TAC: Towered Actor Critic for Handling Multiple Action Types in Reinforcement Learning for Drug Discovery

Sai Krishna Gottipati, Yashaswi Pathak, Boris Sattarov, Sahir, Rohan Nuttall, Mohammad Amini, Mathew E. Taylor, Sarath Chandar

Introduction

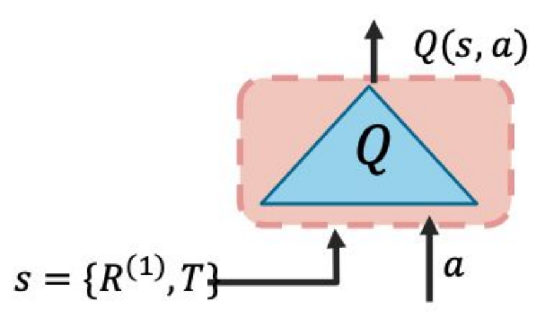
- Reinforcement Learning (RL) algorithms typically deal with only monotonic actions (discrete or continuous).
- However, in several pivotal real world applications like drug discovery, the agent needs to reason over different types of actions.
- To accommodate these needs, we propose a novel framework: **TAC (**towered actor critic) architecture that aims to handle multiple action types.

Contributions

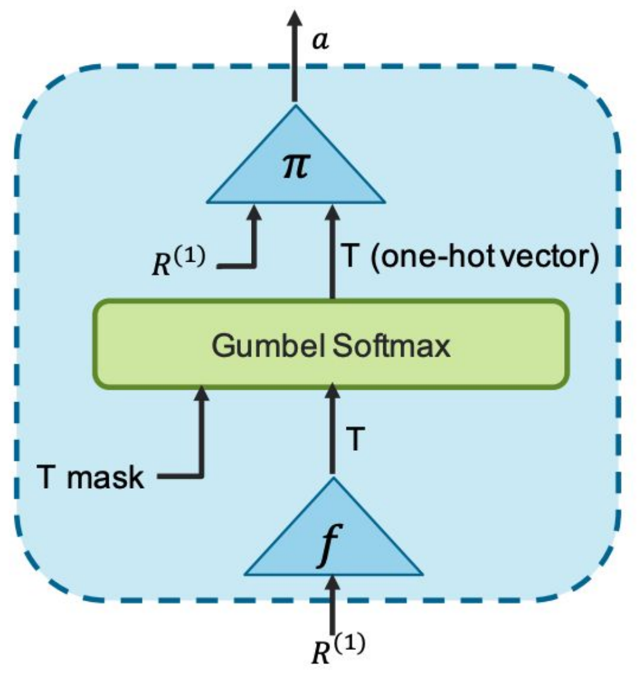
- Introduce a novel mechanism, towered actor critic to train agents on MDPs with multiple action types.
- TAC is applied to the task of reaction based de novo drug design (where there are different action types at every time step) and show significant improvement over the existing state-of-the-art results.
- TAC is also applied to standard RL tasks that do not have an inherent pyramidal/hierarchical structure of actions. Performance is comparable, or improved, relative to TD3 on several continuous action openAI gym environments.

