Reinforcement Learning for Enhanced Targeted Molecule Generation Via Language Models

Salma J. Ahmed **Emad A. Mohammed**
Department of Physics and Computer Science
Wilfrid Laurier University
Waterloo, ON N2L 3C5
{ahme3460,emohammed}@{mylaurier,wlu}.ca

Abstract

Developing new drugs is laborious and costly, demanding extensive time investment. In this study, we introduce an innovative de-novo drug design strategy, which harnesses the capabilities of language models to devise targeted drugs for specific proteins. Employing a Reinforcement Learning (RL) framework utilizing Proximal Policy Optimization (PPO), we refine the model to acquire a policy for generating drugs tailored to protein targets. Our method integrates a composite reward function, combining considerations of drug-target interaction and molecular validity. Following RL fine-tuning, our approach demonstrates promising outcomes, yielding notable improvements in molecular validity, interaction efficacy, and critical chemical properties, achieving 65.37 for Quantitative Estimation of Drug-likeness (QED), 321.55 for Molecular Weight (MW), and 4.47 for Octanol-Water Partition Coefficient (logP), respectively. Furthermore, out of the generated drugs, only 0.041% do not exhibit novelty.

1 Introduction

The journey from conceptualizing a potential drug to its market availability is lengthy and financially demanding [21, 7]. It must navigate through several critical phases to transform a chemical compound or entity into a viable treatment for human diseases. Initially, a specific molecular target (such as a DNA sequence or protein) associated with a disease must be identified [12, 35, 3]. This target serves as the focal point for drug development, offering the potential for therapeutic intervention. Before a drug can be administered to patients, it must undergo rigorous preclinical research trials. These trials, conducted either in vitro or in vivo, aim to evaluate the drug's safety profile and assess its effects on biological systems. This phase is pivotal in determining the drug's potential for therapeutic use and understanding its impact on the body.

The drug progresses to clinical research after successful preclinical trials, where its efficacy and safety are tested on human subjects. This phase involves carefully designed clinical trials conducted in multiple stages to gather comprehensive data on the drug's performance and potential side effects. Subsequently, the accumulated data undergoes thorough scrutiny during the FDA review process, where regulatory authorities assess the drug's safety, efficacy, and overall benefit-risk profile. This extensive journey, while essential for ensuring patient safety and the effectiveness of medications, poses significant challenges [48, 15]. The prolonged timeline and substantial financial investment associated with drug development contribute to the high costs and low success rates in discovering new drugs. As a result, innovative approaches that streamline the drug discovery process and enhance efficiency are crucial for addressing unmet medical needs and advancing therapeutic interventions.

With the rapid advancements in deep learning, many researchers are delving into applying deep learning methodologies to address challenges within the drug discovery domain. These encompass

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

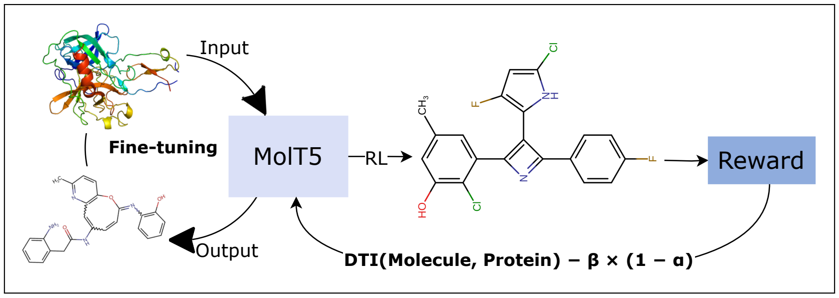


Figure 1: MolT5 [13] is initially fine-tuned for compound generation based on input protein. Subsequently, the fine-tuned model is employed in RL-based policy fine-tuning to acquire the ability for targeted compound generation, with drug-target interaction and validity as rewards.

various areas including Drug-Target Interaction [55, 1, 58, 14, 54], Protein-Protein Interaction [49, 17, 19, 32, 41], Drug-Drug Interaction [11, 28, 59, 31, 45], De-Novo Drug Design [53, 4, 26, 38, 27], and Drug Re-purposing [39, 22, 2, 30, 18]. These approaches leverage the power of deep learning techniques to tackle complex problems in drug discovery, offering promising avenues for accelerating the identification and development of novel therapeutics. This highlights the significant potential of employing deep learning methodologies to expedite the discovery of novel drug candidates, ultimately enhancing our ability to combat emerging diseases. In this study, our primary focus is on harnessing the advancements in Language Models and Reinforcement Learning techniques to facilitate drug discovery by proposing an approach for De-Novo drug design. Our main objective centred around developing a generative model capable of taking a molecule target (protein) as input, representing a disease within the human body. The aim was to utilize this model to generate novel molecules (drugs) specifically designed to treat or mitigate the effects of this protein within the body.

The rest of the paper is organized as follows: Section 2 provides an overview of the related literature. Section 3 discusses the components of the proposed methodology. Section 4 describes the experimental setup and presents the results. Section 5 offers an in-depth analysis and discussion of the findings. Finally, Section 6 concludes the paper, highlighting key insights and directions for future research.

2 Related Work

Numerous approaches documented in the literature address the challenge of de novo drug design, given its significant potential impact when addressed effectively.

In [43], the authors introduced ReLeaSE, a technique for producing novel targeted chemical compounds. This method integrated generative and predictive deep neural networks, where the generative model was trained to generate new chemical compounds, while the predictive model forecasts desired properties. These models are trained independently and combined using a reinforcement learning approach to guide the generation process. Reinvent the Reinforcement Learning (RL) approach described by [8] for generating novel molecules that interact with specific targets by using RL to steer the generative model. This approach employs two Recurrent Neural Networks (RNNs) in an actor-critic setup. The critic, acting as the prior RNN, retains previous knowledge of the SMILES representation. The actor, or agent, is either an identical copy of the prior or a modified version that has undergone some initial training. The agent selects actions by sampling a batch of SMILES (S), which are then evaluated by the prior RNN and scored using a predefined scoring function. The method in [16] introduced a generative RNN-LSTM model for drug synthesis. Initially, the model undergoes training on a dataset of molecular structures to grasp the syntax and patterns inherent in these representations. Subsequently, the RNN-LSTM model undergoes fine-tuning to skew predictions towards specific molecular targets. This adjustment is achieved by leveraging insights gained from the initial training to tailor the model to particular target molecules. Training on a generator RNN model conducted by [60] using molecular data to familiarize it with the syntax of SMILES for molecular representation, enabling the design of novel and effective small molecules. Subsequently, they developed a drug-target interaction model to serve as a reward in a reinforcement learning framework. This model was employed to steer the generation of the RNN model towards specific properties or targeted molecules. A multi-objective approach was introduced by [37] for drug synthesis, emphasizing properties and selectivity tailored to biological targets. Their method employed a transformer decoder to create drugs and a transformer encoder to forecast desired properties and refine learning through a feedback loop.

