EVOLUTIONARY POLICY GRADIENT BASED OPTIMIZATION FOR SMALL MOLECULE DRUG DISCOVERY

Tehemton Khairabadi & Vishal Pagidipally
Applied Research - Life Sciences
Quantiphi
Mumbai, India & Toronto, Canada
{tehemton.khairabadi,vishal.pagidipally}@quantiphi.com

ABSTRACT

Generative AI offers transformative potential for small-molecule drug discovery, enabling faster and more targeted identification of novel therapeutics. Lead discovery and optimization remain pivotal yet challenging, particularly in designing compounds with precise pharmacological profiles and target interactions. Existing gradient-based and gradient-free methods struggle to meet these demands. We introduce EPOSMol, an evolutionary policy gradient framework for lead optimization that refines molecular structures in latent space. Our approach iteratively samples structures, using oracle tools to evaluate fitness based on desired properties, binding affinity, and target interactions. A flexible reward framework enables adaptive policy updates and seamless integration of ranking tools, while dynamic scheduling of population size and exploration parameters optimally balances global search and local refinement. EPOSMol achieves up to 10× improvement over state-of-the-art techniques in generating target-specific hits and exploring high-potential chemical spaces. This work advances AI-driven drug discovery by combining evolutionary algorithms with generative modeling to enhance lead optimization.

1 INTRODUCTION

Drug discovery (DD) seeks to identify and optimize candidate molecules for efficacy, safety, and bioavailability, a process spanning 10–12 years and costing billions, with a success rate of only 0.01–0.001%. Early DD methods include Structure-Based Drug Design (SBDD), ligand-based approaches, and High-Throughput Screening (HTS). SBDD utilizes target protein structures, ligand-based methods find similar actives, and HTS screens large libraries but is costly and computationally intensive. De novo design (Olivecrona M, 2017), (Blaschke T, 2020) generates molecules from scratch but struggles with biological activity and toxicity prediction. Given a chemical space of 10^{23} to 10^{60} of drug-like compounds (Polishchuk, 2013), optimizing compounds remains challenging despite AI advancements.

Generative chemistry models enhance molecular design by refining affinity, toxicity, and pharmacokinetics, using representations like SMILES (Weininger, 1988), SELFIES (Mario Krenn & Aspuru-Guzik, 2019), Fingerprints (Morgan, 1965), or Graphs (Franco Scarselli & Monfardini, 2009). Key frameworks include Variational Autoencoders (VAEs) (Hanjun Dai & Song, 2018; Peter Eckmann & Yu, 2022; Ochiai et al., 2023), Graph Neural Networks (GNNs) (Jiaxuan You & Leskovec, 2018; Chence Shi, 2020; Zang & Wang, 2020), Large Language Models (LLMs) (Ross J, 2022; He J, 2022), and Generative Adversarial Networks (GANs) (Jabbar R, 2022). VAEs support molecular variation but may compromise chemical validity. GNNs and flow-based models excel for molecular graphs but have large latent spaces, limiting optimization. LLMs handle SMILES-based synthesis yet lack detailed 3D structure validity. GANs generate property-focused molecules but require careful training to maintain stability.

Optimization in generative chemistry fine-tunes molecular properties through strategies tailored to specific challenges. Gradient-based optimization in VAEs allows efficient tuning but may suffer in

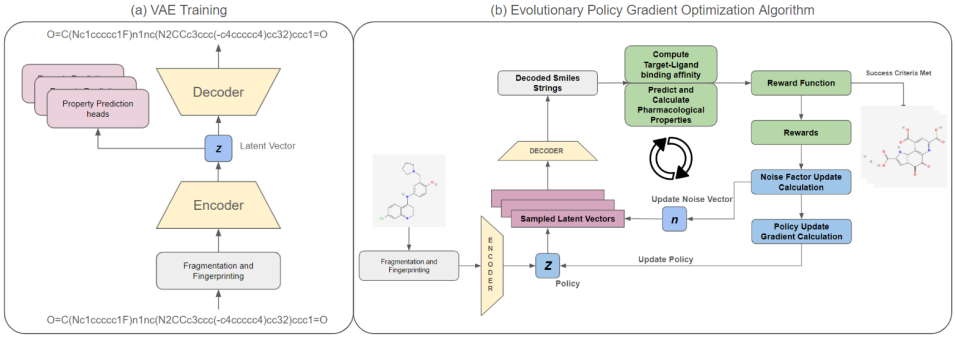

Figure 1: **EPOSMol architecture: (a)** The VAE is trained to reproduce the input compound with a high degree of molecular structural 3d accuracy. The latent space construction of the VAE is further guided by incorporating property losses for QED, SA, LogP and Natural-product likeness scores. **(b)** Our Evolutionary Policy Gradient Algorithm defines a starting policy and iteratively samples a population around the policy to derive rank weighted rewards. This reward is then used to calculate the gradient and guide the policy and noise parameter updates.

