A multi-composition reinforcement learning framework for isomer discovery in 3D

Bjarke Hastrup **François Cornet** **Tejs Vegge** **Arghya Bhowmik**
Technical University of Denmark
{bjaha, frjc, teve, arbh}@dtu.dk

Abstract

We present a generative agent for stoichiometry-constrained isomer search. Our approach trains entirely in 3D using a purely online Reinforcement Learning (RL) framework. Unlike prior approaches, which overfit to specific chemical formulas, we introduce a multi-composition training framework that enables the agent to generalize across a wide range of formulas. This is achieved by leveraging a reference dataset to procedurally define new generation tasks, simultaneously facilitating a more formal evaluation of the agent's discovery capabilities. Combined with new energy- and validity-based rewards, we demonstrate that our approach significantly outperforms previous work, discovering an order of magnitude more valid isomers for unseen test formulas. By addressing these challenges, we aim to reinvigorate progress in self-guided 3D molecular discovery, providing a more robust framework for future innovations in the field.

1 Introduction

The discovery of novel molecules with desired properties is a grand challenge. Effectively exploring the immense chemical space is however a notoriously difficult endeavor that requires innovative search methods. Recently, generative models have emerged as a promising avenue for such task (Anstine and Isayev, 2023). Yet, their training often hinges on the availability of suitable data. Public datasets are usually not curated with the optimization of specific properties in mind, and the property ranges critical for the problem at hand may lie at the extremes or beyond what existing datasets span. This poses a challenge, as generative models must not only interpolate within the provided data but also be capable of extrapolating beyond it. Similarly, constructing an adequate training dataset may prove difficult in some cases, and even when such dataset exists, it inevitably carries inherent chemical and structural biases that can potentially hinder generalization to novel molecular spaces.

A compelling approach to overcoming these limitations is to adopt online (*de-novo*) learning techniques, 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 2D molecular generation (Olivecrona et al., 2017; Bou et al., 2024). However, when 3D geometry is relevant, an additional post-processing step is required to obtain conformations, e.g. via an external software (Riniker and Landrum, 2015). Instead, direct generation in 3D enables molecular structures to be generated and optimized in a fully integrated, end-to-end framework.

Contributions In this paper, we build upon MOLGYM (Simm et al., 2020; 2021) and target *de-novo* isomer discovery, where the agent is tasked to generate 3D conformations given a pre-specified chemical composition. We innovate with a novel multi-composition training scheme and new rewards, and 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 setup is provided in Fig. 1, and we summarize our main contributions as follows:

38th Conference on Neural Information Processing Systems (NeurIPS 2024) – ML4PS Workshop.

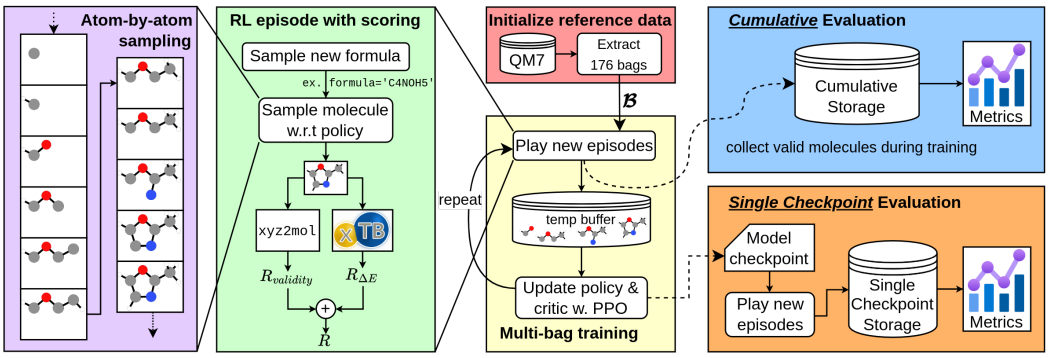

Figure 1: **Multi-composition training and evaluation workflow.** Our framework constructs isomer generation tasks by extracting chemical formulas from a reference dataset and introduces new terminal rewards based on validity and total energy. We evaluate the RL agents' isomer discovery capabilities at just a *single* checkpoint as well as *cumulatively* across the entire discovery campaign.

