Extended Abstract

Motivation According to the WHO, Acquired Immunodeficiency Syndrome (AIDS) has killed around 44 million people and continues to kill around 600,000 people annually according to 2024 data. In the absence of a vaccine or cure, existing treatments require a regiment of drugs to suppress HIV replication and trying to boost the immune system to decrease impact of HIV's ability to destroy white blood cells. However, the drugs used can have side effects and even drug-to-drug effects that can lead to long term side effects. The goal of our project is to try to develop a generative model for preclinical HIV candidate drug discovery with low likelihood of side effects to provide a possible tool that can be used for determining candidate molecules to test.

Method For our experiment, we utilized an existing generative molecule model as a baseline model. From there we will utilize reinforcement learning techniques to train the model towards our desired behavior of generating molecules that have low likelihood of side effects while also having high likelihood of suppressing HIV replication. We will design the reward function to be inline with our desired behavior of the agent. After training, we will utilize evaluation metrics commonly used in evaluating generative molecular models, such as novelty and internal diversity to determine how much the model is able to generate unseen molecules, but also diverse molecules as well.

Implementation We utilized REINVENT, a RNN-based generative model for generating molecules in the SMILES format. With such a large state space, of roughly over 10^{60} possible feasible molecules, we utilized MCTS to help search for paths and actions that will yield high returns, commonly used for large state spaces like Go. Also, we utilized PPO initially as a stable on-policy training method for stable training. Additionally, we added additional penalties to the reward function to deter reward hacking and mechanism support in rewarding molecules that have the associated substructures that helps suppress HIV replication.

Results From our initial setup, we saw that the trained agent achieved high reward, but when evaluating the model, realized the agent developed a policy in utilizing existing seen molecules, and only applying rings to grow and create a uniquely unseen canidate molecule. Additional penalty terms were applied, as discussed above, to penalize the agent from only adding rings to molecules, and promoted differing scaffold configurations, the core framework of a molecule. Eventually, we saw our Hybrid PPO approach lower total training reward, but saw the hacking penalties drop significantly. The resulting model showed high machanism support of 0.69, and diversity of 0.65, while REINVENT itself showed high diversity of 0.81, it showed no mechanism support of only 0.38 and low reward of 0.59.

Discussion The likelihood of the initial setup failing to develop a model that matched our desired behavior is likely due to reward hacking, by recognizing that since the reward function promotes unseen molecules and valid molecules that show high likelihood of stopping HIV replication, the model recognized the easiest method to get maximum reward is to utilize an existing molecule that has high likelihood of suppressing HIV replication, and add rings to the molecule such that it's considered unique, while maintaining the same properties of the original. After adding the additional penalties and mechanism penalties, the new resulting model showed lower overall reward, but more importantly, showed higher diversity that PPO, and higher mechanism support as well, indicating that while it has lower total reward, the newer models using the updated reward functions and hybrid training approach was able to generate molecules more inline with our desired behavior of diverse HIV candidate drug molecules.

Conclusion By the end of the project, we were able to train a general generative molecule model towards a specialized task of generating HIV candidate drug molecules. There are many additional work that can be done on both the approach, such as utilizing a graph-based approach to exploit the structural properties of molecules, but also more recent reinforcement techniques like GRPO and DAPO, which greatly reduce computational complexity. Since drug discovery can be difficult due to the sheer size of possible molecules to test, the goal of the models used in drug discovery aims to act as a tool for experts to help sift through and narrow down possible candidates molecules for experimenting, with full discretion in the hands of the experts.

Preclinical HIV Drug Candidate Discovery with Reinforcement Learning

Kevin Chen
Department of Computer Science
Stanford University
kc2413@stanford.edu

Arda Dastan
Department of Computer Science
Stanford University
adastan@stanford.edu

Elijah Alexander Schacter
Department of Statistics
Stanford University
elisch@stanford.edu

Abstract

According to the WHO, AIDS has killed around 44 million people and continues to kill around 600,000 people annually according to 2024 data. Although developments in medicine have made HIV manageable, there is still no cure and treatment often has nasty adverse effects, drug-drug interactions and long-term efficacy issues. The goal of the project is to utilize existing general molecule generation models as a baseline model and utilize reinforcement learning techniques to refine and specialize the model towards generating HIV candidate drugs that suppress HIV replication. Using PPO and MCTS, along with additional features like exploration bonuses and duplicate penalties, we were able to train an agent to focus on generating novel molecules that are internally diverse. The signficance of the results indicate strong indication of utilizing reinforcement learning in refining general models towards specific tasks wihtin the field of drug discovery, and the growing importance of machine learning models as tools for experts to help explore the vast space of feasable molecules for treatments for diseases such as HIV.

1 Introduction

Within the world of medicine, technology has played a major role in many areas, and have shown to be a strong tool in medical treatment and drug discovery. With more novel diseases being discovered and spread, paired with speed of which diseases can spread in a globalized world, as demonstrated with COVID-19, speedy and effective preclinical drug discovery is necessary to successfully combat compounding issues of diseases in the modern age. With over 10^{60} feasible molecules, discovering new possible candidate drugs can be a hassle and difficult space to explore, giving rise in utilizing machine learning models to optimize drug discovery by pairing experts with machine.