Method

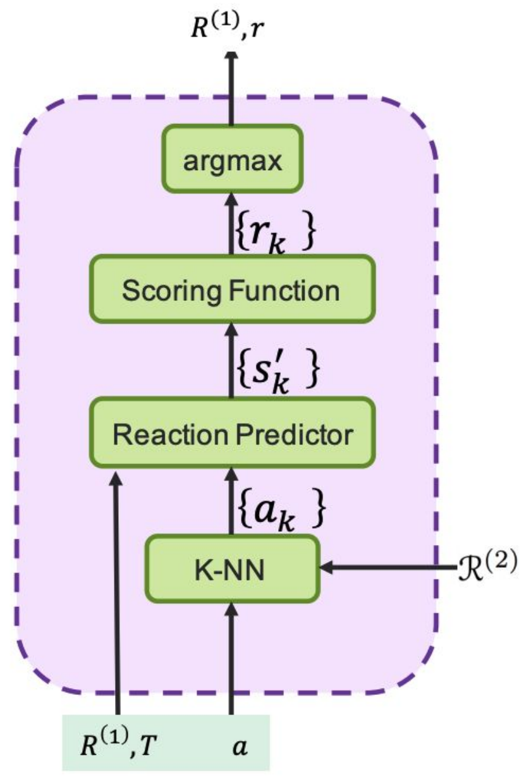
The TAC formulation is developed on top of our previous work for *de novo drug design* called Policy Gradient for Forward Synthesis (PGFS). PGFS is an off-policy algorithm having an actor module that consists of *f* and π networks. The *f* network takes in the current state *s* (reactant R$^{(1)}$) as input and outputs reaction template *T*. The π network uses R$^{(1)}$ and *T* to compute action *a* (feature representation for second reactant). Inside the environment, a KNN uses a and computes the *k* closest valid second reactants among which the maximum rewarding reactant is selected as R$^{(2)}$. R$^{(1)}$, R$^{(2)}$, and T are used to simulate a reaction and transition to the next state. PGFS uses TD3 for updating the actor (*f*, π) and critic (*Q*) networks.



Critic



Actor



Environment

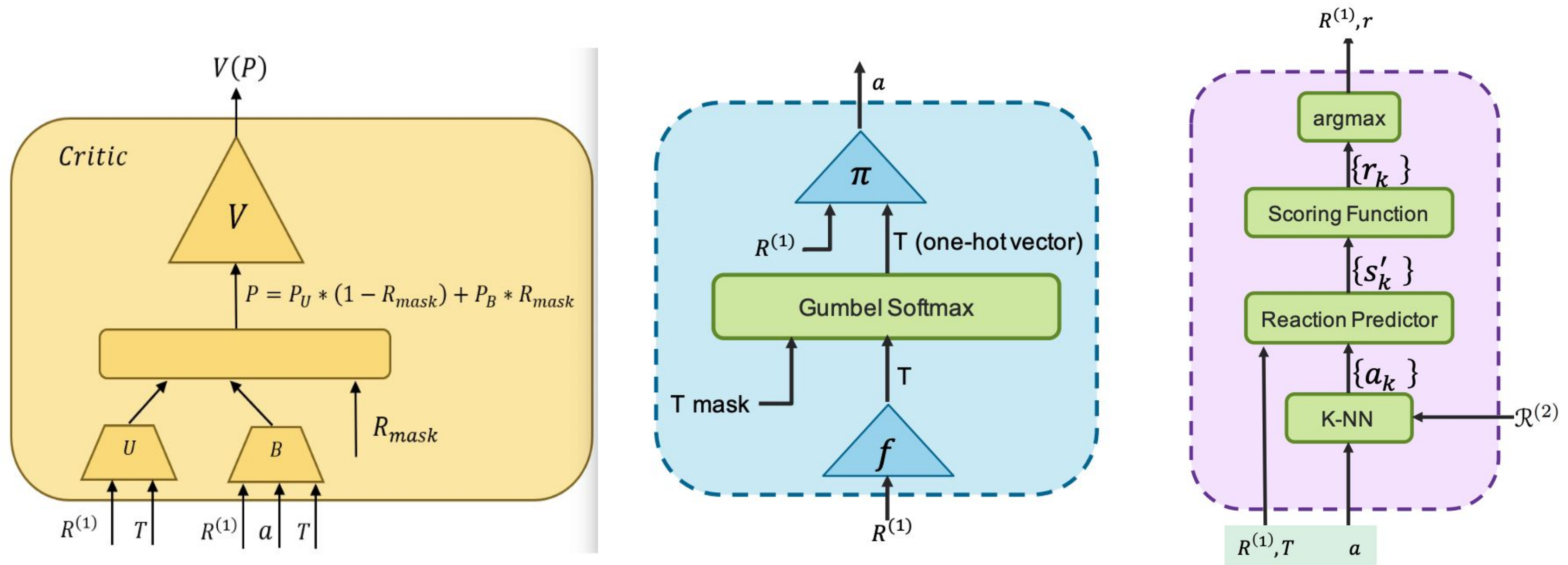
In PGFS, the parameters of the actor and critic networks are updated as follows:

