

INDIAN INSTITUTE OF TECHNOLOGY MADRAS

CONDITIONAL DE NOVO MOLECULAR
GENERATION

AUTHOR

SAICHARAN GANAPATHY

Roll No. ED19B065

SUPERVISOR

DR. NIRAV PRAVINBHAI BHATT

JUNE 2024

ACKNOWLEDGEMENT

I would like to extend my sincere thanks to Professor **Dr. Nirav Pravinbhai Bhatt** for his ongoing support and expertise in my project. His invaluable advice and guidance have been crucial, especially when facing challenges. His continual motivation has significantly contributed to my progress so far.

My gratitude also goes to my PhD student guide, **Roshan M S B**, for his consistent assistance and valuable insights at each step of this project. His contributions have been immensely beneficial to my work.

I am thankful to my **family** and **friends** for their unwavering support and encouragement during this journey. Their belief in me and their support have been a constant source of motivation throughout the course of this project.

Lastly, I would like to express my appreciation to **IIT Madras** and the support staff for providing the necessary resources, environment, and assistance for my research. The facilities and guidance offered by the university have been greatly supportive of my academic efforts.

ABSTRACT

This thesis investigates the application of deep learning techniques, specifically transformer-decoder models, in the realm of inverse molecular design, which holds considerable promise in the field of drug development. The study pivots on the innovative use of advanced natural language processing (NLP) models, adapting strategies from text generation to molecular structure generation. Central to our approach is the use of the SMILES (Simplified Molecular Input Line Entry System) notation, which enables the representation of molecules as sequences of characters, similar to textual data. This alignment allows for the application of techniques originally developed for language models, particularly those based on the Transformer architecture.

Our primary contribution lies in the development and training of a transformer-decoder model, drawing inspiration from the success of generative pre-training (GPT) models in text generation. This model is specifically tailored for the generation of drug-like molecules. A significant aspect of our work involves conditional training, where the model is trained to incorporate additional information such as molecular scaffolds, functional groups, and specific physicochemical properties. This approach enables the generation of molecules that not only resemble drugs but also meet predefined conditions set by the user.

The methodology employed includes advanced techniques such as next token prediction and masked self-attention, fundamental to the Transformer model's ability to handle sequential data effectively. The performance of our model is rigorously evaluated through a variety of metrics. These include the validity of the generated molecules, the Fréchet ChemNet Distance (a measure of similarity to known drug-like molecules), and internal diversity, which assesses the variety within the generated molecular structures.

The results of this study provide insights into the viability and effectiveness of using NLP-inspired models in the context of molecular design. By offering a novel tool that navigates the vast chemical space efficiently under specific conditions, this research could facilitate a more targeted and expedient approach to drug development. This work not only showcases the adaptability of text generation models for applications in chemistry but also sets the stage for future research in the integration of machine learning and molecular design for pharmaceutical advancements.

Keywords: Inverse Molecular Design, Transformer-Decoder Models, SMILES Notation, Drug Development, Natural Language Processing

CONTENTS

Contents	iii
1 Introduction	2
1.1 Context and Challenges in Drug Discovery	2
1.2 Innovative Approach Using Transformer-Decoder Models	2
2 Brief review of literature	4
2.1 MolGenSurvey: A Systematic Survey in Machine Learning Models for Molecule Design	4
2.2 Comparative Study of Deep Generative Models on Chemical Space Coverage	4
2.3 Generative Models as an Emerging Paradigm in the Chemical Sciences .	5
2.4 Searching for High-Value Molecules Using Reinforcement Learning and Transformers	5
2.5 MolGPT: Molecular Generation Using a Transformer-Decoder Model . .	6
2.6 Molecular Sets (MOSES): A Benchmarking Platform for Molecular Gener- ation Models	6
2.7 Sample Efficiency Matters: A Benchmark for Practical Molecular Opti- mization	7
2.8 Reinforced Self-Training (ReST) for Language Modeling	7
2.9 Domain-Agnostic Molecular Generation with Chemical Feedback	8
2.10 SCScore: Synthetic Complexity Learned from a Reaction Corpus	8
2.11 SYBA: Bayesian estimation of synthetic accessibility of organic compounds	9
2.12 Critical assessment of synthetic accessibility scores in computer-assisted synthesis planning	9
3 Datasets	10
3.1 Properties of the Datasets	10
3.1.1 MOSES	10
3.1.2 Guacamol	10
3.1.3 ChemBL	10
3.1.4 Zinc Datasets(250k, 1M, 10M, 270M, 37B)	11
3.1.5 PubChem	11
3.2 Dataset Statistics after Processing	12

3.3	Inter-Dataset Overlap	13
3.4	Data Representations	16
3.4.1	SMILES	16
3.4.2	SELFIES	16
3.4.3	DeepSMILES	17
3.4.4	SAFE	17
4	Methods	19
4.1	Tokenizers	19
4.1.1	Atomwise Tokenization	19
4.1.2	Kmer Tokenization	19
4.1.3	Byte Pair Encoding (BPE)	20
4.1.4	SMILES Pair Encoding (SMILESPE)	20
4.2	Model	20
4.2.1	Motivation for Transformer Architecture	21
4.2.2	Key Equations and Components	21
4.2.3	GPT-2 Architecture	21
4.2.4	Unique Qualities and Advantages	21
4.2.5	Relevance to Molecular Design	22
4.3	Pre-training	22
4.4	Pre-training Evaluation	23
4.4.1	Validity	23
4.4.2	Uniqueness	23
4.4.3	Novelty	23
4.4.4	Internal Diversity	24
4.4.5	Fréchet ChemNet Distance (FCD)	24
4.4.6	Lipophilicity (logP)	25
4.4.7	Penalized LogP	25
4.4.8	Quantitative Estimate of Drug-likeness (QED)	25
4.4.9	Synthetic Accessibility (SA) Score	26
4.4.10	Synthetic Complexity Score (SCScore)	26
4.4.11	SYnthetic Bayesian Accessibility (SYBA) Score	26
4.4.12	Retrosynthetic accessibility score (RAscore)	26
4.4.13	Fragment and Scaffold Similarity	27
4.4.14	Similarity to Nearest Neighbor (SNN)	27
4.4.15	Summary of Model Evaluation Metrics	29
4.4.16	Speed Enhancements	30
4.5	Downstream Fine-Tuning using ReST Framework	31
4.5.1	Key Components of ReST	31
4.5.2	Usefulness in Molecular Generation	31
4.5.3	Fine-Tuning	32

5	Results	33
5.1	Unconditioned Training	33
5.1.1	Training Curves	33
5.1.2	Generation Results	34
5.1.3	Performance on Evaluation Metrics	34
5.2	Conditioned Training - Scaffold	35
5.2.1	Training Curves	35
5.2.2	Generation Results	36
5.2.3	Performance on Evaluation Metrics	36
5.3	Conditioned Training - Properties	37
5.3.1	Training Curves	37
5.3.2	Generation Results	38
5.3.3	Performance on Evaluation Metrics	38
5.4	Conditioned Training - Scaffold, Properties	39
5.4.1	Training Curves	39
5.4.2	Generation Results	40
5.4.3	Performance on Evaluation Metrics	40
5.5	Unconditioned training using GuacaMol Dataset	41
5.5.1	Training Curves	41
5.5.2	Generation Results	42
5.5.3	Performance on Evaluation Metrics	42
5.5.4	Comparison - Effect of Dataset on Model Performance	43
5.6	Downstream Tasks - Alignment to QED	44
6	Conclusion and Future Work	47
7	Declaration of AI-assisted technologies in the writing process	48

INTRODUCTION

1.1 Context and Challenges in Drug Discovery

Drug discovery is an essential yet complex process in the pharmaceutical industry, characterized by high costs, extensive time requirements, and a reliance on traditional methodologies. The conventional approach predominantly involves screening vast libraries of compounds to identify potential drug candidates, a process that is both time-consuming and resource-intensive. Despite the significant investment in these methods, the success rate for finding effective and safe drugs remains relatively low. This challenge is further compounded by the ever-increasing complexity of diseases and the growing demand for more effective treatments.

The concept of inverse molecular design emerges as a novel approach in this context. It represents a paradigm shift from the traditional screening methods to a more proactive design of molecules. Inverse molecular design involves the creation of new molecules, tailored to fit specific therapeutic targets from the outset. However, this approach introduces a new challenge: navigating the vast and largely unexplored chemical space, which contains an innumerable number of potential molecular structures. This immense space presents a significant hurdle, as the manual exploration and design of molecules within it are practically unfeasible with current methodologies.