According to the WHO, Acquired Immunodeficiency Syndrome (AIDS) has killed around 44 million people and continues to kill around 600,000 people annually according to 2024 data. Although developments in medicine have made HIV manageable, there is still no cure and treatment often has nasty adverse effects, drug-drug interactions and long-term efficacy issues. Human Immunodeficiency Virus (HIV), is a virus that attacks the body's white blood cells, a vital component in the immune system's ability to fight of diseases, leading to AIDS, leaving the host vulnerable to life threatening conditions that can be caused by the common cold. Currently, Antiretroviral Therapy (ART) is used to try to combat HIV by utilizing a drug regiment to reduce HIV replication to help the immune

system recover. However, while HIV can be stopped or slowed down with this method, ART can still cause short and long-term side effects.

The goal of our project is to improve early-stage molecular generation for HIV drug discovery using RL. Our project is not trying to directly "cure AIDS" or claim clinical drug discovery. Instead, we are aiming to build an RL-based molecular generator that proposes compounds with high predicted activity against HIV, while penalizing predicted unwanted effects like drug-interaction risk, mutagenicity, cardiotoxicity, and poor solubility. This objective is relevant because the search space of possible molecules is enormous (between 10^{60}-10^{100} feasible molecules). Our goal is to test whether multi-objective RL can produce a better activity–safety tradeoff than simpler baselines like random sampling, supervised fine-tuning, or HIV-only reward optimization.

2 Related Work

Current work on HIV treatment. With HIV being a difficult disease to cure due to the nature of it attacking white blood cells, focus has primarily been split between developing a vaccine and developing treatment to reduce the effects of HIV. Vaccine development has focused on possibly using new mRNA methodology to develop a vaccine Zubair et al. (2025), but until then, ART is commonly used with drugs to suppress HIV replication while trying to boost the immune system to reduce post-disease impacts Volberding and Deeks (2010). With increasing growth in the intersection of technology and medicine, machine learning models have become more integral to research in chemistry, particularly in the area of drug discovery and analysis.

Molecular chemistry and machine learning. MoleculeNet. The intersection of machine learning and molecular chemistry has become an populated field with focus on models to understand and draw inferences from molecules Wu et al. (2018). With the complexity of molecules' behaviors and the large search space of feasable molecules, between 10^{60}-10^{100} feasible molecules, reinforcement learning has become an increasing area of focus to help train models towards specific goals. Common benchmark platforms used to evaluate models focused on molecular properties is MoleculeNet. MoleculetNet provides a platform with not only model evaluation, but a large swath of datasets ranging from electronic properties of molecules, to HIV activity dataset measuring the ability of a molecule to inhibit HIV replication. The HIV dataset will be our main source of data to help train our HIV activity predictor for the reward function, a reference set for training the agent, and a reference set for calculating the novelty of the model.

The most relevant prior work to our project is REINVENT, a baseline recurrent neural network (RNN) model for drug discovery Olivecrona et al. (2017). The core idea is to first train a recurrent neural network on SMILES strings (which are just string representations of molecules) to learn a prior distribution over chemically plausible molecules then use reinforcement learning to fine-tune towards a specific scoring function. This general method gives us a practical template: pretrain a molecular string generator, define a reward based on desired molecular properties, and fine-tune the generator using RL. It also points us towards some really useful open source datasets like ChEMBL that will make our project feasible.

Additionally, in recent years, graph-based approaches have been applied to try to exploit structural features of biomolecular compounds and used for drug discovery. Building off of REINVENT, which provides strong foundational model for molecular compound generation for drug discovery, GraphINVENT utilizes a graph structure to builds molecular compoundsMercado et al. (2021).We focused primarily on REINVENT due to the increase in number of factors to consider if we were to use GraphINVENT, such as node features, edge features, positional embeddings, etc., which won't be feasable given the small timeframe with the main focus on the reinforcement learning portion of the project.

Proximal Policy Optimization (PPO) is a on-policy policy training algorithm that utilizes importance sampling and clipping to provide stable training Schulman et al. (2017). Monte-Carlo Tree Search (MCTS) is a commonly used search algorithm for searching through large state spaces due to it's ability to balance exploration and exploitation in determining paths that shows high promise on return, which is a strong fit for the large state space of feasible molecules to find possible drug candidates Browne et al. (2012). For our project, we aim to start with a simple approach of a stable on-policy algorithm, which PPO provides, and a search algorithm, MCTS, to explore large state spaces to help determine actions of high return towards our goal of generating HIV candidate drugs.

Molecular Sets (MOSES) provides a platform to evaluate model based on multiple evaluation metrics, such as scaffold similarity Polykovskiy et al. (2020). For our project, we will utilize two main evalulation metrics: novelty, internal diversity. Novelty will measure the percentage of generated molecules not found in the reference, with a value within the range of $[0, 1]$. Internal diversity measures how diverse the generated molecules by the model are from one another. Morgan fingerprints is a commonly used bit vector representation of a molecule that combines local substructures that's commonly used in chemistry for similarity measurements between molecules Cereto-Massagué et al. (2015). Tanimoto similarity is used with Morgan finterprints to measure the similarity between two molecules by outputting the ratio of the shared features to total of unique features, generating a value within the range of $[0, 1]$, with 1 indicating identical molecules.

3 Method

3.1 Problem formulation

We frame de novo molecular design as a sequential decision problem over SMILES strings. A molecule m is generated autoregressively as a token sequence $(a_1, \ldots, a_T)$, where the state s_t is the partial SMILES prefix $(a_1, \ldots, a_{t-1})$ and the action a_t selects the next character from a fixed vocabulary. The policy $\pi_\theta(a_t \mid s_t)$ is a recurrent network over this vocabulary. An episode is a complete SMILES string terminated by an end-of-sequence token. Crucially, intermediate steps receive no reward; only the completed, parsed molecule receives a single terminal reward $R(m)$ (Section 3.3). Because the token vocabulary is small and characters are emitted strictly left to right, the partial-prefix representation is a sufficient state for both the policy and the value model.

