Discovering Chemical Space from First Principles with Reinforcement Learning

Bjarke Hastrup[1], François Cornet[1,2], Tejs Vegge[1,3], Arghya Bhowmik[1,3]*
{bjaha, frjc, teve, arbh}@dtu.dk

[1] Dept. of Energy Conversion and Storage, Technical University of Denmark, Denmark
[2] Dept. of Applied Mathematics and Computer Science, Technical University of Denmark, Denmark
[3] Pioneer Center for Accelerating P2X Materials Discovery (CAPeX), Kgs. Lyngby, Denmark

Abstract

Discovering novel stable molecules without training data remains a grand scientific challenge. Current molecular generative models are trained on large, pre-curated datasets, which introduce biases and limit exploration of novel chemistry. In contrast, we propose a new paradigm: autonomous, generalized agents capable of mapping vast, unknown chemical spaces without any pretraining. For the first time, we present a self-guided agent that autonomously constructs valid 3D isomers under stoichiometric constraints and is trained exclusively online using reinforcement learning. Unlike existing approaches that generally overfit to a specific chemical formula, we establish a multi-composition training scheme that enables a broad generalization across diverse chemistry, guided by energy- and validity-based rewards. Our agent can discover up to an order of magnitude more valid isomers on unseen test formulas than the baseline. These results fulfil the promise of online RL as a powerful paradigm for scalable *tabula rasa* exploration of the chemical configuration space.

1 Introduction

Autonomous discovery of novel molecules with bespoke properties is the new frontier in computational chemistry. Effectively navigating the vast chemical space requires innovative and data-efficient search strategies. Generative models have emerged as a promising avenue for this task (Anstine and Isayev, 2023), yet their performance often hinges on the availability of suitable training data. Large public datasets are rarely curated with specific property optimization in mind, and the relevant property regimes may lie at the fringes - or entirely outside - the observed data distribution (Brown et al., 2019). This poses a fundamental challenge: models must not only interpolate but also extrapolate beyond known examples (Schrier et al., 2023). Constructing task-specific datasets is likewise non-trivial and, even when feasible, introduces chemical and structural biases that may limit exploration of novel chemical spaces.

A compelling approach to overcoming these limitations is to adopt online (*tabula rasa*) learning techniques—frameworks that learn from scratch without relying on pre-curated datasets-such as Reinforcement Learning (RL) (Sutton and Barto, 2018), where an agent learns to explore the chemical space through trial and error (Sridharan et al., 2024). This has proved very successful at SMILES based molecular generation (Olivecrona et al., 2017; Popova et al., 2018; Bou et al., 2024). However, for 3D geometry an additional post-processing step is needed to generate conformer ensembles (e.g., with the ETKDG method (Riniker and Landrum, 2015)). These automatically generated conformer ensembles add a costly extra step and often miss out on conformers with the highest stability or best

*Corresponding author: arbh@dtu.dk

European Conference on Neural Information Processing Systems (EurIPS) Workshop: SimBioChem 2025.

property. Instead, directly generating molecules in 3D enables molecular structures to be constructed and optimized within a fully integrated, end-to-end framework.

In the supervised setting, some of the most promising directions for molecule generation in 3D are either based on denoising diffusion (Hoogeboom et al., 2022; Le et al., 2024; Cornet et al., 2024a), flow matching (Song et al., 2023; Irwin et al., 2025), or auto-regressive models that build molecules in an atom-by-atom fashion (Gebauer et al., 2019; 2022; Roney et al., 2022; Daigavane et al., 2023; Ochoa et al., 2024). Although diffusion models could potentially be integrated into a pretraining-finetuning framework (Black et al., 2024), it remains unclear whether they can effectively be used for *tabula rasa* learning for navigating a 3N dimensional energy surface. In the purely online setup, RL has been used for conformer (Jiang et al., 2022; Volokhova et al., 2024) and isomer (Simm et al., 2020; 2021) generation. Flam-Shepherd et al. (2022) extended MOLGYM to place fragments instead of individual atoms, improving scalability and the size of the generated molecules. Meldgaard et al. (2021) used online RL but only after an offline pretraining phase. Whereas their pretraining was multi-compositional, their online finetuning was for single compositions only and further relied on result aggregation from 64 parallel fine-tunings spawned after pertaining. A general-purpose RL algorithm for *tabula rasa* 3D molecular structure discovery has yet to be demonstrated. Existing results show only limited success on simple organic molecules or metal clusters of fixed composition (Modee et al., 2023), with poor generalization across stoichiometries. These studies offer useful technical insights—such as the role of final-reward training (Elsborg and Bhowmik, 2023) and the constraints of current RL formulations for 3D structure search—but they have limited relevance for identifying genuinely novel molecules.

The evaluation of RL algorithms for molecular structure discovery presents a fundamental challenge. In the standard generative modeling paradigm, where models are trained via supervised learning to approximate the training distribution, performance is typically assessed through stochastic rollouts at a single final checkpoint, reflecting how well the converged model captures the underlying data (Gómez-Bombarelli et al., 2018; Gebauer et al., 2019; Cornet et al., 2024b). In contrast, evaluation in RL is considerably more nuanced (Henderson et al., 2018; Dulac-Arnold et al., 2020). The performance of an RL agent can vary significantly depending on the checkpoint selected, as the underlying policy evolves throughout training (Islam et al., 2017). A single-checkpoint evaluation, though convenient and widely used (Xia et al., 2022), often fails to capture the broader behavioral dynamics and exploration strategies adopted at different training stages (Colas et al., 2019). This evolving behavior, combined with the open-ended nature of RL tasks, renders traditional metrics — such as distance to a reference dataset — largely inadequate. Instead, meaningful evaluation often requires the direct involvement of a chemist to quantify the utility of individually generated structures through score/reward functions (Brown et al., 2019; Polykovskiy et al., 2020; Schwalbe-Koda and Gómez-Bombarelli, 2020). This reliance on task-specific metrics complicates automation and has hindered the establishment of universally recognized benchmark tasks, making it difficult to objectively compare algorithms and slowing methodological convergence in the field (Olivecrona et al., 2017; Xie et al., 2021; Nie et al., 2024).

A core challenge in online RL for 3*D molecular* discovery in particular, is balancing the delicate trade-off between exploration and physical stability. While policy stochasticity is essential for escaping locally optimal behavior (Haarnoja et al., 2018; Schulman et al., 2017), excessive spatial noise can corrupt energy based evaluations, which depend sensitively on the atomic coordinates (Smith et al., 2017; Gastegger et al., 2021). This degrades the reward signals and destabilizes the learning. The problem is compounded in 3D atomistic environments, where high-reward actions lie in a multimodal and discontinuous space. Here, small perturbations rarely improve the objective, but often disrupt chemically valid structures, rendering local exploration ineffective (Rose et al., 2021).

A true paradigm shift in molecular discovery requires the ability to explore the full chemical space from first principles—without relying on hand-crafted rules, curated datasets, or preconceived feature biases. Such an approach would learn viable chemistry entirely through exploration of chemical motifs leading to possible discovery beyond current human knowledge and intuition. As a key step toward this grand vision, we have been successful in training *composition-generalizable* RL agents capable of discovering stable 3D molecules across diverse chemical formulas.

Inspired by the MOLGYM (Simm et al., 2020; 2021) framework, we take a significant step forward toward training self-guided RL agents that can generalize across chemical space. Specifically, we target isomer discovery, where the agent is tasked with generating 3D conformations given a pre-

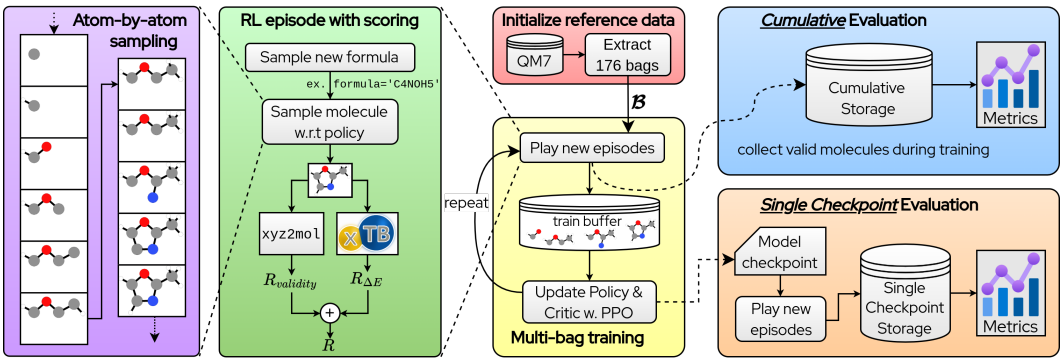

Figure 1: **Multi-composition training and evaluation workflow.** Our framework constructs isomer generation tasks by extracting chemical formulas from a reference dataset and introduces new terminal rewards based on validity and total energy. We evaluate the RL agents' isomer discovery capabilities at just a *single* checkpoint, as well as *cumulatively* across the entire discovery campaign. **Crucially, no 3D structures are shown to the agent during training.**

specified chemical composition. Our success originated from a novel multi-composition training scheme and new reward schemes. We demonstrate that RL can be effectively applied to isomer discovery, without overfitting to a fixed set of atoms as in prior work (Simm et al., 2020; 2021). A visual abstract of our workflow is provided in Fig. 1. This resolves long standing limitations and stagnation in RL for *tabula rasa* 3D atomic structure discovery and we summarize our main contributions as follows:

- We introduce new terminal rewards based on energy and chemical validity, thereby training the agent to build stable and *valid* molecules.
- We propose a groundbreaking multi-composition training setup based on chemical compositions drawn from a broad chemical space derived from the QM7 reference dataset, facilitating generalization across stoichiometries.
- We design a broader multi-bag evaluation scheme to facilitate benchmarking of online isomer discovery agents and assess various combinations of the proposed reward terms.