unconstrained spaces. For external properties like binding affinity, gradient-free methods such as Covariance Matrix Adaptation Evolution Strategy (CMA-ES) (Hansen et al., 2003), Genetic Algorithms (Annu Lambora & Chopra, 2019), Differential Evolution (Rainer Storn, 1997), Bayesian Optimization (Snoek et al., 2012), and swarm optimization are preferred. These methods navigate complex fitness landscapes but struggle with high dimensions and slow convergence. Bayesian optimization (Snoek et al., 2012) is effective for multi-objective tasks but computationally demanding, while Evolutionary Algorithms require extensive generations. Reinforcement Learning (RL) (Zhenpeng Zhou & Riley, 2019; Loeffler et al., 2024) enforces property-specific constraints but demands significant tuning.

Our approach integrates an evolutionary policy gradient method (Sehnke et al., 2010) with dynamic schedulers for population size and noise factor, optimizing molecular search within latent space for properties like binding affinity. Unlike traditional gradient-free methods that struggle with high dimensions, our approach efficiently navigates latent space using gradient information, ensuring comprehensive sampling of drug-like compounds. By incorporating binding affinity into the reward function and leveraging docking tools, we enhance structure-based optimization, balancing local and global search.

Combining evolutionary population-based sampling (Sehnke et al., 2010) with reward function (Loeffler et al., 2024), our method defines the fitness landscape while adjusting population size and noise factor to balance exploration and exploitation. This prevents convergence to local optima, steadily guiding optimization toward a global optimum while considering both short- and long-term molecular design objectives.

2 METHODOLOGY

Our work, EPOSMol, is an evolutionary policy gradient-based optimization algorithm for small molecule lead discovery and optimization. We enhance the vanilla policy gradient method for faster convergence by parameterizing sampling noise and population size, along with ablation studies on scheduler-based approaches. Figure 1b illustrates the EPOSMol workflow. Additionally, we incorporate strategies like random restarts, structural diversity penalties, and parallel initialization of multiple populations with varied initial conditions to improve robustness and search diversity.

Evolutionary Policy Gradient combines evolutionary algorithms with policy gradient methods, balancing reward-driven updates with population-based search. Each iteration evaluates multiple candidate policies, assigning rewards and updating parameters based on a weighted average favoring higher-performing solutions. This approach enhances exploration while refining high-reward regions. In each episode, N solutions are sampled, rewarded, ranked, and used to update the policy via the ADAM optimizer. Algorithm 1 provides the pseudo-code for our EPOSMol method. The policy

update is governed using:

$$\mu_{t+1} = \mu_t + \alpha \frac{1}{N} \sum_{i=1}^{N} r_i z_i \tag{1}$$

Where $\mu_{t+1}$is the updated policy, μ_t is the current policy, α is the learning rate for policy update, N is the population size, r_i are the rewards from the i-th sampled solution, $r_i \in R$, z_i is the sampled latent variable representing the explored solutions.

This update mechanism allows for adaptive learning of parameters based on their performance, promoting both exploration and exploitation within the search space.

Sampling the solutions in every episode for a given policy is done using:

$$\mu^{'} = \mu + S(0, \sigma^2) \tag{2}$$

Where $\mu^{'}$ is the sampled solution latent vector, μ is the current policy, S(0, σ^2) is the noise factor parameterized as sigma vector the same size as the policy.