$$\min L(\theta^Q) = \frac{1}{N}\sum_i |y_i - Q(R_i^{(1)}, \{T_i, a_i\})|^2 \quad (1)$$

$$\min L(\theta^{f,\pi}) = -\sum_i \text{Critic}(R_i^{(1)}, \text{Actor}(R_i^{(1)})) \quad (2)$$

$$\min L(\theta^f) = -\sum_i (T_i^{(1)}, log(f(R_i^{(1)}))) \quad (3)$$

- PGFS uses two types of reaction templates (actions): uni-molecular and bi-molecular.
- uni-molecular reactions do not require a second reactant but, the parameters of the pi-network are updated.
- We thus propose a novel critic architecture to independently handle both the action types (uni-molecular and bi-molecular).
- We first predict the product of the chemical reaction (i,e, next state) and then compute the value function for the predicted product:



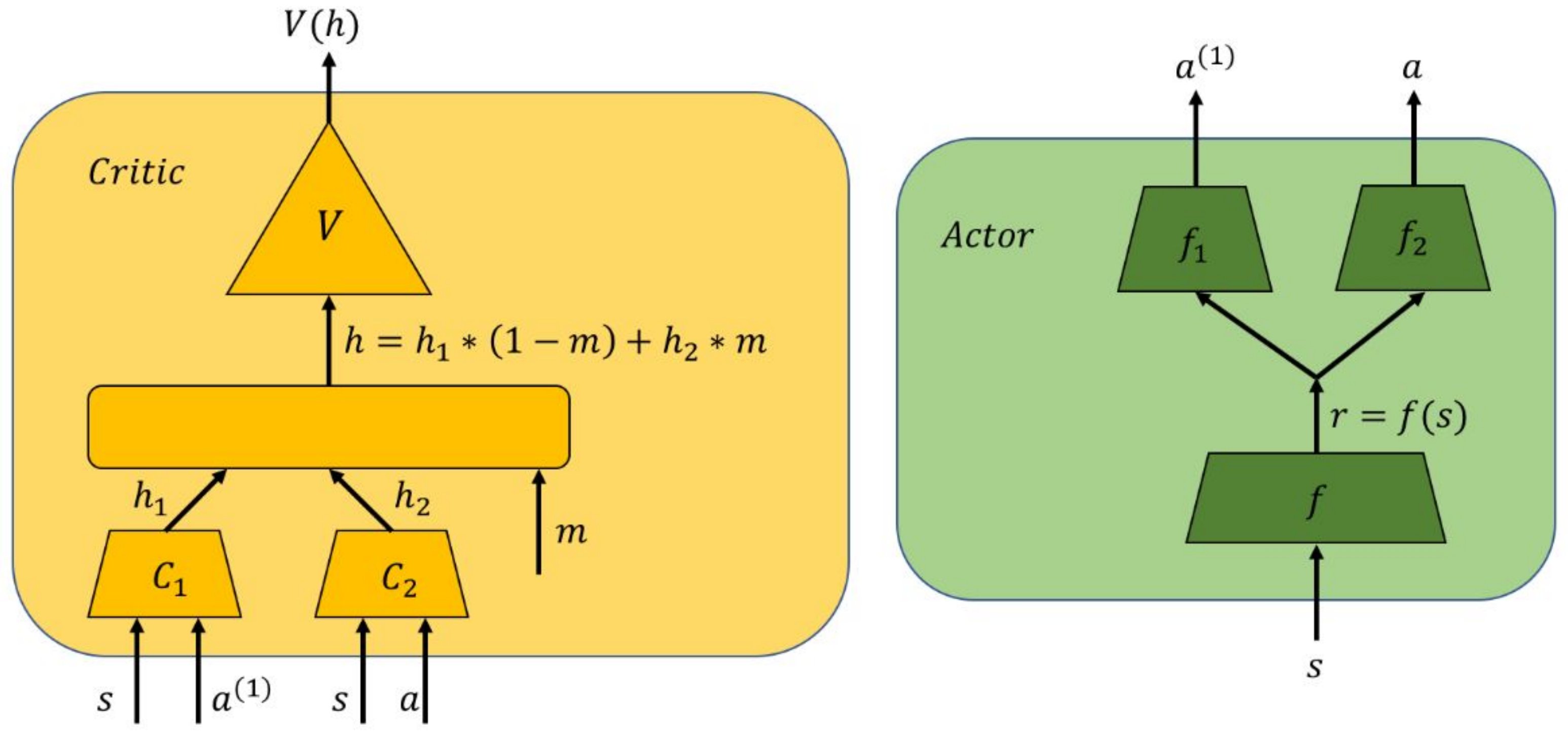
Environment

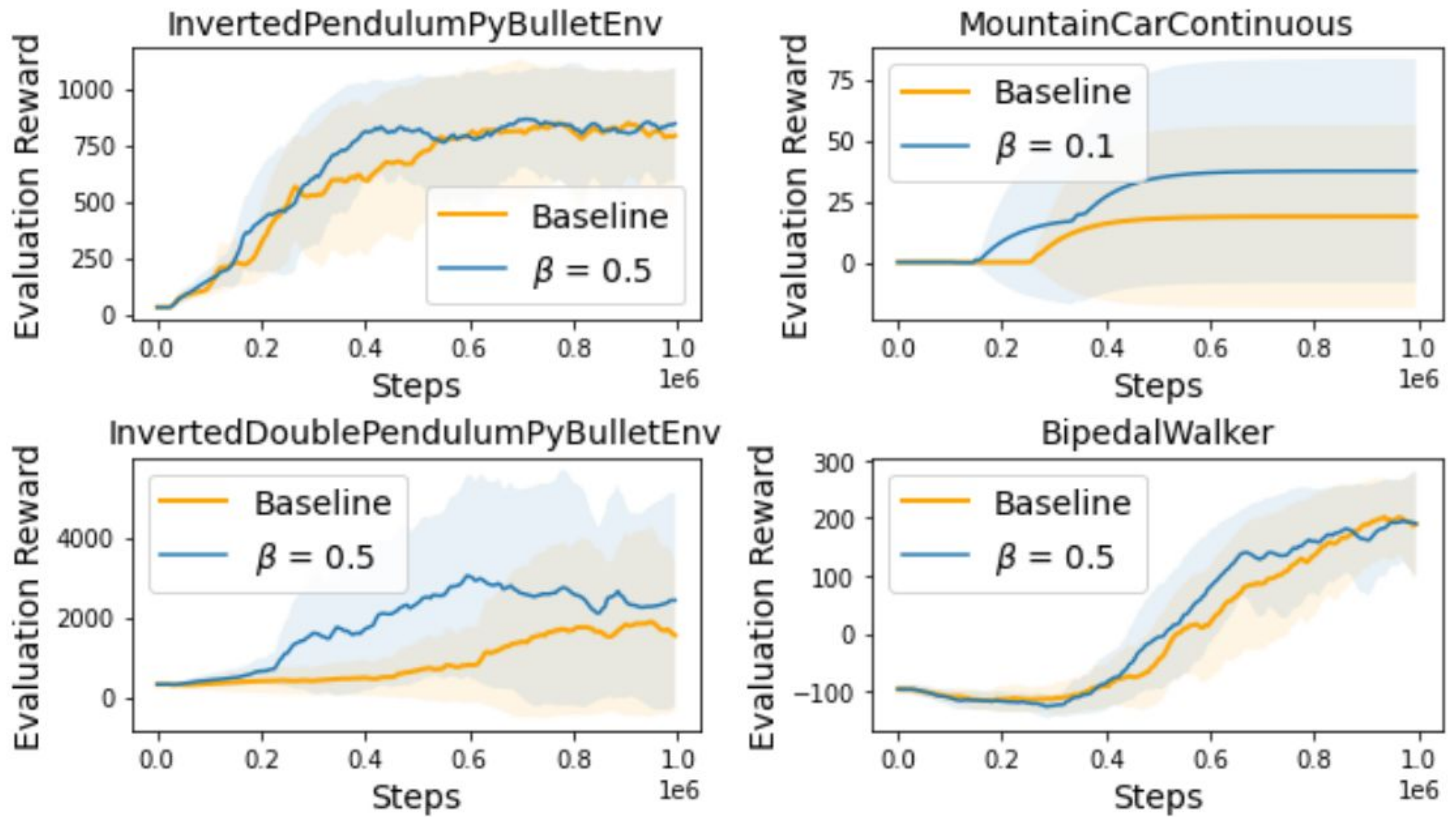
```
procedure BACKWARD(buffer minibatch)