3 Methods

In this section, we thoroughly examine the methodology components, spanning from the dataset utilized to the generative model and the elements of the reinforcement learning paradigm.

3.1 Dataset

In our approach, we utilize BindingDB [34], a publicly available database encompassing binding affinities of protein-ligand complexes, across all our experiments. The dataset selection specifically targeted proteins serving as drug targets, with their corresponding structural information sourced from the Protein Data Bank [6]. Leveraging this dataset facilitated the achievement of our objectives. Additionally, we implemented a filtering criterion within this dataset to exclude complexes featuring protein amino acid lengths exceeding 500, primarily due to computational resource constraints.

3.2 Generative Model

An essential aspect of the methodology involves utilizing a generative model, which serves as the vital tool for generating molecular compounds, aligning with the stated objective of the method to generate a molecule drug based on a given protein. We employed MolT5 [13], a self-supervised learning framework built upon an encoder-decoder transformer architecture [51]. Initially, MolT5 underwent pretraining on a substantial volume of unlabeled molecule compound strings and natural language text. Subsequently, the model underwent fine-tuning on two distinct tasks(i.e., molecule captioning and text-based de novo molecule generation). In the first task, the model receives a molecule string prompt and aims to generate a caption, while in the second task, it generates a molecule string based on a provided textual description. Encouraging outcomes were achieved on both tasks, motivating the adoption of the model for our objective. Given a protein amino acid, the model is designed to produce a targeted molecule drug. We utilize the base version of MolT5 (molt5-base) and conducted further fine-tuning employing the protein-ligand complexes sourced from BindingDB to enhance the model's knowledge for our specific task, as depicted on the left side of Figure 1. The fine-tuning process is evaluated using the Bilingual Evaluation Understudy metric (BLEU) [40], which measures the similarity between the generated molecule and the actual molecule in the protein-ligand complexes, with the protein serving as input to the model. We intend for the model to replicate the actual molecule compound within the complex by instilling in the model the ability to generate molecule compounds based on a protein string. Following fine-tuning, the model demonstrated proficiency in generating molecules given proteins as inputs. We integrate the molt5-fine-tuned model into RL paradigm to steer the learning process toward generating molecules that target specific proteins.

3.3 Drug-Target Interaction (DTI)

Our objective is to enable the generative model [13] to generate molecule compounds tailored to specific proteins. To assess the activity of these molecules, we integrated a Drug Target Interaction (DTI) model using DeepPurpose [20], a PyTorch toolkit for molecular modeling and prediction. It provides encodings for molecules and protein sequences, which are fed into a multi-layer perceptron (MLP) decoder to predict binding scores. We tested two encoding methods: one using CNNs for both SMILES strings and protein sequences and the other using a CNN for protein sequences combined with transformer encoders for substructure fingerprints. Focusing on IC50 scores, we trained the model using BindingDB data and adopted an ensemble learning approach, averaging predictions from both models. This method produced promising results and was selected for system integration.

3.4 Reinforcement Learning Fine-tuning

We implemented the Proximal Policy Optimization (PPO) algorithm [46], a widely utilized method for training policies in reinforcement learning tasks [56, 9, 57, 25], within our Reinforcement Learning (RL) framework. Our goal was to leverage this algorithm to guide the model in learning a policy

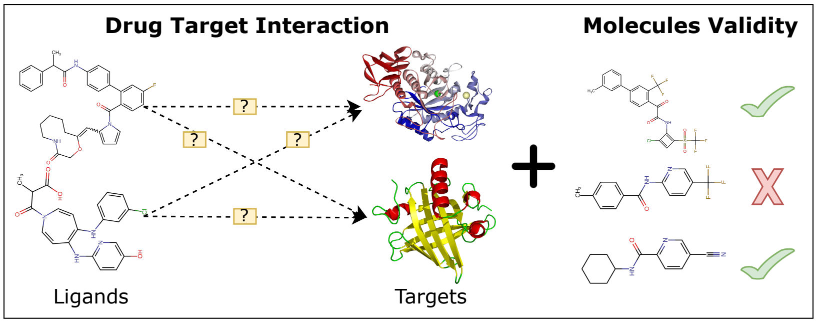


Figure 2: The lefthand side depicts the methods used for reward calculation, focusing on drug-target interaction (DTI), while the righthand side illustrates the evaluation of molecule validity.

where, upon receiving a protein amino acid sequence as input, the model would generate a molecule compound (drug) tailored to the particular protein, as depicted on the right side of Figure 1. We employed the molt5-base model, which we had previously fine-tuned, as the initial model. We then proceeded to update and assess the model's performance before and after the fine-tuning process. We employed a combined reward function to optimize the model and align it with the desired performance, which will be elaborated upon in the subsequent subsection.

3.4.1 Rewards

The reward selection holds significant importance within the realm of RL, as it serves as the feedback signal given to an agent, reflecting both its actions and the state of the environment. It serves as a measure of the agent's performance, indicating its effectiveness in accomplishing tasks. Hence, we opted for a composite reward strategy as illustrated in Figure 2 to attain the desired level of performance.

DTI: We integrated the Drug Target Interaction (DTI) model into our RL methodology to serve as the reward mechanism. This decision aligned with our objective of training the model to generate molecules tailored to specific proteins. Consequently, when the model produces a molecule compound unrelated to the protein target, it receives a low reward score. Conversely, it gets a higher score when the generated molecule is targeted to the protein.

Validity: In addition to targeting specific proteins, we aim for the generated molecule to meet chemical validity criteria, ensuring its viability as a potential drug. Hence, we incorporated validity as the second reward within the RL paradigm. To assess validity, we employed the rdkit toolkit [29] to determine whether the generated molecule complies with established chemical standards.

$$Reward = DTI(M_G, P) - \beta \times (1 - \alpha) \tag{1}$$

$$\alpha = Validity(M_G) \in \{0, 1\} \tag{2}$$

Equation 1 delineates the reward computation process. Initially, it computes the interaction between the protein (P) and the generated molecule (M_G) via the fine-tuned Drug Target Interaction model, denoted as DTI. Subsequently, the validity of the molecule is evaluated, with β representing the penalty applied when the molecule is invalid ($\beta = 0.30$). When the molecule is valid, indicated by $\alpha = 1$, the penalty term effectively reduces to zero. Equation 2 elucidates the definition of α, a binary variable with either 0 or 1 values, signifying the generated molecule's validity status.