- We introduce new terminal rewards based on energy and chemical valency, thereby teaching the agent to build stable and *valid* molecules.
- We propose a multi-bag training setup based on a *reference dataset* to facilitate generalization across stoichiometries.
- We design a broader multi-bag evaluation scheme to facilitate benchmarking of online isomer discovery and evaluate various combinations of the presented reward terms.

Related Work In the supervised setting, the most promising directions for molecule generation in 3D are currently either based on diffusion models (Hoogeboom et al., 2022), 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). While these models could potentially be integrated into a pretraining-finetuning framework (Black et al., 2024), it remains unclear whether they can effectively be used for *de-novo* learning. 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-composition, their online finetuning was for single compositions only and further relied on result aggregation from 64 parallel finetunings spawned after pretraining. In contrast, we aim to train stoichiometry-agnostic RL agents.

2 Methods

We train an RL agent to build stable and valid molecules autoregressively in an atom-by-atom fashion, 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 different evaluation schemes used in Section 3.

Episode and terminal scoring We generate molecules similarly to MOLGYM, where at each step, our agent observes a state $s_t = (\mathcal{C}_t, \mathcal{B}_t)$ consisting of the current canvas $\mathcal{C}_t$ (i.e. molecule built so far) and the current atom bag $\mathcal{B}_t$ (i.e. remaining atoms to be placed). 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 a deterministic transition to the next state $s_{t+1} = (\mathcal{C}_{t+1}, \mathcal{B}_{t+1})$, where $\mathcal{C}_{t+1} = \mathcal{C}_t \cup \{(e_t, x_t)\}$ and $B_{t+1} = B_t \setminus \{e_t\}$. This process continues until the bag is empty and a complete molecule $\mathcal{C}_\mathcal{T}$ has been formed.

The agent optimizes its stochastic policy $\pi_\theta(a_t|s_t)$ in search of the optimal parameters θ that maximize the expected discounted sum of future rewards (known as *return*) from any given state, $V^\pi(s_t) = \mathbb{E}_{\pi_\theta}\left[\sum_{t'=t}^{T} \gamma^{t'} r(s_{t'}, a_{t'})\right]$, 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 specifying the distribution over bags (see Multibag setup below). Unlike MOLGYM however, our agent only receives a reward at

the terminal state. This reward is determined based on quantum mechanical energy using `GFN2-xTB` (Bannwarth et al., 2019) and chemical valency rules via `xyz2mol` (Kim and Kim, 2015). See Appendix A.1 for more details on the autoregressive sampling scheme and the reward functions.

Agent policy and optimization We parametrize our neural network policy as the original MOLGYM *internal* agent, but we replace the *invariant* backbone based on SCHNET (Schütt et al., 2017) with its *equivariant* counterpart PAINN (Schütt et al., 2021). Similarly to MOLGYM, we optimize the agent's policy with PPO (Schulman et al., 2017).

Multibag setup We leverage a *reference dataset* from which we solely extract formulas to construct a bag set, $\boldsymbol{\mathcal{B}}$, used for multi-composition training. 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 $\boldsymbol{\mathcal{B}}_w = \text{permutation}_w(\boldsymbol{\mathcal{B}})$. 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.

3 Experiments

In this section, we present three distinct evaluation scenarios to assess our agent's performance. First, in Section 3.1, we evaluate our agent's discovery capabilities in the single-bag generation paradigm, directly comparing it to previous online RL methods. Next, in Section 3.2, we broaden the scope by aggregating results across a random split of chemical formulas from the QM7 dataset, enabling a comprehensive comparison of reward signals. Finally, in Section 3.3, we examine the complete pool of molecules generated during training, with a particular focus on the breadth of discovery achieved.