1.2 Innovative Approach Using Transformer-Decoder Models

The application of advanced computational techniques, particularly those inspired by the field of natural language processing (NLP), offers a promising solution to these challenges. This section introduces the use of transformer-decoder models, a groundbreaking adaptation from NLP, to the realm of molecular design. These models, which have shown remarkable success in text generation and understanding, are now being repurposed to address the complexities of chemical structure generation.

Central to this approach is the use of SMILES (Simplified Molecular Input Line Entry System) notation, which allows for the representation of chemical structures as sequences of characters. This notation enables the application of transformer-decoder

models to molecular design, treating chemical structures in a manner akin to linguistic sequences. The unique aspect of this methodology lies in its ability to generate novel molecular structures that are not just random assortments of atoms but are chemically valid and potentially efficacious as drug candidates.

Further, this research incorporates conditional training into the transformer-decoder models. This technique enables the models to generate molecules based on specified conditions, such as desired biological activity, molecular scaffolding, or pharmacokinetic properties. Such targeted molecule generation could be particularly transformative for personalized medicine, where treatments need to be tailored to individual patient profiles. The conclusion of this section underscores the potential of this research to significantly expedite the drug discovery process, reduce associated costs, and open new frontiers in the understanding and exploration of chemical space.

BRIEF REVIEW OF LITERATURE

2.1 MolGenSurvey: A Systematic Survey in Machine Learning Models for Molecule Design

The paper [Du et al., 2022](#) is a detailed exploration of machine learning applications in molecular design. It comprehensively covers various molecule representation methods, such as 1D strings, 2D graphs, and 3D geometries. These representations are crucial for different machine learning models to accurately interpret and generate molecular structures. The paper also systematically reviews generative models and combinatorial optimization methods used in molecular design. The methods described in the paper give insights into generating new molecules and optimizing their properties.

Additionally, the paper categorizes molecule design problems and outlines their setups, inputs, outputs, and objectives. This categorization is beneficial for understanding how different machine-learning techniques can be applied to specific molecular design tasks. The review’s focus on the broad spectrum of machine learning applications in molecular design, including challenges and future opportunities, offers valuable insights and context for our work.

2.2 Comparative Study of Deep Generative Models on Chemical Space Coverage

The paper [Zhang et al., 2021](#) proposes a novel metric for evaluating deep molecular generative models based on the chemical space coverage of a reference dataset, GDB-13. The performance of the models was compared by calculating what fraction of the structures, ring systems, and functional groups could be reproduced from the largely unseen reference set when using only a small fraction of GDB-13 for training. The results show that the performance of the generative models studied varies significantly using the benchmark metrics introduced herein, such that the generalization capabilities of the generative models can be clearly differentiated. The paper also discusses the validity and repetition rate of the sampled molecules and the analysis of the GDB-13 database. The models benchmarked in this study are recurrent neural networks (RNNs),

autoencoder (AE)-based networks, generative adversarial networks (GANs), and graph neural networks (GNNs). The paper provides a useful new metric that can be used for evaluating and comparing generative models.

2.3 Generative Models as an Emerging Paradigm in the Chemical Sciences

The paper [Anstine et al., 2023](#) highlights the limitations of traditional computational approaches to chemical species design, which are often limited by the need to compute properties for a vast number of candidates. In contrast, generative models aim to start from the desired property and optimize a corresponding chemical structure. The paper provides an overview of popular generative algorithms, including generative adversarial networks, variational autoencoders, flow, and diffusion models. It highlights key differences between each of the models and provides insights into recent success stories.

The authors also discuss outstanding challenges for realizing generative modeling discovered solutions in chemical applications. The paper emphasizes the potential of generative models in the chemical sciences, driven by the widespread adoption of machine learning and data-driven research, as well as advances in accelerated computational power and a well-developed software ecosystem of ML tools. The authors anticipate that generative models will be crucial for overcoming challenges across the chemical sciences, leading to a reallocation of human scientific creativity and accelerating the rate at which solutions to pressing issues are found.

2.4 Searching for High-Value Molecules Using Reinforcement Learning and Transformers

The study [Ghugare et al., 2023](#) presents ChemRLformer, an innovative RL-based algorithm for molecular design, exploring the effects of text representation and algorithmic training choices in reinforcement learning (RL). The research involved rigorous experimentation to understand how different text grammars and training methodologies impact the RL policy’s effectiveness in generating molecules with specific properties. ChemRLformer is analyzed across 25 molecular design tasks, including complex protein docking simulations, providing valuable insights into the molecular design problem space and demonstrating its superior performance compared to previous methods.

ChemRLformer’s development is guided by several key findings: using SMILES notation is more effective than SELFIES, the quality of pretraining molecules is crucial, and both transformer and RNN architectures exhibit comparable performance. The study also highlights the benefits of incorporating a hill-climb buffer and Log P regularization, while cautioning against the use of overly complex methods like KL regularization or intricate actor-critic algorithms, which may not yield proportional benefits. These insights

provide a roadmap for future molecular design efforts, emphasizing the importance of molecule quality metrics.

2.5 MolGPT: Molecular Generation Using a Transformer-Decoder Model

This paper [Bagal et al., 2022](#) presents technical details on the implementation and evaluation of the MolGPT model. MolGPT, based on the transformer-decoder architecture, is designed to process SMILES strings representing molecular structures. The model leverages a masked self-attention mechanism, enabling it to learn complex patterns in molecular data. The authors assess MolGPT’s performance by its ability to generate molecules that are not only valid and diverse but also adhere to specified chemical properties, demonstrating its potential for targeted molecular design.

Key experiments in the paper include assessing the model’s capacity to control multiple properties of the generated molecules, and using saliency maps to interpret the model’s decision-making process. These saliency maps provide insight into which parts of the input SMILES strings are most influential in determining the structure of the generated molecules. This interpretability is crucial for practical applications in drug discovery and material science, where understanding the rationale behind molecular design is essential. The study’s results show MolGPT’s effectiveness in generating molecules that meet specific criteria, marking a significant step in computational chemistry and molecular modeling.

2.6 Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models

The research paper [Polykovskiy et al., 2018](#) proposes a dataset and evaluates several baseline models for generating molecules. The dataset is based on the ZINC Clean Leads collection and contains 1,936,963 molecules with internal diversity of 0.857. The baseline models include character-level recurrent neural networks, variational autoencoders, adversarial autoencoders, junction tree variational autoencoders, and non-neural baselines. The models are evaluated based on several metrics, including validity, uniqueness, novelty, internal diversity, fragment and scaffold similarity, similarity to a nearest neighbor, and Fréchet ChemNet Distance. The results show that the neural network-based models successfully capture the statistics of the dataset, while the non-neural baselines fail to produce valid molecules. The study provides a useful benchmark for future research in generative models for molecules. Technical concepts highlighted in the paper include SMILES strings, Bemis-Murcko scaffolds, BRICS fragments, Morgan fingerprints, Kullback-Leibler divergence, Wasserstein-1 distance, and Fréchet ChemNet Distance.

2.7 Sample Efficiency Matters: A Benchmark for Practical Molecular Optimization

The paper [Gao et al., 2022](#) presents an exploration of key technical elements in molecular design, encompassing an array of string representations for molecules, including the Simplified Molecular-Input Line-Entry System (SMILES) and SELF-referencing Embedded Strings (SELFIES), as well as graph-based and synthesis-based strategies for molecular design. The paper also delves into a variety of optimization algorithms such as screening, genetic algorithms, Monte-Carlo Tree Search (MCTS), Bayesian optimization, variational autoencoders (VAEs), and reinforcement learning (RL) techniques. The study further examines the benchmark framework, which encompasses oracles, metrics, and the utilized dataset. The primary measure of efficacy is the area under the curve (AUC) of the top-K average property value in relation to the number of oracle calls (AUC top-K). An extensive analysis is conducted on the effectiveness of diverse molecular optimization methodologies, evaluated against the established metrics and oracles.

2.8 Reinforced Self-Training (ReST) for Language Modeling

The paper [Gulcehre et al., 2023](#) introduces Reinforced Self-Training (ReST), a novel approach in language modeling, particularly focusing on machine translation. ReST combines reinforcement learning from human feedback (RLHF) with large language models (LLMs) to align the model outputs more closely with human preferences. The process involves two distinct steps: Grow and Improve. In the Grow step, ReST generates a new dataset by sampling outputs from the current model policy. This is critical for expanding the range of data the model is exposed to. Subsequently, in the Improve step, the model undergoes fine-tuning using offline reinforcement learning algorithms. This step is designed to refine the model's performance based on the newly generated dataset.