Algorithm 1 Evolutionary Policy Gradient for Optimization

Require: VAE encoder f_{enc}, VAE decoder f_{dec}, Reward function ***R***, Noise factor σ, Population size N

Initialize policy μ = f_{enc}(initial SMILES)
while success criteria not met **do**
  Sample $\mu^{'}$ ⟵ Sample around the policy based on the noise factor σ and population size N
  f_{dec} (u$^{'}$) ⟵ Decode the sampled latent vectors into **Solutions**
  R ⟵ Reward each entity of the sampled population
  Normalize and ***Rank*** the rewards
  calculate the ***Gradient***
  μ_{t+1} ⟵ Update the ***Policy***
  σ_{t+1} ⟵ Update the ***Standard Deviation*** sampling noise vector
end while

3 RESULTS

In Table 1, the REINVENT 4.0 benchmark shows our approach achieved a hit rate of 18.125% (1,160 hits out of 6,400 unique compounds), significantly outperforming the vanilla (1.85%) and transfer learning methods (3.46%) in REINVENT, with 9.7x and 5.5x increases, respectively. Our method also generated a broader variety of unique scaffolds and identified compounds with higher predicted binding affinities, while also producing more compounds with core structures similar to bioassay-tested compounds, demonstrating its robustness in generating promising candidates.

Hit Rate Improvements: The results from our unrestricted, extended optimization run were highly encouraging. As shown in Figure 2, the optimization process consistently improved the number of

Table 1: Comparison of EPOSMol to various other methods to generate valid hits for the PDK-1 target (PDB ID 2XCH). The average top-10 binding affinity scores are reported Valid hit defined as BA <= -8.0 & QED >=0.7

Method	Valid Hits	Vina top 10% (kcal/mol)	Hit Rate	Baseline Improvement
GCPN (Jiaxuan You & Leskovec, 2018)	2	-8.1 ±0.09	0.03%	-
GraphAF (Chence Shi, 2020)	5	-8.1 ±0.1	0.07%	-
MolDQN (Z. Zhou & Riley, 2019)	47	-8.2 ±0.49	0.73%	0.39x
MARS (Y. Xie & Li, 2021)	49	-8.7 ±0.71	0.77%	0.41x
REINVENT 4.0 (Baseline) (Loeffler et al., 2024)	119	-10.1 0±.63	1.85%	1x
REINVENT 4.0 + TL (Loeffler et al., 2024)	222	-10.1 ±0.42	3.46%	1.8x
NP-VAE + CMA-ES	279	-10.3 ±0.9	4.36%	2.36x
Ours	**1160**	**-11.4 ±1.1**	**18.12%**	**9.7x**
Ours (extended run)	**4137**	**-12.1 ±1.7**	**18.59%**	**10x**

Figure 2: Episode wise Hit Rate plot: The blue line represents the population size that varies between 150-50 samples. the red line represents the invalid compounds being generated that do not meet the success criteria defined in the experiment. The green line represents the number of valid hits generated and we can see that this number goes up as we go through the optimization process.

Figure 3: Binding Affinity distribution between the molecules present in ZINC250k dataset (orange) and ones generabted by our EPOSMol method (blue) for PDK1 (PDB ID 2XCH). It is evident that our method is successfully optimizing compounds towards having better bidning affinity. Red dashed line represents the Binding Affinity acceptance criteria of <= -8.0.
* smaller value indicates better binding affinity

successful hits, eventually exceeding the rate of non-hits. This trend highlights EPOSMol's capacity to explore a broad chemical space and increase the number of high-quality hits over time. The overall population of compounds varied cyclically between 150 and 50 across 100 episodes, while the noise factor ranged from 2.0 to 1.0 over a span of 50 episodes. This approach performed comparably to strategies with dynamically adjusted population size and noise factor, offering effective results with minimal overhead. Additionally, the varying population sizes and noise factors along with the diversity penalty helped prevent the model from becoming trapped in local optima, promoting thorough exploration of the fitness landscape and uncovering alternative optima.

In total, the optimization generated 36,420 compounds, of which 22,245 were unique molecules. When applying a hit criterion of binding affinity (BA) <= -8.0, we observed 11,890 hits among the generated compounds, achieving a 32.64% hit rate across all compounds or 53.45% when considering only unique compounds and excluding duplicates. Narrowing the hit criteria to BA <= -8.0 and QED >= 0.7, we identified 4,137 unique hits, yielding an 11.35% hit rate overall, or 18.59% among unique compounds. The optimization produced a total of 14,175 duplicate compounds. In practice one would let this run for several hundred more episodes to search through the chemical space even more thoroughly, but in our tests we had identified the pocket of known binders and several other pockets of very good binders by the 350-500 episode mark based on the target being studied.