The optimization target is multi-objective: we seek molecules with high predicted HIV activity while simultaneously penalizing predicted toxicity and poor ADMET behavior, structural liabilities, distributional implausibility, and the statistical signatures of reward hacking. At the same time, we reward proximity to known active mechanism classes and enforce scaffold diversity so that the final candidate set does not collapse onto a single high-scoring motif.

3.2 Generative backbone

Following REINVENT (Olivecrona et al., 2017), we use a recurrent prior over SMILES as the generative backbone. The prior RNN is trained to model a distribution over chemically plausible molecules; reinforcement learning then fine-tunes a copy of this network (the agent) toward our reward function. We elected to operate directly on SMILES rather than on graphs as in GraphINVENT (Mercado et al., 2021), because the graph formulation introduces additional node-, edge-, and positional-embedding machinery that was not feasible to tune within the project timeframe; SMILES let us concentrate effort on the reinforcement learning component itself.

3.3 Reward function

For any syntactically valid molecule m, the terminal reward is the composite

$$\begin{aligned} R_{\text{valid}}(m) = \alpha \min\{\hat{p}_{\text{HIV}}(m),\, c_{\text{HIV}}\} - \beta P_{\text{ADMET}}(m) - \delta P_{\text{RDKit}}(m) \\ - \eta_{\text{OOD}} P_{\text{OOD}}(m) + \omega_{\text{mech}} G_{\text{mech}}(m) - \omega_{\text{proxy}} P_{\text{proxy}}(m) \\ - \rho\, \mathbf{1}\{m \in \mathcal{M}_{\text{seen}}\} - \tau\, \mathbf{1}\{N_B(s(m)) > C_{\text{scaf}}\}. \end{aligned} \tag{1}$$

Invalid SMILES (strings that RDKit cannot parse) receive a fixed negative reward. We describe each term in turn.

HIV activity. $\hat{p}_{\text{HIV}}(m)$ is a validation-rank-normalized random forest score for HIV inhibition trained on the MoleculeNet HIV dataset (Wu et al., 2018). Rank normalization against held-out validation predictions is necessary because the dataset is extremely class-imbalanced, which otherwise renders the raw probability scale uninformative. We cap the activity term at c_{HIV} so that the agent cannot drive total reward arbitrarily high by pushing a single proxy score past the range supported by the training data.

ADMET penalty. $P_{\mathrm{ADMET}}(m)$ aggregates a panel of predicted absorption, distribution, metabolism, excretion, and toxicity endpoints into a single penalty, with a hard escalation when the most dangerous endpoints become likely:

$$P_{\mathrm{ADMET}}(m) = P^{\mathrm{avg}}_{\mathrm{ADMET}}(m) + \gamma_{\mathrm{deadly}}\, I_{\mathrm{deadly}}(m), \tag{2}$$

$$P^{\mathrm{avg}}_{\mathrm{ADMET}}(m) = \hat{p}_{\mathrm{hERG}}(m) + \hat{p}_{\mathrm{AMES}}(m) + 0.5\,\hat{p}_{\mathrm{Caco2}}(m) + 0.5\,\hat{p}_{\mathrm{HIA}}(m) + 0.5\,\hat{p}_{\mathrm{solubility}}(m), \tag{3}$$

$$I_{\mathrm{deadly}}(m) = \mathbf{1}\big\{\max(\hat{p}_{\mathrm{hERG}}(m),\, \hat{p}_{\mathrm{AMES}}(m)) > 0.9\big\}. \tag{4}$$

The two highest-stakes endpoints (hERG cardiotoxicity and AMES mutagenicity) are weighted fully, while the remaining permeability and solubility heads are down-weighted by a factor of 0.5, reflecting that they describe developability rather than acute liability. The indicator I_{deadly} applies a large fixed penalty $\gamma_{\mathrm{deadly}} = 10$ whenever either fatal-class endpoint exceeds probability 0.9, effectively excluding such molecules from the elite set regardless of their predicted activity.

Structural-alert penalty. $P_{\mathrm{RDKit}}(m)$ is a deterministic structural-alert penalty computed with RDKit as the average of the two recommended filter families, PAINS and BRENK, clipped to $[0, 1]$:

$$P_{\mathrm{RDKit}}(m) = \min\Big\{1,\ \tfrac{1}{2}\big(\mathbf{1}_{\mathrm{PAINS}}(m) + \mathbf{1}_{\mathrm{BRENK}}(m)\big)\Big\}. \tag{5}$$

This term is independent of any learned model and therefore cannot itself be hacked, providing a stable floor on chemical reasonableness.

Out-of-distribution penalty. $P_{\mathrm{OOD}}(m)$ penalizes molecules that drift away from the support of the MoleculeNet training set. Let $f(\cdot)$ denote the Morgan fingerprint and $s_{\max}(m) = \max_{x \in \mathcal{D}_{\mathrm{train}}} \mathrm{Tanimoto}(f(m), f(x))$ be the maximum Tanimoto similarity to any training molecule. Then

$$P_{\mathrm{OOD}}(m) = \frac{[\, s_0 - s_{\max}(m)\,]_+}{s_0}, \qquad s_0 = 0.40. \tag{6}$$

This is a deliberately loose penalty: it activates only once a molecule's nearest training neighbor falls below similarity s_0, and it was introduced chiefly to discourage the pathologically large, low-similarity molecules observed in early runs.