The significance of ReST lies in its ability to improve translation quality significantly, which has been validated through both automated metrics and human evaluation. This method stands out for its efficient use of computational resources and sample usage. The results from the paper suggest that ReST can serve as a powerful tool in enhancing the alignment of language models with human preferences, thereby improving their efficacy in real-world applications. This methodology could potentially revolutionize the way machine translation and other language processing tasks are approached, offering a more nuanced and human-aligned performance.

2.9 Domain-Agnostic Molecular Generation with Chemical Feedback

In this paper, Fang et al., 2024 introduced MOLGEN, a novel pre-trained molecular language model. This model stands out due to its domain-agnostic pre-training approach and its integration of a chemical feedback paradigm. This framework helps ensure the generation of chemically valid molecules that are structurally sound and exhibit expected chemical activity.

MOLGEN employs the SELFIES (Self-referencing Embedded Strings) molecular language, which guarantees chemically sound molecular structures, unlike the traditional SMILES notation which can lead to syntactically incorrect strings. It also introduces the "Chemical Feedback Paradigm", which addresses the issue of "molecular hallucinations," where structurally correct molecules generated by the model do not exhibit the anticipated chemical properties. By aligning the model's generative probabilities with real-world chemical preferences, MOLGEN can effectively rectify its outputs, enhancing both their chemical validity and practical utility.

For both targeted molecule discovery and constrained molecular optimization tasks, chemical feedback paradigm was employed to align the PLM with the optimization objectives. In the molecule discovery experiments, MOLGEN set new benchmarks by optimizing properties like penalized logP and QED (quantitative estimate of drug-likeness).

2.10 SCScore: Synthetic Complexity Learned from a Reaction Corpus

This study by Coley et al., 2018 introduces the Synthetic Complexity Score (SCScore), a pioneering metric developed to evaluate the synthetic complexity of molecules. This metric is derived through a neural network that has been trained on a vast dataset from the Reaxys database, comprising over 12 million reactions. The SCScore quantifies synthetic complexity based on the number of steps required from standard starting materials. A key feature of the SCScore is its reaction-centric evaluation methodology, which scores molecules within the context of their synthetic pathways rather than in isolation.

The model underpinning the SCScore employs molecular fingerprints and is trained using a hinge loss function, ensuring that products are consistently evaluated as more complex than their reactants. The efficacy of the model was tested using a specific subset from the Reaxys database, validating its ability to accurately differentiate between the complexities of reactants and products. By grounding the model in such a robust dataset, the SCScore reduces the subjective biases often associated with expert evaluations, thus providing a scalable and objective metric.

2.11 SYBA: Bayesian estimation of synthetic accessibility of organic compounds

Voršilák et al., 2020 introduced the SYBA (SYnthetic Bayesian Accessibility) model, an innovative fragment-based scoring system for rapidly classifying organic compounds as either easy-to-synthesize (ES) or hard-to-synthesize (HS). Unlike traditional complexity-based metrics that often misjudge the synthetic accessibility based on structural complexity alone, SYBA utilizes a Bernoulli naïve Bayes classifier that computes synthetic accessibility based on the presence and absence of molecular fragments.

SYBA showed an improvement over random forest classification and outperformed SAScore and SCScore with default threshold settings. For instance, SYBA achieved a classification accuracy of 0.844 and an area under the ROC curve (AUC) of 0.903 when compared against manually curated test sets. However, SYBA's assumption that molecular fragments are independent can be a simplification that doesn't always hold true, potentially affecting the accuracy in more complex molecular structures. While SYBA performs well with default thresholds, the performance of other methods like SAScore significantly improves upon threshold optimization, suggesting that SYBA might face competition if other methods are optimally tuned.

2.12 Critical assessment of synthetic accessibility scores in computer-assisted synthesis planning

In this paper, Skoraczynski et al., 2023 perform an analysis of the effectiveness of synthetic accessibility scores (SAscore, SYBA, SCScore, and RAscore) in predicting and enhancing the efficiency of computer-assisted synthesis planning (CASP). It critically assesses whether these scores can reliably predict the outcomes of retrosynthesis planning, and if they can enhance the efficiency of retrosynthesis by prioritizing viable synthetic routes, thereby reducing the search space. The study employs AiZynthFinder, a CASP tool, to test the accuracy of these scores in predicting the feasibility of chemical syntheses.

Results indicate that synthetic accessibility scores are generally effective in discriminating between feasible and infeasible molecules, with implications for improving the rapidity and precision of CASP tools. The study reveals that synthetic accessibility scores, particularly SAscore and RAscore, effectively predict synthesizability with high accuracy. SAscore and RAscore demonstrate AUC values of 0.90 and 0.85, respectively, and accuracy rates of 0.81 and 0.85, suggesting that these models can significantly influence the prioritization process within CASP tools. However, the effectiveness of integrating these scores into the selection process within AiZynthFinder did not markedly improve performance, indicating that further refinement and integration methods might be necessary.

DATASETS

3.1 Properties of the Datasets

3.1.1 MOSES

The MOSES dataset [Polykovskiy et al., 2018](#) is a curated collection from the ZINC database, focusing specifically on the ZINC Clean Leads collection. It comprises 4,591,276 molecules, each selected based on specific criteria: a molecular weight between 250 and 350 Daltons, no more than 7 rotatable bonds, and an XlogP value of 3.5 or less. The dataset excludes molecules with charged atoms or atoms other than C, N, S, O, F, Cl, Br, and H. It also omits molecules with cycles longer than 8 atoms. Additionally, the selection process involved the application of medicinal chemistry filters (MCFs) and PAINS filters, ensuring the dataset’s relevance for benchmarking in medicinal chemistry and drug discovery.

3.1.2 Guacamol

The GuacaMol dataset [Brown et al., 2019](#) is derived from the ChEMBL 24 database, known for its synthesized and biologically tested molecules. This dataset offers a more realistic representation of drug-like molecules compared to others like ZINC or QM9. The refining process includes removing salts, neutralizing charges, excluding molecules with overly long SMILES strings or less frequently occurring elements, and filtering based on similarity to a set of known drugs. The result is a dataset tailored for benchmarking in drug discovery, available for download with reproducible creation through a provided docker container.

3.1.3 ChemBL

The ChEMBL database [Zdrazil et al., 2023](#) is a comprehensive resource for drug discovery, offering detailed bioactivity data, chemical structures, and target information for a wide range of drug-like compounds. It includes quantitative measurements such as IC50 and EC50, data on approved drugs, and is regularly updated. Widely accessible

to researchers, ChEMBL is invaluable for medicinal chemistry and pharmacological research.

3.1.4 Zinc Datasets(250k, 1M, 10M, 270M, 37B)

The ZINC database [Tingle et al., 2023](#) is a comprehensive collection of commercially available chemical compounds for virtual screening and drug discovery. ZINC-22 is a vast database of small molecules for ligand discovery, featuring a user-friendly interface, CartBlanche, for efficient analog searching. It efficiently handles the vast chemical space by using scalable search methods and rapid data access techniques. Despite its rapid growth, ZINC-22 continues to show increasing chemical diversity, particularly in complex compounds. The database, anticipating expansion to over a trillion molecules, is freely accessible online and is pivotal for future molecule docking and discovery.

3.1.5 PubChem

PubChem [Kim et al., 2023](#) is a widely-used public repository for chemical molecules and their activities, primarily aimed at supporting the fields of drug discovery and chemical biology. This extensive database is maintained by the National Center for Biotechnology Information (NCBI) and features a vast collection of chemical compounds, substances, and bioactivity data.

3.2 Dataset Statistics after Processing

The initial step in processing the datasets involved cleansing them to eliminate any redundant records found within each dataset. Subsequently, we implemented the standardization of the SMILES (Simplified Molecular Input Line Entry System) strings. Additionally, a new column was introduced, displaying the SELFIES (Self-referencing Embedded Strings) corresponding to each molecule. In the final phase of data preparation, we removed all extraneous columns from the dataset that were not pertinent to our analysis. The statistics of the datasets is highlighted in 3.1 and Table 3.2, and visualised in Figure 3.1, 3.2 and 3.3.