4 Experimental Setup and Results

This section delves into the comprehensive exploration of our approach, detailing each experimental procedure and its corresponding outcomes. Through systematic refinement of each component, we aimed to optimize performance and achieve superior results.

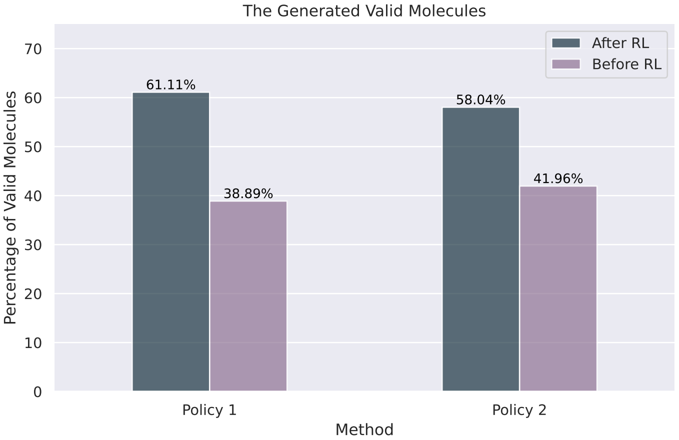


Figure 3: The percentage of valid molecules before and after fine-tuning a model with reinforcement learning using two policies. Policy 1 combines Drug-Target Interaction (DTI) and molecule validity for rewards, while policy 2 only considers molecule validity. This evaluation assesses the method's effectiveness and adaptability to different objectives and policies.

4.1 Drug Generation

Our initial step involved configuring the generative model to produce molecular compounds when provided with a protein amino acid sequence as input. This is a crucial aspect of our experimental pipeline, as illustrated in the lefthand side of Figure 1. Through meticulous fine-tuning of BindingDB complexes, the model demonstrated proficiency in generating molecular compounds given protein inputs. However, our objectives extend beyond mere molecule generation; we aim for targeted compounds tailored to the specific protein. To achieve this, we will proceed with additional fine-tuning iterations to imbue the model with a refined policy for learning protein-targeted molecule generation.

4.2 Reinforcement Learning Fine-tuning Setup

We employed the Transformer Reinforcement Learning (TRL) library [52] for fine-tuning our reinforcement learning model, illustrated in 1. This library, built upon the transformers framework, is designed explicitly for refining transformer models through various techniques, including supervised fine-tuning (SFT), Reward Modeling (RM), Direct Preference Optimization (DPO), and Proximal Policy Optimization (PPO). We opted for the PPO method to fine-tune our model due to its recognized efficiency and effectiveness in training complex policies. PPO stands out for its ability to learn with fewer samples than alternative methods. Moreover, its utilization of a trust region optimization approach helps prevent drastic policy changes, ensuring stable learning dynamics. Furthermore, PPO demonstrates scalability, making it suitable for addressing large-scale problems and high-dimensional action spaces. During experimentation, we explored different values for the policy optimization parameters, such as top_k and top_p. Our analysis revealed that the most favorable results were achieved with top_k set to 50 and top_p to 0.95.

4.3 Reward Optimization

As outlined in the previous sections, the reward utilized for fine-tuning the generative model in reinforcement learning consisted of two key components(i.e.,the Drug Target Interaction model and the validity assessment of the generated molecule.) In the subsequent sections, we delve into the specifics of our experimentation with various setups, elaborating on the details and outcomes.

4.3.1 Ensemble Learning for DTI

To elevate the performance of the Drug Target Interaction model as a reward, we embarked on various experiments, ranging from employing individual models to amalgamating them into groups to explore diverse outcomes. These efforts culminated in the final experiment: an ensemble learning strategy

Chemical Properties	**Before**$_{RL}$	**After**$_{RL}$
Quantitative Estimation of Drug-likeness (QED)	0.5705	0.6537
Molecular Weight (MW)	387.84	321.55
Octanol-Water Partition Coefficient (logP)	4.75	4.47

Table 1: The resulting values of analyzing the chemical properties of the generated molecules both before and after fine-tuning the generative model with reinforcement learning.

involving two distinct DeepPurpose modelsCNN and transformer-based. After rigorous evaluation, we determined that the optimal approach was to merge the predicted affinities, as depicted in Equation 3. In this equation, *C* represents CNN, *T* represents Transformer, and *P* represents predictions. We assigned weights (C_w = 0.25 and T_w = 0.75) to each model's predictions to reflect their impact on the final affinity score, as these values were determined to yield the best performance. This fusion technique allowed us to leverage the strengths of both models, resulting in a more resilient and efficient improvement.

$$Affinity = (P_C \times C_w) + (P_T \times T_w) \tag{3}$$

4.3.2 Integration Validity

Initially, our objective was for the model to generate targeted molecule compounds tailored to a specific protein exclusively. However, we analyzed the outcomes and discovered that specific generated molecules were chemically invalid. Consequently, we introduced another factor into the reward calculation(i.e, the validity of the generated compounds.) This adjustment allowed us to optimize the generative model to learn a singular policy and acquire a hybrid or mixed policy. This enhancement ensures a more comprehensive optimization approach, accommodating the generated compounds' efficacy and chemical validity considerations. The selection of the (β) value in equation 1, which denotes the penalty applied to the reward for invalid molecules, was determined through trial and error. As we fine-tuned the model and monitored its learning and optimization, we systematically experimented with values ranging from 0.1 to 0.7. After rigorous evaluation, we found that 0.3 yielded the most optimal results. This iterative approach ensured that the penalty value was finely calibrated to strike a balance between penalizing invalid molecules sufficiently while maintaining effective model optimization. In Figure 3 for Policy 1, we depict the percentage of valid molecules generated before and after applying RL fine-tuning with the combined reward.

4.4 Validating the Proposed Method Efficacy

To ascertain the effectiveness of our methodology, we undertook experiments to refine the model's focus solely on generating valid molecules. This involved modifying the reward calculation to prioritize validity over assessing drug-target interactions. The notable optimization observed in the model's performance under this refined policy, as depicted in Figure 3 (policy 2), underscores the efficacy of this approach. These experiments offer compelling evidence of our methodology's ability to generate valid molecular compounds and highlight its adaptability to diverse learning policies.