While Section 3.1 and Section 3.2 focus on evaluating a single final checkpoint, Section 3.3 examines the agent's performance throughout the entire training process. Despite these differences in evaluation scope, all three cases are based on the same training runs illustrated in Fig. 2. Notably, our newly introduced terminal reward terms, **A** and **V**, enable significantly more stable training dynamics.

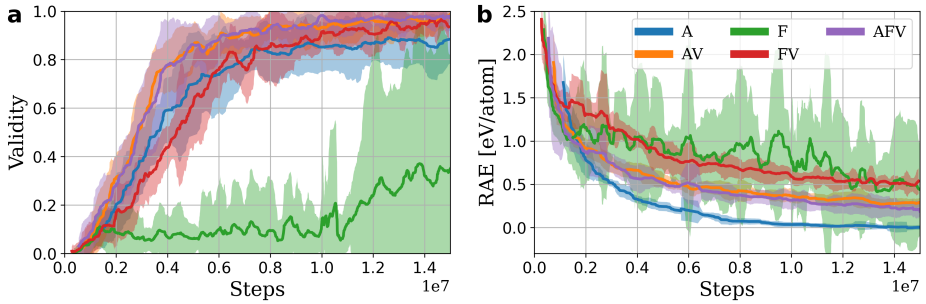

Figure 2: **Learning Curves.** *(a)* Validity and *(b)* Relative Atomic Energy (RAE) plotted as a function of the number of single-atom placements on the canvas. The RAE metric quantifies the excess energy relative to the average energies of QM7 molecules with the same chemical formula (see Appendix C for detailed metric definitions). Results are aggregated across three independent training runs with different random seeds, with shading representing ± 2 standard deviations.

3.1 Single-bag discovery

We adopt the setup from Simm et al. (2021), counting the number of valid constitutional isomers[1] discovered by our agent when deployed on a single bag. Table 1 compares our results with those of previous work, highlighting the effectiveness of our new training setup, reward scheme, and data collection procedure. Remarkably, despite not being explicitly trained on certain formulas[2], our agent frequently discovers up to an order of magnitude more constitutional isomers than baseline agents. However, it is important to note several key differences between our approach and prior work, which we discuss in detail in Appendix B.

[1]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.

[2]The smaller bags $\{C_4H_7N, C_3H_8O\}$ were part of the training set while the bigger bags $\{C_3H_5NO_3$, $C_7H_{10}O_2$, $C_7H_8N_2O_2\}$ were not.

Table 1: **Single-bag discovery.** Comparison of our ***atomisation*** and ***validity*** guided ***multi-bag*** agent (**MB-AV**) against prior work, taken directly from Simm et al. (2021). Our agent outperforms previous work, discovering an order of magnitude more valid isomers, even for unseen formulas.

Training type:	Single-bag training (on eval bag)		QM7 multibag
Collection type:	Cumulative argmax × 10 seeds		Single CP stochastic
Agent:	INTERNAL	COVARIANT	MB-AV **(ours)**
$C_3H_5NO_3$	35	65	**246** ± 40
C_4H_7N	18	25	**30** ± 4
C_3H_8O	4†	**8**†	3 ± 0
$C_7H_{10}O_2$	21	85	**716** ± 281
$C_7H_8N_2O_2$	58	118	**2662** ± 1583

†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^- on the first carbon atom, an OH^- 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 would only emphasize our improved discovery capabilities.

3.2 Multi-bag aggregated evaluation

A broader evaluation scheme is crucial for benchmarking future innovations in self-guided 3D molecular discovery. To achieve this, we aggregate results from a random split of 20 holdout formulas in the QM7 dataset, offering a more comprehensive assessment compared to single-bag evaluation.

Setup: For this experiment, we evaluate the final model checkpoints at 15 million steps. For each test bag $\mathcal{B}_i$, we sample $N_i = P \cdot N_i^{\text{ref}}$ 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. Metrics are calculated individually for each bag and then aggregated using a weighted average based on N_i. The results, including standard deviations across three seeds, are presented in Table 2 (see Appendix C for metric details). Note that for the rRAE energy measure, we first performed structural relaxation of the sampled molecules using the same `GFN2-xTB` calculator.