Dataset	Number of Rows	File Size
zinc_250k	249,455	18.18 MB
zinc_1m	999,998	72.13 MB
moses	1,936,962	93.42 MB
guacamol	1,591,011	140.94 MB
chembl	2,066,232	189.25 MB
zinc_10m	9,999,971	722.37 MB
pubchem	114,850,452	2.54 GB
zinc_270m	269,536,671	12.5 GB

Table 3.1: *Datasets Information Sorted by File Size*

3.3 Inter-Dataset Overlap

Dataset Pair	Overlap (Absolute)	Overlap (%)
moses & guacamol	71,675	2.07%
moses & zinc_250k	13,907	0.64%
moses & zinc_1m	1,868	0.06%
moses & zinc_10m	18,579	0.16%
moses & chembl	73,316	1.87%
guacamol & zinc_250k	2,528	0.14%
guacamol & zinc_1m	204	0.01%
guacamol & zinc_10m	2,135	0.02%
guacamol & chembl	1,093,236	42.64%
zinc_250k & zinc_1m	82	0.01%
zinc_250k & zinc_10m	908	0.01%
zinc_250k & chembl	3,162	0.14%
zinc_1m & zinc_10m	966,612	9.63%
zinc_1m & chembl	285	0.01%
zinc_10m & chembl	2944	0.02%

Table 3.2: Overlap of SMILES Data between Various Datasets

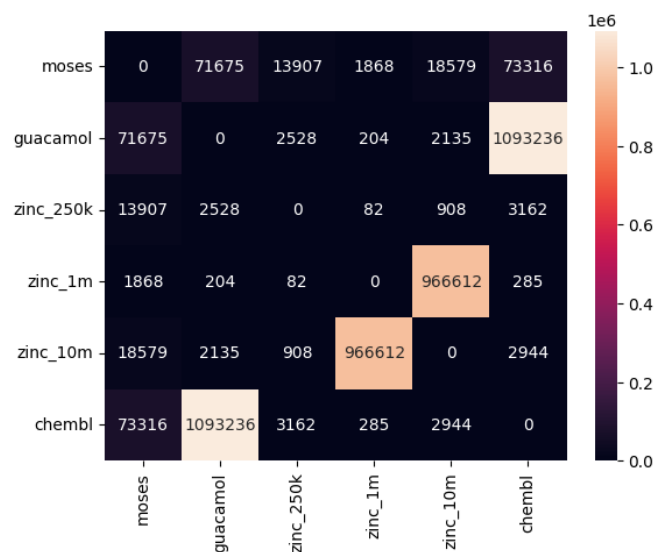


Figure 3.1: Heatmap showing the total overlap of SMILES strings between all datasets.

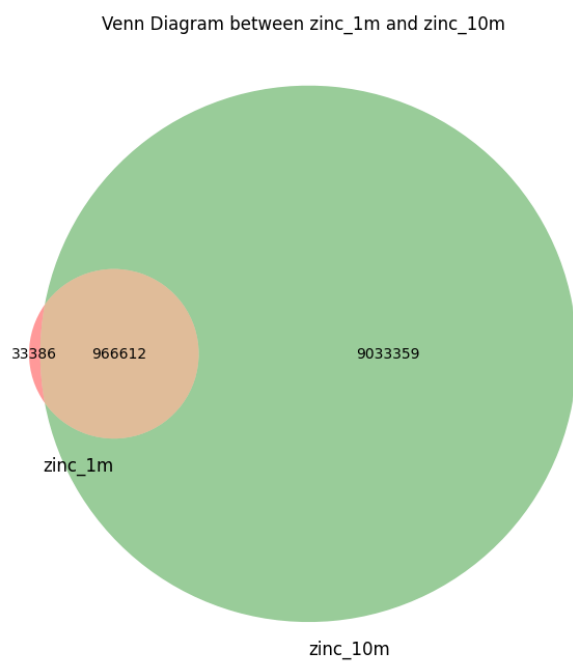


Figure 3.2: Venn diagram showing the absolute overlap of SMILES strings between Zinc 1M and Zinc 10M datasets.

Venn Diagram between chembl, guacamol, and moles

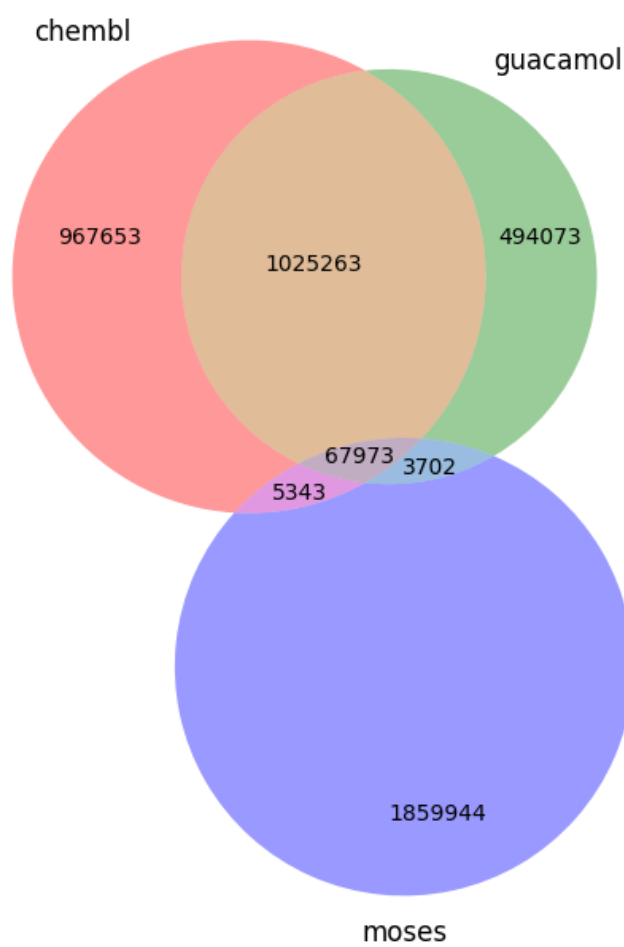


Figure 3.3: Venn diagram showing the absolute overlap of SMILES strings between GuacaMol and ChemBL datasets.

3.4 Data Representations

The choice of molecular representation significantly impacts the effectiveness of machine learning models in molecular generation. Various representations capture different aspects of molecular structures, and each has its advantages and disadvantages. Below, we describe the four primary types of molecular representations used in this project.

3.4.1 SMILES

The Simplified Molecular Input Line Entry System (SMILES) [Nath et al., 2021](#) is a widely used notation that encodes molecular structures as linear strings. Each string uniquely represents a molecule, detailing atoms and the bonds between them. SMILES strings are compact and human-readable, making them convenient for various computational applications.

Advantages:

- Human-readable and compact.
- Widely supported and used in cheminformatics tools and databases.
- Efficient for storing and processing large datasets.

Disadvantages:

- Sensitive to syntax errors; small mistakes can lead to invalid structures.
- Multiple valid SMILES representations for the same molecule can complicate model training and evaluation.

Example: For benzene, the SMILES representation is:

`"c1ccccc1"`

3.4.2 SELFIES

SELFIES (Self-Referencing Embedded Strings) [Krenn et al., 2020](#) are a more robust alternative to SMILES, designed to ensure that every possible string is a valid molecule. SELFIES address the syntactical issues inherent in SMILES by using a different encoding mechanism that inherently guarantees valid chemical structures.

Advantages:

- Ensures syntactic validity of all generated strings.
- Can be converted to and from SMILES without loss of information.

Disadvantages:

- Not as human-readable as SMILES.
- Slightly longer strings compared to SMILES.

Example: For benzene, the SELFIES representation might be:

"[[C][=C][C][=C][C][=C][Ring1][=Branch1]"

3.4.3 DeepSMILES

DeepSMILES [O’Boyle et al., 2018](#) modifies the SMILES notation to improve its suitability for machine learning applications by addressing common syntactical issues. It simplifies the representation by using a single character for ring closures and postfix notation for branches, thus avoiding unbalanced parentheses and unmatched ring closure digits.

Advantages:

- Reduces common syntactical errors found in SMILES.
- Maintains compactness while being more suitable for machine learning models.

Disadvantages:

- Requires understanding of the modified notation.
- Conversion to and from traditional SMILES is necessary for compatibility with existing tools.

Example: For benzene, the DeepSMILES representation is:

"cccccc6"

3.4.4 SAFE

SAFE (Structure-Aware Fragment Embeddings) [Noutahi et al., 2024](#) is a newer approach that encodes molecular structures by considering both atomic and substructural information. This representation aims to capture more chemical context and improve the quality of generated molecules.

