AdsorbRL: Deep Multi-Objective Reinforcement Learning for Inverse Catalysts Design

Romain Lacombe
Stanford University

Lucas Hendren
Stanford University

Khalid El-Awady
Stanford University

{rlacombe, hendren, kae}@stanford.edu

Abstract

A central challenge of the clean energy transition is the development of catalysts for low-emissions technologies. Recent advances in Machine Learning for quantum chemistry drastically accelerate the computation of catalytic activity descriptors such as adsorption energies. Here we introduce *AdsorbRL*, a Deep Reinforcement Learning agent aiming to identify potential catalysts given a multi-objective binding energy target, trained using offline learning on the *Open Catalyst 2020* and *Materials Project* data sets. We experiment with Deep Q-Network agents to traverse the space of all $\sim$160,000 possible unary, binary and ternary compounds of 55 chemical elements, with very sparse rewards based on adsorption energy known for only between 2,000 and 3,000 catalysts per adsorbate. To constrain the actions space, we introduce Random Edge Traversal and train a single-objective DQN agent on the known states subgraph, which we find strengthens target binding energy by an average of 4.1 eV. We extend this approach to multi-objective, goal-conditioned learning, and train a DQN agent to identify materials with the highest (respectively lowest) adsorption energies for multiple simultaneous target adsorbates. We experiment with Objective Sub-Sampling, a novel training scheme aimed at encouraging exploration in the multi-objective setup, and demonstrate simultaneous adsorption energy improvement across all target adsorbates, by an average of 0.8 eV. Overall, our results suggest strong potential for Deep Reinforcement Learning applied to the inverse catalysts design problem.

1 Introduction: Challenges of Catalysts Design

A central challenge of the clean energy transition is the development of high-performance catalysts for electrochemical and thermocatalytic energy conversion [1] [2]. The problem of catalysts design [3], which seeks to identify high-performance materials with increased intrinsic activity for any desired reaction of interest, is a critical technology enabler for low-emissions technologies, from H_2 production from water and renewable energy, to the transformation of waste CO_2 into valuable non-fossil fuels and feed-stock [4]. Rapid progress in Machine Learning (ML) methods for computational quantum chemistry [5], along with recently available datasets of catalysts [6] and their properties [7], could significantly accelerate the identification of materials with enhanced catalytic activity.

A key descriptor of catalytic activity is **adsorption energy**, or the energy with which the reagent species or reaction intermediate (**'adsorbate'**), a small molecule on a surface site (e.g. $H_2O\star$), binds to the surface of the heterogeneous **catalyst** (here an inorganic compound nano-particle). A core tenet of heterogeneous catalysis science, the Sabatier Principle [8], holds that the optimal catalyst should have a binding energy with the reactants that is neither too weak nor too strong.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Identifying materials that best match a target energy profile for multiple adsorbates, for instance strong binding with $\star$OH but weaker adsorption of $H_2O\star$, is thus particularly helpful for catalysts design [9]. Computational approaches to catalyst design have traditionally leveraged advances in computational chemistry to screen large chemical spaces for materials with optimal adsorption energies for key intermediates [10]. Despite advances in Density Functional Theory (DFT) [5] computation, estimating adsorption energy for a single (catalyst, adsorbate) pair still requires heavy computational resources, which makes *in silico* high-throughput screening of catalysts costly [11].

Inverse design adopts the opposite approach: starting from the desired property, the task aims to design catalysts from first principles so they best fit that objective [12]. While ML techniques for inverse materials design show strong promise, synthesizability and physical realization of discovered materials is a challenge [13] compared to high-throughput screening of known materials.

We turn to Reinforcement Learning (RL) [14] to propose a third way: training an agent to traverse a space of materials, not by exhaustive search, but by gravitating towards optima for the target property.

We introduce *AdsorbRL*, a Deep Reinforcement Learning (DRL) [15] agent trained to traverse a space of materials and identify promising catalysts given a multi-objective binding energy target. Specifically, we use Offline RL [16] on the *Materials Project* [7] and *Open Catalyst 2020* [6] datasets of adsorption energies to train a DRL agent to identify catalysts which bind the strongest (lowest adsorption energy) or the weakest (highest adsorption energy) with an array of target adsorbates, chosen for their importance for the clean energy transition.

We hypothesize that RL can be especially helpful to navigate chemical space in the multi-objective setting. By learning to traverse a sparse rewards environment in a multi-objective goal-conditioned setup, our agent could help identify materials with desirable properties for the multi-objective target at hand, and serve as an *in silico* rapid screening mechanism to identify leads on which to further focus computational chemistry resources (DFT computations, MD simulations, etc.).

This paper presents our experiments with several Deep RL setups to traverse a large space of compounds, and identify materials with the desired adsorption energies profile for a set of adsorbates of interest. We introduce Multi-Objective DQN with Sub-Sampling, and a novel algorithm to train such a generalized multi-objective agent.

Our findings indicate the promise of Deep RL in navigating complex chemical spaces, and present novel approaches to tackle goal-conditioned multi-objective Reinforcement Learning for materials design. These methods could serve as a foundation for more complex computational challenges in large chemical spaces, and the development of novel materials for a wide range of applications in heterogeneous catalysis, electrochemical energy conversion, and low-emissions technologies.

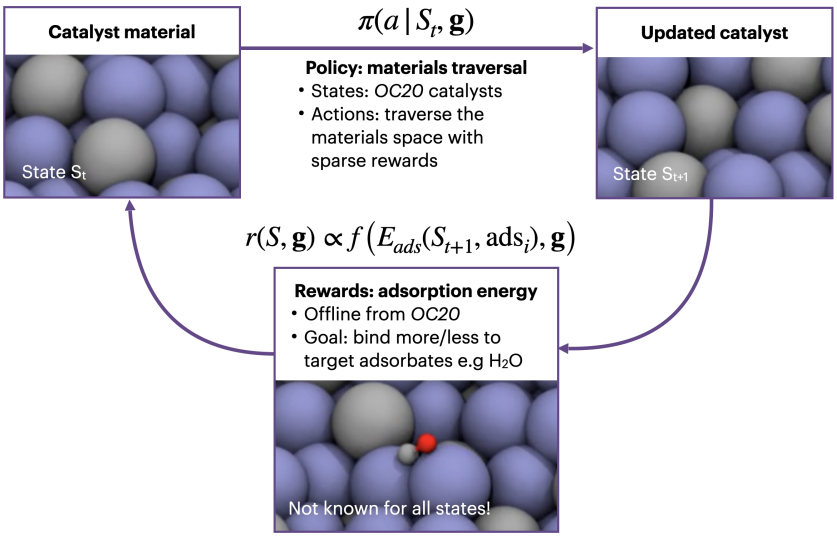

Figure 1: Deep RL approach to the catalyst design problem. Images: *Open Catalyst Project.*