5 Results Analysis and Discussion

In this section, we delve into the comprehensive analysis of the proposed method, which includes assessing the chemical properties of the generated compounds and examining their interactions with proteins.

5.1 Chemical Properties

To evaluate the efficacy of our proposed approach, we analyzed the chemical properties of the generated molecules or drugs both before and after fine-tuning the generative model with reinforcement learning. We assessed a range of widely recognized molecular properties or descriptors commonly employed in drug discovery and computational chemistry, including Quantitative Estimation of

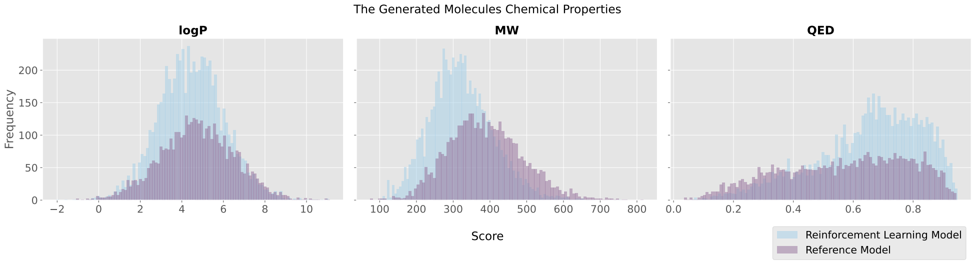


Figure 4: This figure illustrates the distributions of the chemical properties of the generated molecules, including Octanol-Water Partition Coefficient (log P), Molecular Weight, and Quantitative Estimation of Drug-likeness (QED), both before and after fine-tuning the model with Reinforcement Learning.

Drug-likeness (QED), Molecular Weight (MW), and Octanol-Water Partition Coefficient (logP). These metrics are typically utilized to gauge the potential of a compound to progress into a viable drug candidate.

1. **Quantitative Estimation of Drug-likeness (QED):** This metric is a tool for evaluating and assessing the drug-likeness of chemical compounds or molecules. Integrating various molecular descriptors provides a numerical estimation indicating the likelihood that a molecule possesses favorable drug properties. We subjected the generative model to the same set of proteins to gauge the QED of the generated molecules, both before and after fine-tuning with reinforcement learning. The outcomes revealed that the Mean QED of the generated molecules was 0.5705 before RL fine-tuning and increased to 0.6537 after RL fine-tuning, as depicted in table 1. This observation underscores that the molecules generated after RL fine-tuning exhibit enhanced desirable drug properties.
2. **Molecular Weight (MW):** This characteristic plays a pivotal role in understanding a compound's pharmacokinetics, formulation, and toxicity behavior. It represents the sum of the atomic weights of all atoms within a molecule. Our evaluation encompassed MW values of all generated molecules, conducted both before and after RL fine-tuning, employing an identical set of proteins for assessment. Our analysis revealed that the mean MW value was 387.84 before fine-tuning, which subsequently decreased to 321.55 after RL fine-tuning, as shown in table 1. This reduction in molecular weight carries potential advantages, as compounds with lower MW are often more readily absorbed and metabolized, and they may exhibit reduced complexity, facilitating more straightforward and more accessible synthesis pathways.
3. **Octanol Water Partition Coefficient (logP):** Solubility and permeability play pivotal roles in predicting crucial drug properties such as absorption, distribution, metabolism, and excretion (ADME). Thus, logP serves as a fundamental metric to assess the partitioning behavior of a generated compound between an organic solvent and water. Our investigation unveiled that prior to fine-tuning, the mean logP value stood at 4.7539, which subsequently decreased to 4.4766 after RL fine-tuning, as illustrated in table 1. Notably, a log P value of 5 or less is often considered optimal for drug candidates, aligning with Lipinski's Rule of Five for oral bioavailability [33]. Therefore, we further scrutinized the percentage of compounds with logP values less than or equal to 5, revealing figures of 62.972% after RL fine-tuning and 56.106% before RL fine-tuning. These findings suggest that after RL fine-tuning, a more significant proportion of the generated compounds exhibited logP values within the optimal range, which indicates favorable solubility and permeability characteristics. Thus, the fine-tuning process positively impacted the fundamental properties essential for effective drug development.

As depicted in Figure 4, the distribution of the chemical properties of the generated molecules, including Octanol-Water Partition Coefficient (log P), Molecular Weight, and Quantitative Estimation of Drug-likeness (QED), highlights the significant improvements and effectiveness achieved through the Reinforcement Learning fine-tuning of the model. This enhancement results in the molecules exhibiting the desired and optimal chemical properties.

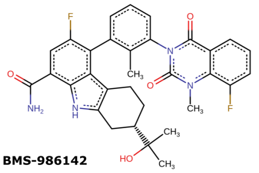
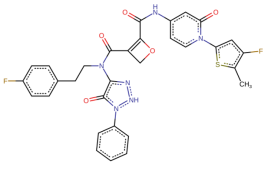
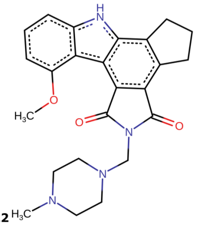
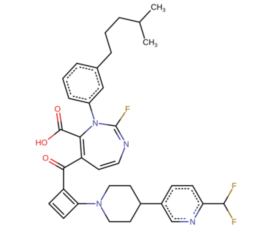
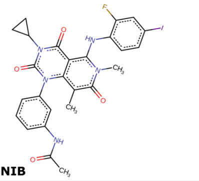
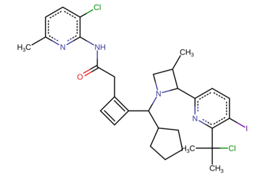
Input Protein	Reference Molecules	Generated Molecules	TS
BTK	 BMS-986142		0.6711
PARP	 CEP-9722		0.5327
BRAF	 TRAMETINIB		0.5850

Figure 5: Illustrative instances of the molecules generated through inputting different proteins into the model, juxtaposed with samples of protein inhibitors and the Tanimoto similarity (TS) between the generated and inhibitor compounds.

5.2 Assessing Molecular Novelty

We assessed our methodology by inputting proteins into our RL fine-tuned model and examining the novelty of the generated molecules compared to the training data. This analysis aimed to determine whether the model merely memorized molecules. Our findings revealed that a mere 0.041% of molecules were memorized from the training dataset. This underscores the model's capability to generate genuinely novel molecules.