Advantages:

- Encodes both atomic and substructural information, capturing more chemical context.
- Potentially improves the quality and validity of generated molecules.

Disadvantages:

- More complex encoding process.
- Requires specific decoding mechanisms to revert to standard representations.

Example: For benzene, the SAFE representation would use a structure-aware encoding that is not as straightforward as a simple string but would involve more detailed chemical context capturing fragments and their relationships.

Tokenization and data representation are pivotal in molecular generation projects as they determine how chemical information is fed into models. The choice of representation affects the model's ability to learn and generate novel, valid, and drug-like molecules effectively.

METHODS

4.1 Tokenizers

Tokenization is a crucial step in the preprocessing of molecular data for generative models. In the context of molecular generation, tokenization involves converting the chemical structure of a molecule, typically represented by a SMILES (Simplified Molecular Input Line Entry System) string, into a sequence of tokens that can be processed by deep learning models. Effective tokenization can capture the chemical structure and properties of molecules, thereby improving the performance of generative models. Different tokenization methods can impact the efficiency and accuracy of the model, as they determine how the molecular information is represented.

4.1.1 Atomwise Tokenization

Atomwise tokenization involves breaking down a SMILES string into individual atoms and special characters, which represent bonds and branching. Each character in the SMILES string is treated as a separate token. For example, the SMILES string for benzene, "c1ccccc1", would be tokenized as:

`['c', '1', 'c', 'c', 'c', 'c', 'c', '1']`

This method captures the atomic structure of the molecule but might lose some information about the connectivity and substructures.

4.1.2 Kmer Tokenization

Kmer tokenization breaks down the SMILES string into overlapping substrings of length k . This method captures more contextual information than atomwise tokenization. For example, using a k-mer length of 2 (bi-gram) and 3 (tri-gram) for the SMILES string "c1ccccc1", the tokenized output would be:

`['c1', '1c', 'cc', 'cc', 'cc', 'c1']`

For a k-mer length of 3 (tri-gram):

$$['c1c', '1cc', 'ccc', 'ccc', 'cc1']$$

Kmer tokenization helps in preserving local structures and patterns within the molecule.

4.1.3 Byte Pair Encoding (BPE)

Byte Pair Encoding (BPE) [Gallé, 2019](#) is a subword tokenization method that iteratively merges the most frequent pairs of characters or substrings. This method effectively compresses the SMILES string into fewer tokens, capturing more information per token. BPE can capture common substructures and recurring patterns in the molecular data. For example, the SMILES string "c1cccc1" might be tokenized by BPE as:

$$['c', '1', 'cccc', '1']$$

This result depends on the dataset used to train the tokenizer, as the most commonly occurring patterns can vary between different datasets.

4.1.4 SMILES Pair Encoding (SMILESPE)

SMILES Pair Encoding (SMILESPE) [X. Li et al., 2021](#) is inspired by the byte-pair-encoding (BPE) tokenization and is a variant of BPE specifically designed for tokenizing SMILES strings. It merges pairs of characters or substrings based on their frequency in a training corpus of SMILES strings. SMILESPE is designed to optimize the tokenization process for chemical structures, capturing both common substructures and chemical motifs.

4.2 Model

In our work, we use an adaptation of the Generative Pre-Training-2 (GPT-2) Transformer, with 345M parameters. It features an architecture of stacked decoder blocks, each containing a masked self-attention layer and a fully connected neural network. The self-attention layers produce 256-sized vectors, processed by the neural network with a hidden layer outputting 1024-sized vectors, followed by a GELU activation layer. The final output of each block is a 256-sized vector, fed into the subsequent decoder block, with a total of eight such blocks in the model. The model assigns position value embeddings to track input sequence order and uses separate embeddings for condition and SMILES tokens during conditional training, distinguishing between the two. These embeddings are combined into a 256-dimensional vector for each token in the SMILES string, which then serves as the model's input. This architecture enables MolGPT to efficiently process and generate molecular structures.

4.2.1 Motivation for Transformer Architecture

The transformer model, introduced by Vaswani et al., 2017 in their seminal paper *Attention Is All You Need*, addressed limitations in sequence-to-sequence models dependent on recurrent neural networks (RNNs) and convolutional neural networks (CNNs). The transformer model overcomes issues like vanishing gradients and inefficient parallelization through a novel self-attention mechanism, enabling simultaneous processing of input data sequences.

4.2.2 Key Equations and Components

Self-Attention Mechanism

The self-attention mechanism in the transformer model allows each position in the encoder to attend to all positions in the previous layer of the encoder. It is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (4.1)$$

where Q , K , and V are the query, key, and value matrices, respectively, and d_k is the dimension of the key vectors.

Multi-Head Attention

Multi-head attention allows the model to jointly attend to information from different representation subspaces:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O \quad (4.2)$$

where each head is defined as:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \quad (4.3)$$

and where the projections are parameter matrices $W_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$, and $W^O \in \mathbb{R}^{hd_v \times d_{\text{model}}}$.

4.2.3 GPT-2 Architecture

Developed by OpenAI, Generative Pre-trained Transformer 2 (GPT-2) Radford et al., 2019 builds upon the transformer architecture for text generation. It features an autoregressive model trained on a large corpus of text, enabling it to generate coherent and contextually relevant sequences.

4.2.4 Unique Qualities and Advantages

- **Parallel Processing:** The transformer model processes elements of input data simultaneously, leading to efficient training.
- **Handling of Long-Range Dependencies:** Through self-attention, the model effectively captures dependencies, regardless of their position in the input sequence.
- **Versatility:** Adaptable to a wide range of tasks beyond NLP.

4.2.5 Relevance to Molecular Design

In molecular design, the transformer’s ability to process sequential data and its powerful attention mechanism make it ideal for handling SMILES notation. GPT-2’s autoregressive nature (Radford et al., 2019) and extensive pre-training allow for effective generation of novel molecular structures, crucial for exploring the vast chemical space.

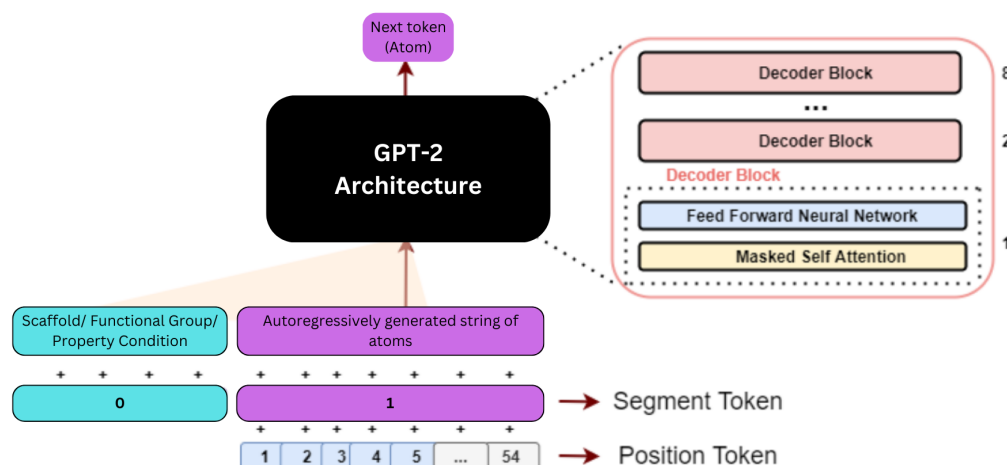


Figure 4.1: Model Architecture (Bagal et al., 2022)

4.3 Pre-training

Each model underwent training for 20 epochs using the Adam optimizer with a learning rate set at $6e-4$. The use of the Adam optimizer was chosen for its effectiveness in computational efficiency and fast convergence. The learning rate was determined to offer an optimal balance between convergence speed and accuracy.

The model (as shown in Figure 4.1) was both trained and tested using the MOSES bench-marking dataset Polykovskiy et al., 2018 to provide a comprehensive basis for evaluating the model’s performance.

For the generation process, the approach involved the use of a start token, selected randomly from the initial tokens of molecules in the training dataset. This strategy was employed to ensure that the generation of molecular structures began from a diverse range of starting points, reflecting the variability in the dataset.

Additionally, experiments were conducted to test the model’s ability to control molecular properties as well as the model’s effectiveness in generating desired scaffold and functional group structures within molecules. The training of these models was performed using an NVIDIA GeForce GTX 1080 Ti graphics card.