2 Results

2.1 RL environment: Isomer search

We trained an RL agent to build stable and valid *isomers* (i.e. different molecules with the same *pre-specified* chemical formula) autoregressively (atom-by-atom), using a linear combination of reward terms based on quantum chemical energy evaluations and validity checks. Our training framework is illustrated in Fig. 1, along with the two separate evaluation schemes.

Multi-composition training We leveraged the QM7 dataset (Blum and Reymond, 2009; Rupp et al., 2012) as a *reference dataset*, and used it to derive a bag set, $\mathcal{B}$, of molecular compositions that was used for multi-composition training[2]. In practice, training rollouts are performed synchronously by a collection of N_w workers, each endowed with a uniquely randomized iterable of the bag set $\mathcal{B}_w = \text{permutation}_w(\mathcal{B})$. When worker w has generated a molecule for a particular bag (or failed to do so), it simply proceeds to the next bag in its bag set.

Autoregressive molecule sampling The molecule construction process is framed as a sequential decision-making task, where, after sampling an initial bag of atoms $\mathcal{B}_0$, an agent iteratively selects

[2]Notice that "formula" and "bag" are used interchangeably throughout this article, as they carry the same physical meaning. So, to clarify, a *bag set*, $\mathcal{B}$, is simply a collection of chemical formulas. Similarly, to describe how our agent generalizes across chemical formulas, we use words such as "multi-bag", "multi-composition", "stoichiometry-agnostic", etc., depending on the context.

and places atoms in 3D space to incrementally build the molecule. In RL terms, the agent observes the state $s_t = (\mathcal{C}_t, \mathcal{B}_t)$ consisting of the current molecular canvas $\mathcal{C}_t$ (i.e. the molecule built so far) and the remaining atom bag $\mathcal{B}_t$. The agent's action $a_t = (e_t, x_t)$ involves choosing an atom $e_t \in \mathcal{B}_t$ and assigning its 3D position $x_t \in \mathbb{R}^3$ leading to the deterministic transition to the next state $s_{t+1} = (\mathcal{C}_{t+1}, \mathcal{B}_{t+1})$, where

$$\mathcal{C}_{t+1} = \mathcal{C}_t \cup \{(e_t, x_t)\}, \quad B_{t+1} = B_t \setminus \{e_t\}.$$

This process continues until the bag is empty and a complete molecule $\mathcal{C}_T$ has been formed. The distribution over molecules constructed in this autoregressive process is given by

$$p(\mathcal{C}_T|\mathcal{B}_0) = \prod_{t=0}^{T-1} \pi_\theta(a_t|s_t), \tag{1}$$

where $\pi_\theta(a_t|s_t)$ is the agent's probabilistic policy governing the placement of atom e_t at position x_t, given the current molecular state s_t. This formulation captures the conditional nature of molecule construction starting from the initial bag $\mathcal{B}_0$.

Notably, in this environment, the agent must implicitly learn to construct valid molecules, as no explicit validity constraints are imposed during generation. Also, atoms are sampled without replacement, and their positions remain fixed after placement. The randomness in the generation process comes solely from the agent's policy, as the environment transitions are fully deterministic. As such, the molecule-building task can be formulated as a fully observable, finite-horizon Markov Decision Process (MDP) with a hybrid discrete-continuous action space, where the episode length is determined by bag size.

Reinforcement Learning objective The agent's stochastic policy $\pi_\theta(a_t|s_t)$ is optimized in search of the optimal parameters θ that maximize the expected discounted sum of future rewards (known as *return*) from any given state,

$$V^\pi(s_t) = \mathbb{E}_{\pi_\theta}\left[\sum_{t'=t}^{T} \gamma^{t'} r(s_{t'}, a_{t'})\right], \tag{2}$$

where $\gamma \in (0, 1]$ is the discount factor and $r(s_t, a_t)$ is the reward received at time step t for taking action a_t in state s_t. So starting with an empty canvas at $t = 0$, the agent must learn to maximize $J(\theta) = \mathbb{E}_{s_0 \sim \mu_0}[V^\pi(s_0)]$ with μ_0 denoting the distribution over bags.

A new terminal reward structure In RL, reward design is often the single most critical factor determining success or failure. Whereas the original MOLGYM frameworks uses *per-step* rewards as shown in Box 2.1b, we train agents which only receive reward at the terminal state, i.e. when the molecule is completed (Box 2.1a). These rewards are determined based on quantum mechanical energy using `GFN2-xTB` (Bannwarth et al., 2019) and chemical valency checks via `xyz2mol` (Kim and Kim, 2015). The choice of terminal rewards stems from the fact that the temporal structure of our RL episodes is an artificial construct, introduced solely to enable the factorization of the agent's molecular sampling policy (Eq. (1)), and the intermediate molecular states $\{\mathcal{C}_t\}_{t<T}$ visited during the episode are not necessarily chemically or energetically meaningful. To prevent training from being obscured by misleading or noisy signals from these partial, often non-physical intermediates, we introduce new terminal rewards that are only queried once the molecule is fully constructed, as shown in Box 2.1a.

Box 2.1a: Terminal Rewards (ours)

In this work, we introduce the following two terminal rewards:

- **Atomization energy ($\mathcal{A}$)**: This reward is based on the negative difference between the potential energy of the final molecule $\mathcal{C}_T$ and the sum of potential energies of each of its constituent atoms in isolation:

$$\Delta E = \left(\sum_{t=1}^{T} E(e_t)\right) - E(\mathcal{C}_T), \quad r_A(s_T) = \begin{cases} \Delta E + \frac{1}{2}(\Delta E)^2 & \text{if } \Delta E > 0, \\ \Delta E & \text{if } \Delta E < 0, \end{cases} \tag{3}$$

where ΔE is the binding strength (positive is better) and the polynomial transformation provides extra resolution around high scoring molecules, as this allows the agent to differentiate between "good" and "really good" molecules.

- **Validity** ($\mathcal{V}$): A boolean validity check based on the requirement that generated molecules can be successfully parsed by the `xyz2mol` function, which converts arbitrary 3D point clouds into `rdkit` mol objects (Landrum, 2024):

$$r_V(s_T) = \begin{cases} 1 \text{ if } \mathcal{C}_T \text{ is a valid molecule,} \\ 0 \text{ else.} \end{cases} \tag{4}$$

  In particular, we verify that the molecule is not fragmented (consisting of smaller isolated molecules) and that no atom is charged. This validity reward term was introduced because we observed that energy-based rewards alone were insufficient to guide the agent effectively. The reason is that the 3D molecules generated by the *stochastic* agent policy inevitably contain spatial noise - both to facilitate exploration, but also because the agent is unaware of the correct low-energy configurations. As a result, the agent may produce molecules that are chemically valid (in terms of valency) but significantly distorted, thus getting penalized too harshly by the energy-based reward signal. To avoid this, our validity bonus encourages the agent to generate molecules that are first and foremost chemically valid.

Box 2.1b: Baseline Reward

For comparison, the MOLGYM baseline used the following reward:

- ***Per-step Formation energy*** ($\mathcal{F}$): In contrast to the terminal reward, reward can be assigned at every step throughout the episode and is given by the negative difference in energy between the resulting molecule $\mathcal{C}_{t+1}$ and the sum of energies of the previous molecule $\mathcal{C}_t$ and a new atom of element e_t

$$r_F(s_t, a_t) = (E(\mathcal{C}_t) + E(e_t)) - E(\mathcal{C}_{t+1}), \quad t = 0, ..., T-1. \tag{5}$$

2.2 Experiments

Figure 2 illustrates our training and evaluation scheme. Through linear combinations of the 3 fundamental reward components ($\mathcal{A}$, $\mathcal{V}$, $\mathcal{F}$) introduced in Box 2.1(a+b),, we define 5 distinct reward functions **A**, **AV**, **F**, **FV**, and **AFV**, each corresponding to a separate *agent* that is trained independently three times using different random seeds (the linear coefficients are shown in Table 3 in the Appendix). Generally, our analysis places particular emphasis on the agents **A** and **AV**, as these incorporate our newly proposed terminal rewards $\mathcal{A}$ and $\mathcal{V}$. Specifically, we address the following research questions:

- **Q1: Comparison to previous work.** *How do our agents (**A** & **AV**) perform in comparison to previous work in online molecular discovery in 3D?* This initial experiment evaluates their discovery capabilities in the single-bag generation paradigm where baselines are available.
- **Q2: Generalization ability.** *Which reward functions generalize to our multi-bag setting and to out-of-sample (unseen chemical composition) generation in particular?* Here we broaden the evaluation scope relative to **Q1** by aggregating results across a random held-out split of bags represented in QM7 (see Fig. 2b), enabling a comprehensive comparison of reward signals at various stages of training.
- **Q3: Exploration of chemical space.** *Did our tabula rasa agents rediscover the molecules from the QM7 reference dataset? Did they go beyond and even expand on this dataset?* Here we will interpret the training run as a discovery campaign and examine the complete pool of molecules obtained.