2 Related Work

2.1 Machine Learning for Adsorption Energy Predictions

The *Open Catalyst Project* [6] is a research effort between Meta Fundamental AI Research (FAIR) and Carnegie Mellon University using Artificial Intelligence to model and discover new catalysts for renewable energy conversion and storage. Quantum mechanical computational methods such as Density Functional Theory [5] can help evaluate new catalysts, but are severely limited by the very high computational cost of the higher precision levels of theory.

For that purpose, *Open Catalyst Project* has developed and released two public datasets, *Open Catalyst 2020 (OC20)* [6] and *Open Catalyst 2022 (OC22)* [17], to train ML models to efficiently approximate these calculations. These datasets together contain 1.3 million molecular relaxations with results from over 260 million DFT calculations. Importantly, Lan et al. [11] recently released *AdsorbML*, an hybrid ML and computational DFT model trained to predict the adsorption energy of an adsorbate onto the sorbent surface. Inference greatly accelerates computation and achieves final energies within $\sim$0.1 eV of ground truth.

2.2 Machine Learning for Catalysts Design

Rapid progress in Machine Learning has lead to increased interest in applying data-driven learning techniques to the identification of catalysts. Zitnick et al. [3] introduce the general problem of catalysts design, and provide an overview of Machine Learning approaches to the problem to motivate the introduction of the *Open Catalyst* dataset. Seh et al. [1] provide an overview of how chemists in the laboratory can combine experiment and theory to design catalysts in order to increase their intrinsic activity or their number of active sites.

Freeze et al. [12] review recent advances in inverse design for catalysts, including approaches based on Machine Learning, as well as more general optimization (gradient ascent, genetic algorithms), but do not point to Reinforcement Learning. Zhu et al. [18] and Noh et al. [13] provide a comprehensive review of the published literature on Machine Learning for electrocatalysts and inorganic solid materials, respectively, including approaches using *OC20* data. While many different techniques have been applied to this problem space, none based on Reinforcement Learning are cited.

2.3 Reinforcement Learning for Molecular and Materials Design

In the more general area of inverse materials design, a number of Deep RL-based approaches have reported success. Zhou, et al. [19] apply the Deep Q-Network (DQN) algorithm [20] to molecular optimization. Their work demonstrates the effectiveness of a multi-objective reinforcement learning network for organic materials design, via a step-by-step addition or removal of elements or bonds.

Sridharan, et al. [21] explore the application of Deep RL to molecules generation by training a model to reconstruct 2D molecular graph based Nuclear Magnetic Resonance (NMR) spectroscopy. They designed an innovative framework using Monte-Carlo-Tree-Search (MCTS) and Graph Convolutional Networks to reconstruct the most likely molecular structure based on the chemical information extracted from NMR spectroscopy output. They demonstrated that applying their model in a RL setting helped consistently identify exact molecular structures based only on spectrograms.

Pan, et al. [22] apply Deep Q-Networks [20] to inverse inorganic materials design. This paper demonstrates the use of Deep RL for the identification of new materials based on desired properties, and was especially novel in its approach of inorganic materials design modeling, and how to integrate chemical constraints through Lagrange multipliers to impose charge neutrality and electro-negativity balance. This approach demonstrated an RL agent navigating trajectories in state space via successive additions of elements to an inorganic compound for the inverse design task.

Finally Sui, et al. [23] demonstrate the effectiveness of Deep RL for general digital materials design. They specifically show how RL improves on traditional approaches to find new design patterns even in vast design spaces, and show an application to additive manufacturing via the selection of soft or stiff voxels.

To our knowledge, Deep RL techniques have not been tried on sorbents and electrocatalyst design to date, which is what we propose in this paper.

3 Dataset and Features

Our dataset is sourced from the *Materials Project*, an open-access database for discovery of inorganic materials, which provides pre-computed properties for a large number of candidate catalysts [7]. Specifically, we use the *Catalysis Explorer*, an online application that provides pre-computed adsorption energies for various catalysts in the *Materials Project* under different configurations, sourced from the *Open Catalyst Project OC20* [6] dataset. We explore catalysts associated with the following 6 adsorbates, which are of major significance for clean energy:

- $\star\mathbf{OH_2}$: adsorbed water, a key reagent for the Oxygen Evolution Reaction (OER) in H_2 generation from water electrolysis [24], and a key product of the Oxygen Reduction Reaction (ORR) reaction for H_2 fuel cells [25];
- $\star\mathbf{OH}$: adsorbed hydroxyl radical, a key intermediate for OER [24] and ORR [25], often a rate-limiting step requiring the application of electrochemical over-potentials as its adsorption energy tends to scale linearly with those of $\star$O and $\star$OOH [9];
- $\star\mathbf{CH_4}$: adsorbed methane, of importance for direct air capture of natural gas, and for the CO_2 Reduction Reaction (CO_2RR) for e.g. methanation of captured carbon dioxide [26];
- $\star\mathbf{CH_2}$: adsorbed methylene radical, another important intermediate for CO_2 Reduction Reaction (CO_2RR) for the ethylene electro-catalystic production pathway, a key building block of the modern chemicals industry [4];
- $\star\mathbf{N_2}$: adsorbed molecular nitrogen, a key reagent for the Nitrogen Reduction Reaction (NRR) [27] for ammonia production, an essential ingredient of synthetic fertilizers without which half the world population would not be fed [28];
- $\star\mathbf{NH_3}$: adsorbed ammonia, the desired product of Nitrogen reduction (NRR) [27].

For each adsorbate compound, the *Materials Project Catalysis Explorer* provides a set of properties including the compound, bulk formula, adsorption energy, and Miller indices describing a particular lattice structure and orientation. For tractability purposes, and reasoning that the lowest energy configuration may be most representative of overall activity, we filter for all unique 1, 2, and 3-element compounds and select the **lowest adsorption energy configuration** for the target adsorbate among all combinations of stoichiometry and Miller indices, ignoring lattice orientation, and coordination environment.