Property Improvements: The best binding affinity predicted by Vina for the BioAssay compounds was -10.9, for the PDK1 target while our method generated compounds with an improved binding affinity of -12.1. Additionally, 91 of our generated compounds achieved a binding affinity of -10.9 or better. This improvement is a direct result of our reward function and hit criteria, which guide the algorithm toward regions in chemical space with enhanced binding affinity. Consequently, the algorithm not only captures and exploits promising regions of chemical space with strong hits but also continues to explore the entire chemical space of the training distribution. This dual capability allows it to identify diverse groups of compounds with improved binding affinity to the protein target, demonstrating both the local and global optimization strengths of our approach for molecular optimization. Figure 3 showcases Binding Affinity distribution and the efficacy of our method to generate more hits than the ZINC250k dataset that was used to train our underlying VAE.

The majority of the hits identified exhibited high drug-likeness, with a mean QED score of 0.615. Similar to previous experiments, we relaxed the QED requirements in the hit criteria to allow exploration of chemical space regions closely related to the bioassay compounds, which typically show a lower QED score with a mean of 0.502. This adjustment enabled the model to investigate

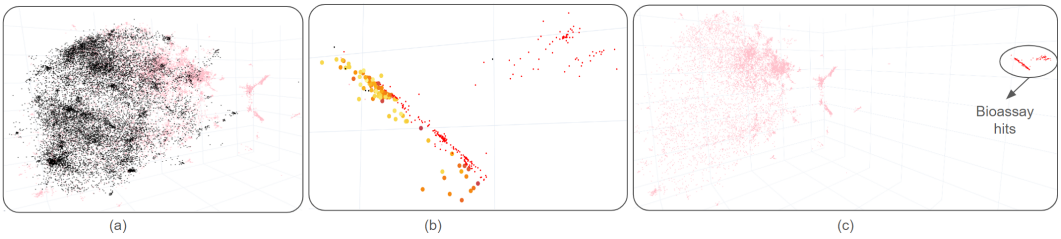

Figure 4: **(a)** the overall generated (black) chemical landscape coverage with DrugBank (pink) compounds. **(b)** Larger colored circles represent valid hits, with their color indicating Binding Affinity values with deeper shades of red indicating better BA. Smaller red dots denote bioassay compounds within the same chemical space. **(c)** Structural difference between the DrugBank compounds in pink and the bioassay hits shown on the extreme right in red.

and generate compounds within chemical space regions relevant to the bioassay dataset, while still prioritizing overall drug-like properties.

Structural Diversity: Our method efficiently explores the drug-like chemical space while rapidly identifying promising regions, demonstrating its optimization power. To map this landscape, we computed 2048-bit Morgan fingerprints for DrugBank compounds, hits, rejected compounds, bioassay compounds, and ZINC250k. These high-dimensional fingerprints were reduced to three dimensions using UMAP, revealing compound clusters and highlighting the extent of our optimization process.

Figure 4a shows extensive coverage across the drug-like chemical landscape. This highlights the algorithm's ability to explore vast spaces and uncover regions with potential drug-like properties. Notably, our method quickly identified a promising region containing bioassay compounds, a specialized and less-represented section of the landscape.

Since most bioassay hits had QED scores below 0.7, we relaxed the QED requirement, allowing the model to explore regions similar to bioassay compounds. Despite the vast search space, our method identified compounds in this region and generated hits with significantly improved binding affinities compared to previous bioassay compounds. This highlights our approach's ability to generalize across the complex drug-like landscape, producing diverse, potent candidates while optimizing globally.

Figure 4c shows that bioassay compounds are distinct from typical drug-like compounds. Figure 4b zooms into the bioassay compound space, where our method's generated hits and rejected compounds are plotted. These hits exhibited stronger binding affinities than tested compounds, highlighting the model's precision in identifying potent candidates in challenging regions.

Figure 5 showcases pairs of our designed compounds and bioassay compounds, demonstrating that our method preserves structural similarity while significantly improving binding affinity, emphasizing its optimization capabilities.

Our method generated compounds across 570+ generic Bemis-Murcko scaffolds. After filtering hits with a binding affinity <= -8, we found over 340 unique scaffold buckets, showcasing the model's strong exploration and ability to identify diverse regions of chemical space with high binding affinities.

4 Conclusion

We developed an Evolutionary Policy Gradient approach for molecule optimization, outperforming state-of-the-art methods in hit rate and binding affinity. It generated up to 32.64% valid, unique compounds—10× better than REINVENT 4.0—demonstrating superior exploration and optimization. The model-agnostic framework applies to any VAE-like model with a continuous latent space. By integrating binding affinity, drug-likeness, and diversity penalties, it efficiently explores chemical space while avoiding local optima.