While Sections 2.3 and E (**Q1**+**Q2**) focus on evaluating single checkpoints, Section 2.4 (**Q3**) examines agent performance throughout the entire training process. Despite these differences in evaluation scope, all three cases are based on the same training runs visualized in Fig. 3.

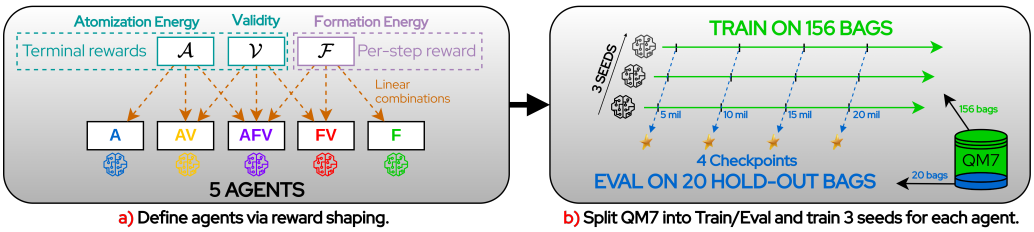

Figure 2: **Overview of experiments.** *(a)* Agents are defined in terms of the reward functions with which they are trained. *(b)* Training data comprises 156 QM7 bags, with 20 remaining bags held out for evaluation. During training (green) we save 4 checkpoints for each seed and perform out-of-sample evaluations (blue) for all checkpoints (5 agents, 3 seeds, 4 ckpts $\Rightarrow$ 60 evaluations in total).

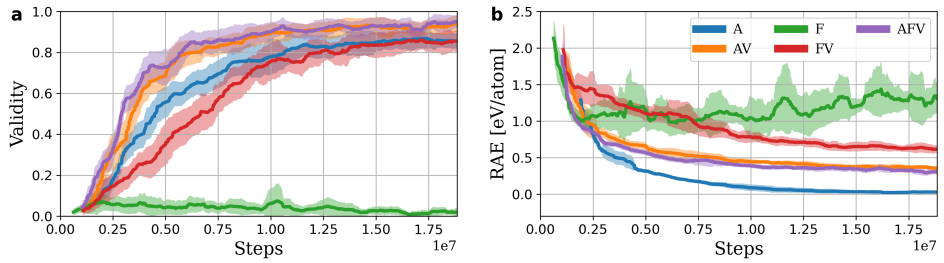

Figure 3: **Learning Curves (in-sample).** *(a)* Validity and *(b)* unrelaxed Relative Atomic Energy (RAE) of continuously collected training rollouts, plotted against total number of single-atom placements (environment steps). The RAE metric quantifies the excess energy relative to the average energies of QM7 molecules with the same chemical formula (see page 17 for detailed metric definitions). Shading represents ± 1 standard deviation between the 3 seeds. **Notably, our newly introduced terminal reward terms, $\mathcal{A}$ and $\mathcal{V}$, enable significantly more stable training dynamics.**

2.3 Q1: Comparison againts previous work: *Single-bag discovery task*

Table 1: **Q1: Single-bag discovery.** Our AV agent outperforms previous work (numbers taken from Simm et al. (2021)) by discovering an order of magnitude more valid isomers for evaluation bags *beyond* its training set. In contrast, the baseline agents were trained explicitly on the presented bags.

	Training type:	Single-bag training (on eval bag)		QM7 multibag	
	Collection type:	Cumulative argmax $\times$ 10 seeds		Single CP stochastic	
	Bag: \ *Agent:*	INTERNAL	COVARIANT	**A (ours)**	**AV (ours)**
IN QM7 (TRAIN)	C_3H_8O	$4^\dagger$	$\mathbf{8}^\dagger$	3.0 $\pm$ 0.0	3.0 $\pm$ 0.0
	C_4H_7N	18	25	13.0 $\pm$ 0.8	**36.7** $\pm$ 1.3
BEYOND QM7 DATASET	$C_3H_5NO_3$	35	65	49.0 $\pm$ 7.5	**544** $\pm$ 46
	$C_7H_{10}O_2$	21	85	198.0 $\pm$ 24.9	**808** $\pm$ 122
	$C_7H_8N_2O_2$	58	118	145.7 $\pm$ 39.7	**1213** $\pm$ 212

$^\dagger C_3H_8O$ is a small and fully saturated chemical formula and we only see 3 feasible positions for an oxygen atom on a 3-membered carbon chain: an OH group on the first carbon atom, an OH group on the central carbon atom, or an O between carbon atoms 1 and 2. Since both baseline agents reportedly discovered strictly more than 3 isomers without providing code for their uniqueness check, we suspect their numbers are mistakenly reported in all 5 cases, which only further emphasizes the improved discovery capabilities of our approach.

We adopt the evaluation protocol from Simm et al. (2021), counting the number of valid constitutional isomers[3] discovered by our agents (**A** and **AV**) when deployed on a single bag. Table 1 compares our

[3]Isomer counts are determined following the standard convention: unique SMILES strings are generated using `RDKit` (Landrum, 2024), expressed in canonical form, and exclude isomeric information.

results with those reported in prior work. Notably, although **AV** is not explicitly trained on certain formulas, it consistently discovers up to an order of magnitude more constitutional isomers than baseline agents on the last three formulas that exceed the scope of QM7, containing more than seven heavy atoms. In contrast, the **A** agent performs comparably to the baselines in terms of isomer count.

In Fig. 6, we visualize the high reward molecules, ranked by formation energy per atom after structural relaxation, and compare their energy distributions in the center column. While the **AV** excelled in breadth of discovery, the **A** agent - trained solely with energy-based rewards - tends to sample molecules with significantly better formation energies.

While the discovery statistics in Table 1 highlight the effectiveness of our training setup, reward formulation, and data collection strategy, it is important to note several key differences between our approach and the baseline methods:

Baselines: The **INTERNAL** and **COVARIANT** agents from MOLGYM use a single-bag training paradigm. This is a costly approach that requires a separate training run for each conceivable molecular formula (bag). Additionally, the discovered isomer count is aggregated over 10 independent runs using different seeds. The molecules used for isomer counting are collected throughout the training (referred to here as *cumulative* data collection), and the molecules are always generated by selecting the most likely action (i.e. $\arg\max_{\{a_t\}} \pi_\theta(a_t|s_t)$), resulting in just a single molecule at every checkpoint during training, thus relying solely on the gradual drift of the agent policy to achieve diverse sampling.

Proposed scheme: We adopt a multi-bag training strategy, using compositions derived from molecules in the QM7 reference dataset, and evaluate discovery performance on the same test bags used by the baseline methods. Unlike the baselines however, we report results based on molecules sampled stochastically from the learned agent policy at a single checkpoint (CP). Concretely, we use the third checkpoint—taken after 15 million training steps—and sample 10,000 molecules per random seed for each test-time formula listed in Table 1 and Fig. 6.

2.4 Q3: Chemical space exploration - *Training as a discovery campaign*

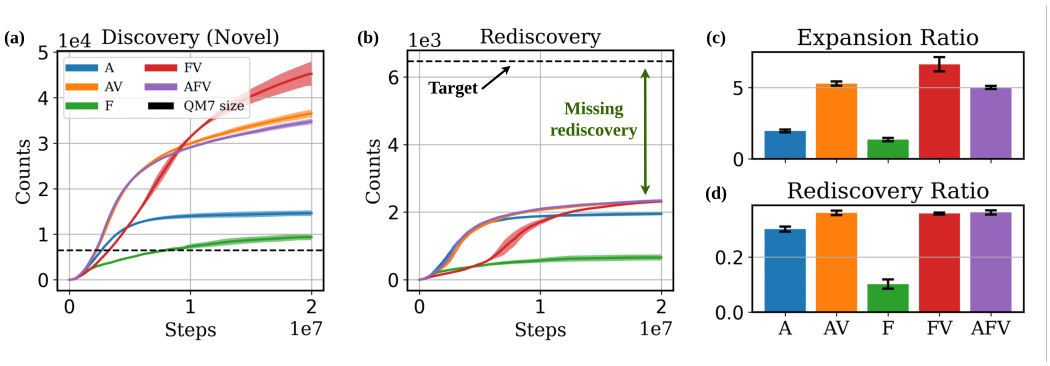

Figure 4: **Q3. Cumulative discovery campaign.** (a) Number of *novel* SMILES discovered during training. (b) Number of QM7 SMILES rediscovered. (c)-(d) Total expansion and rediscovery *relative* to the size of QM7. Although the agents are able to discover many novel molecules and expand on the QM7 dataset by several multiples, their rediscovery ratios are remarkably consistently capped around 40%, thus indicating a subclass of molecular structures inaccessible to our RL agents.