4.4 Pre-training Evaluation

4.4.1 Validity

The validity metric is crucial in the evaluation of de-novo molecular generation models as it measures the proportion of generated molecules that are chemically valid. A molecule is considered valid if it adheres to chemical rules and can be successfully parsed and sanitized by cheminformatics tools like RDKit. In our approach, the validity is calculated by converting each generated SMILES string to an RDKit molecule object and checking if the conversion is successful. The formula for validity can be expressed as:

$$\text{Validity} = \frac{N_{\text{valid}}}{N_{\text{total}}} \quad (4.4)$$

where N_{valid} is the number of valid molecules and N_{total} is the total number of generated molecules. A higher validity score indicates that the model generates more chemically plausible molecules, which is essential for further downstream tasks and practical applications in drug discovery. Typically, a validity score close to 1 is desirable, indicating that nearly all generated molecules are valid.

4.4.2 Uniqueness

The uniqueness metric evaluates the diversity of the generated molecules by measuring the proportion of unique molecules within the set of generated SMILES strings. This metric ensures that the model does not produce redundant or identical molecules, which is important for discovering novel chemical compounds. To calculate uniqueness, we first canonicalize the SMILES strings to a standard form and then count the number of distinct canonical SMILES. The formula for uniqueness is given by:

$$\text{Uniqueness} = \frac{N_{\text{unique}}}{N_{\text{total}}} \quad (4.5)$$

where N_{unique} is the number of unique canonical SMILES and N_{total} is the total number of generated molecules. A higher uniqueness score indicates a greater diversity in the generated molecules, which is desirable for exploring a wide chemical space. Typically, a uniqueness score close to 1 is desirable, indicating that most of the generated molecules are unique.

4.4.3 Novelty

The novelty metric assesses the ability of the model to generate new and previously unseen molecules by comparing the generated molecules to a set of molecules from the training data. Novelty is crucial for de-novo molecular generation as it measures the extent to which the model can explore new areas of the chemical space beyond the training data. To calculate novelty, we first canonicalize the SMILES strings of both

the generated molecules and the training set. The novelty score is then determined by identifying the proportion of generated molecules that are not present in the training set. The formula for novelty is:

$$\text{Novelty} = \frac{|\text{gen_smiles_set} \setminus \text{train_set}|}{|\text{gen_smiles_set}|} \quad (4.6)$$

where $|\text{gen_smiles_set} \setminus \text{train_set}|$ represents the number of unique generated molecules not found in the training set, and $|\text{gen_smiles_set}|$ is the total number of unique generated molecules. A higher novelty score indicates that the model is capable of generating a larger proportion of new molecules, which is essential for innovation in drug discovery. A novelty score close to 1 is desirable, indicating that most of the generated molecules are novel compared to the training data.

4.4.4 Internal Diversity

The internal diversity metric assesses the variability among the generated molecules by measuring the pairwise dissimilarity between them. This metric is important for evaluating whether the generated molecules cover a broad chemical space, which is desirable for discovering a wide range of novel compounds. Internal diversity is calculated using the Tanimoto similarity between the fingerprint representations of the generated molecules. The internal diversity score is given by:

$$\text{Internal Diversity} = 1 - \frac{1}{|G|^2} \sum_{x,y \in G} \text{Tanimoto}(x, y) \quad (4.7)$$

where $|G|$ is the number of generated molecules, and $\text{Tanimoto}(x, y)$ is the Tanimoto similarity between the fingerprints of molecules x and y . A higher internal diversity score indicates a greater variety among the generated molecules, which is beneficial for exploring different chemical structures. Typically, an internal diversity score close to 1 is desirable, indicating that the generated molecules are highly diverse.

4.4.5 Fréchet ChemNet Distance (FCD)

The Fréchet ChemNet Distance (FCD) [Preuer et al., 2019](#) is a metric used to compare the distribution of generated molecules to that of a reference set, typically the training set. FCD evaluates the chemical realism and diversity of the generated molecules by computing the Fréchet Distance between feature vectors obtained from a pre-trained neural network, such as ChemNet. This metric is analogous to the Fréchet Inception Distance (FID) used in image generation tasks. The FCD score is given by:

$$\text{FCD} = \|\mu_g - \mu_t\|^2 + \text{Tr}(\Sigma_g + \Sigma_t - 2(\Sigma_g \Sigma_t)^{1/2}) \quad (4.8)$$

where μ_g and μ_t are the mean feature vectors of the generated and training sets, respectively, and Σ_g and Σ_t are their corresponding covariance matrices. A lower FCD score indicates a higher similarity between the generated and training set distributions,

suggesting that the generated molecules are more chemically realistic and diverse. Typically, a lower FCD score is desirable, indicating that the model generates molecules with a distribution close to the real molecules in the training set.

4.4.6 Lipophilicity (logP)

[Mannhold et al., 2009](#) The LogP metric measures the hydrophobicity of molecules, defined as the logarithm of the partition coefficient between octanol and water. It is a crucial property in drug discovery, as it affects the absorption, distribution, metabolism, and excretion (ADME) of a molecule. For a small molecule drug to be a candidate for oral administration, the LogP value typically should be between 0 and 5. The LogP value is computed using RDKit’s Crippen estimation method. A higher LogP value generally indicates higher hydrophobicity. For drug-like molecules, a LogP value within the range of 0 to 5 is considered optimal, balancing solubility and permeability.

4.4.7 Penalized LogP

The penalized LogP metric combines the hydrophobicity of a molecule with penalties for synthetic accessibility (SA) and the presence of long cycles, providing a more comprehensive evaluation of the molecule’s drug-likeness. Penalized LogP is calculated as the LogP value minus the SA score and a penalty for the number of long cycles. This metric helps identify molecules that are not only hydrophobic but also synthetically feasible and structurally desirable. The formula for penalized LogP is given by:

$$\text{Penalized LogP} = \text{LogP} - \text{SA score} - \text{Long Cycle Penalty} \quad (4.9)$$

where LogP is the hydrophobicity measure, SA score evaluates the ease of synthetic accessibility, and Long Cycle Penalty accounts for the presence of large ring structures which are typically less drug-like. A higher penalized LogP score indicates that the molecule has a good balance of hydrophobicity, synthetic accessibility, and structural properties, making it a better candidate for drug development.

4.4.8 Quantitative Estimate of Drug-likeness (QED)

The Quantitative Estimate of Drug-likeness (QED) [Bickerton et al., 2012](#) metric evaluates how likely a molecule is to be a viable candidate for a drug based on certain desirable traits that successful drug molecules tend to possess. The QED score ranges from 0 to 1, where a score closer to 1 indicates a higher likelihood that the molecule is drug-like. The QED score combines multiple molecular properties such as molecular weight, lipophilicity (LogP), polar surface area, and the number of hydrogen bond donors and acceptors, among others. A higher average QED score indicates that the generated molecules possess characteristics commonly associated with successful drugs, making them more promising candidates for further development.

4.4.9 Synthetic Accessibility (SA) Score

The Synthetic Accessibility (SA) score [Murthy, 2021](#) estimates the ease of synthesizing a molecule, ranging from 1 (easy) to 10 (hard). It combines fragment contributions, derived from a large dataset of molecules in PubChem, with penalties for molecular complexity, including large rings, non-standard ring fusions, stereocomplexity, and molecule size. The SA score helps identify molecules that are not only drug-like but also synthetically feasible. A lower average SA score is preferable, indicating that the molecules are easier to synthesize.

4.4.10 Synthetic Complexity Score (SCScore)

The Synthetic Complexity Score (SCScore) [Coley et al., 2018](#) rates the synthetic complexity of molecules on a scale from 1 to 5, where a higher score indicates greater synthetic complexity. This metric is based on the premise that, on average, the products of published chemical reactions are more synthetically complex than their corresponding reactants. The SCScore helps in evaluating how challenging it would be to synthesize a molecule, considering both its structural features and the likelihood of successful synthesis. A lower average SCScore is preferable, indicating that the molecules are less synthetically complex and therefore easier to synthesize.

4.4.11 SYnthetic Bayesian Accessibility (SYBA) Score

The SYnthetic Bayesian Accessibility (SYBA) score [Voršilák et al., 2020](#) is a fragment-based method used to rapidly classify organic compounds as easy-to-synthesize (ES) or hard-to-synthesize (HS). This score is based on a Bernoulli naïve Bayes classifier that assigns SYBA score contributions to individual fragments based on their frequencies in databases of ES and HS molecules. The SYBA score is trained on ES molecules from the ZINC15 database and HS molecules generated by the Nonpher methodology. The SYBA score provides a measure of the synthetic accessibility of a molecule, with a higher score indicating easier synthesis. A higher average SYBA score indicates that the generated molecules are generally easier to synthesize.