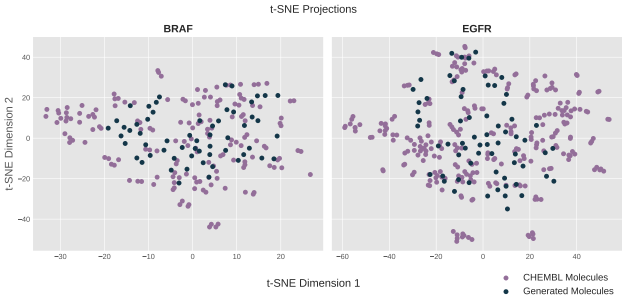


Figure 6: T-SNE visualizations of fingerprint descriptors show that many generated compounds closely match or resemble reference molecules.

5.3 Investigation of Model-Generated Molecules

To delve deeper into the outcomes produced by the model, we meticulously examined its outputs using a diverse range of proteins, such as human Bruton's tyrosine kinase (BTK), poly ADP-ribose polymerase (PARP), v-Raf murine sarcoma viral oncogene homologue B (BRAF), G-protein coupled receptor 6 (GPR6), and epidermal growth factor receptor (EGFR), for a comprehensive evaluation. Subsequently, we compiled a collection of Inhibitors known to interact with each protein from the ChEMBL database [10] and performed Tanimoto coefficient [5] similarity calculations between the generated compounds and reference compounds. Figure 5 presents a visual representation of a subset

Method	**Protein**	**Metrics**			
		Novel	Unique	Diversity	Filters
AAE [36]	-	0.793	1.0	0.855	0.996
JTN-VAE [23]	-	0.914	1.0	0.855	0.976
VAE [24]	-	0.694	1.0	0.855	0.997
CharRNN [47]	-	0.8419	1.0	0.856	0.994
latentGAN [44]	-	0.949	1.0	0.856	0.973
FSM-DDTR [37]	-	0.9596	0.998	0.871	-
Zhang et al. [60]	BTK	0.990	-	0.674	0.308
	BRAF	**0.989**	-	0.666	0.413
	EGFR	0.979	-	0.702	**0.793**
	PARP	0.992	-	**0.979**	0.398
Proposed Method	BTK	**1.0**	**1.0**	**0.853**	**0.666**
	BRAF	0.982	**1.0**	**0.858**	**0.672**
	EGFR	**1.0**	**1.0**	**0.853**	0.630
	PARP	**1.0**	**1.0**	0.8481	**0.614**
	GPR6	**1.0**	**1.0**	**0.849**	**0.638**

Table 2: Evaluation of the molecular generation models utilizing MOSES.

of generated and reference compounds and their corresponding Tanimoto Similarity scores, enhancing comprehensibility. Further visual representations are available in appendix A. Notably, although the model was fine-tuned on protein amino acids of length 500, it effectively generated targeted compounds for proteins with longer sequences, such as BTK, BRAF, PARP, and EGFR. Moreover, we utilized the t-distributed Stochastic Neighbor Embedding (t-SNE) algorithm [50] to visualize samples of the generated compounds' fingerprint descriptors and reference compounds' fingerprint descriptors in two dimensions, aiming to investigate whether the properties of the generated compounds aligned with the ChEMBL data. Figure 6 illustrates the results for BRAF and EGFR, indicating that several generated compounds match or closely resemble reference molecules. Further visualizations are available in appendix B.

5.4 Performance Analysis

We compared our approach with alternative molecular generation models using Molecular Sets (MOSES) [42] as a benchmark. Our evaluation focused on various criteria: novelty, uniqueness, filters, and internal diversity. We utilized the molecules generated by inputting five selected proteins into the model for this assessment. The results, presented in Table 2, demonstrate the effectiveness of our proposed method.

6 Conclusion and Futurework

In this paper, we propose a targeted De Novo drug design strategy. Our approach leverages a reinforcement learning (RL) policy to generate compounds specifically tailored to target or interact with proteins. Given any target protein, the method can generate a targeted molecular compound. Through evaluations across diverse scenarios, we demonstrate the effectiveness of our approach. Our framework efficiently produces tailored chemical compounds for proteins while maintaining favorable chemical properties, such as Molecular Weight (MW), Quantitative Estimation of Drug-likeness (QED), and Octanol-Water Partition Coefficient (logP). Looking ahead, we plan to explore larger generative models and experiment with different reward functions to achieve even better outcomes.

References

[1] Karim Abbasi, Parvin Razzaghi, Antti Poso, Saber Ghanbari-Ara, and Ali Masoudi-Nejad. Deep learning in drug target interaction prediction: current and future perspectives. *Current Medicinal Chemistry*, 28(11):2100–2113, 2021.

[2] Alexander Aliper, Sergey Plis, Artem Artemov, Alvaro Ulloa, Polina Mamoshina, and Alex Zhavoronkov. Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data. *Molecular pharmaceutics*, 13(7):2524–2530, 2016.

[3] Mark A Ator, John P Mallamo, and Michael Williams. Overview of drug discovery and development. *Current Protocols in Pharmacology*, 35(1):9–9, 2006.

[4] Qifeng Bai, Shuo Liu, Yanan Tian, Tingyang Xu, Antonio Jesús Banegas-Luna, Horacio Pérez-Sánchez, Junzhou Huang, Huanxiang Liu, and Xiaojun Yao. Application advances of deep learning methods for de novo drug design and molecular dynamics simulation. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 12(3):e1581, 2022.

[5] Dávid Bajusz, Anita Rácz, and Károly Héberger. Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of cheminformatics*, 7:1–13, 2015.

[6] Protein Data Bank. Protein data bank. *Nature New Biol*, 233(223):10–1038, 1971.

[7] Nurken Berdigaliyev and Mohamad Aljofan. An overview of drug discovery and development. *Future medicinal chemistry*, 12(10):939–947, 2020.

[8] Thomas Blaschke, Josep Arús-Pous, Hongming Chen, Christian Margreitter, Christian Tyrchan, Ola Engkvist, Kostas Papadopoulos, and Atanas Patronov. Reinvent 2.0: an ai tool for de novo drug design. *Journal of chemical information and modeling*, 60(12):5918–5922, 2020.

[9] Rupayan Das, Angshuman Khan, and Gunjan Paul. A proximal policy optimization with curiosity algorithm for virtual drone navigation. *Engineering Research Express*, 6(1):015057, 2024.

[10] Mark Davies, Michał Nowotka, George Papadatos, Nathan Dedman, Anna Gaulton, Francis Atkinson, Louisa Bellis, and John P Overington. Chembl web services: streamlining access to drug discovery data and utilities. *Nucleic acids research*, 43(W1):W612–W620, 2015.