    $T_{i+1}, a_{i+1}, R_{\text{mask}_{i+1}} \leftarrow \text{Actor-target}(R_{i+1}^{(1)})$
    $a_{i+1} \leftarrow a_{i+1} + \epsilon, \quad \epsilon \sim \text{clip}(\mathcal{N}(0, \tilde{\sigma}), -c, c)$
    $A_{i+1} \leftarrow \{T_{i+1}, R_{mask_{i+1}}, a_{i+1}\}$
    $P_{i+1}, Q_{i+1} \leftarrow \min_{j=1,2} \text{Critic-target}_j(R_{i+1}^{(1)}, A_{i+1})$
    $y_i \leftarrow r_i + \gamma Q_{i+1}$
    $P_i, Q_i \leftarrow \textsc{Critic}(R_i^{(1)}, \{T_i, R_{mask_i}, a_i\})$
    $\mathcal{L}_{\text{value}} \leftarrow \sum_i |y_i - Q_i)|^2$
    $\mathcal{L}_{\text{auxil}} \leftarrow |P_i - R_{i+1}^{(1)}|^2$
    $\mathcal{L}_{\text{critic}} \leftarrow \mathcal{L}_{\text{value}} + \alpha \mathcal{L}_{\text{auxil}}$
    $\mathcal{L}_{\text{policy}} \leftarrow -\sum_i \textsc{Critic}(R_i^{(1)}, \textsc{Actor}(R_i^{(1)}))$
    $\mathcal{L}_{\text{auxil-actor}} \leftarrow -\sum_i (T_i^{(1)}, log(f(R_i^{(1)})))$
    $\mathcal{L}_{\text{actor}} \leftarrow \mathcal{L}_{\text{policy}} + \beta \mathcal{L}_{\text{auxil-actor}}$
    $\min \mathcal{L}_{\text{actor}}, \mathcal{L}_{\text{critic}}$
```

Experimental Results:

Method	QED	clogP	RT	INT	CCR5
ENAMINEBB	**0.948**	5.51	7.49	6.71	8.63
RS	**0.948**	8.86	7.65	7.25	8.79 (8.86)
GCPN	**0.948**	7.98	7.42(7.45)	6.45	8.20(8.62)
JT-VAE	0.925	5.30	7.58	7.25	8.15 (8.23)
MSO	**0.948**	26.10	7.76	7.28	8.68 (8.77)
PGFS	**0.948**	27.22	7.89	7.55	9.05
TAC-FS	**0.948**	28.97	7.92	**7.75**	9.17
α-TAC-FS	**0.948**	**31.13**	**8.02**	**7.75**	**9.49***

TAC-generic





Conclusion

We propose a novel framework for handling multiple action types and demonstrated significant improvements over SOTA in de novo drug design. We further extended the framework to all gym environments and demonstrated better or comparable performance to TD3.