In the previous experiments, we evaluated agent performance based on stochastic rollouts from a single checkpoint - a simple and general scheme widely used in generative modeling (e.g., supervised distribution learning) - but one that overlooks the evolving nature of an RL agent's policy and introduces arbitrariness due to checkpoint selection. To fully leverage this behavioral drift over the course of training, we instead store every molecule generated in a cumulative storage buffer as illustrated in Fig. 1 (blue), which allows us to track the discovery process across time.

The cumulative discovery campaign is summarized in Fig. 4. During training, each agent discovers between 10,000 and 45,000 unique SMILES strings across the 156 training formulas (Fig. 4a). For comparison, the QM7 training subset contains only 6,465 molecules, so the number of generated

molecules far exceeds the number of known reference structures. This relationship is captured by the expansion ratio shown in Fig. 4c, quantifying how many novel molecules the agent generates relative to the original QM7 set.

To assess the agent's ability to reproduce known chemistry, we also count how many of the original QM7 molecules were rediscovered during training (Fig. 4b), with final rediscovery statistics shown in Fig. 4d. The rediscovery curves clearly plateau, indicating that further training does not yield additional rediscovered molecules. This plateau behavior across agents raises an important question:

Are there particular molecular substructures that our agents systematically fail to learn or explore?

To investigate this, we analyze the rediscovery performance of our two most promising agents, **A** and **AV**, by comparing the sets of rediscovered molecules to the full QM7 training subset. The analysis is presented in Section F in the appendix. Overall, our findings suggest that the agent's discovery policy is biased toward constructing small, aliphatic, and less topologically complex fragments. Meanwhile, more exotic, strained, or electronically diverse motifs are significantly underexplored.

Finally, in Fig. 5, we examine the distribution of formation energies for the rediscovered molecules. Despite recovering less than 50% of the QM7 training set, both agents preferentially rediscover molecules with lower-than-average (i.e., more negative) formation energies. Notice again that this comparison is restricted to molecules within QM7, despite the agents having vastly expanded beyond it.

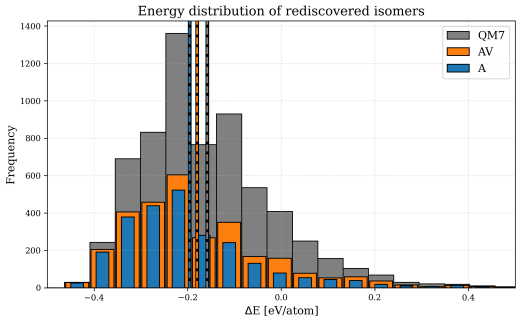

Figure 5: **Rediscovery energy distributions.** The figure shows the formation energy distribution of all QM7 training molecules (grey), together with the energy distribution of rediscovered molecules for the two agents **A** and **AV**, with mean values shown vertically. Despite rediscovering less than 50%, the RL rediscovered energies are actually better (more negative) than the QM7 average.

3 Discussion

We presented an autoregressive, multi-composition reinforcement learning (RL) agent for 3D isomer discovery, trained purely online across a large set of molecular formulas derived from the QM7 dataset. This represents a significant advancement over composition-specific RL agents for 3D structure discovery that have been developed in recent years. Our method enables, for the first time, the autonomous exploration of broad chemical spaces without reliance on curated datasets, paving the way for serendipitous molecular discoveries. Our method used smaller learning rates, higher entropy coefficients, and a multi-bag training scheme that improved chemical and geometric diversity to achieve generalization unlike previous methods. These design choices prevented premature convergence to locally optimal policies and enabled broader exploration of molecular space. As a result, the agent learned to generate a diverse set of valid isomers—even for unseen formulas—and significantly outperformed single-bag agents. We attribute this performance boost to the limited exploration and poor representation learning in single-bag settings.

Notably, we found that terminal rewards yielded more stable learning than per-step rewards, despite common assumptions favoring the latter for better credit assignment. This likely stems from the chemical implausibility of intermediate structures, which makes stepwise rewards noisy or misleading. However, a major limitation of terminal rewards—and RL more broadly—is the sparse and delayed nature of the learning signal, leading to inefficient training. We also observed diminishing returns with extended training, suggesting limited scalability in the current setup.

Future work should address these issues, e.g., by introducing mechanisms that penalize structural redundancy across rollouts to avoid exploration collapse. Reducing spatial noise from the stochastic policy, which interferes with energy-based evaluation, is also crucial. Finally, reframing the task as online finetuning of a pretrained model could substantially accelerate training and improve sample efficiency, making RL more practical for real-world molecular and materials discovery.

Software and Data

Code and instructions will be made available at ...

Acknowledgements

We acknowledge financial support from the Independent Research Fund Denmark with project DELIGHT, Grant No. 0217-00326B, and the Pioneer Center for Accelerating P2X Materials Discovery (CAPeX), DNRF grant number P3. We also acknowledge support from the Novo Nordisk Foundation Data Science Research Infrastructure 2022 Grant: A high-performance computing infrastructure for data-driven research on sustainable energy materials, Grant No. NNF22OC0078009. Finally, we thank Jan Jensen and Magnus Strandgaard for useful discussions on molecular scoring using `xTB-GFN2` and `xyz2mol`.

References

Dylan M Anstine and Olexandr Isayev. Generative models as an emerging paradigm in the chemical sciences. *Journal of the American Chemical Society*, 145(16):8736–8750, 2023.

Nathan Brown, Marco Fiscato, Marwin HS Segler, and Alain C Vaucher. Guacamol: benchmarking models for de novo molecular design. *Journal of chemical information and modeling*, 59(3): 1096–1108, 2019.

Joshua Schrier, Alexander J Norquist, Tonio Buonassisi, and Jakoah Brgoch. In pursuit of the exceptional: research directions for machine learning in chemical and materials science. *Journal of the American Chemical Society*, 145(40):21699–21716, 2023.

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

Bhuvanesh Sridharan, Animesh Sinha, Jai Bardhan, Rohit Modee, Masahiro Ehara, and U Deva Priyakumar. Deep reinforcement learning in chemistry: A review. *Journal of Computational Chemistry*, 2024.

Marcus Olivecrona, Thomas Blaschke, Ola Engkvist, and Hongming Chen. Molecular de-novo design through deep reinforcement learning. *Journal of cheminformatics*, 9:1–14, 2017.

Mariya Popova, Olexandr Isayev, and Alexander Tropsha. Deep reinforcement learning for de novo drug design. *Science advances*, 4(7):eaap7885, 2018.

Albert Bou, Morgan Thomas, Sebastian Dittert, Carles Navarro, Maciej Majewski, Ye Wang, Shivam Patel, Gary Tresadern, Mazen Ahmad, Vincent Moens, et al. Acegen: Reinforcement learning of generative chemical agents for drug discovery. *Journal of Chemical Information and Modeling*, 2024.

Sereina Riniker and Gregory A Landrum. Better informed distance geometry: using what we know to improve conformation generation. *Journal of chemical information and modeling*, 55(12): 2562–2574, 2015.

Emiel Hoogeboom, Vıctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d. In *International conference on machine learning*, pages 8867–8887. PMLR, 2022.

Tuan Le, Julian Cremer, Frank Noe, Djork-Arné Clevert, and Kristof T Schütt. Navigating the design space of equivariant diffusion-based generative models for de novo 3d molecule generation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=kzGuiRXZrQ.

François Cornet, Grigory Bartosh, Mikkel Schmidt, and Christian Andersson Naesseth. Equivariant neural diffusion for molecule generation. *Advances in Neural Information Processing Systems*, 37: 49429–49460, 2024a.

Yuxuan Song, Jingjing Gong, Minkai Xu, Ziyao Cao, Yanyan Lan, Stefano Ermon, Hao Zhou, and Wei-Ying Ma. Equivariant flow matching with hybrid probability transport for 3d molecule generation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=hHUZ5V9XFu.

Ross Irwin, Alessandro Tibo, Jon Paul Janet, and Simon Olsson. Semlaflow – efficient 3d molecular generation with latent attention and equivariant flow matching. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025. URL https://openreview.net/forum?id=bee2G6pEh0.