Mechanism reward. $G_{\mathrm{mech}}(m)$ rewards similarity to known HIV-active compounds grouped by inhibitory mechanism. For each mechanism we assemble a reference set of ChEMBL molecules deemed active at $\mathrm{pChEMBL} \geq 6$, following the threshold convention of Sturm et al. (2020). Let $S_{\mathrm{mech}}(m)$ be the maximum Tanimoto similarity to any reference molecule across the mechanism sets; the reward is a soft gate

$$G_{\mathrm{mech}}(m) = \sigma\left(\frac{S_{\mathrm{mech}}(m) - 0.45}{0.05}\right). \tag{7}$$

This provides an orthogonal, structure-grounded signal for activity that is much harder to game than the single learned classifier, since it requires genuine resemblance to molecules with established mechanism support.

Proxy anti-hacking penalty. $P_{\mathrm{proxy}}(m)$ explicitly penalizes the failure mode of high predicted activity coupled with no mechanism support:

$$P_{\mathrm{proxy}}(m) = \sigma\left(\frac{\hat{p}_{\mathrm{HIV}}(m) - 0.85}{0.05}\right) \cdot \sigma\left(\frac{0.45 - S_{\mathrm{mech}}(m)}{0.05}\right). \tag{8}$$

The first factor is near 1 only when the activity score is very high; the second is near 1 only when mechanism similarity is very low. (The second factor equals $1 - G_{\mathrm{mech}}(m)$, so the penalty can equivalently be read as "confident activity with low mechanism reward.") Together the mechanism reward and the proxy penalty steer the agent away from regions where $\hat{p}_{\mathrm{HIV}}$ is inflated but no orthogonal evidence agrees.

Memory and diversity penalties. The final two terms enforce exploration and diversity. The indicator $\mathbf{1}\{m \in \mathcal{M}_{\text{seen}}\}$ penalizes molecules already generated in any previous run, discouraging the agent from over-indexing on a small set of repeatedly rediscovered hits. The scaffold indicator $\mathbf{1}\{N_B(s(m)) > C_{\text{scaf}}\}$ triggers when more than C_{scaf} molecules sharing the same Bemis–Murcko scaffold appear in the current batch, following the diversity mechanism of REINVENT (Olivecrona et al., 2017).

3.4 Policy optimization

From sequence-level to per-token credit assignment. A vanilla policy-gradient variant assigns the single terminal reward $R(m)$ to the entire generated sequence. We ran this variant only as an early baseline to establish whether per-token credit assignment was worthwhile; it is not part of the final method and we therefore do not develop it in detail here, reporting it solely as a point of comparison in Section 4. The final method instead uses a per-timestep actor-critic update. For each action a_t we compute an advantage

$$A_t = G_t - V_\phi(s_t), \qquad G_t = \sum_{k=t}^{T} \gamma^{\,k-t} r_k. \tag{9}$$

Because intermediate rewards are zero and only the completed molecule is scored, $G_t = \gamma^{\,T-t} R(m)$. With $\gamma = 1$, every action in a trajectory receives the same raw return $R(m)$, yet the advantages differ across tokens because the critic $V_\phi(s_t)$ depends on the partial SMILES prefix. The update can therefore credit or penalize individual token decisions relative to what the prefix predicted, which is the key benefit of the actor-critic formulation over sequence-level reward assignment.

PPO objective. The policy is updated with a clipped PPO objective (Schulman et al., 2017),

$$\mathcal{L}_{\text{PPO}}(\theta) = -\mathbb{E}_t\big[\min\big(r_t(\theta)A_t,\ \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)A_t\big)\big], \qquad r_t(\theta) = \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)}, \tag{10}$$

with clipping range $\epsilon = 0.20$.

Value model. The critic $V_\phi(s_t)$ is a GRU followed by an MLP layer that emits a scalar value estimate. The GRU consumes the learned embedding of each SMILES character in sequence; this is a natural fit because tokens are added one at a time and the character vocabulary is small. The value model is trained by regression to the observed returns,

$$\mathcal{L}_V(\phi) = \mathbb{E}_t\Big[\big(V_\phi(s_t) - G_t\big)^2\Big], \tag{11}$$

and is updated every iteration alongside the policy.

REINVENT elite-replay term. To keep the agent anchored to the chemically plausible prior, we add the REINVENT replay loss over an elite buffer $\mathcal{B}$ of high-reward molecules. Writing the negative log-likelihood of a sequence under a model as $\text{NLL}(m) = -\sum_{t=1}^{T} \log \pi(a_t \mid s_t)$, the augmented prior target and replay loss are

$$\text{NLL}^*(m) = \text{NLL}_{\text{prior}}(m) - \zeta R(m), \quad \zeta = 20; \qquad \mathcal{L}_{\text{replay}}(\theta) = \mathbb{E}_{m\sim\mathcal{B}}\Big[\big(\text{NLL}_\theta(m) - \text{NLL}^*(m)\big)^2\Big]. \tag{12}$$

Intuitively, this is an L_2 penalty that pulls the agent's likelihood of elite molecules toward the prior's likelihood, offset by the reward: a higher $R(m)$ permits the agent to drift further from the prior for that molecule. Reward thus appears twice in the overall objective; we retain this to honor the REINVENT formulation exactly, and because the replay term is applied only to the elite buffer it concentrates the prior anchoring on the molecules we ultimately care about selecting.