Dynamic scheduling of population size and noise enables a smooth transition from broad exploration to focused optimization. Our method identified 570+ unique scaffolds and structurally diverse, high-

affinity hits. Extended runs explored fringe drug-like space, further improving compound quality. Future work will enhance sampling efficiency to reduce duplicates.

REFERENCES

Kunal Gupta Annu Lambora and Kriti Chopra. Genetic algorithma literature review. in 2019 international conference on machine learning. *Big Data, Cloud and Parallel Computing*, 1(1): 380–384, 2019.

G Richard Bickerton, Gaia V Paolini, Jérémy Besnard, Sorel Muresan, and Andrew L Hopkins. Quantifying the chemical beauty of drugs. *Nature chemistry*, 4(2):90–98, 2012.

Chen H Margreitter C Tyrchan C Engkvist O Papadopoulos K Patronov A Blaschke T, Arús-Pous J. Reinvent 2.0: an ai tool for de novo drug design. *J Chem Inform Model*, 60(12):5918–5922, 2020.

Zhaocheng Zhu Weinan Zhang Ming Zhang Jian Tang Chence Shi, Minkai Xu. Iclr, 2020.

P. Ertl and A. Schuffenhauer. Journal of cheminformatics, 1, 1–11, 2009.

Ah Chung Tsoi Markus Hagenbuchner Franco Scarselli, Marco Gori and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009.

Bo Dai Steven S. Skiena Hanjun Dai, Yingtao Tian and Le Song. Syntax-directed variational autoencoder for structured data, 2018.

Nikolaus Hansen, Sibylle D Müller, and Petros Koumoutsakos. Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (cma-es). *Evolutionary computation*, 11(1):1–18, 2003.

Tyrchan C Czechtizky W Patronov A Bjerrum EJ Engkvist O He J, Nittinger E. Transformer-based molecular optimization beyond matched molecular pairs. *J Cheminform*, 14(1):18, 2022.

Kamoun S Jabbar R, Jabbar R. Recent progress in generative adversarial networks applied to inversely designing inorganic materials: a brief review. computat mater sci, 2022.

Sung Ju Hwang Jaehyeong Jo, Seul Lee. Score-based generative modeling of graphs via the system of stochastic differential equations,icml, 2022.

Zhitao Ying Vijay Pande Jiaxuan You, Bowen Liu and Jure Leskovec. Graph convolutional policy network for goal-directed molecular graph generation. *Advances in Neural Information Processing Systems*, 31(1), 2018.

Craig Knox, Mike Wilson, Christen M Klinger, Mark Franklin, Eponine Oler, Alex Wilson, Allison Pon, Jordan Cox, Na Eun Chin, Seth A Strawbridge, et al. Drugbank 6.0: the drugbank knowledgebase for 2024. *Nucleic acids research*, 52(D1):D1265–D1275, 2024.

Hannes H Loeffler, Jiazhen He, Alessandro Tibo, Jon Paul Janet, Alexey Voronov, Lewis H Mervin, and Ola Engkvist. Reinvent 4: Modern ai–driven generative molecule design. *Journal of Cheminformatics*, 16(1):20, 2024.

AkshatKumar Nigam Pascal Friederich Mario Krenn, Florian Hase and Alan Aspuru-Guzik. Self-referencing embedded strings (selfies): A 100% robust molecular string representation, machine learning: Science and technology, 1, 2019.

H. L. Morgan. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *Journal of Chemical Documentation*, 5(2):107–113, 1965.

C A James C Morley T Vandermeersch N M O'Boyle, M Banck and G R Hutchison. Open babel: An open chemical toolbox. *J. Cheminf*, 3(1):33, 2011.

Toshiki Ochiai, Tensei Inukai, Manato Akiyama, Kairi Furui, Masahito Ohue, Nobuaki Matsumori, Shinsuke Inuki, Motonari Uesugi, Toshiaki Sunazuka, Kazuya Kikuchi, et al. Variational autoencoder-based chemical latent space for large molecular structures with 3d complexity. *Communications Chemistry*, 6(1):249, 2023.

Engkvist O Chen H Olivecrona M, Blaschke T. Molecular de-novo design through deep reinforcement learning. *J Cheminform*, 9(1):48, 2017.