Niklas Gebauer, Michael Gastegger, and Kristof Schütt. Symmetry-adapted generation of 3d point sets for the targeted discovery of molecules. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Niklas WA Gebauer, Michael Gastegger, Stefaan SP Hessmann, Klaus-Robert Müller, and Kristof T Schütt. Inverse design of 3d molecular structures with conditional generative neural networks. *Nature communications*, 13(1):973, 2022.

James P. Roney, Paul Maragakis, Peter Skopp, and David E. Shaw. Generating realistic 3d molecules with an equivariant conditional likelihood model, 2022. URL https://openreview.net/forum?id=Snqhqz4LdK.

Ameya Daigavane, Song Kim, Mario Geiger, and Tess Smidt. Symphony: Symmetry-equivariant point-centered spherical harmonics for molecule generation. *arXiv preprint arXiv:2311.16199*, 2023.

Raul Ortega Ochoa, Tejs Vegge, and Jes Frellsen. Molminer: Transformer architecture for fragment-based autoregressive generation of molecular stories. *arXiv preprint arXiv:2411.06608*, 2024.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=YCWjhGrJFD.

Runxuan Jiang, Tarun Gogineni, Joshua Kammeraad, Yifei He, Ambuj Tewari, and Paul M Zimmerman. Conformer-rl: A deep reinforcement learning library for conformer generation. *Journal of Computational Chemistry*, 43(27):1880–1886, 2022.

Alexandra Volokhova, Michał Koziarski, Alex Hernández-García, Cheng-Hao Liu, Santiago Miret, Pablo Lemos, Luca Thiede, Zichao Yan, Alán Aspuru-Guzik, and Yoshua Bengio. Towards equilibrium molecular conformation generation with gflownets. *Digital Discovery*, 3:1038–1047, 2024.

Gregor Simm, Robert Pinsler, and José Miguel Hernández-Lobato. Reinforcement learning for molecular design guided by quantum mechanics. In *International Conference on Machine Learning*, pages 8959–8969. PMLR, 2020.

Gregor Simm, Robert Pinsler, Gábor Csányi, and José Miguel Hernández-Lobato. Symmetry-aware actor-critic for 3d molecular design. In *International Conference on Learning Representations*, 2021.

Daniel Flam-Shepherd, Alexander Zhigalin, and Alán Aspuru-Guzik. Scalable fragment-based 3d molecular design with reinforcement learning. *arXiv preprint arXiv:2202.00658*, 2022.

Søren Ager Meldgaard, Jonas Köhler, Henrik Lund Mortensen, Mads-Peter V Christiansen, Frank Noé, and Bjørk Hammer. Generating stable molecules using imitation and reinforcement learning. *Machine Learning: Science and Technology*, 3(1):015008, 2021.

Rohit Modee, Ashwini Verma, Kavita Joshi, and U Deva Priyakumar. Megen-generation of gallium metal clusters using reinforcement learning. *Machine Learning: Science and Technology*, 4(2): 025032, 2023.

Jonas Elsborg and Arghya Bhowmik. Equivariant graph-representation-based actor–critic reinforcement learning for nanoparticle design. *Journal of Chemical Information and Modeling*, 63(12): 3731–3741, 2023.

Rafael Gómez-Bombarelli, Jennifer N Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, Ryan P Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 4(2):268–276, 2018.

François Cornet, Bardi Benediktsson, Bjarke Hastrup, Mikkel N Schmidt, and Arghya Bhowmik. Om-diff: inverse-design of organometallic catalysts with guided equivariant denoising diffusion. *Digital Discovery*, 3(9):1793–1811, 2024b.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Gabriel Dulac-Arnold, Nir Levine, Daniel J Mankowitz, Jerry Li, Cosmin Paduraru, Sven Gowal, and Todd Hester. An empirical investigation of the challenges of real-world reinforcement learning. *arXiv preprint arXiv:2003.11881*, 2020.

Riashat Islam, Peter Henderson, Maziar Gomrokchi, and Doina Precup. Reproducibility of benchmarked deep reinforcement learning tasks for continuous control. *arXiv preprint arXiv:1708.04133*, 2017.

Jun Xia, Yanqiao Zhu, Yuanqi Du, and Stan Z Li. A systematic survey of chemical pre-trained models. *arXiv preprint arXiv:2210.16484*, 2022.

Cédric Colas, Olivier Sigaud, and Pierre-Yves Oudeyer. A hitchhiker's guide to statistical comparisons of reinforcement learning algorithms. *arXiv preprint arXiv:1904.06979*, 2019.

Daniil Polykovskiy, Alexander Zhebrak, Benjamin Sanchez-Lengeling, Sergey Golovanov, Oktai Tatanov, Stanislav Belyaev, Rauf Kurbanov, Aleksey Artamonov, Vladimir Aladinskiy, Mark Veselov, et al. Molecular sets (moses): a benchmarking platform for molecular generation models. *Frontiers in pharmacology*, 11:565644, 2020.

Daniel Schwalbe-Koda and Rafael Gómez-Bombarelli. Generative models for automatic chemical design. In *Machine Learning Meets Quantum Physics*, pages 445–467. Springer, 2020.

Yutong Xie, Chence Shi, Hao Zhou, Yuwei Yang, Weinan Zhang, Yong Yu, and Lei Li. Mars: Markov molecular sampling for multi-objective drug discovery. *arXiv preprint arXiv:2103.10432*, 2021.

Dou Nie, Huifeng Zhao, Odin Zhang, Gaoqi Weng, Hui Zhang, Jieyu Jin, Haitao Lin, Yufei Huang, Liwei Liu, Dan Li, et al. Durian: A comprehensive benchmark for structure-based 3d molecular generation. *Journal of Chemical Information and Modeling*, 65(1):173–186, 2024.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. Pmlr, 2018.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Justin S Smith, Olexandr Isayev, and Adrian E Roitberg. Ani-1, a data set of 20 million calculated off-equilibrium conformations for organic molecules. *Scientific data*, 4(1):1–8, 2017.

Michael Gastegger, Kristof T Schütt, and Klaus-Robert Müller. Machine learning of solvent effects on molecular spectra and reactions. *Chemical science*, 12(34):11473–11483, 2021.

Dominic C Rose, Jamie F Mair, and Juan P Garrahan. A reinforcement learning approach to rare trajectory sampling. *New Journal of Physics*, 23(1):013013, 2021.

L. C. Blum and J.-L. Reymond. 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. *J. Am. Chem. Soc.*, 131:8732, 2009.

M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld. Fast and accurate modeling of molecular atomization energies with machine learning. *Physical Review Letters*, 108:058301, 2012.

Christoph Bannwarth, Sebastian Ehlert, and Stefan Grimme. Gfn2-xtb—an accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *Journal of chemical theory and computation*, 15 (3):1652–1671, 2019.

Yeonjoon Kim and Woo Youn Kim. Universal structure conversion method for organic molecules: from atomic connectivity to three-dimensional geometry. *Bulletin of the Korean Chemical Society*, 36(7):1769–1777, 2015.

Greg Landrum. *RDKit: Open-source cheminformatics*, 2024. URL https://www.rdkit.org/.

Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. *Advances in neural information processing systems*, 30, 2017.

Kristof Schütt, Oliver Unke, and Michael Gastegger. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In *International Conference on Machine Learning*, pages 9377–9388. PMLR, 2021.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8:229–256, 1992.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

Vijay Konda and John Tsitsiklis. Actor-critic algorithms. *Advances in neural information processing systems*, 12, 1999.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015.

Heta A. Gandhi and Andrew D. White. Explaining molecular properties with natural language. *ChemRxiv*, 2022. doi: 10.26434/chemrxiv-2022-v5p6m-v3.

A Agent policy

A.1 Hierarchical action scheme

We parametrize the agent's neural network policy following the *internal* agent structure in previous work Simm et al. (2020), where at each decision step the agent breaks its action up into a hierarchical cascade of subactions, with each selected subaction being explicitly used as condition for the following. The three subactions are as follows:

- *a)* selection of focal atom on the current canvas, around which a new atom will be placed,
- *b)* selection of an element from the remaining bag,
- *c)* 3D placement of the new atom using a spherical coordinate system (d, ψ, α), where the coordinate axes are derived from distance vectors from the focal atom to - and among - its nearest neighbors.

In summary, the total agent policy is represented by the following factorization

$$\pi_\theta(d, \alpha, \psi, e, f|s) = p(d, \alpha, \psi|e, f, s)p(e|f, s)p(f|s), \tag{6}$$

where the element e and focal atom f are sampled from categorical distributions and each of the coordinates (ψ, α, d) (which together define a unique mapping to a 3D location for the new atom) are sampled from continuous distributions (univariate Gaussians).