Composite objectives and MCTS variant. We compare five training configurations that differ in how these losses are combined. The two strongest combine the PPO with the elite replay (REINVENT-style) loss:

$$\mathcal{L}_{\text{hybrid}} = \mathcal{L}_{\text{PPO}} + 0.25\,\mathcal{L}_{\text{replay}} \tag{13}$$

$$\mathcal{L}_{\text{kitchen}} = \mathcal{L}_{\text{PPO}} + 0.5\,\mathcal{L}_{\text{replay}} \tag{14}$$

The "Kitchen Sink" configuration additionally drives exploration with Monte-Carlo Tree Search (Browne et al., 2012), which is well suited to the enormous SMILES state space because it balances exploration against exploitation when expanding promising partial molecules. The two configurations differ in only two respects: Kitchen Sink uses a Monte Carlo-estimated value target and a heavier replay weight (0.50 vs. 0.25), whereas the hybrid uses a single sampled terminal reward. This Monte Carlo target is super simple multi-rollout (no complex sampling). We considered more complex MCTS methods but deemed them not worth it given the results of the simple rollout method.

4 Experimental Setup

4.1 Data

The MoleculeNet HIV dataset (Wu et al., 2018) serves three roles: it trains the HIV-activity reward model, it provides the reference set $\mathcal{D}_{\text{train}}$ for the out-of-distribution penalty, and it is the reference distribution against which novelty is measured. The dataset is highly imbalanced (only $\approx 3.7\%$ active), which motivates the rank normalization of $\hat{p}_{\text{HIV}}$. For the mechanism reward we draw active compounds ($\text{pChEMBL} \geq 6$) from ChEMBL, grouped by HIV inhibitory mechanism (reverse transcriptase, integrase, protease, and CCR5), selecting mechanisms with sufficient dataset support to be used reliably.

4.2 Reward and evaluation models

Training-time activity model. The reward signal during RL training uses a random forest HIV classifier, chosen for fast, stable scoring across the many thousands of molecules generated per iteration.

Held-out reranking model. To control for model-selection bias, the final candidate set is reranked with a GraphGPS graph neural network held out from training, which achieves test ROC-AUC 0.802. When reranking, we replace the random-forest score with the rank-normalized ensemble

$$\hat{p}^{\text{raw}}_{\text{HIV}}(m) = \tfrac{1}{2}\big(\hat{p}_{\text{RF}}(m) + \hat{p}_{\text{GNN}}(m)\big), \tag{15}$$

where both components are rank-normalized against their respective validation predictions before averaging—normalization is essential here since the random forest and GNN are not guaranteed to share a probability scale.

ADMET and structural models. The ADMET penalty draws on five predicted endpoints (hERG, AMES, Caco-2, HIA, and aqueous solubility). The RDKit structural-alert penalty uses the PAINS and BRENK filter catalogs and requires no training.

4.3 Training configuration

All agents are fine-tuned for 300 iterations from the same pretrained REINVENT prior. We compare five strategies of increasing complexity: (1) REINVENT elite replay, (2) a value-baseline policy gradient, (3) PPO, (4) the PPO + REINVENT hybrid, and (5) the Kitchen Sink (MCTS PPO + REINVENT). Reward-function hyperparameters, are listed in Table 1.

Table 1: Reward-function and optimization hyperparameters.

Symbol	Value	Symbol	Value
α	1.0	ω_{mech}	0.10
c_{HIV}	0.85	ω_{proxy}	0.30
β	0.15	ρ	0.20
γ_{deadly}	10	τ	0.20
δ	0.40	C_{scaf}	2
η_{OOD}	0.25	ϵ (PPO clip)	0.20
s_0	0.40	ζ	20

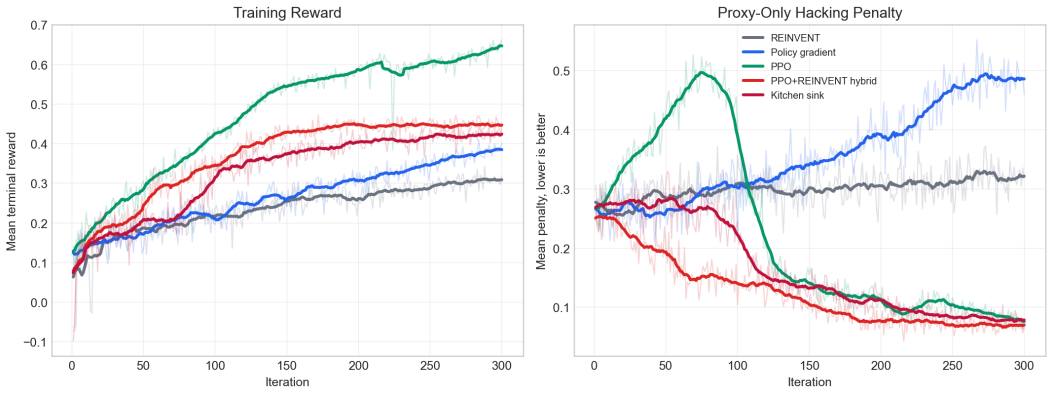

Figure 1: Training Curves

4.4 Candidate selection

We select the final top-100 molecules with a procedure designed to balance quality against scaffold diversity. From the trained agent we sample $N = 5000$ molecules, keep only valid SMILES, and deduplicate. Each molecule is scored with the reward function *without* the batch scaffold penalty. We then group molecules by Bemis–Murcko scaffold and sort within each group by reward: the top two molecules per scaffold are kept penalty-free, and the duplicate-scaffold penalty is applied to the remainder. Finally we re-sort globally by adjusted reward and take the top 100. All reranking in this stage uses the RF/GNN ensemble $\hat{p}_{\mathrm{HIV}}^{\mathrm{raw}}$ in place of the training-time random-forest score.