Bo Zhao Mudong Feng Michael Gilson Peter Eckmann, Kunyang Sun and Rose Yu. Limo: Latent inceptionism for targeted molecule generation. *Proceedings of the 39th International Conference on Machine Learning*, 162(1):5777–5792, 2022.

Madzhidov T.I. & Varnek A. Polishchuk, P.G. Estimation of the size of drug-like chemical space based on gdb-17 data. *J Comput Aided Mol*, 27(1):675–679, 2013.

Kenneth V. Price Rainer Storn. Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, 11(1):341–359, 1997.

RDKit. Rdkit: Open-source cheminformatics.

Chenthamarakshan V Padhi I Mroueh Y Das P Ross J, Belgodere B. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intell*, 4(12): 1256–1264, 2022.

Frank Sehnke, Christian Osendorfer, Thomas Rückstieß, Alex Graves, Jan Peters, and Jürgen Schmidhuber. Parameter-exploring policy gradients. *Neural Networks*, 23(4):551–559, 2010.

Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. neurips, 2012.

O. Trott and A. J. Olson. Vina, journal of computational chemistry,31, 455–461, 2010.

David Weininger. Smiles, a chemical language and information system. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, 1988.

Regina Barzilay Wengong Jin and Tommi Jaakkola. Junction tree variational autoencoder for molecular graph generation. *Proceedings of the 35th International Conference on Machine Learning*, 80(1):2323–2332, 2018.

H. Zhou Y. Yang W. Zhang Y. Yu Y. Xie, C. Shi and L. Li. nternational conference on learning representations,, 2021.

L. Li R. N. Zare Z. Zhou, S. Kearnes and P. Riley. Scientific reports,9,10752, 2019.

Chengxi Zang and Fei Wang. Moflow: An invertible flow model for generating molecular graphs. *In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 26(1):617–626, 2020.

Li Li Richard Zare Zhenpeng Zhou, Steven Kearnes and Patrick Riley. Optimization of molecules via deep reinforcement learning. scientific reports, 9:10752, 07, 2019.

A Appendix

A.1 Reward Modeling

The reward function in our optimization framework incorporates multiple components to define the fitness landscape and achieve the desired optimization goals. For practical purposes, we focus on three core components: binding affinity, diversity penalty, and drug-likeness. Each component represents a distinct aspect of molecule evaluation, collectively shaping the reward landscape and guiding the optimization toward chemically diverse and effective compounds.

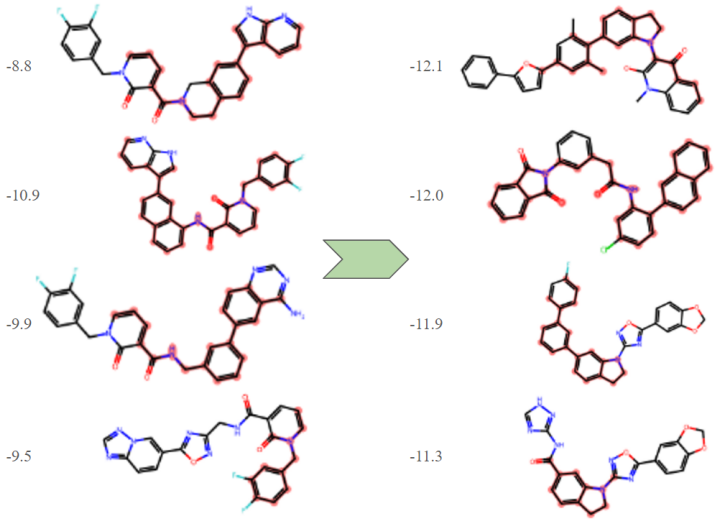

Figure 5: Structural similarity and Binding Affinity improvements: On the left are the bioassay compounds and the EPOSMol generated compounds on the right. The atoms and bonds highlighted in orange indicate the Maximum Common Substructure. The binding affinity is mentioned on the left of each molecule. It is evident that the generated compounds share a high degree of scaffold similarity with the bioassay compounds while drastically improving upon the Binding Affinity of these compounds. * smaller value indicates better binding affinity

A.1.1 Diversity Penalty

To prevent over-exploration of specific regions in chemical space, we developed a diversity penalty mechanism. This penalty is triggered when generated compounds are overly similar to those already explored. We implemented this by setting a similarity threshold of 0.6, over the scaffolds of generated compounds. Any new compound with a similarity score exceeding this threshold with an existing compound was grouped into the same "bucket," while a unique compound formed a new bucket.