While not immediately evident from the equations above, we must highlight the following potential conceptual flaws in the specific code implementation of this policy:

- *a)* The 3D position distribution $p(d, \alpha, \psi|e, f, s)$ from Eq. (6) actually imposes an assumption of independence between the internal spherical coordinates:

$$p(d, \alpha, \psi|e, f, s) = p(d|e, f, s)p(\alpha|e, f, s)p(\psi|e, f, s), \tag{7}$$

  since these variables are sampled independently rather than sequentially with mutual conditioning.
- *b)* Even an infinitesimally small change in atomic positions can cause permutations in the nearest-neighbor ordering, thereby altering how prediction spherical coordinates (ψ, α, d) map to Cartesian canvas space.

In combination, these issues limit the agent's ability to make intentional and coordinated placement decisions with predictable outcomes. However, despite their potential significance, we proceed with this baseline implementation and leave the development of mitigation strategies for these policy limitations to future work.

A.2 Backbone message passing

Instead of the *invariant* backbone (Schütt et al., 2017) used in the baseline work Simm et al. (2020), we use its *equivariant* counterpart (Schütt et al., 2021), which additionally propagates vectorial features through its message-passing layers. Although the spatial actions themselves only need rotational *invariance*, given the nearest-neighbor-based internal coordinate system, invariant GNN architectures are known to have limited spatial and orientational awareness compared to equivariant architectures. This limitation has been underscored in subsequent work (Simm et al., 2021) that demonstrated that equivariant architectures more effectively capture geometric relationships. Thus, enhancing the message-passing backbone by adopting the equivariant PaiNN architecture is well justified. Note, however, that within the neural network computation graph, the agent's spatial subactions (d, α, ψ) remain dependent exclusively on the invariant scalar tensor outputs rather than directly utilizing equivariant vectorial features.

B Training

B.1 Proximal Policy Optimization

As shown in Eq. (6), action probabilities are explicitly modeled through a stochastic policy, making the framework amenable to the broader class of policy gradient methods (Williams, 1992; Sutton et al., 1999). These methods optimize the expected return by following the gradient of the policy's parameters and are particularly effective when combined with a learned value function to form an Actor-Critic architecture (Konda and Tsitsiklis, 1999), which maintains both a stochastic policy (the actor) and a value function $V_\theta(s_t)$ (the critic). Specifically, we employ Proximal Policy Optimization (PPO) (Schulman et al., 2017), a widely used Actor-Critic algorithm designed to stabilize learning by preventing excessively large policy updates, as we shall see shortly.

At each iteration of PPO training, the agent first samples molecules according to its *current* action policy, i.e. we record trajectory rollouts consisting of transition tuples $(s_t, a_t, r_t, s_{t+1}) \sim \pi_\theta$. Based on this data buffer of freshly sampled rollouts, $\mathcal{B}_\theta$, PPO then enters a sequence of optimization steps, $k = 1, 2, 3.., K$, where it performs gradient ascent on the following combined objective:

$$\mathcal{L}^{PPO}(\theta) = \mathcal{L}^{\text{CL}}(\theta) - c_1\mathcal{L}^{\text{V}}(\theta) + c_2\mathcal{L}^{\mathcal{H}}(\theta), \qquad c_1, c_2 \geq 0, \tag{8}$$

where $\mathcal{L}^{\text{CL}}$ denotes the PPO *clipped* surrogate objective, $\mathcal{L}^{\text{V}}$ is the value function loss, and $\mathcal{L}^{\mathcal{H}}$ represents the entropy regularization term.

The first objective term, $\mathcal{L}^{\text{CL}}(\theta)$, plays a central role in policy optimization, as it directly guides the policy toward selecting high-scoring actions. Rather than naïvely maximizing raw returns, PPO uses the Generalized Advantage Estimator (GAE) (Schulman et al., 2015) to compute a smoothed signal of how advantageous an action was, relative to the expected value:

$$\hat{A}_t = \sum_{k=0}^{T-t-1} (\gamma\lambda)^k \delta_{t+k}, \qquad \text{with} \quad \delta_t = r_t + \gamma V_\theta(s_{t+1}) - V_\theta(s_t), \tag{9}$$

where $\gamma \in [0, 1]$ is the discount factor and $\lambda \in [0, 1]$ is a smoothing parameter that controls the bias-variance trade-off. The resulting advantage term $\hat{A}_t$ captures the degree of *positive surprise*, i.e. how much better (or worse) the observed return was compared to what the critic predicted.

After a few policy updates, the optimized policy π_θ can deviate significantly from the original policy that generated the data, which we from now on denote $\pi_{\theta_{\text{old}}}$. This mismatch can lead to unstable learning, as the objective is now being evaluated under a different distribution. To mitigate this, PPO introduces a clipped surrogate objective that limits how much the policy is allowed to change in a single update. The objective is defined as:

$$\mathcal{L}^{\text{CL}}(\theta) = \mathop{\mathbb{E}}_{(s_t,a_t)\sim\mathcal{B}_{\theta_{\text{old}}}} \left[\min\left(r_t^{\spadesuit}(\theta)\hat{A}_t; \text{clip}(r_t^{\spadesuit}(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t \right) \right], \quad r_t^{\spadesuit}(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}. \tag{10}$$

Here, the policy ratio $r_t^{\spadesuit}(\theta)$ quantifies how the probability of taking action a_t under the updated policy compares to that under the old policy. The `clip` operation ensures that this ratio stays within a trust region defined by $[1-\epsilon, 1+\epsilon]$, where ϵ is a small constant (commonly 0.2). This conservative update rule prevents excessively large policy shifts, thereby improving training stability. Intuitively, it balances learning progress - encouraging updates when $\hat{A}_t$ is large - with policy trustworthiness, by suppressing updates that would cause the policy to change too aggressively.

Next, the value loss term $\mathcal{L}^{\text{V}}$ is implemented as the mean squared error (MSE) between the predicted state-value estimates $V_\theta(s_t)$ and the empirical returns R_t. This term ensures that the critic network effectively predicts future returns, enabling more accurate advantage estimation in future iterations, while simultaneously distilling better representations into the shared backbone:

$$\mathcal{L}^{\text{V}}(\theta) = \mathop{\mathbb{E}}_{s_t\sim\mathcal{B}_{\theta_{\text{old}}}} \left[\left(V_\theta(s_t) - R_t \right)^2 \right]. \tag{11}$$

Finally, the entropy term encourages exploration by penalizing overly certain action distributions:

$$\mathcal{L}^{\mathcal{H}}(\theta) = \mathcal{H}[\pi_\theta(e_t, f_t|s_t)] = \mathop{\mathbb{E}}_{(s_t,f_t,e_t)\sim\mathcal{B}_{\theta_{\text{old}}}} \left[-\log \pi_\theta(e_t, f_t|s_t) \right]. \tag{12}$$

Note that **entropy maximization is applied only to the two categorical subactions**: choosing the focal atom f and selecting a new element e. The continuous spatial subactions (d, α, ψ), parameterized as univariate Gaussians, are exempt from entropy regularization to avoid inadvertently disrupting precise spatial predictions. Furthermore, to prevent premature convergence and encourage sustained exploration, we linearly increase the entropy coefficient, c_2, during training from 0.15 to 0.25.

B.2 Hyper parameters

In Table 2 we show the hyper parameters used for model training. Additionally, we highlight the individual hyper parameters whose values most crucially facilitated the training of a *generalizable* agent (multi-component agent), as well as those unique to our setup. In addition to our new reward structures and training schemes, the most noticeable difference from previous work is the use of larger data collections, smaller learning rates and higher exploration factors.

Table 2: Hyperparameters.

Category	Hyperparameter	Value	Code variable name
ROLLOUT	Range $[d_{\text{min}}, d_{\text{max}}]$ (Å)	[0.8, 1.8]	`[min,max]_mean_distance`
	Workers	8	`num_envs`
	Env steps per PPO batch	512	`num_steps_per_iter`
	Fail reward	-3	`min_reward`
OPTIMIZATION	Discount factor γ	1.0	`discount`
	GAE parameter λ	0.97	`lam`
	Value coefficient	0.5	`vf_coef`
	Advantage clipping ϵ	0.2	`clip_ratio`
	Gradient Clipping	0.5	`gradient_clip`
	Learning rate (PPO-Adam)	5e-5	`learning_rate`
	Minibatch size	256	`mini_batch_size`
	Entropy coef start	0.15	`start_entropy`
	Entropy coef end	0.25	`final_entropy`
	Entropy increase steps	30,000	`total_steps`
PAINN EMBEDDING	Layers	3	`num_interactions`
	Network width	128	`network_width`
	r_{cutoff} (Å)	5	`cutoff`
QM7 SPLITS	Num training bags	156	`..`
	Num eval bags	20	`n_test`

B.3 Reward coefficients

For each trained agent presented in Section 2.2, we used the reward coefficients listed in Table 3. As shown in the table, we assigned a higher coefficient to validity compared to energy based reward components. This was decided based on the observed magnitude and variance of the reward components.

Table 3: Coefficients used to construct the reward functions (agents) as linear combinations of the basic reward components $\mathcal{A}$, $\mathcal{F}$ and $\mathcal{V}$. Empty corresponds to zero.