4.5 Evaluation metrics

We evaluate the top-100 set along four axes: (i) *mean reward*, the average composite reward; (ii) *mean mechanism support*, the average maximum mechanism similarity G_{mech}, indicating whether selected molecules resemble known actives; (iii) *mean proxy penalty* (lower is better), which flags drift into high-$\hat{p}_{\mathrm{HIV}}$ / low-G_{mech} regions and so serves as a direct measure of reward hacking; and (iv) *internal diversity*, computed as 1 minus the mean pairwise Tanimoto similarity over the final 100 molecules. We also report novelty (the fraction of generated molecules absent from the reference set), following the MOSES benchmark conventions (Polykovskiy et al., 2020). However, in all experiments the fraction of novel molecules is 100%; as such, we will not discuess it further. We adopt this multi-metric evaluation deliberately: because drug-discovery rewards score fundamentally unknown molecules, they are unusually susceptible to reward hacking, and a single activity number is insufficient evidence that a candidate is genuinely promising. Reporting mechanism support, proxy penalty, and diversity alongside reward lets us distinguish honest improvement from proxy exploitation and guards against collapse onto a narrow region of chemical space.

5 Results

5.1 Results Proper

The chronological order/evolution of experiments can be seen in the figures above moving left to right. The first method, REINVENT, is lifted directly from the main reference for this paper, Olivecrona et al. (2017). We did not expect this method to perform very well because the paper indicated that it often needed to be supplemented with other methodologies in specific tasks. However, we wanted to see how well it could perform in isolation to use as a benchmark as we increased model complexity.

Our second approach was to use vanilla policy gradient method. As can be seen in the figures (Figure 2), this simple policy gradient method with terminal rewards and no discounting did slightly better in terms of reward, mechanism activity, and proxy penalty, but sacrificed some amount of diversity.