This simplifies the materials space exploration problem to **traversal of the space of all $\sim$160,000 possible unary, binary and ternary compounds of 55 chemical elements** (Figure 4). Adsorption energies are known for only between 2,000 and 3,000 catalysts per adsorbate, providing a Given the **very sparse reward** for our agent, with only 7,386 total unique catalysts in our dataset of $\sim$160,000 possible compounds with known adsorption energy for at least one adsorbate (Table 4). We use the $\star\mathrm{OH}_2$ adsorbate for our single objective experiments, and the 6 target adsorbates above for the multi-objective setup. Finally, in experiments (3), (4) and (5), we **further constrain this problem to the *OC20*-subgraph** of compounds with known adsorption energy for at least one target adsorbate.

Our dataset is available for download on the AdsorbRL GitHub repository.

4 Model: Reinforcement Learning Setting

Our Reinforcement Learning model comprises of the following elements, illustrated in Figure 1.

States: S. States are unary, binary or ternary compounds of 55 elements matching to catalysts from the *Materials Project* dataset, comprising of 1 to 3 atomic elements forming the compound (e.g. SiC: Silicon carbide). Catalysts are represented by a 55 dimensional one-hot vector in our full generality setup (experiments (1), (3), (4), and (5)). In experiment (2), states are simply a one-element square on the Periodic Table of Elements grid, represented by their atomic number Z.

Actions: a. These are steps the agent can take to traverse the dataset of materials. They are defined differently for each experimental setup:

- Experiment (1): in the full states/full actions setup (section 5.1), actions can be: addition of an element (subject to the total elements in the catalyst being less than or equal to 3); removal

of an element (subject to there being at least one element in the catalyst); 'do-nothing', an action that leaves the catalyst as is; and a terminate action that ends the episode.

- Experiment (2): on the Periodic Table setup (section 5.2), only 5 actions are allowed: move left by one element, move right by one element, move up by one element, move down by one element, or stay in place. Trajectories automatically terminate after 9 steps.
- Experiment (3)—(5): in the full states/constrained actions setup (section 5.3 and 5.4), there are five possible actions: add a random element ('random edge'), remove the first, second, or third element in the catalyst, or terminate the episode.

Reward: $r \propto f\left(E_{ads}(S_i)\right)$**.** The reward for arriving in a state is a function of its adsorption energy (E_{ads}) for the target adsorbate. In single-objective experiments, we use $r = -E_{ads}$ and $r = E_{ads}^2$ to target terminal states with the strongest binding possible (measured by a large, negative energy value). For multi-objective setup we use $r = -E_{ads}^3$ to encourage the agent to find extrema with strongest (respectively, weakest) adsorption. Rewards functions for each experiment are reported in Table 5.

Termination: we either offer termination as an action option to the agent, or upon completion of a number of steps (e.g 9 steps).

Policy: $\pi(a|S)$**.** We train a policy to choose the next action at each state, using the Q-Learning [29] and Deep Q-Network algorithms [20] with the following Bellman equation [30]:

$$\mathrm{Q}^*(a|S) = r(a|S) + \gamma \max_a \left(\mathrm{Q}^*(a|S')\right)$$

Multi-objective goal-conditioning. In this setup, we train our policy to follow multiple objectives at once. We define an objective vector $\mathbf{g}$ as an array of +1 or -1 for each of the 6 adsorbates in our dataset, encoding whether we seek to bind stronger (+1, minimize energy) or weaker (-1, maximize energy). The reward becomes a function of E_{ads} for each adsorbate and the objective vector $\mathbf{g}$:

$$\pi(a|S) = \pi(a|S, \mathbf{g})$$

$$\mathbf{g} = (g_i) = (\dots, \{-1, +1\}_i, \dots)$$

$$r(S, \mathbf{g}) = f\left(E_{ads}(S_{t+1}, \mathrm{ads}_i)\right), \mathbf{g})$$

The Bellman equation [30] in the goal-conditioned, multi-objective setup becomes:

$$\mathrm{Q}^*(a|S, \mathbf{g}) = r(a|S, \mathbf{g}) + \gamma \max_a \left(\mathrm{Q}^*(a|S', \mathbf{g})\right)$$

Evaluation metrics. To evaluate our trained policies, we run a number of roll-outs from random initial states and compute the average adsorption energy of all final states reached by the policy. The delta with the average energy of initial states (respectively its inverse for weak binding targets in the multi-objective setup) measures how well our agent solves the problem:

$$\Delta = \frac{1}{N} \sum_{i \in \text{final}} -\left(E_{ads}(S_i)\right) - \frac{1}{N} \sum_{j \in \text{initial}} -\left(E_{ads}(S_j)\right)$$

5 Experiments & Results

We present the setup for each of the five experiments we report in Table 5.

Exp.	Target adsorbates	States	Actions	Algorithm	Reward
(1)	Single ($\star OH_2$)	All states	Graph traversal	DQN	$-E_{ads}$
(2)	Single ($\star OH_2$)	Periodic table	$\leftarrow \mid \rightarrow \mid \uparrow \mid \downarrow$	Q-Learning	E_{ads}^2
(3)	Single ($\star OH_2$)	*OC20*-subgraph	Random edges	DQN	$-E_{ads}^3$
(4)	Multi-objective	*OC20*-subgraph	Random edges	DQN	$-E_{ads} \cdot \mathbf{g}$
(5)	Multi-objective	*OC20*-subgraph	Random edges	DQN with Sub-Sampling	$-E_{ads_i} g_i \quad i \sim \mathcal{U}(i = 1 \dots 6)$

Table 1: Summary of experiments (1) to (5).

5.1 Offline RL on Full State & Actions Space

Experiment (1). Our first experiment attempts to train an agent to traverse the full compounds space from a random initial catalyst to the lowest energy catalyst using reward shaping [31]. Our offline data represents the set of all possible tuples, (s, a, r, s') where s and s' are existing catalysts in the dataset ('valid' states) and s' is reachable from s via a valid action, a (resulting in approximately 81,000 offline training tuples). The reward for a valid state is its inverse adsorption energy ($r = -E_{ads}$). We deem states with unknown adsorption energy 'invalid' and assign them a penalty ($-\lambda$) reward. We use a DQN model [20] with 2 hidden layers of sizes 512 and 64 and hyperparameters similar to those used later in sections 5.3 and 5.4 and train our network for 100,000 training steps.