Agents \ *Components*	$\mathcal{A}$	$\mathcal{F}$	$\mathcal{V}$
A	1		
AV	1		3
F		1	
FV		1	3
AFV	1	1	3

C Q1: Molecule visualizations

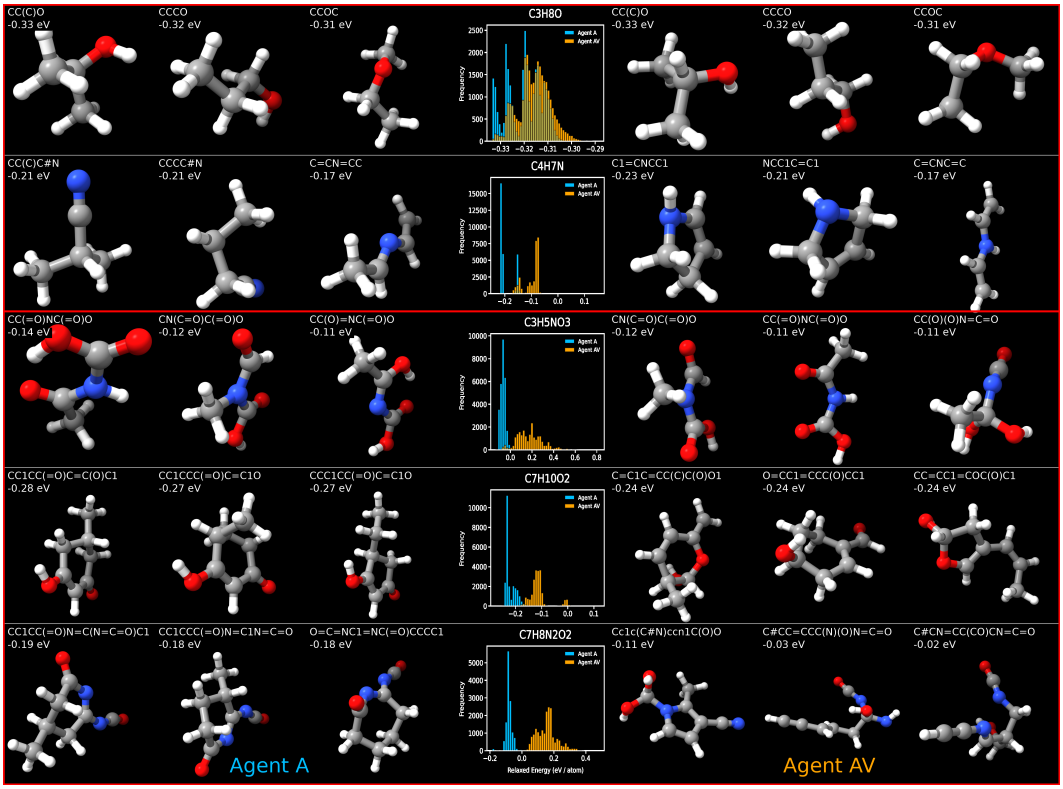

Figure 6: **Q1 visualizations.** Histograms of formation energy per atom after structural relaxation (center column) together with top 3 best scoring molecules (lowest energy) for Agent **A** on the left and Agent **AV** on the right. **Agent A samples molecules of significantly lower energies.**

D Metric definitions

We split the evaluation metrics of Fig. 7 into two distinct categories. The *discovery metrics* are purely count-based and contain the following metrics:

- **Validity**: Validity is not directly built into the molecular generation procedure used in our framework[4]. Instead we incentivize the agent to create valid molecules based on a simple discrete reward term $r_{\text{valid}} = 1$ if valid and $r_{\text{valid}} = 0$ if invalid. The validity metric is straightforwardly defined as

$$\text{Validity} = \frac{\#\text{valid molecules}}{\#\text{sampled molecules}}. \tag{13}$$

- **Rediscovery & Expansion Ratios**:
  Relating the discovery counts to our reference dataset (QM7) helps to probe whether the agent explores broadly or if there are large gaps in its exploration. For each formula, we therefore construct the set of uniquely discovered SMILES from the RL generated molecules. Each discovered SMILES will then either be in the reference set already or correspond to a "novel" molecule, i.e. $N^{\text{gen}}_{\text{unique}} = N_{\text{rediscovered}} + N_{\text{novel}}$. The rediscovery and expansion ratios are calculated by relating $N_{\text{rediscovered}}$ and N_{novel} to the number of reference molecules

$$\text{Rediscovery Ratio} = \frac{N_{\text{rediscovered}}}{N^{\text{QM7}}_{\text{unique}}}, \qquad \text{Expansion Ratio} = \frac{N_{\text{novel}}}{N^{\text{QM7}}_{\text{unique}}}. \tag{14}$$

The *energy metrics* pertain to the *quality* of the discovered geometries rather than their sheer *quantity*. Since our agent was trained on energy based reward terms, it should be able to generate low energy isomers. However, as our PPO agent uses 3D-spatial noise on the atomic positions in order to facilitate exploration, we must first perform structural relaxation on the generated molecules using the same `xTB-GFN2` calculator that was employed for reward calculations during training. Specifically, we calculate the following energy based metrics:

- **Relaxed Relative Atomic Energy (rRAE)**: This measure is defined w.r.t. our reference dataset QM7 and is calculated (at the individual molecule level) as the energy difference between our RL generated molecule and the mean energy of all the QM7 molecules of the same chemical formula (bag)

$$\Delta E_{\text{rRAE}}(\mathcal{C}_T) = E(\mathcal{C}_T) - \bar{E}^{\mathcal{B}(\mathcal{C}_T)}_{\text{QM7}} = E(\mathcal{C}) - \frac{1}{|\mathcal{B}(\mathcal{C}_T)|} \sum_{i=1}^{|\mathcal{B}(\mathcal{C}_T)|} E(\mathcal{C}^{\text{QM7}}_i). \tag{15}$$

  It measures the agent's joint ability to discover both low energy isomers (2D connectivity) as well as sampling low energy conformers (3D positions) for the connectivity matrix of that isomer.

- **Root-Mean-Square Deviation (RMSD):** To quantify how far the generated 3D structures deviate from their corresponding relaxed geometries, we compute the Root-Mean-Square Deviation (RMSD) between each generated molecule and its structure after geometry optimization using the `xTB-GFN2` method. This metric measures the average atomic displacement required to reach a local energy minimum and is defined as:

$$\text{RMSD}(\mathcal{C}_T) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left|\mathbf{x}_i - \mathbf{x}^{\text{relaxed}}_i\right|^2}, \tag{16}$$

  where $\mathbf{x}_i$ and $\mathbf{x}^{\text{relaxed}}_i$ denote the 3D positions of atom i before and after end-of-episode relaxation, and N is the number of atoms. A low RMSD indicates that the generated geometry was already close to a local minimum, suggesting a physically meaningful placement of atoms by the agent. In contrast, a high RMSD implies the presence of significant strain or artifacts in the initial structure that required substantial correction during optimization.

[4] A word on *uniqueness*: The typically reported uniqueness measure which relates the number of unique molecules to the number of sampled molecules would be a misleading metric to use in our case, since we are sampling molecules constrained to yield a pre-specified chemical formula, thereby increasing the probability of generating identical molecules compared to unconstrained sampling. As an example, we found just 3 isomers out of 10,000 generated molecules for C_3H_8O in Table 1 (3 is actually the maximal number of unique molecules for this particular chemical formula). Thus, we decided to leave out this metric from Fig. 7.

E Q2: Reward term comparison - *Multi-bag aggregated evaluation*

While the single-bag discovery task demonstrated the superior discovery capabilities of our agents, a much broader evaluation scheme is necessary for a robust comparison of reward signals. To achieve this, we also carried out the experiments that was outlined in Fig. 2b (blue). This experiment aggregates results from a random split of 20 holdout formulas in the QM7 dataset, offering a more comprehensive assessment compared to single-bag evaluation.

For each test bag $\mathcal{B}_i$, $i = 1, 2, ..., 20$ we sampled $N_i = P \cdot N_i^{\text{ref}}$ molecules, where N_i^{ref} is the number of isomers in the reference dataset for $\mathcal{B}_i$, and $P = 100$ is a proportionality factor. This scaling ensures that the number of sampled molecules reflects the expected isomer diversity for each bag. The results, including standard deviations across three seeds, are presented in Fig. 7. Note that to obtain these metrics we first calculated these statistics for each bag individually and then aggregated across all hold-out bags using a weighted average according to N_i.