For all agents, except for the agent which is rewarded per-step as in MOLGYM, we observe a very high rate of validity of 94% and higher. Although high validity is preferred, we want our RL agents to take risks in order to discover new chemistries. The atomization agent **MB-A** is purely guided by terminal energy and is thus able to sample molecule that are lower in energy compared to the molecules from QM7. However, this agent is also the one most probe to exploration collapse.

Table 2: **Multi-bag evaluation.** Discovery and geometry metrics in the multi-bag evaluation case.

	Discovery metrics [↑]			Energy metric [eV/atom] [↓]
Agents:	Validity	Rediscovery-ratio	Expansion-ratio	ΔE_{rRAE}
MB-A	0.94 ± 0.05	0.12 ± 0.02	0.36 ± 0.07	**-0.03** ± 0.00
MB-AV	0.96 ± 0.00	0.14 ± 0.02	1.73 ± 0.32	0.04 ± 0.01
MB-F (MOLGYM rew)	0.30 ± 0.28	**0.19** ± 0.06	1.59 ± 0.66	0.02 ± 0.02
MB-FV	0.94 ± 0.01	0.15 ± 0.02	**2.12** ± 0.10	0.09 ± 0.00
MB-AFV	**0.98** ± 0.00	0.13 ± 0.03	1.37 ± 0.31	0.01 ± 0.02

3.3 Cumulative discovery

In the previous experiments, we evaluated the agent's performance based on the quality of stochastic rollouts at a single checkpoint, a method commonly used in generative modeling (e.g., supervised distribution learning). However, this approach introduces the arbitrariness of checkpoint selection, as performance can vary significantly across different stages of training (see training progression plots in Fig. 2). While the single-checkpoint evaluation scheme is simple and general, it overlooks the evolving nature of an RL agent's policy, which explores different strategies at various training stages.

To fully leverage this behavioral "drift," we store every molecule generated throughout training in a cumulative storage buffer, as illustrated in Fig. 1. In Fig. 3(a), we analyze a single training run of the **MB-AV** agent, visualizing the space of rediscovered molecules using a t-SNE projection (Van der Maaten and Hinton, 2008) of SOAP representations (Bartók et al., 2013). These are extracted from the QM7 geometries of the rediscovered SMILES. While exploration doesn't entirely collapse, the **MB-AV** agent fails to discover new molecules after a few million steps. In fact, we observe the same tendency among all agents with the atomization reward component (Fig. 3(d)-(f)).

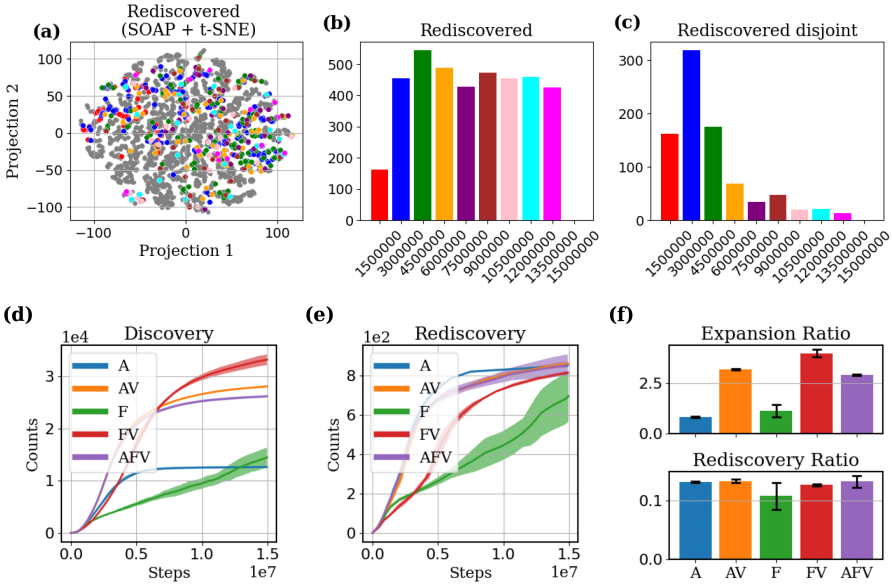

Figure 3: **Cumulative (re)discovery (in-sample).** (a)-(c): Rediscovery abilities of **MB-AV** agent. (a) SOAP space with "time of discovery" colored according to the batched time axis of the adjacent subfigures. (b) Rediscovery per batch. (c) Marginal rediscovery gain per batch. (d)-(e) Cumulative discovery and rediscovery in absolute terms (note the difference in magnitude). (f) Rediscover and Expansion *ratios*.