We find that this approach does not yield a useful agent, and fails to noticeably converge towards desired states (Figure 3). A high penalty is needed to coax the agent to stay away from invalid states: adsorption energies of valid states are in the range of $(0, -10)$ eV and we use a penalty $\lambda \in (10, 200)$. We find the agent mostly learns to move quickly to an invalid state and terminate with an overall reward of $-\lambda$, and avoids longer episodes that would accumulate multiple penalties.

This can be understood by noting the sparsity of the state space. Referring to Figure 4, we see that the cardinality of the 'valid' set is around 2,379 catalysts, while it is $\simeq$160,000 for the whole space (all possible 3-element combos using one of 55 elements). This implies that over 98% of possible states are invalid and we hypothesize that the sparsity of the data makes it too hard to learn to navigate to the optimal state. Figure 3 reports the low success rate for varying values of λ.

5.2 Simplified State & Actions Space: Periodic Table of Elements

Experiment (2). To test our hypothesis about the sparse reward issues faced in experiment (1), we aim to simplify our RL setup as much as possible to test whether agents trained in a more tractable setup do exhibit lower average terminal state energies. We implement the simplest environment for an agent to learn fundamental chemical knowledge: *GridWorld* [29] on the Periodic Table of Elements [32]. This setup is novel to our knowledge and can serve as a building block to learn more complex cheminformatics problems.

We define our Markov Decision Process (MDP) as follows:

- 86 single element states (all elements from atomic number $Z = 1$ (H) to $Z = 86$ (Rn);
- 5 actions: $\{\_| \leftarrow | \rightarrow | \downarrow | \uparrow\}$ to move between elements on the periodic table (see figure 4);
- Episodes last for a set 9 steps duration (long enough to reach the optimal element from any random starting point on the table).

We find that the Q-learning algorithm [29] on single elements with a reward defined as E_{ads}^2 consistently reaches the lowest energy states. Table 7 in Appendix reports final states for 20 random roll-outs: the agent reaches a -7.4 eV average terminal energy vs -1.5 eV for starting elements – see figure (4) – and terminates on top-2 states for $\sim$ 95% of roll-outs starting from random states. We report 20 random roll-outs from a policy learned on this environment in Table 7 in appendix.

5.3 Simplified Actions Space: Random Edge Traversal

Experiment (3): Single Objective DQN with Random Edge Traversal on *OC20*-Subgraph. In light of the results from experiment (2), we hypothesize that the periodic table environment works better than the original Section 5.1 setup for two reasons: (i) learning a limited set of actions is more tractable compared to the full 55 possible actions in experiment (1), and (ii) valid states are too sparse, and a negative reward on those invalid states leads to shorter episodes which never reach valid states. In other words, **we now traverse the subgraph of *OC20*-subgraph materials with known adsorption energy for our target adsorbate**.

To that end, we reduce the original action space (from Section 5.1) from 60 actions to 5 actions, and introduce the following modified setup: **an action is only taken if it leads to a 'valid' state**. If it is invalid, the episode does not end and no reward is given; the state just remains unchanged. Finally, we introduce Random Edge Traversal: instead of separately enumerating each element that could be added or removed, we add a random element if the agent chooses to expand the chemical compound.

This method of constraining the action space is, to our knowledge, novel in the literature, and introduces a hybrid between bandits (pull a lever to add a random element) and reinforcement learning (the agent can backtrack, removing elements previously added). Formally, the Markov Decision Process (MDP) is defined as follows:

- A state is still a 55-dimensional 1-hot vector representing the catalyst chemical composition (up to three non-zero elements at any given time).
- 5 possible actions: `<add>` a random element (that hasn't already been added to the catalyst), remove the `<first>`, `<second>`, or `<third>` element in the catalyst (if they exist), or `<stop>` the episode.
- Rewards are returned when the agent terminates the episode or if we hit maximum episode length (between 20 and 75 steps). The reward function is $-E_{ads}^3$ to magnify the rewards towards the strongest (respectively weakest) binding elements.
- Transitions are guided by the actions as in Section 5.1, with the caveat that the state remains unchanged change when an invalid action is selected. Initial state is chosen at random, and discount factor is 0.9.

Our agent is a Double-DQN with two hidden layers with layer sizes 512 and 64. To avoid exploding Q-values explode when the target update period is too low, we use an update period of 300 steps, along with a learning rate of 10^{-3}, and we implement an epsilon-greedy policy with $\epsilon = 0.1$.

Experimental results reveal that trajectory roll-outs tend to be much longer than in Section 5.1. Common trajectory lengths range from 10-40 steps, while trajectories between 40 and 75 steps (the maximum length) are slightly less common. We report 9 random roll-outs from a policy learned on this environment in Table 7 in appendix. **We find that final states improve target adsorption energy compared to random initial states by an average of 4.1 eV.**

We report average terminal state energies over 50 roll-outs for single-objective agents trained in experiments (1), (2) and (3) in table 2.

Experiments	Initial state	Exp (1)	Exp (2)	Exp (3)
Avg. E_{ads} (eV)	-1.5	-2.2	-7.4	-5.6
Δ (eV)	-	0.7	5.9	4.1

Table 2: Experiments (1), (2) & (3) (single-objective). Average energies over 50 single-objective roll-outs. Objective: minimize adsorption energy (higher Δ is better).

5.4 Multi-Element Multi-Objective DQN on *OC20*-Subgraph

Experiment (4): Multi-Objective DQN. In practical catalysis experiments, a wide variety of potential adsorbates and possible reaction intermediates are present in the environment. Selectivity to a given reaction product is a major challenge [10] [4], and as a result, ideal catalyst design agents should learn to identify catalysts with stronger binding energy with certain adsorbates, and weaker adsorption with others. This is particularly important to try and break the 'scaling relations' such as between $\star$O, $\star$OH, and $\star$OOH in ORR [24].