[11] Yifan Deng, Xinran Xu, Yang Qiu, Jingbo Xia, Wen Zhang, and Shichao Liu. A multimodal deep learning framework for predicting drug–drug interaction events. *Bioinformatics*, 36(15):4316–4322, 2020.

[12] Amol B Deore, Jayprabha R Dhumane, Rushikesh Wagh, and Rushikesh Sonawane. The stages of drug discovery and development process. *Asian Journal of Pharmaceutical Research and Development*, 7(6):62–67, 2019.

[13] Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. Translation between molecules and natural language. *arXiv preprint arXiv:2204.11817*, 2022.

[14] Qingyuan Feng, Evgenia Dueva, Artem Cherkasov, and Martin Ester. Padme: A deep learning-based framework for drug-target interaction prediction. *arXiv preprint arXiv:1807.09741*, 2018.

[15] Lalji K Gediya and Vincent CO Njar. Promise and challenges in drug discovery and development of hybrid anticancer drugs. *Expert opinion on drug discovery*, 4(11):1099–1111, 2009.

[16] Anvita Gupta, Alex T Müller, Berend JH Huisman, Jens A Fuchs, Petra Schneider, and Gisbert Schneider. Generative recurrent networks for de novo drug design. *Molecular informatics*, 37(1-2):1700111, 2018.

[17] Somaye Hashemifar, Behnam Neyshabur, Aly A Khan, and Jinbo Xu. Predicting protein–protein interactions through sequence-based deep learning. *Bioinformatics*, 34(17):i802–i810, 2018.

[18] Seyed Aghil Hooshmand, Mohadeseh Zarei Ghobadi, Seyyed Emad Hooshmand, Sadegh Azimzadeh Jamalkandi, Seyed Mehdi Alavi, and Ali Masoudi-Nejad. A multimodal deep learning-based drug repurposing approach for treatment of covid-19. *Molecular diversity*, 25:1717–1730, 2021.

[19] Xiaotian Hu, Cong Feng, Tianyi Ling, and Ming Chen. Deep learning frameworks for protein–protein interaction prediction. *Computational and Structural Biotechnology Journal*, 20:3223–3233, 2022.

[20] Kexin Huang, Tianfan Fu, Lucas M Glass, Marinka Zitnik, Cao Xiao, and Jimeng Sun. Deeppurpose: a deep learning library for drug–target interaction prediction. *Bioinformatics*, 36(22-23):5545–5547, 2020.

[21] James P Hughes, Stephen Rees, S Barrett Kalindjian, and Karen L Philpott. Principles of early drug discovery. *British journal of pharmacology*, 162(6):1239–1249, 2011.

[22] Naiem T Issa, Vasileios Stathias, Stephan Schürer, and Sivanesan Dakshanamurthy. Machine and deep learning approaches for cancer drug repurposing. In *Seminars in cancer biology*, volume 68, pages 132–142. Elsevier, 2021.

[23] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction tree variational autoencoder for molecular graph generation. In *International conference on machine learning*, pages 2323–2332. PMLR, 2018.

[24] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[25] Lukáš Klein, Ivan Zelinka, and David Seidl. Optimizing parameters in swarm intelligence using reinforcement learning: An application of proximal policy optimization to the isoma algorithm. *Swarm and Evolutionary Computation*, 85:101487, 2024.

[26] Sowmya Ramaswamy Krishnan, Navneet Bung, Gopalakrishnan Bulusu, and Arijit Roy. Accelerating de novo drug design against novel proteins using deep learning. *Journal of Chemical Information and Modeling*, 61(2):621–630, 2021.

[27] Sowmya Ramaswamy Krishnan, Navneet Bung, Sarveswara Rao Vangala, Rajgopal Srinivasan, Gopalakrishnan Bulusu, and Arijit Roy. De novo structure-based drug design using deep learning. *Journal of Chemical Information and Modeling*, 62(21):5100–5109, 2021.

[28] Prashant Kumar Shukla, Piyush Kumar Shukla, Poonam Sharma, Paresh Rawat, Jashwant Samar, Rahul Moriwal, and Manjit Kaur. Efficient prediction of drug–drug interaction using deep learning models. *IET Systems Biology*, 14(4):211–216, 2020.

[29] Greg Landrum. Rdkit documentation. *Release*, 1(1-79):4, 2013.

[30] Chun Yen Lee and Yi-Ping Phoebe Chen. New insights into drug repurposing for covid-19 using deep learning. *IEEE Transactions on Neural Networks and Learning Systems*, 32(11):4770–4780, 2021.

[31] Geonhee Lee, Chihyun Park, and Jaegyoon Ahn. Novel deep learning model for more accurate prediction of drug-drug interaction effects. *BMC bioinformatics*, 20:1–8, 2019.

[32] Minhyeok Lee. Recent advances in deep learning for protein-protein interaction analysis: A comprehensive review. *Molecules*, 28(13):5169, 2023.

[33] Christopher A Lipinski, F Lombardo, Beryl W Dominy, and Paul J Feeney. Cas: 528: Dc% 2bd3mxitvohs7o% 3d: Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. vol. 46, issue 1-3. *Adv Drug Deliv Rev*, pages 3–26, 2001.

[34] Tiqing Liu, Yuhmei Lin, Xin Wen, Robert N Jorissen, and Michael K Gilson. Bindingdb: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic acids research*, 35(suppl_1):D198–D201, 2007.

[35] Georg Lurje and Heinz-Josef Lenz. Egfr signaling and drug discovery. *Oncology*, 77(6):400–410, 2010.

[36] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

[37] Nelson RC Monteiro, Tiago O Pereira, Ana Catarina D Machado, José L Oliveira, Maryam Abbasi, and Joel P Arrais. Fsm-ddtr: End-to-end feedback strategy for multi-objective de novo drug design using transformers. *Computers in Biology and Medicine*, 164:107285, 2023.

[38] Ferruccio Palazzesi and Alfonso Pozzan. Deep learning applied to ligand-based de novo drug design. *Artificial intelligence in drug design*, pages 273–299, 2022.

[39] Xiaoqin Pan, Xuan Lin, Dongsheng Cao, Xiangxiang Zeng, Philip S Yu, Lifang He, Ruth Nussinov, and Feixiong Cheng. Deep learning for drug repurposing: Methods, databases, and applications. *Wiley interdisciplinary reviews: Computational molecular science*, 12(4):e1597, 2022.

[40] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[41] Yifan Peng and Zhiyong Lu. Deep learning for extracting protein-protein interactions from biomedical literature. *arXiv preprint arXiv:1706.01556*, 2017.