4 Conclusion and Discussion

We presented an autoregressive multi-composition RL agent for 3D isomer discovery, trained purely online on a large set of bags derived from the QM7 reference dataset. Compared to previous approaches, we employed smaller learning rates, higher exploration factors (entropy coefficients), and achieved greater training diversity through our multi-bag framework. This approach helped prevent the agent from "memorizing" rewarding actions or getting trapped in local minima. As a result, our generalist agent successfully learned to sample a wide range of valid molecules, even for unseen chemical formulas. It significantly outperformed single-bag agents in isomer discovery, which we hypothesize struggle due to limited exploration and insufficient geometric diversity to learn meaningful molecular representations. Furthermore, we found that, contrary to common assumptions about credit assignment, terminal rewards led to much more stable learning compared to per-step rewards. This is likely because partially constructed molecules in the atom-by-atom case are often chemically and energetically unrealistic. However, a significant drawback of terminal rewards—and RL in general—is the limited reward signal per episode, which results in inefficient training. We also observed diminishing marginal gains with extended training, indicating that the current setup does not scale effectively with increased computational resources. For purely online discovery, future work should address exploration collapse by introducing mechanisms to measure and penalize structural similarity to past rollouts. It will also be crucial to mitigate spatial noise arising from the stochastic agent policy, as this interferes with the evaluation of energy-based reward terms. Depending on the target application and data availability, reframing the learning task as online finetuning of a pretrained model could provide a more efficient approach, significantly accelerating experimentation and enhancing the applicability of RL in molecular and material discovery.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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

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

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Albert P Bartók, Risi Kondor, and Gábor Csányi. On representing chemical environments. *Physical Review B—Condensed Matter and Materials Physics*, 87(18):184115, 2013.

Greg Landrum et al. Rdkit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum*, 8:31, 2013.

A Training details

A.1 Reinforcement learning environment

Autoregressive molecule sampling The molecule construction process is modeled as a sequential decision-making task, where, after sampling a bag of atoms $\mathcal{B}_0$, an agent iteratively selects and places atoms in 3D space to incrementally build the molecule. In reinforcement learning (RL) terms, the agent observes the state $s_t = (\mathcal{C}_t, \mathcal{B}_t)$ consisting of the current molecular canvas $\mathcal{C}_t$ 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 all atoms from the bag are placed, forming a complete molecule. 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 inital 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 (see Section A.1). 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 the bag size.

Penalization of unrealistic molecules Whenever an atom is positioned dangerously close to any of the existing canvas atoms the episode is terminated and the agent is given a fixed failure penalty R_{kill}. As a consequence of the terminal reward scheme and the early termination of unpromising rollouts, an untrained agent will most likely not obtain any reward whatsoever. We therefore decided to give the agent a small constant and positive per-step reward to incentivize it to reach the end of the episode at which point the energy and validity dependent terminal reward is given. Theoretically however, this should be unnecessary whenever the death penalty is $R_{\text{kill}} < 0$ and the RL discount factor $\gamma < 1$.

Reward structure The temporal dimension in our RL episode is an artificial construction intended solely to facilitate the factorization of the agent's molecular sampling policy. As such, the partially completed molecules $\{\mathcal{C}_t\}_{t<T}$ are not guaranteed to make much sense from a chemical/energetic point of view. As a consequence, we only reward the agent once it has placed all atoms from the bag[3]. Specifically, we implement the following two terminal rewards:

- **Atomisation energy (A)**: The negative difference between potential energy of the final molecule and the sum of potential energies of each of its constituent atoms in isolation

$$r_A(s_T) = -\Delta E = \sum_{t=1}^{T} E(e_t) - E(\mathcal{C}_T) \tag{2}$$

  To evaluate energies we make use of the `GFN2-xTB` calculator.