We extend our experimental setup to a **multi-objective reinforcement learning problem**, where each adsorbate is a separate objective. In this setup, we train an agent to find catalysts that minimize (respectively maximizes) adsorption energies for each objective adsorbate. **We traverse the *OC20*-subgraph of materials with known adsorption energy to at least one target adsorbate.**

We first implement a standard approach whenever we have multiple objectives: summing weighted rewards from each objective. In this scenario, the reward from each objective is the energy, and the weights are either -1 or +1, based on whether we seek to minimize or maximize the energy for that adsorbate. We train a DQN (same hyperparameters as previously), with the MDP defined as follows:

- A state is the same 55-dimensional 1-hot vector representing the catalyst compound.

- 5 possible actions: <add> a random element, remove the <first>, <second>, or <third> element in the catalyst (if they exist), or <stop> the episode.
- Rewards are returned when we either hit the maximum episode length, which is now set to 20. The reward is the dot product of goal vector and adsorption energies (we don't use negative cube rewards since average adsorption energies matter for the maximization cases):

$$r = -E_{ads} \cdot \mathbf{g}$$

Experiment (5): Multi-Objective DQN with Sub-Sampling. We hypothesize that as the number of objectives increases in the previous setup, there's a higher chance that the agent gets stuck in local minima, where a step in the environment may help optimize for one or two adsorbates but hurt the rest, which might discourage exploration.

Considering that policies learned on a random mixture of objectives might encourage exploration, we introduce **Multi-Objective DQN with Sub-Sampling**, a new method by which we randomly sample one objective among the six for each training roll-out, :

$$r = -E_{ads_i} \times g_i \quad i_{\text{roll-out}} \sim \mathcal{U}(i = 1 \dots 6)$$

This roll-out-level objective sampling approach is, to our knowledge, a novel contribution to literature. We evaluate this approach in our experimental setup, by comparing sub-sampling pairs of objectives to a baseline from the previous experimental setup where all objectives are used in the reward computation for all roll-outs.

We report average final state energies over 50 roll-outs for experiments (4) and (5) in table 3 (stable over several runs). Final states found with sub-sampling are a better fit with their respective objectives for adsorbates 2, 3, and 4, but slightly worse for 1, 5, and 6. We notice that convergence with sub-sampled objective rewards takes longer than the baseline, as one may expect. Experimental results (reported in Table 3 and Figure (5) in appendix) **support the hypothesis that sub-sampling objectives helps encourage exploration, with longer average trajectories than baseline.**

Overall, we find that, in the difficult multi-objective setting, **both baseline and sub-sampling approaches improve final state adsorption energy** towards the desired direction (increase or decrease), simultaneously across all 6 adsorbates, by an average of **0.8 eV**.

Adsorbate Objective	1: *CH2 Increase	2: *CH4 Increase	3: *N2 Increase	4: *NH3 Decrease	5: *OH2 Decrease	6: *OH Decrease
Initial state	-2.2	-3.3	-1.8	-1.6	-1.9	-1.9
Exp (4): Baseline	**-2.2**	-3.0	-1.5	-1.9	**-3.9**	**-3.9**
Exp (5): Sub-Sampling	-2.3	**-3.0**	**-1.6**	**-2.1**	-3.8	-3.8

Table 3: Experiment (4) and (5) (multi-objective). Average energies over 50 multi-objective roll-outs for the three experimental setups (rows) on each of the 6 objectives (columns).

6 Analysis

6.1 Challenges of Sparse Known States Graph Traversal

A particular challenging aspect of or our setup is the very sparse nature of the problem. The limited number of states for which adsorption energies are known ($\sim 2,000$–3000) compared to the larger number of possible states ($55 + 55 \times 54 + 55 \times 54 \times 53 = 160,345$), makes learning to converge to low energy states difficult for our agent in the initial full state/actions setup.

We found success after drastically simplifying our setup with the Periodic Table *GridWorld* environment, which reinforced the intuition that limiting the number of actions is paramount to obtaining helpful results in a sprase rewards setup. This encouraged us to explore Random Edge Traversal of the *OC20*-subgraph to significantly limit the number of actions a model must learn, and our results show this may prove a helpful general principle to traverse complex material spaces of sparsely documented properties.

6.2 Generalizing Inverse Catalyst Design with RL

Our results raise the question of whether the performance we report warrants the cost and complexity of training Deep RL models, where more standard optimization techniques would be more straightforward for offline learning on known energy datasets, even in the multi-objective setting.

Using Deep RL to solve this class of problems despite its inherent complexity enables us to address problems linear solvers cannot. While these experiments use offline learning on a dataset where adsorption energies are known, we envision using ML-based adsorption energy estimators as a critic in an actor-critic RL setup, and use exploitation of states with known energy to direct computational resources where exploring new unknown states is likely to be most beneficial.

Another enticing approach enabled by RL would be to include additional sparse signals to our reward model. For example, hard to model physical properties whenever they are known (stability, selectivity), materials cost, patents, or even human feedback (RLHF [33]), where candidate catalysts are ranked by scientists based on their experience (e.g. manufacturing cost, experimental complexity, industry preferences, etc.). Such a complex reward function or multiple set of criteria would be hard to encode in a tractable way for a direct optimization setup to solve, but would be a good fit for multi-objective Deep RL approaches.

7 Conclusion

This paper presents our experiments with various Deep RL setups. We introduce Multi-Objective DQN with Sub-Sampling and Random Edge Traversal, a novel method to train a generalized multi-objective agent to traverse a large space of possible catalysts with sparse known properties, and identify materials with the desired adsorption energies profile for a set of target adsorbates of interest.

We demonstrated that in the goal-conditioned multi-objective setting, Deep RL can identify promising materials for any combination of target adsorbates binding energies. In practice, conducting a large number of roll-outs and identifying the most common terminal states would point to promising materials on which to focus computational and experimental resources.

Our findings indicate the promise of Deep RL in navigating complex chemical spaces, and present novel approaches to tackle goal-conditioned multi-objective Reinforcement Learning for materials design. These methods could serve as a foundation for more complex computational challenges in large chemical spaces, and the development of novel materials for a wide range of applications in heterogeneous catalysis, electrochemical energy conversion, and low-emissions technologies.

Known Limitations and Future Work

First, we report results on a single combination of 6 adsorbates, and a single objective vector in the multi-objective setup. Further experimentations on a larger set of adsorbates objective vectors would help results robustness, especially for the novel objective sub-sampling approach we introduce.