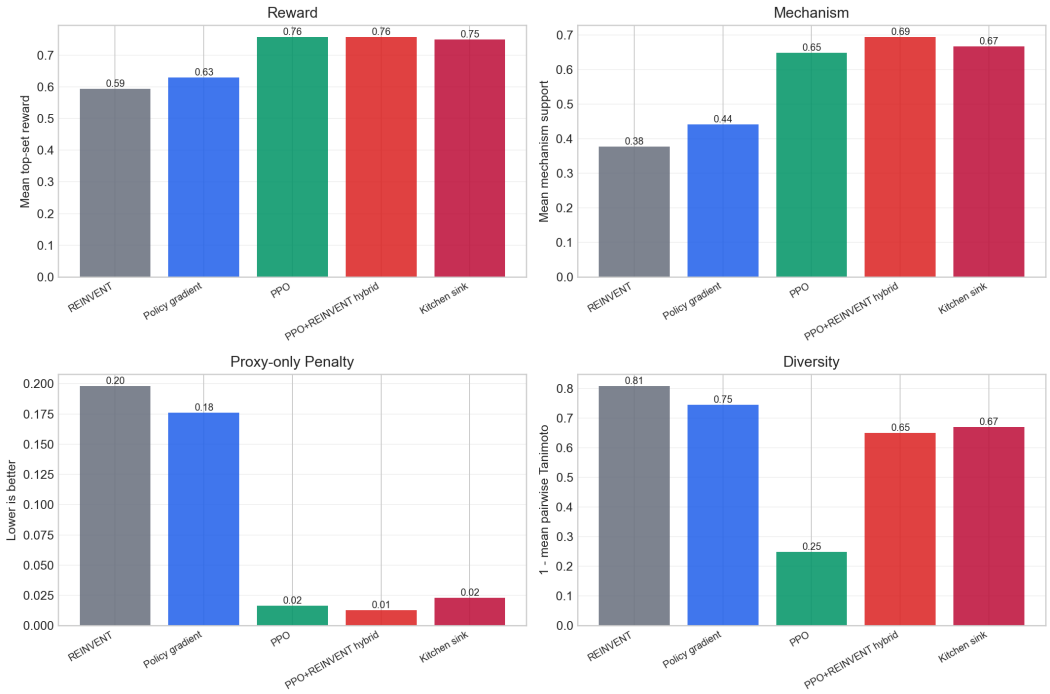

Figure 2: Top 100 Metrics

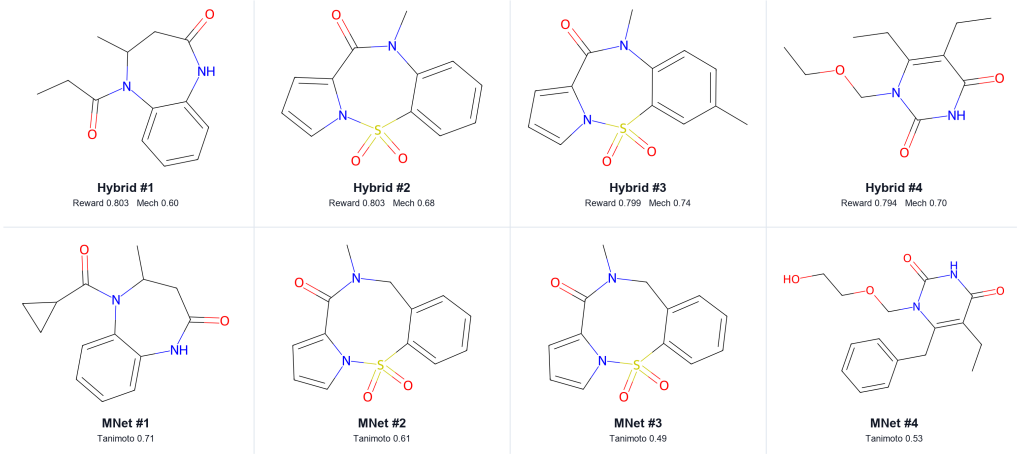

Figure 3: Top Molecules

Seeing positive results we moved onto using PPO as the next attempt at better performance. In addition to the usual change with clipping and probability weighting you would expect in a PPO upgrade, we moved away from the sequence-level policy-gradient update that assigned the same reward to the entire generated SMILES sequence. As shown in the method section, in our PPO implementation we compute an advantage with the reward-to-go instead. This change brought about a serious increase in performance across reward, mechanism, and proxy metrics, but really hurt us in terms of diversity. From here we were fairly sure that PPO was going in the right direction, but needed something to augment diversity. Creating a hybrid model with REINVENT seemed like the logical choice here because REINVENT loss would help to anchor the model closer to the RNN prior, thus preventing over-concentration around one family of molecules.

Our last two models were both based on this hybrid method. Both achieved great performance: strong reward, mechanism and proxy penalty with reasonable diversity. Between the two models, the only difference is that in the "kitchen sink" implementation we used MCTS to hopefully capture a better, more stable, value function in training. Sadly, this did not generate better results and its considerably slower training time is a serious drawback. As is, the best method we uncovered was the basic hybrid method; its struck a strong balance between reward and diversity without strong evidence of egregious hacking.

Based on the results, we see that PPO achieved the highest reward of 0.76, with the hybrid models achieving the same reward of 0.76 and 0.75. However, we see that PPO performed significantly worse in terms of 0.25 diversity, with hybrid approaches achieving a higher diversity of 0.65 and 0.67, respectively (Figure 2). Also, we see that even though REINVENT itself achieved high diversity of 0.81, it fails to achieve high mechanism support, which the hybrid models do achieve with 0.69 and 0.67. This indicates that the hybrid models are able to generate molecules that have the proper substructures associated with HIV replication suppression. Across both reward, mechanism support, proxy-only penalties, and diversity, the hybrid models are able to achieve high reward, high mechanism support, low proxy-only penalties, and modest diversity, showing how the hybrid approaches achieve an overall better balance of our desired behavior relative to REINVENT and PPO models.

5.2 Figure 1 Anomaly

While not central to our experiment, one odd result appeared in the Proxy Penalty training curve. The PPO method saw a massive increase, followed by a massive decrease in the hacking penalty in the first 100 training iterations. It is hard to explain exactly why this occurred, but it looks like the PPO method tended towards high $\hat{p}_{\mathrm{HIV}}$ values regardless of active mechanism success in the first 50 iterations and then suddenly started optimizing towards higher mechanism areas afterword. It is certainly an interesting result that should be analyzed more in a different context.

5.3 Discussion

In our research, the number one problem that we kept circling back to is how to prevent reward hacking. In drug discovery, we need to try to find a way to measure the value of molecules that are out of distribution by definition; if they were in distribution they wouldn't be a discovery. The issue with evaluating molecules that are out of distribution is that it is very difficult to trust the model-based scores. For example, in our first few attempts, the top molecules were achieving $\hat{p}_{\mathrm{HIV}}$ scores of $.98 \sim 1$. Meanwhile the average $\hat{p}_{\mathrm{HIV}}$ scores of active molecules in the validation set was only around $.85$. While simply accepting these results would make our reward function look nice, it would be unreasonable to assume that these activity numbers were real. As such, we adjusted our approach to be solely targeted towards combating hacking.

We added in the orthogonal mechanism heads to act as a second criteria rather than simply a high $\hat{p}_{\mathrm{HIV}}$ score and tacked on the proxy penalty to further penalize purely $\hat{p}_{\mathrm{HIV}}$-maximizing behavior. The out-of-distribution penalty also helped significantly in preventing extremely unrealistic-looking molecules from making there way into the top 100. In addition, the REINVENT elite replay played a key role at increasing diversity. This diversity not only gave us the desirable result of having a more diverse discovery space, but added yet another layer to combat hacking towards one specific

molecule-type that happened to score very high out of distribution.

In the end, the hybrid PPO + REINVENT model got us to a point where we could reasonably say that these candidate molecules could be possible at combating HIV. However, due to the out of distribution nature of the task, there is no way to know for sure without expert filtering and then experimentation. We believe that we accomplished the main goal we set out in the introduction; our model successfully narrowed the search space down to a reasonable top 100 that a researcher in the field could use to heavily simplify the task of drug exploration.

6 Conclusion

The project shows promising results in utilizing reinforcement learning techniques and algorithms to specialize a general generative molecular model towards specific tasks, such as preclinical HIV candidate drug discovery. Utilizing existing research in the area with REINVENT, we were able to combine with multiple RL techniques and algorithms, such as PPO, replay buffers, and design a reward function to allow the agent to learn a policy that better fits our desired behavior instead of reward hacking. We were able to iterate over our initial design to help better avoid reward hacking with additional terms for the reward function and experimenting with different hyperparameter values to reach our desired behavior. All in all, our results show high novelty and high diversity among the trained model to generate HIV candidate drugs, but can be taken further with newer techniques and knowledge from experts in the field.