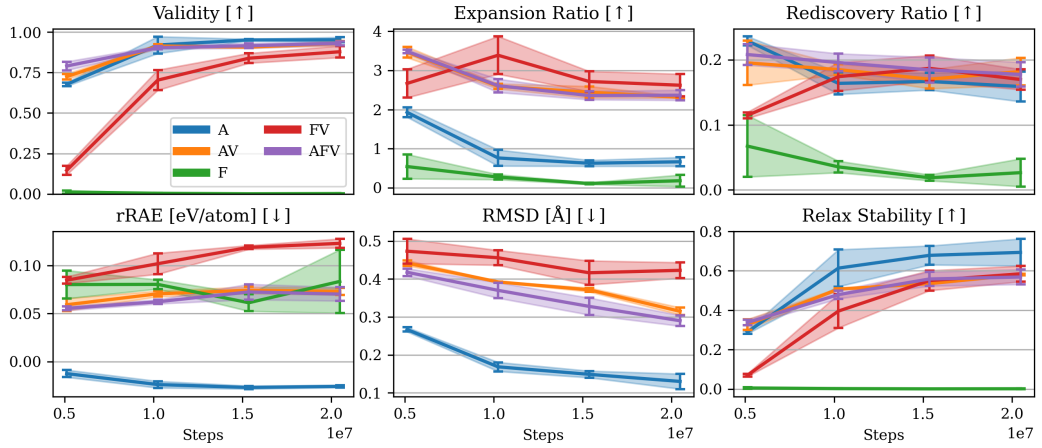

Figure 7: **Q2: Out-of-sample agent comparison.** We report discovery metrics (top row) and geometry metrics (bottom row) in the multi-bag evaluation setting outlined in Fig. 2b (see page 17 for detailed metric definitions). Each point reflects a weighted average across 20 test bags. Error bars denote standard deviation across three random seeds. Results show that agent **A** consistently outperforms on 3D metrics, while **AV** and **AFV** perform identically—highlighting the redundancy of $\mathcal{F}$ in our setup. Agent **F** fails to discover valid molecules due to excessive intra-episode rewards. Flat rediscovery and expansion metrics suggest no mode collapse, but agents fail to turn this into continued improvement toward more stable molecules.

From Fig. 7 we make the following observations:

- **>90% validity** for agents that use either $\mathcal{A}$ or $\mathcal{V}$, with only **FV** being slow to reach these levels. This is more than sufficient in an RL context, as agents must take exploratory moves in order to drive discovery, and an agent that always reaches 100% validity is possibly too conservatively. Despite these high validities, we note that only $60\% - 70\%$ of the generated molecules were "Relax Stable" (bottom right), meaning that relaxation did not alter the geometries significantly enough to change their bond connectivity and thus their SMILES representation.
- **Agent A dominates on 3D metrics (same as Q1).** This confirms our observation from **Q1** that the terminal atomization energy signal, $\mathcal{A}$, best facilitates discovery of low energy structures. In fact, not only is it better than all other agents across all checkpoints, it is also better than the molecules present in the QM7 reference dataset *on average*.
- **Agent F does not learn to create valid molecules.** Most likely it receives too strong intra-episode reward signals and fails to plan for the full horizon, which is a crucial requirement as molecules are only collected at the end of the episode when the bag is empty.
- **AV and AFV perform identically.** This shows that the per-step formation energy, $\mathcal{F}$, is a redundant signal in our setup and does not contribute significantly to justify the increase in energy queries within each episode.

- **Relaxed energy metrics don't improve (rRAE).** While the raw energies of the generated molecules do improve during training (not shown here), this effect is washed out upon relaxation as shown in the *rRAE* plot. Despite the generated molecules becoming less noisy, as seen from the *RMSD* and *Relax Stability* curves, the agents' abilities to select low-energy isomers appears stagnant. And while it seems positive that the RMSD continues decreasing throughout training, this is merely a bi-product of the gradual narrowing of the width of the univariate Gaussians in their action policies from Eq. (6) (recall that in order to facilitate exploration, the agents add Gaussian noise to the three spherical coordinates (r, α, ψ) as they sample the 3D position of the new atom).
- **Rediscovery & Expansion metrics are flat.** On the positive side, this means that the agents don't display severe mode collapse, which would cause these numbers to decrease during training. The sustained discovery is likely a consequence of the entropy regularization term (Eq. (12)) on the two discrete actions that ultimately determine the connectively and hence the SMILES object.

F Q3: Functional group analysis

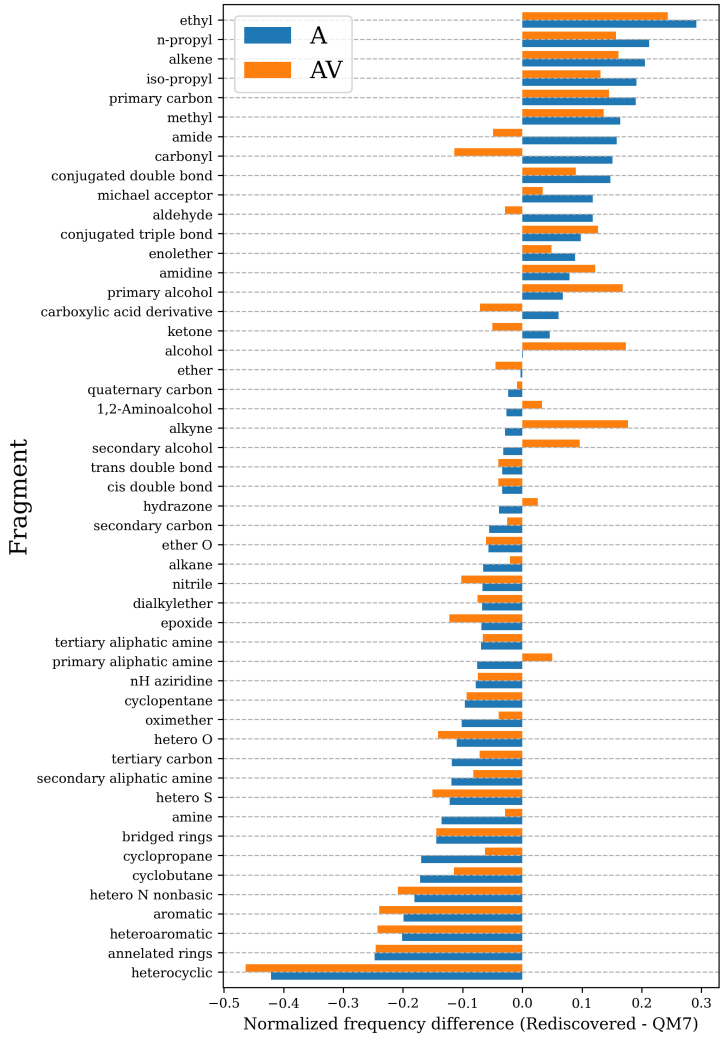

Figure 8: **Functional group analysis.** Frequency difference of the 50 most common QM7 functional groups between rediscovered and reference molecules. In particular, to avoid rare groups populating the extreme ends of the spectrum, we present a normalized frequency difference calculated as $\tilde{\Delta}_f^i = (f_{\text{RL}}^i - f_{\text{QM7}}^i)/\sqrt{f_{\text{QM7}}^i}$ with $f_{\mathcal{D}}^i$ representing the rate of occurence of fragment i in dataset $\mathcal{D}$. Positive values indicate overrepresentation by the RL agent; negative values indicate underrepresentation.

We first pool rediscoveries across all three seeds to ensure robust statistics. Then, using the `exmol` package (Gandhi and White, 2022), we extract the most frequent functional groups in QM7 and compare their occurrence in rediscovered molecules vs. the full dataset. The results are shown in Fig. 8 where several key trends emerge. Among the most strongly underrepresented functional groups, we find

- **Heterocycles and aromatic systems** (e.g., *heterocyclic*, *heteroaromatic*, *aromatic*, *annelated rings*): These involve complex ring topologies and delocalized bonding, which are difficult to construct via sequential atom placement.

- **Strained and fused rings** (e.g., *cyclopropane*, *cyclobutane*, *cyclopentane*, *bridged rings*, *nH aziridine*): Geometrically strained or topologically complex rings are less favored due to their instability and the precise coordination required to assemble them.

- **Heteroatoms and functionalized amines** (e.g., *hetero N nonbasic*, *hetero S*, *hetero O*, *amine*, *primary/secondary aliphatic amine*, *oximether*): These groups introduce electronic and geometric diversity, making them harder to learn and reproduce, especially when their placement significantly affects molecular stability.

In contrast, several functional groups are overrepresented, indicating that the agents preferentially discover molecules with these features:

- **Simple alkyl groups** (e.g., *methyl*, *ethyl*, *n-propyl*, *iso-propyl*, *primary carbon*): These groups are structurally simple and frequently encountered in organic molecules, making them easy for the agent to generate and overrepresented in the rediscovered set.
- **Carbonyl-containing groups** (e.g., *carbonyl*, *amide*): These planar, well-defined functional groups may be favored by the energy model used during training, and are commonly found in stable molecules.
- **Unsaturated hydrocarbons** (e.g., *alkene*): These motifs are structurally simple and energetically favorable, leading to their frequent appearance in generated molecules.