Second, we report goal-conditioning only on extrema, and train agents seeking to maximize or minimize adsorption energy. Other approaches to train our agent to seek states with intermediate adsorption energy, such as scalar conditioning (defining a target value for E_{ads}), would be particularly helpful to find optimal catalysts, which usually have intermediate binding strength on the activity-binding energy 'volcano plots' (Sabatier Principle [10]).

Relabeling targets is a potent way to improve goal-conditioned agents, and future work could focus on improving the performance of our multi-objective agent through Hindsight Experience Replay [34], as well as Prioritized Experience Replay (PER) [35]. Other offline algorithms such as Conservative Q-Learning [36] may also prove helpful.

Lastly, our overall objective was to train an RL agent to traverse a vast space of possible materials with multiple target adsorbates. While we used offline RL on known adsorption energies datasets to facilitate experimentation, we envision an actor-critic setup where an actor is trained using our graph traversal setup, and a critic uses ML-based adsorption energy evaluation models. An ML-based, lightweight critic such as *AdsorbML* [11] could perform approximate but fast energy evaluations for unknown states traversed at roll-out time, and more exact DFT computations could be reserved for frequent final states revealed by accumulated roll-outs as strong catalyst candidates.

Code & Data Access

The code implementation for our experiments, as well our datasets compiled from the *Materials Project* [7] and *Open Catalyst 2020* [6] datasets, are made available for download for reproducibility purposes on the AdsorbRL GitHub repository.

Acknowledgements

The authors wish to thank Prof. Chelsea Finn, Dr. Karol Hausman, and Jonathan Yang at Stanford University for guidance on this project, as well as Ajay Kannan for his contribution to Random Edge Traversal and the multi-element, multi-objective setup. We are grateful to Meta AI and the chemical engineers, materials scientists and AI researchers who collaborated on the *Open Catalyst Project* and *Materials Project*, and everyone whose research informed these data sets. This work would not have been possible without them.

References

[1] Zhi Wei Seh, Jakob Kibsgaard, Colin F. Dickens, Ib Chorkendorff, Jens K. Nørskov, and Thomas F. Jaramillo. Combining theory and experiment in electrocatalysis: Insights into materials design. *Science*, 355(6321):eaad4998, 2017.

[2] Steven Chu and Arun Majumdar. Opportunities and challenges for a sustainable energy future. *Nature*, 488(7411):294–303, 2012.

[3] C. L. Zitnick, L. Chanussot1, A. Das, S. Goyal, J. Heras-Domingo, C. Ho, W. Hu, T. Lavril, A. Palizhati, M. Rivière, M. Shuaibi, A. Sriram, K. Tran, B. Wood, J. Yoon, D. Parikh, and Z. Ulissi. An introduction to electrocatalyst design using machine learning for renewable energy storage, 2020. arXiv:2010.09435v1.

[4] Phil De Luna, Christopher Hahn, Drew Higgins, Shaffiq A Jaffer, Thomas F Jaramillo, and Edward H Sargent. What would it take for renewably powered electrosynthesis to displace petrochemical processes? *Science*, 364(6438):eaav3506, 2019.

[5] S. Kurth, M.A.L. Marques, and E.K.U. Gross. Density-functional theory. *Encyclopedia of Condensed Matter Physics 2005*, pages 395–402, 2005.

[6] Lowik Chanussot, Abhishek Das, Siddharth Goyal, Thibaut Lavril, Muhammed Shuaibi, Morgane Riviere, Kevin Tran, Javier Heras-Domingo, Caleb Ho, Weihua Hu, Aini Palizhati, Anuroop Sriram, Brandon Wood, Junwoong Yoon, Devi Parikh, C. Lawrence Zitnick, and Zachary Ulissi. Open catalyst 2020 (oc20) dataset and community challenges. *ACS Catalysis*, 11(10):6059–6072, May 2021.

[7] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin A. Persson. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, 07 2013.

[8] Paul Sabatier. The method of direct hydrogenation by catalysis. *Nobel Lecture*, 1912.

[9] Javier Pérez-Ramírez and Núria López. Strategies to break linear scaling relationships. *Nature Catalysis*, 2:971–976, 2019.

[10] Jens K Nørskov, Thomas Bligaard, Jan Rossmeisl, and C. H. Christensen. Towards the computational design of solid catalysts. *Nature Chemistry*, 1:37–46, 2009.

[11] J. Lan, A. Palizhati, M. Shuaibi, B.M. Wood, B. Wander, A. Das, M. Uyttendaele, C. L. Zitnick, and Z. W. Ulissi. Adsorbml: Accelerating adsorption energy calculations with machine learning, 2023. arXiv:2211.16486.

[12] Jessica G. Freeze, H. Ray Kelly, and Victor S. Batista. Search for catalysts by inverse design: Artificial intelligence, mountain climbers, and alchemists. *Chemical Reviews*, 119(11):6595–6612, 2019.

[13] Juhwan Noh, Geun Ho Gu, Sungwon Kim, and Yousung Jung. Machine-enabled inverse design of inorganic solid materials: promises and challenges. *Chemical Science*, 11:4871–4881, 2020. Minireview.

[14] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285, 1996.

[15] Kai Arulkumaran, Marc Peter Deisenroth, Miles Brundage, and Anil Anthony Bharath. Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine*, 34(6):26–38, 2017.

[16] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

[17] Richard Tran, Janice Lan, Muhammed Shuaibi, Brandon M. Wood, Siddharth Goyal, Abhishek Das, Javier Heras-Domingo, Adeesh Kolluru, Ammar Rizvi, Nima Shoghi, Anuroop Sriram, Félix Therrien, Jehad Abed, Oleksandr Voznyy, Edward H. Sargent, Zachary Ulissi, and C. Lawrence Zitnick. The open catalyst 2022 (oc22) dataset and challenges for oxide electrocatalysts. *ACS Catalysis*, 13(5):3066–3084, February 2023.

[18] S. Zhu, K. Jiang, B. Chen, and S. Zheng. Data-driven design of electrocatalysts: principle, progress, and perspective. *J. Mater. Chem. A*, 11:3849, 2023.

[19] Z. Zhou, S. Kearnes, and L. Li et al. Optimization of molecules via deep reinforcement learning. *Sci Rep*, 9:10752, 2019.