[42] Daniil Polykovskiy, Alexander Zhebrak, Benjamin Sanchez-Lengeling, Sergey Golovanov, Oktai Tatanov, Stanislav Belyaev, Rauf Kurbanov, Aleksey Artamonov, Vladimir Aladinskiy, Mark Veselov, Artur Kadurin, Simon Johansson, Hongming Chen, Sergey Nikolenko, Alan Aspuru-Guzik, and Alex Zhavoronkov. Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models. *Frontiers in Pharmacology*, 2020.

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

[44] Oleksii Prykhodko, Simon Viet Johansson, Panagiotis-Christos Kotsias, Josep Arús-Pous, Esben Jannik Bjerrum, Ola Engkvist, and Hongming Chen. A de novo molecular generation method using latent vector based generative adversarial network. *Journal of Cheminformatics*, 11:1–13, 2019.

[45] Yang Qiu, Yang Zhang, Yifan Deng, Shichao Liu, and Wen Zhang. A comprehensive review of computational methods for drug-drug interaction detection. *IEEE/ACM transactions on computational biology and bioinformatics*, 19(4):1968–1985, 2021.

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

[47] Marwin HS Segler, Thierry Kogej, Christian Tyrchan, and Mark P Waller. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS central science*, 4(1):120–131, 2018.

[48] Sandeep Sinha and Divya Vohora. Drug discovery and development: An overview. *Pharmaceutical medicine and translational clinical research*, pages 19–32, 2018.

[49] Tanlin Sun, Bo Zhou, Luhua Lai, and Jianfeng Pei. Sequence-based prediction of protein protein interaction using a deep-learning algorithm. *BMC bioinformatics*, 18:1–8, 2017.

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

[51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[52] Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, and Shengyi Huang. Trl: Transformer reinforcement learning. `https://github.com/huggingface/trl`, 2020.

[53] Mingyang Wang, Zhe Wang, Huiyong Sun, Jike Wang, Chao Shen, Gaoqi Weng, Xin Chai, Honglin Li, Dongsheng Cao, and Tingjun Hou. Deep learning approaches for de novo drug design: An overview. *Current opinion in structural biology*, 72:135–144, 2022.

[54] Yan-Bin Wang, Zhu-Hong You, Shan Yang, Hai-Cheng Yi, Zhan-Heng Chen, and Kai Zheng. A deep learning-based method for drug-target interaction prediction based on long short-term memory neural network. *BMC medical informatics and decision making*, 20:1–9, 2020.

[55] Ming Wen, Zhimin Zhang, Shaoyu Niu, Haozhi Sha, Ruihan Yang, Yonghuan Yun, and Hongmei Lu. Deep-learning-based drug–target interaction prediction. *Journal of proteome research*, 16(4):1401–1409, 2017.

[56] Shuo Yang and Gjergji Kasneci. Is crowdsourcing breaking your bank? cost-effective fine-tuning of pre-trained language models with proximal policy optimization. *arXiv preprint arXiv:2402.18284*, 2024.

[57] Her-Terng Yau, Ping-Huan Kuo, Po-Chien Luan, and Yung-Ruen Tseng. Proximal policy optimization-based controller for chaotic systems. *International Journal of Robust and Nonlinear Control*, 34(1):586–601, 2024.

[58] Jiaying You, Robert D McLeod, and Pingzhao Hu. Predicting drug-target interaction network using deep learning model. *Computational biology and chemistry*, 80:90–101, 2019.

[59] Tianlin Zhang, Jiaxu Leng, and Ying Liu. Deep learning for drug–drug interaction extraction from the literature: a review. *Briefings in bioinformatics*, 21(5):1609–1627, 2020.

[60] Yunjiang Zhang, Shuyuan Li, Miaojuan Xing, Qing Yuan, Hong He, and Shaorui Sun. Universal approach to de novo drug design for target proteins using deep reinforcement learning. *ACS omega*, 8(6):5464–5474, 2023.

A Exploring Similarity of Generated Molecules and Protein Inhibitors

We showcase further instances of molecules generated by inputting different proteins into the model. These molecules are matched with protein inhibitors sourced from the ChEMBL database. Compared to the inhibitors, their similarity is assessed using the Tanimoto similarity (TS) metric. These instances are visually represented in Figures 7 through 11.

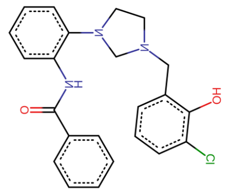
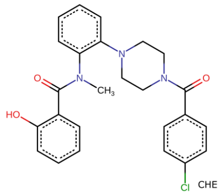
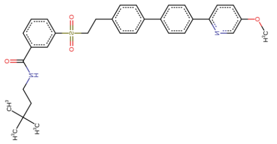
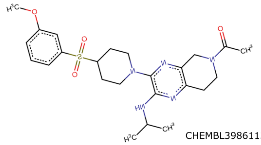
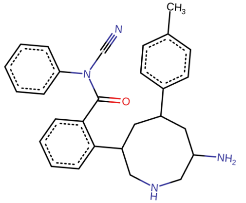
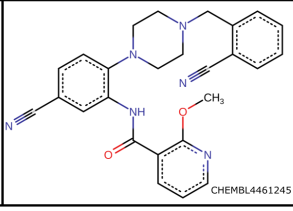
Generated Molecules	Reference Molecules	Tanmito Similarity
	 CHEMBL4450709	**0.6390**
	 CHEMBL3986113	**0.4053**
	 CHEMBL4461245	**0.4254**

Figure 7: Examples of molecules generated by inputting the G-protein coupled receptor 6 (GPR6) protein to the model, alongside samples of protein inhibitors, with the Tanimoto similarity (TS) measured between the generated and inhibitor compounds.

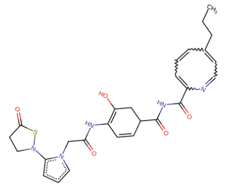
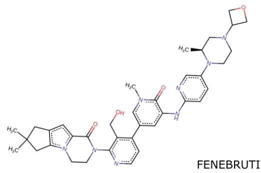
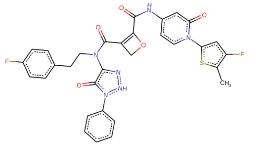
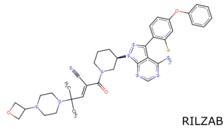
Generated Molecules	Reference Molecules	Tanmito Similarity
	 FENEBRUTINIB	**0.5823**
	 RILZABRUTINIB	**0.5975**

Figure 8: Examples of molecules generated by inputting the Bruton's tyrosine kinase (BTK) protein to the model, alongside samples of protein inhibitors, with the Tanimoto similarity (TS) measured between the generated and inhibitor compounds.