4.4.12 Retrosynthetic accessibility score (RAscore)

The RAscore metric [Thakkar et al., 2021](#) evaluates the synthetic accessibility of molecules using a machine learning model trained on a large dataset of reaction data. This score predicts the ease of synthesizing a molecule based on its structure, with the aim of identifying compounds that are more feasible to produce in a laboratory setting. The RAscore ranges from 0 to 1, where a higher score indicates easier synthesis. A higher average RAscore suggests that the generated molecules are easier to synthesize, making them more suitable for practical applications in drug discovery and development.

4.4.13 Fragment and Scaffold Similarity

The Fragment and Scaffold Similarity metrics evaluate how closely the fragments and scaffolds of the generated molecules resemble those of a reference set, typically the training set. These metrics help ensure that the generated molecules are chemically relevant and maintain structural features common to known compounds.

Fragment similarity is calculated using the BRICS fragmentation method, which decomposes molecules into smaller, synthetically relevant fragments. The similarity score is determined by comparing the frequency of fragments in the generated molecules to those in the reference set. The average fragment similarity can be computed as:

$$\text{Average Fragment Similarity} = \frac{1}{N} \sum_{i=1}^N \text{Similarity}(F_i, F_{\text{ref}}) \quad (4.10)$$

where F_i represents the fragments of the i -th generated molecule and F_{ref} represents the fragments of the reference set.

Scaffold similarity measures the resemblance of the core structural frameworks (scaffolds) of the generated molecules to those of the reference set. The Bemis-Murcko scaffold extraction method is used to identify these scaffolds. The similarity score is calculated by comparing the frequency of scaffolds in the generated molecules to those in the reference set. The average scaffold similarity can be computed as:

$$\text{Average Scaffold Similarity} = \frac{1}{N} \sum_{i=1}^N \text{Similarity}(S_i, S_{\text{ref}}) \quad (4.11)$$

where S_i represents the scaffold of the i -th generated molecule and S_{ref} represents the scaffolds of the reference set. A higher similarity score indicates that the generated molecules maintain structural characteristics similar to those in the training set, enhancing their chemical relevance.

4.4.14 Similarity to Nearest Neighbor (SNN)

The Similarity to Nearest Neighbor (SNN) metric [Ertoz et al., 2002](#) evaluates how similar each generated molecule is to its nearest neighbor in a reference set, typically the training set. This metric helps ensure that the generated molecules are relevant and within the chemical space of known compounds.

SNN is calculated by first computing the molecular fingerprints of both the generated and reference molecules. For each generated molecule, the Tanimoto similarity to every molecule in the reference set is calculated, and the highest similarity score (i.e., the nearest neighbor) is recorded. The average SNN score is given by:

$$\text{Average SNN} = \frac{1}{N} \sum_{i=1}^N \max_j \text{Tanimoto}(F_i, F_j) \quad (4.12)$$

where N is the number of generated molecules, F_i is the fingerprint of the i -th generated molecule, and F_j is the fingerprint of the j -th reference molecule. A higher SNN score indicates that the generated molecules are more similar to known compounds, suggesting they are within a familiar and relevant chemical space.

4.4.15 Summary of Model Evaluation Metrics

Table 4.1 summarizes the evaluation metrics prepared, their significance, the range of values they can take and the target values for obtaining drug-like behaviour in the generated molecules.

Metric	Description	Range of Values	Desired Range
Validity	Proportion of chemically valid molecules	0 to 1	Close to 1
Uniqueness	Proportion of unique molecules	0 to 1	Close to 1
Novelty	Proportion of molecules not in the training set	0 to 1	Close to 1
Internal Diversity	Pairwise dissimilarity among generated molecules	0 to 1	Close to 1
FCD	Distribution similarity between generated and reference molecules	0 to ∞	Lower values
LogP	Hydrophobicity of molecules	∞ to ∞	0 to 5
Penalized LogP	LogP adjusted for synthetic accessibility and structural penalties	∞ to ∞	Higher values
QED	Drug-likeness score combining multiple molecular properties	0 to 1	Close to 1
SA Score	Ease of molecule synthesis	1 to 10	Lower values
SCScore	Synthetic complexity of molecules	1 to 5	Lower values
SYBA Score	Synthetic accessibility using Bayesian classifier	0 to 1	Higher values
RAScore	Synthetic accessibility using reaction data	0 to 1	Higher values
Fragment Similarity	Similarity of fragments to reference set	0 to 1	Higher values
Scaffold Similarity	Similarity of scaffolds to reference set	0 to 1	Higher values
SNN	Similarity to nearest neighbor in reference set	0 to 1	Higher values

Table 4.1: Summary of Model Evaluation Metrics

4.4.16 Speed Enhancements

To facilitate the selection of high-quality molecules from a large corpus of molecular data, it is essential to score them using the aforementioned metrics and filter out the best set of molecules for model training. However, some of these metrics can be computationally intensive and time-consuming. For a sample of 5000 molecules evaluated, the computational times are recorded in table 4.2.

Metric	Time Taken (seconds)
FCD	88.19
SCScore	21.71
Penalized LogP	20.99
SA	19.53
Internal Diversity	10.67
QED	4.54
SYBA	4.52
LogP	1.53
Uniqueness	0.88
Novelty	0.84
Validity	0.55

Table 4.2: Time Taken to Compute Various Metrics for 5000 Molecules

To address this potential challenge, we have implemented several metrics in Rust/C++ to enhance the computation speed. Specifically, we utilized the C++ implementation of the RDKit library. The following table 4.3 highlights the speed-up results achieved:

Metric	Compute Speed Enhancement
FCD, SCScore	Can be run on GPUs to speed up the computation
LogP	9x speed up achieved on C++
Internal Diversity	6x speed up achieved on C++
Validity	Similar performance in Python and C++
Uniqueness	Similar performance in Python and C++
Novelty	Similar performance in Python and C++

Table 4.3: Metric Computation Speed Enhancement

4.5 Downstream Fine-Tuning using ReST Framework

Reinforced Self-Training (ReST) [Gulcehre et al., 2023](#) is a cutting-edge algorithm designed to optimize large language models (LLMs) by leveraging principles of reinforcement learning (RL) from feedback. This framework integrates dataset generation with policy improvement, offering a robust method to align models with specific desired outcomes, even in domains requiring high-quality outputs like molecular generation.

4.5.1 Key Components of ReST

The ReST method consists of two main components (Figure 4.2):

Grow Step: In the Grow Step, the current policy (model) generates a new dataset by sampling multiple output sequences for each context from the original dataset. For molecular generation, this involves creating diverse molecular structures based on existing ones. This step enriches the dataset with varied samples, enabling the model to learn from a broader range of molecular configurations.

Improve Step: The generated samples are evaluated using a reward model that scores each sample based on desired properties (e.g., drug-likeness, specific therapeutic activities). Samples with high scores are selected to fine-tune the model. This iterative process, with increasingly stringent thresholds, progressively refines the model’s output quality.

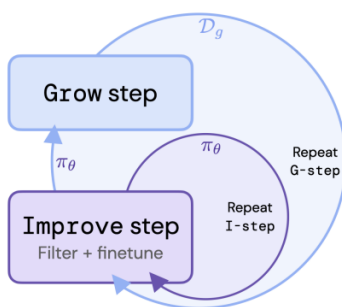


Figure 4.2: ReST method ([Gulcehre et al., 2023](#))

4.5.2 Usefulness in Molecular Generation

ReST’s framework is particularly beneficial for optimizing molecular generation models to meet specific objectives. Here’s how it can be leveraged:

Optimization of Desired Properties: ReST can use oracles that score molecules based on drug-likeness metrics, fine-tuning the model to generate molecules that are more likely to be viable drug candidates. Additionally, ReST can optimize models to generate

molecules with specific therapeutic properties, such as anti-malarial or anti-fungal activities, by using oracles tailored to these properties.

Effective Use of Small Datasets: ReST is particularly powerful when only small datasets of molecules with desired properties are available. The Grow Step allows the model to generate new samples based on the small dataset, effectively augmenting the training data. By iteratively refining the model using high-scoring samples, ReST ensures that the model learns effectively even from a limited amount of data. This process helps in extracting the maximum value from small datasets, aligning the model closely with the desired properties.