[20] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

[21] Sridharan B, Mehta S, Pathak Y, and Priyakumar UD. Spectra to structure: Deep reinforcement learning for molecular inverse problem, 2021. ChemRxiv. Cambridge: Cambridge Open Engage; 2021. doi:10.26434/chemrxiv-2021-4hc7k.

[22] Elton Pan, Christopher Karpovich, and Elsa Olivetti. Deep reinforcement learning for inverse inorganic materials design, 2022.

[23] F. Sui, R. Guo, Z. Zhang, G. Gu, and L. Lin. Deep reinforcement learning for digital materials design. *ACS Materials Lett.*, 3:1433–1439, 2021.

[24] Isabela C Man, Hai-Yan Su, Federico Calle-Vallejo, Heine A Hansen, José I Martínez, Nilay G Inoglu, John Kitchin, Thomas F Jaramillo, Jens K Nørskov, and Jan Rossmeisl. Universality in oxygen evolution electrocatalysis on oxide surfaces. *ChemCatChem*, 3(7):1159–1165, 2011.

[25] Minhua Shao, Qiaowan Chang, Jean-Pol Dodelet, and Regis Chenitz. Recent advances in electrocatalysts for oxygen reduction reaction. *Chemical reviews*, 116(6):3594–3657, 2016.

[26] Stephanie Nitopi, Erlend Bertheussen, Soren B Scott, Xinyan Liu, Albert K Engstfeld, Sebastian Horch, Brian Seger, Ifan EL Stephens, Karen Chan, Christopher Hahn, et al. Progress and perspectives of electrochemical co2 reduction on copper in aqueous electrolyte. *Chemical reviews*, 119(12):7610–7672, 2019.

[27] Aayush R Singh, Brian A Rohr, Jay A Schwalbe, Matteo Cargnello, Karen Chan, Thomas F Jaramillo, Ib Chorkendorff, and Jens K Nørskov. Electrochemical ammonia synthesis: The selectivity challenge, 2017.

[28] Jan Willem Erisman, Mark A Sutton, James Galloway, Zbigniew Klimont, and Wilfried Winiwarter. How a century of ammonia synthesis changed the world. *Nature Geoscience*, 1(10):636–639, 2008.

[29] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[30] Richard Bellman. The theory of dynamic programming. *Bulletin of the American Mathematical Society*, 60(6):503–515, 1954.

[31] Y. Wu, G. Tucker, and O. Nachum. Behavior regularized offline reinforcement learning, 2019. arXiv:1911.11361v1.

[32] IUPAC. Periodic table of elements.

[33] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[34] Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, OpenAI Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. *Advances in neural information processing systems*, 30, 2017.

[35] Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. *arXiv preprint arXiv:1511.05952*, 2015.

[36] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1179–1191, 2020.

Adsorbate	Dataset size (# of known catalysts)
$\star OH_2$	2,379
$\star CH_2$	2,759
$\star CH_4$	2,409
$\star N_2$	2,111
$\star NH_3$	2,473
$\star OH$	2,655
All adsorbates	7,386 (unique)

Table 4: Sparse rewards: number of catalysts for which adsorption energy is known for the corresponding adsorbates, out of ∼160,000 possible compounds. Does not sum up due to duplicates.

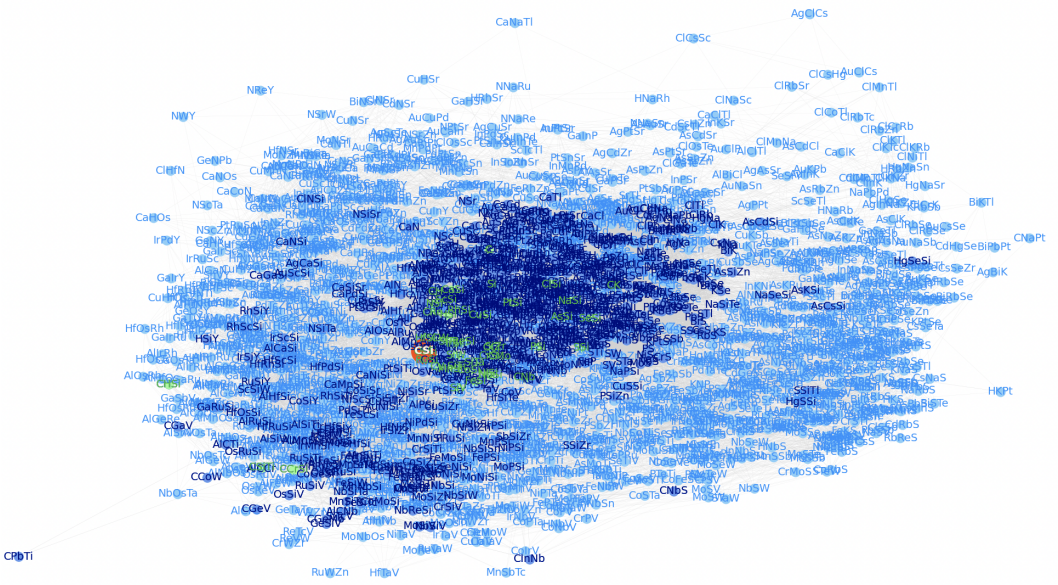

Figure 2: 3-hop ego graph of the lowest energy state for $\cdot OH_2$ adsorbate (SiC). Red: SiC. Green: 1-hop ego network. Dark blue: 2-hope ego network. Light blue: 3-hop ego network.

