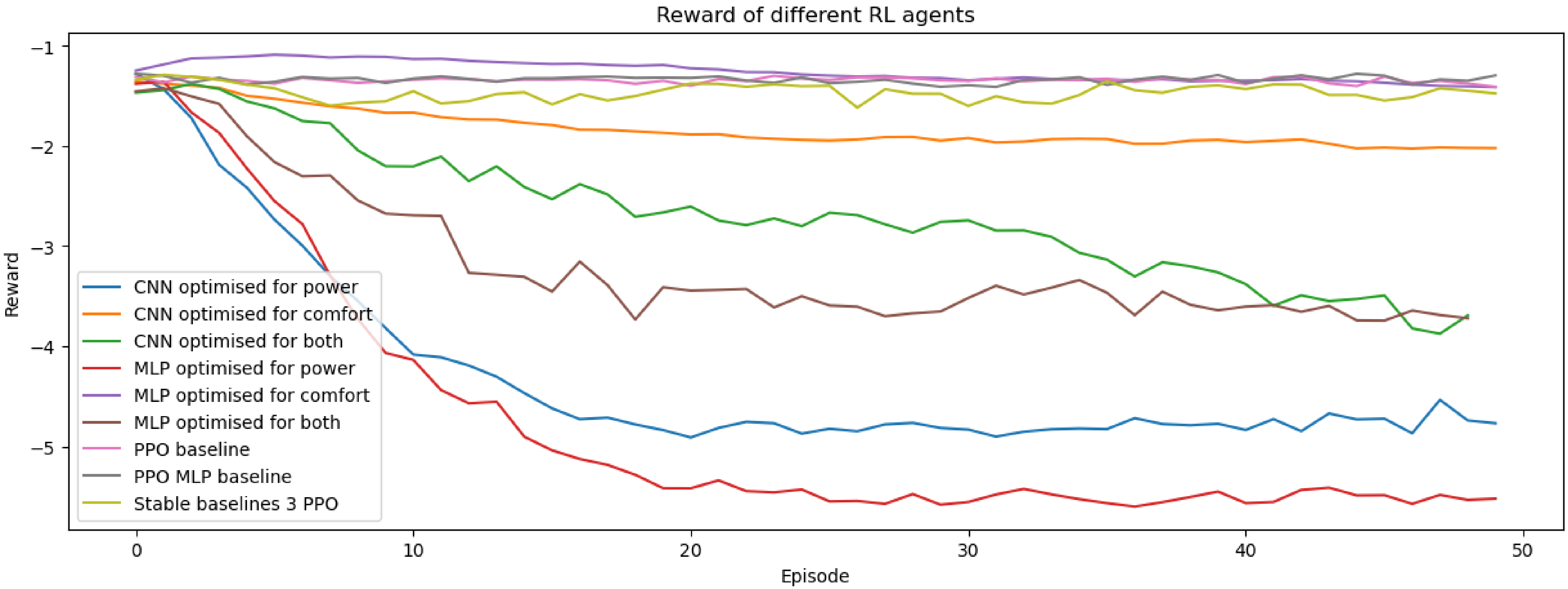
Nisha Menon, Shantanu Saboo, Tanmay Ambadkar, & Umesh Uppili (2022). Discrete Sequencing for Demand Forecasting: A novel data sampling technique for time series forecasting. In *3rd International Conference on Intelligent Data Science Technologies and Applications, IDSTA 2022, San Antonio, TX, USA, September 5-7, 2022* (pp. 61–67). IEEE.

Results



Power consuption of different RL agents
Power consumption (Lower is better)
40000
35000
30000
25000
20000
15000
10000
5000
0
10
20
30
40
50
Episode
CNN optimised for power
CNN optimised for comfort
CNN optimised for both
MLP optimised for power
MLP optimised for comfort
MLP optimised for both
PPO CNN baseline
PPO MLP baseline
Stable baselines 3 PPO


Comfort Violation (%) of different RL agents
CNN optimised for power
CNN optimised for comfort
CNN optimised for both
MLP optimised for power
MLP optimised for comfort
MLP optimised for both
PPO baseline
PPO MLP baseline
Stable baselines 3 PPO
Comfort Violation (%) (Lower is better)
Episode
0
10
20
30
40
50
60
70
80


Reward of different RL agents
Reward
Episode
−1
−2
−3
−4
−5
0
10
20
30
40
50
CNN optimised for power
CNN optimised for comfort
CNN optimised for both
MLP optimised for power
MLP optimised for comfort
MLP optimised for both
PPO baseline
PPO MLP baseline
Stable baselines 3 PPO

Thank You!