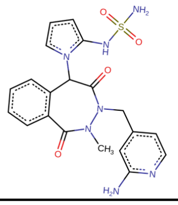
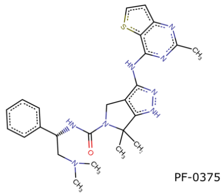
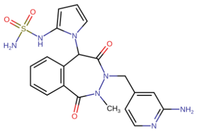
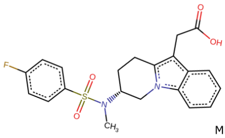
Generated Molecules	Reference Molecules	Tanmito Similarity
	 PF-03758309	**0.6390**
	 MK-7246	**0.5401**

Figure 9: Examples of molecules generated by inputting the v-Raf murine sarcoma viral oncogene homologue B (BRAF) protein to the model, alongside samples of protein inhibitors, with the Tanimoto similarity (TS) measured between the generated and inhibitor compounds.

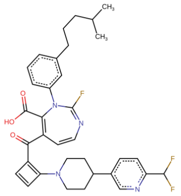
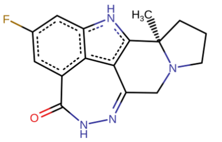
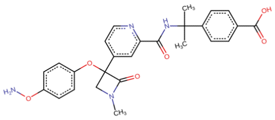
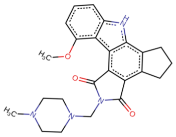
Generated Molecules	Reference Molecules	Tanmito Similarity
	 PAMIPARIB	**0.6058**
	 CEP-9722	**0.4815**

Figure 10: Examples of molecules generated by inputting the poly ADP-ribose polymerase (PARP) protein to the model, alongside samples of protein inhibitors, with the Tanimoto similarity (TS) measured between the generated and inhibitor compounds.

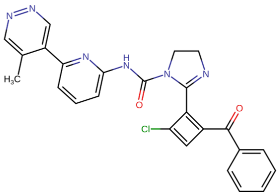
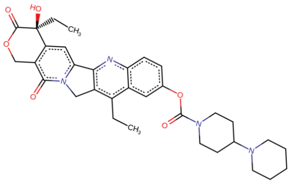
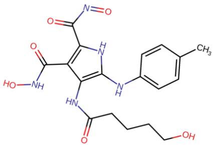
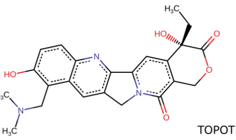
Generated Molecules	Reference Molecules	Tanmito Similarity
	 IRINOTECAN	**0.6400**
	 TOPOTECAN	**0.5520**

Figure 11: Examples of molecules generated by inputting the epidermal growth factor receptor (EGFR) protein to the model, alongside samples of protein inhibitors, with the Tanimoto similarity (TS) measured between the generated and inhibitor compounds.

B Visualizing Molecular Similarity with t-SNE

We employed molecular fingerprint descriptors to visually represent the generated molecules and protein inhibitors utilizing the Stochastic Neighbor Embedding (t-SNE) algorithm. This visualization highlights their similarity, further affirming the efficacy of our proposed method. Figures 12 to 14 elucidate these findings.

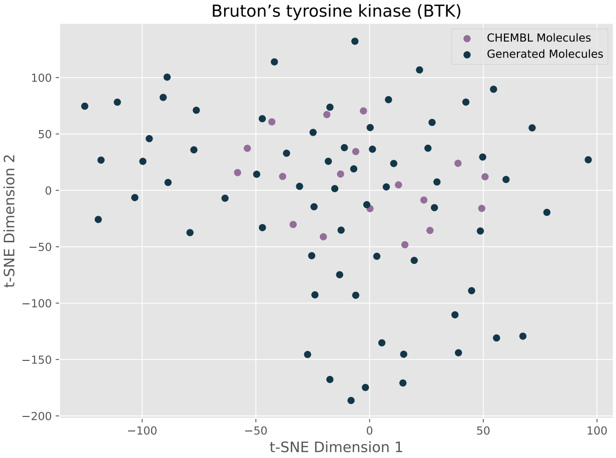


Figure 12: Visual representations using t-SNE projections illustrating the fingerprint descriptors of both generated and reference molecules for BTK.

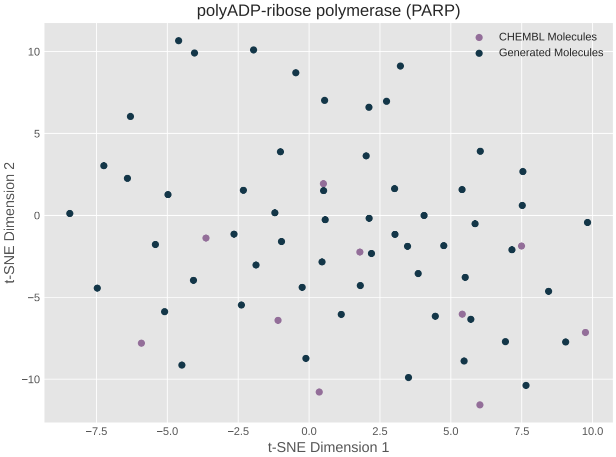


Figure 13: Visual representations using t-SNE projections illustrating the fingerprint descriptors of both generated and reference molecules for PARP.

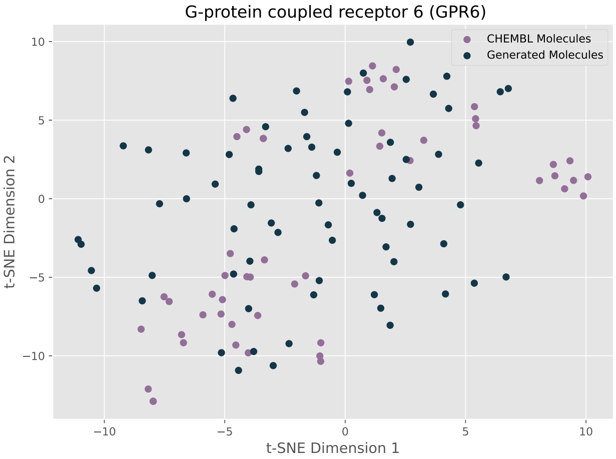


Figure 14: Visual representations using t-SNE projections illustrating the fingerprint descriptors of both generated and reference molecules for GPR6.