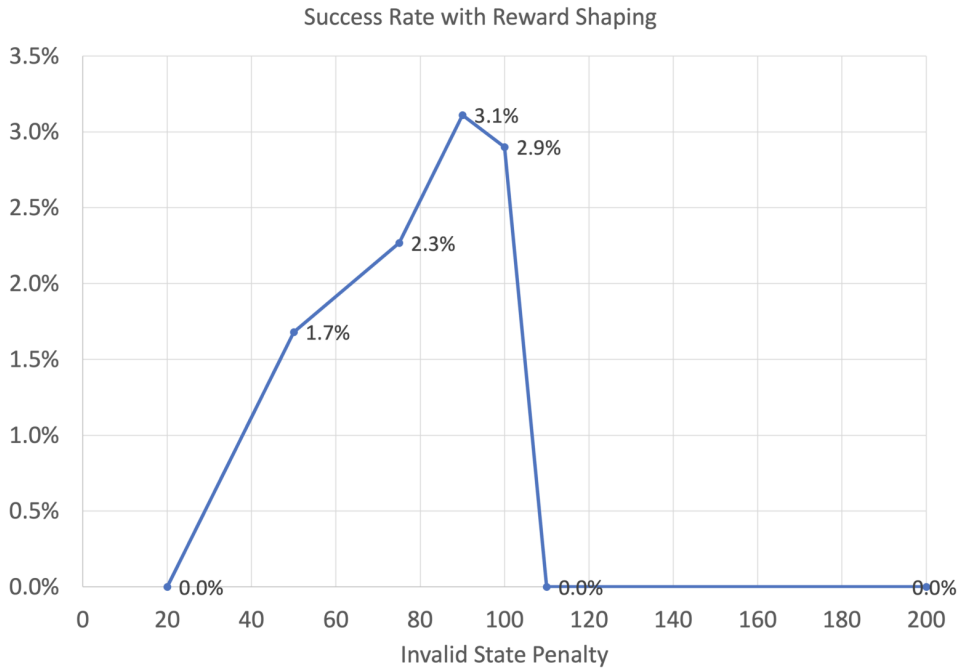

Figure 3: **Experiment (1).** Success rate in reaching the optimal valid state using reward shaping for each value of λ. The graph illustrates the difficulty in training a DQN agent that converges in our dataset with sparse rewards.

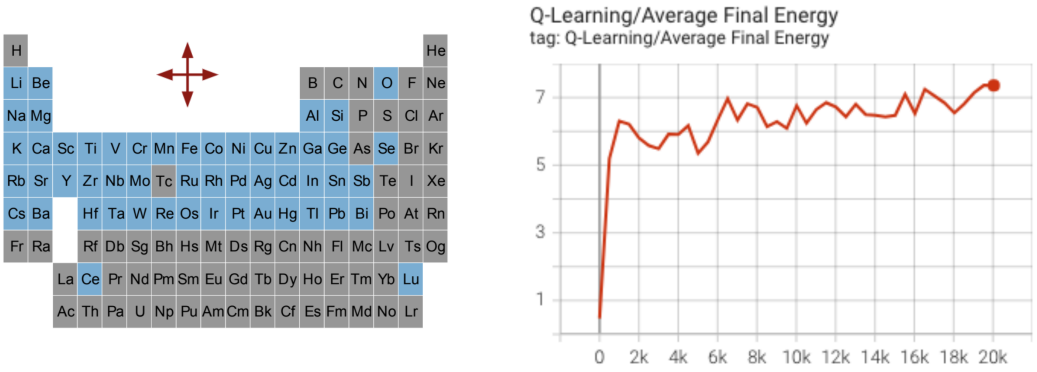

Figure 4: **Experiment (2).** Left: Periodic table of elements, traversed in a *GridWorld* setting (stay | left | right | up | down). Elements present in *OC20* are highlighted in blue. Right: average final energy after evaluation roll-outs while Q-learning the policy (NB: plot represents $-E_{ads}$, final adsorption energy is negative).

Final state	Trajectory length	Energy (eV)
Carbon	9	-7.9
Carbon	9	-7.9
Iron	9	-9.1
Carbon	9	-7.9
Iron	9	-9.1
Carbon	9	-7.9
Iron	9	-9.1
Manganese	9	-0.9
Carbon	9	-7.9
Carbon	9	-7.9
Carbon	9	-7.9
Carbon	9	-7.9
Carbon	9	-7.9
Carbon	9	-7.9
Carbon	9	-7.9
Iron	9	-9.1
Iron	9	-9.1
Carbon	9	-7.9
Carbon	9	-7.9

Table 5: **Experiment (2).** Twenty random roll-outs from a policy learned on this environment. Lower energy is better. Average random single-element reward is -1.5eV.

Final state	Trajectory length	Energy (eV)
Carbon	24	-7.9
Iron	40	-9.1
Sulfur and Vanadium	5	-6.3
Iron	10	-9.1
Iron	24	-9.1
Cesium and Hydrogen	1	-0.8
Tantalum and Vanadium	9	-2.1
Iron	24	-9.1
Iron	18	-9.1

Table 6: **Experiment (3).** Nine random roll-outs from a policy learned on this environment. Lower energy is better. Average random single-element reward is -1.5eV.

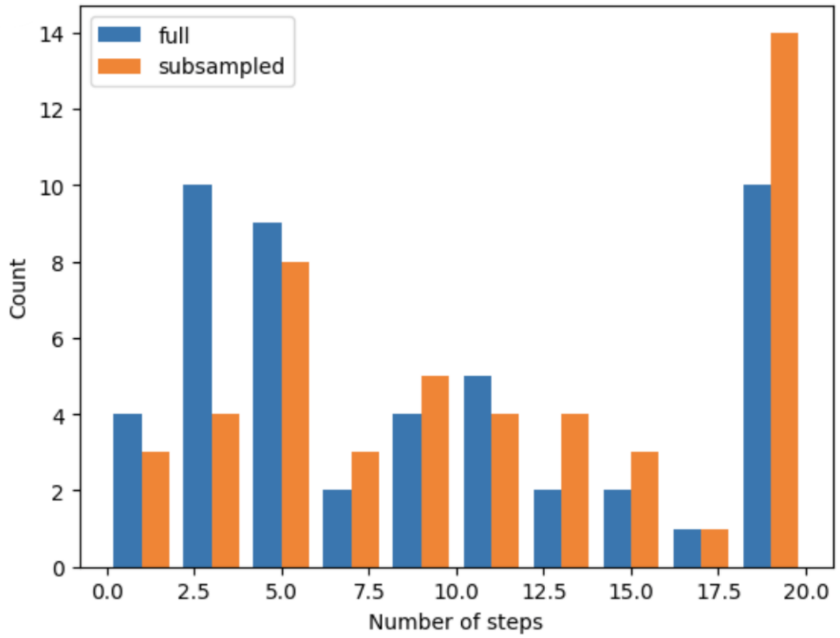

Figure 5: **Experiments (4) & (5): multi-objective DQN on OC20-Subgraph.** Number of episode steps at roll-out, full objective vs objective sub-sampling.