A bucket is capped at a maximum capacity (e.g. 800 compounds), and a diversity penalty is applied when the unique compound count in any bucket exceeds half of this size (i.e., 400). Specifically, duplicate compounds incur a diminished reward, decreasing proportionally with the number of duplicates within that bucket. The penalty formula is as follows:

$$R_{dp} = \begin{cases} 1, & \text{if } bucket\_count \leq \frac{dp\_factor}{2} \\ \frac{dp\_factor - bucket\_count}{\frac{dp\_factor}{2}}, & \text{otherwise} \end{cases} \tag{3}$$

$$R_{dp} = \frac{R_{dp}}{duplicate\_count} \tag{4}$$

Where

- R_{dp}: Diversity penalty reward.
- bucket_count: current size of the matched scaffold bucket.
- dp_factor: the maximum allowed compounds per bucket, currently set to 800
- duplicate_count : the duplicate counter tracking each generated compound

This penalization mechanism effectively flattens or even inverts the reward landscape in over-explored regions, pushing the policy away from local optima, similar to the annealing process. This approach encourages exploration across diverse regions in the chemical space, reducing the likelihood of stagnation in suboptimal regions.

A.1.2 DOCKING SCORE

Docking studies were conducted using Vina2.2_GPU (Trott & Olson, 2010), with ligand preparation performed using RDKit (RDKit) and OpenBabel (N M O'Boyle & Hutchison, 2011) to convert structures into pdbqt format. We then docked these ligands to generate binding affinities. To integrate binding affinity into the reward function, we normalized its values to a 0-1 range. This was achieved by dividing the square of the binding affinity (BA) by the square of an optimal target affinity, taken as -13. Specifically, the scaling formula for binding affinity

$$\mathrm{R_{ba}} = \frac{BA^2}{-13^2} \tag{5}$$

A binding affinity threshold of -8.0 was applied, aligning with the REINVENT protocol (Loeffler et al., 2024), to classify "successful" hit molecules. This criterion helps prioritize compounds with stronger binding affinity within the optimization process.

A.1.3 QED

We used the Quantitative Estimate of Drug-likeness (QED) descriptor, implemented in RDKit, to evaluate the drug-likeness of compounds. Molecules with a QED score of 0.7 or higher are considered to have high drug-likeness, scoring the full value of 1 in the reward function. For compounds below the 0.7 threshold, scores were scaled linearly between 0 and 1 to reflect intermediate levels of drug-likeness. The QED component is defined as follows:

$$R_{qed} = \begin{cases} 1, & \text{if } QED \geq 0.7 \\ \frac{QED}{0.7}, & \text{otherwise} \end{cases} \tag{6}$$

A.1.4 OVERALL REWARD CALCULATION

The total reward R is the product of the binding affinity component, the QED component, and the diversity penalty, formulated as:

$$\mathrm{R} = R_{ba} * R_{dp} * R_{qed} \tag{7}$$

This reward formulation ensures a balanced optimization by encouraging high binding affinity, drug-likeness, and diversity, preventing convergence on any one aspect at the expense of the others, and thus guiding the search toward a broader exploration of chemical space.

A.2 NOISE SAMPLING AND POPULATION SIZE

To optimize convergence and balance exploration and exploitation, we implemented dynamic schedulers for both population size (ranging from 150 to 50) and noise factor (from 2.0 to 1.0). These schedulers adaptively reduce population size and noise when high-reward regions are identified, allowing the model to exploit these areas more efficiently. By limiting exploration in high-reward regions, this approach minimizes duplicate compound generation and refines optimization around promising leads. Conversely, when the reward declines due to the diversity penalty signaling sufficient exploitation of a high-reward area, the schedulers increase both noise factor and population size. This enables the model to expand its search, quickly moving away from local minima by sampling more points farther from the current policy and making significant policy adjustments to seek new optimal regions.

In ablation studies, we also explored simpler cosine schedulers to parameterize the noise factor (with an episodic length of 50) and the population size (episodic length of 100). While this approach showed a slight decrease in finding hits with the highest binding affinities, it proved more flexible in exploring high-reward regions and transitioning out of them without excessive sampling. When noise and diversity penalties increased, the cosine scheduler naturally shifted exploration focus, preventing oversampling. This method's simplicity and reproducibility offered a practical alternative, and we utilized cosine schedulers for all downstream analyses to parameterize population size and sampling noise factor effectively.

$$Population_t = 50 + 50(1 + \cos\frac{\pi t}{100}) \tag{8}$$

$$Sampling\_Noise_t = 1 + 0.5(1 + \cos\frac{\pi t}{50}) \tag{9}$$

The learning rate too was on a cosine scheduler varying between 0.3-0.1 with an episodic period of 150 episodes.