6.1 Future Work

While we utilized PPO for our main method of policy training, more recent techniques and methods that are more computationally effective have come out. Group Relative Policy Optimization (GRPO), unlike PPO, avoids training a value function, but instead utilizes group average and standard deviation to normalize group returns, helping reduce computational complexity while being able to determine positive and negative returns within each group Shao et al. (2024). Recently, DeepSeek has built on top of GRPO and created Decoupled Clip and Dynamic Sampling Policy Optimization (DAPO)Yu et al. (2026). Unlike GRPO, DAPO removes the KL divergance term, and increases the clipping ceiling to promote increased diversity in exploring samples and paths that return positive returns and only sampling data points that have a non-zero and non-one accuracy to promote effective gradients.

Furthermore, as discussed extensively, a key factor of the complexity of the project was designing a proper reward function to train our agent towards our desired behavior, but often encountered reward hacking that focused on optimizing total reward resulting in unsuccessful behavior for the task. One method that can help both simplify the reward function and avoid reward hacking is utilizing Reinforcement Learning from Human Feedback (RLHF) to utilize preference pairs instead to help train a reward model to align the rewards with what we expect Ouyang et al. (2022). Also, due to the nature of the project, utilizing RLHF can require additional time in labeling and requiring expert opinions in the field.

Additionally, as discussed, GraphINVENT takes a different approach to the problem with utilizing graph representations instead of SMILES representation. Attempting the project with GraphINVENT can be interesting by leveraging node and edge level features along with exploiting substructures within the graph representation of molecules.

7 Team Contributions

- **Kevin Chen:** Worked on HIV activity predictor. Worked on poster and poster presentation. Worked on RL finetuning. Worked on final writeup. Worked on Extended Abstract, Abstract, Introduction, Related Works.
- **Arda Dastan:** Worked on ADMET and RDKit predictors. Worked on poster. Worked on RL finetuning. Worked on final writeup. Worked on Methods and Experiment Setup.
- **Elijah Alexander Schacter:** Worked on HIV activity predictor. Worked on RL models. Worked on poster an poster presentation. Worked on final writeup. Worked on Results and Discussion.

8 AI Disclosure

Codex and Claude Code were used to pull datasets, pull comparable HIV drugs on the market, parse the RDKit library, build skeleton models for the heads in the reward function and value function, implement modal infrastructure including configuration (hyperparameter) sweeps, write READMEs and other md files to keep track of results and make team communication easier, and create graphs and images. The essential core RL algorithms and update rules were implemented by the team in RL/core.

Changes from Proposal We used REINVENT instead of GraphINVENT to focus more time on reinforcement learning section of our approach instead of graph-related work that would be necessary if we used GraphINVENT instead. Additionally, we had to modify our reward function as time went on to better align the trained agent with our desired behavior to avoid reward hacking.

References

Cameron B Browne, Edward Powley, Daniel Whitehouse, Simon M Lucas, Peter I Cowling, Philipp Rohlfshagen, Stephen Tavener, Diego Perez, Spyridon Samothrakis, and Simon Colton. 2012. A survey of monte carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in games* 4, 1 (2012), 1–43.

Adrià Cereto-Massagué, María José Ojeda, Cristina Valls, Miquel Mulero, Santiago Garcia-Vallvé, and Gerard Pujadas. 2015. Molecular fingerprint similarity search in virtual screening. *Methods* 71 (2015), 58–63.

Rocío Mercado, Tobias Rastemo, Edvard Lindelöf, Günter Klambauer, Ola Engkvist, Hongming Chen, and Esben Jannik Bjerrum. 2021. Graph networks for molecular design. *Machine Learning: Science and Technology* 2, 2 (2021), 025023.

Marcus Olivecrona, Thomas Blaschke, Ola Engkvist, and Hongming Chen. 2017. Molecular de-novo design through deep reinforcement learning. *Journal of cheminformatics* 9, 1 (2017), 48.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. arXiv:2203.02155 [cs.CL]

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

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300* (2024).

Noé Sturm, Andreas Mayr, Thanh Le Van, Vladimir Chupakhin, Hugo Ceulemans, Joerg Wegner, Jose-Felipe Golib-Dzib, Nina Jeliazkova, Yves Vandriessche, Stanislav Böhm, Vojtech Cima, Jan Martinovic, Nigel Greene, Tom Vander Aa, Thomas J Ashby, Sepp Hochreiter, Ola Engkvist, Günter Klambauer, and Hongming Chen. 2020. Industry-scale application and evaluation of deep learning for drug target prediction. *Journal of Cheminformatics* 12, 1 (2020), 26.

Paul A Volberding and Steven G Deeks. 2010. Antiretroviral therapy and management of HIV infection. *The Lancet* 376, 9734 (2010), 49–62.

Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. 2018. MoleculeNet: a benchmark for molecular machine learning. *Chemical science* 9, 2 (2018), 513–530.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. 2026. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems* 38 (2026), 113222–113244.

Akmal Zubair, Hanbal Ahmad, Muhammad Muaz Arif, and Muhammad Ali. 2025. mRNA vaccines against HIV: Hopes and challenges. *HIV medicine* 26, 6 (2025), 824–838.