- **Validity (V)**: A boolean validity check based on the requirement that generated molecules can be successfully parsed by the `xyz2mol` function which attempts to read an arbitrary 3D point cloud into an `rdkit` mol object (Landrum et al., 2013):

$$r_V(s_T) = \begin{cases} 1 \text{ if } \mathcal{C}_T \text{ is a valid molecule,} \\ 0 \text{ else.} \end{cases} \tag{3}$$

[3] See supplementary material for penalization of unrealistic molecules.

In particular, we verify that the molecule is not fragmented (consisting of smaller isolated molecules) and that no atom is charged. This reward term was introduced since we found evidence that atomisation energy alone simply isn't a good predictor for validity when the agent generates molecules containing spatial noise.

For comparison, the original MolGym frameworks used the following reward term only:

- **Per-step formation energy (F)**: In contrast to the terminal rewards above, this reward is 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

$$r_F(s_t, a_t) = (E(\mathcal{C}_t) + E(e_t)) - E(\mathcal{C}_{t+1}), \quad t = 0, ..., T-1. \tag{4}$$

B Details on Single-bag discovery experiment

Baseline 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 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. $\arg\max$), 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 instead use a multi-bag training scheme with QM7 as a reference dataset and report discovery results on the same bags as the compared baselines, but based on molecules sampled stochastically according to the learned agent policy at just a single checkpoint (CP) (see the orange box in Fig. 1). We sample 10,000 molecules for each test formula in Table 1.

C Evaluation metrics

All of the metrics reported in Table 2 and explained below are first calculated *per formula* and are then aggregated using a weighted average, with weight coefficients proportional to the number of isomers (SMILES) in the reference dataset.

Validity (& Uniqueness) Validity is not directly built into the molecular generation procedure used in our framework. Instead we incentivize the agent to create valid molecules based on a simple discrete reward term $r_{\text{valid}} = 1$ if valid, $r_{\text{valid}} = 0$ if invalid, and the validity is straightforwardly defined as

$$\text{validity} = \frac{\#\text{valid molecules}}{\#\text{sampled molecules}}. \tag{5}$$

A word on uniqueness: Notice that the typically reported uniqueness measure

$$\text{uniqueness} = \frac{\#\text{unique molecules}}{\#\text{sampled molecules}}$$

would be a misleading metric to use in our case, since we are generating molecules conditioned on a specific formula and the probability of generating identical molecules is therefore much higher than for unconstrained search. 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 Table 2.

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. Each discovered SMILES will then either be in the reference set or correspond to a novel molecule:

$$N_{\text{unique}} = N_{\text{rediscovered}} + N_{\text{novel}}. \tag{6}$$

The rediscovery and expansion ratios are calcululated straightforwardly as

$$\text{Rediscovery Ratio} = \frac{N_{\text{rediscovered}}}{N_{\text{unique}}^{\text{QM7}}}, \tag{7}$$

$$\text{Expansion Ratio} = \frac{N_{\text{novel}}}{N_{\text{unique}}^{\text{QM7}}}. \tag{8}$$

Energy metric The following metric pertains to the **quality** of the discovered molecules 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. To probe the quality of these relaxed molecules, we calculate the following energy based metric:

- **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)

$$\Delta E_{\text{rRAE}}(\mathcal{C}_T) = E(\mathcal{C}_T) - \bar{E}_{\text{QM7}}^{\mathcal{B}(\mathcal{C}_T)} = E(\mathcal{C}) - \frac{1}{|\mathcal{B}(\mathcal{C}_T)|} \sum_{i=1}^{|\mathcal{B}(\mathcal{C}_T)|} E(\mathcal{C}_i^{\text{QM7}}). \tag{9}$$

  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.