B EXPERIMENTS

B.1 MODEL TRAINING

For the generative model we chose a small 12M parameter model based on the NP-VAE (Ochiai et al., 2023) architecture with a latent space of 256 dimensions because of its favorable trade-off between model performance and latent space size. We first break down the molecules into a non-cyclic tree structure by identifying functional fragments in the molecule and linking them in the graph. The number of nodes in this composed tree is equal to the number of fragments identified from the original molecule with edges drawn between connecting nodes. The 3D ECFP is then calculated using RDKit for each captured node. The encoder consists of a Child Sum Tree-LSTM which generates a latent vector. This latent vector is then decoded back into a compound using a depth-first algorithm. Further to construct the latent space to represent bioactivity as well, a fully connected prediction head is trained over the latent vector. In our study we found the best performance to be when we trained multiple property prediction heads that comprised of logP (Wengong Jin & Jaakkola, 2018), QED (Bickerton et al., 2012) and SA score (Ertl & Schuffenhauer, 2009).

The pre-trained weights provided on the NP-VAE GitHub repository showed limited capacity to generate structures outside the original training dataset. To overcome this, we retrained the NP-VAE model using an expanded dataset comprising ZINC250k (Jaehyeong Jo, 2022), DrugBank (Knox et al., 2024), and a custom library containing 50k compounds resembling underrepresented subsets of DrugBank structures. This broader dataset improved the model's generalizability, allowing it to generate structures closely aligned with certain experimentally identified compounds that are notably diverse from those in DrugBank and ZINC250k. The models were trained for 100 epoch over 36 hours using four Nvidia Tesla T4 GPUs. We produced two versions of the model: one guided by the Natural Product (NP) likeness score to structure the latent space and another using a combination of LogP, Synthetic Accessibility (SA), and QED scores. Testing showed no significant differences between these versions within the policy optimization framework, as both generated similar clusters of optimized hits. This outcome indicates that the policy optimization process is capable of effectively navigating noisy reward landscapes to achieve robust, global optimization.

B.2 TARGET

For this study, we selected Phosphoinositide-dependent kinase-1 (PDK1) as the protein target, following the REINVENT 4.0 protocol to establish a concise benchmark for comparison. PDK1, with PDB ID 2XCH, is a critical enzyme that regulates various cellular functions, including growth, survival, and metabolism. It exerts its effect by activating several key protein kinases, such as AKT, through phosphorylation at specific sites. PDK1 is a core component of the PI3K/AKT signaling pathway, which is heavily implicated in cancer development and progression, making it a significant target in cancer therapy. Known binders for this protein target have been characterized in PubChem assay AID1798002, containing 315 compounds with a common Pyrroloquinazoline or similar core structure. These existing binders provide a basis for assessing the model's capability to identify and optimize compounds with potential therapeutic relevance.

B.3 REINVENT 4.0 BENCHMARKING

In our comparative study with REINVENT 4.0, we maintained the same study parameters, generating 6,400 molecules and applying identical reward function and hit classification criteria. As in the REINVENT 4.0 study, compounds were classified as successful hits if they met a binding affinity threshold of <= -8.0 and a QED score above 0.7. The only deviation from the REINVENT 4.0 protocol was in the docking tool used: we opted for Vina2.2_GPU due to its open-source nature and accessibility without licensing requirements. This approach, however, remains adaptable, allowing for any scoring methodology that may be available or preferred in other contexts. We also ran this same benchmark with several existing methods for comparison which are captured in Table 1

B.4 EXTENDED RUNS

Beyond the comparative study, we extended the optimization run to approximately 350 episodes, generating a total of 36,420 molecules. To fully explore the potential of our approach, we relaxed the QED score requirement for successful hit classification, allowing for a more comprehensive investigation of the fringe regions of drug-like chemical space, which often exhibit lower drug-likeness. This adjustment was also informed by the observation that assay compounds with strong binding affinities to the target frequently had low QED scores, with over half falling within the 0.26–0.5 range. This flexible hit classification enabled our method to thoroughly explore diverse chemical landscapes, capturing promising compounds even in regions with traditionally lower drug-likeness.