4.5.3 Fine-Tuning

The fine-tuning of our models was conducted using the Reinforced Self-Training (ReST) framework, which integrates reinforcement learning to align model outputs with desired properties. The models underwent 15 epochs of training, with 10 grow steps and 5 improve steps, employing the Adam optimizer with a learning rate set at $5e-4$. This rate was selected to ensure an optimal balance between convergence speed and accuracy.

The fine-tuning process was performed using an NVIDIA GeForce GTX 1080 Ti graphics card, providing the necessary computational power for efficient training.

For the fine-tuning, a Llama-small model pre-trained on the PubChem dataset was chosen. The string 'CCO' was used as a start token to initiate the generation process. This approach ensured that the generation of molecular structures began from the same starting points each time, helping us better observe the role of the ReST method.

The fine-tuning process focused on improving the average Quantitative Estimate of Drug-likeness (QED) score of the generated samples. The results indicated over 10% improvement, from 0.41 in the base model to 0.453 in the fine-tuned model, demonstrating the effectiveness of the ReST framework in generating molecules with desirable properties.

RESULTS

5.1 Unconditioned Training

In this section, we look at the results of training a model without any molecular conditions. The model is trained to generate valid molecules autoregressively by understanding the grammar of the chemical space.

5.1.1 Training Curves



Figure 5.1: Plots for Unconditioned Training of the Model

From the plots shown in Figure 5.1, we see that the validation loss and training loss have reduced over the epochs. This indicates that the model is not overfitting as it is able to perform reasonably well even on unseen data points.

5.1.2 Generation Results

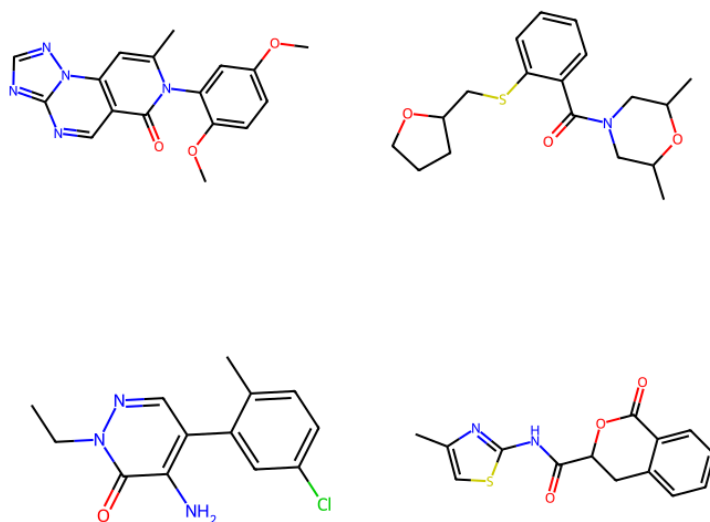


Figure 5.2: *Unconditioned Generation Samples*

5.1.3 Performance on Evaluation Metrics

Metric	Value	Metric	Value
Valid	0.994	FCD/TestSF	1.185
Unique@1000	1.000	SNN/TestSF	0.582
Unique@10000	0.998	Frag/TestSF	0.993
FCD/Test	0.559	Scaf/TestSF	0.059
SNN/Test	0.633	IntDiv	0.849
Frag/Test	0.997	IntDiv2	0.843
Scaf/Test	0.898	Filters	0.998
logP	0.017	QED	0.003
SA	0.010	Weight	1.423
		Novelty	0.749

Table 5.1: *Performance of Unconditioned Generation on Evaluation Metrics*

5.2 Conditioned Training - Scaffold

We now train the model by feeding every molecule along with its scaffold information. This enables us to prompt the resultant model with the desired scaffold to generate molecules containing them.

5.2.1 Training Curves



Figure 5.3: *Plots for Scaffold Conditioned Training of the Model*

Figure 5.3, illustrates a decrease in both training and validation loss across the epochs. This trend suggests that the model is learning effectively without overfitting, as evidenced by its consistent performance on new, unseen data.

5.2.2 Generation Results

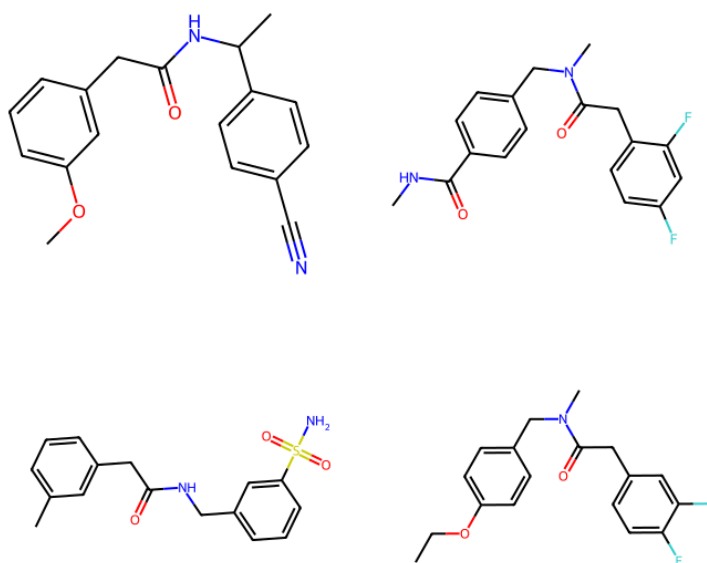


Figure 5.4: Scaffold Conditioned Generation Samples

5.2.3 Performance on Evaluation Metrics

Metric	Value	Metric	Value
Valid	0.985	FCD/TestSF	20.161
Unique@1000	0.894	SNN/TestSF	0.740
Unique@10000	0.702	Frag/TestSF	0.855
FCD/Test	18.911	Scaf/TestSF	0.244
SNN/Test	0.576	IntDiv	0.779
Frag/Test	0.857	IntDiv2	0.763
Scaf/Test	0.000	Filters	0.999
logP	0.194	QED	0.036
SA	0.194	Weight	6.689
		Novelty	0.999

Table 5.2: Performance of Scaffold Conditioned Generation on Evaluation Metrics

5.3 Conditioned Training - Properties

We then explore the model's ability to understand and control some chemical properties of the molecules generated. We train the molecule with the **synthetic accessibility score** (Measurement of the difficulty of synthesizing a compound) and the **logarithm of the partition coefficient** (the partition coefficient compares the solubilities of the solute in two immiscible solvents at equilibrium). The trained model should be able to generate molecules the showcase desired SAS and logP values.

5.3.1 Training Curves

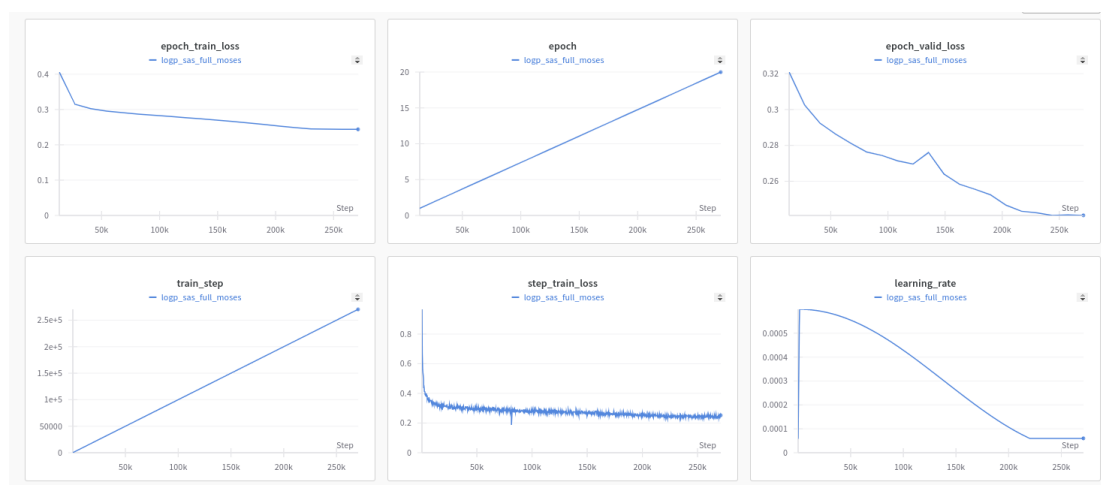


Figure 5.5: *Plots for Property Conditioned Training of the Model*

The graphs in Figure 5.5 demonstrate a consistent reduction in both training and validation loss over successive epochs. This pattern indicates effective learning by the model and a lack of overfitting, as it maintains good performance on data it has not previously encountered.

5.3.2 Generation Results

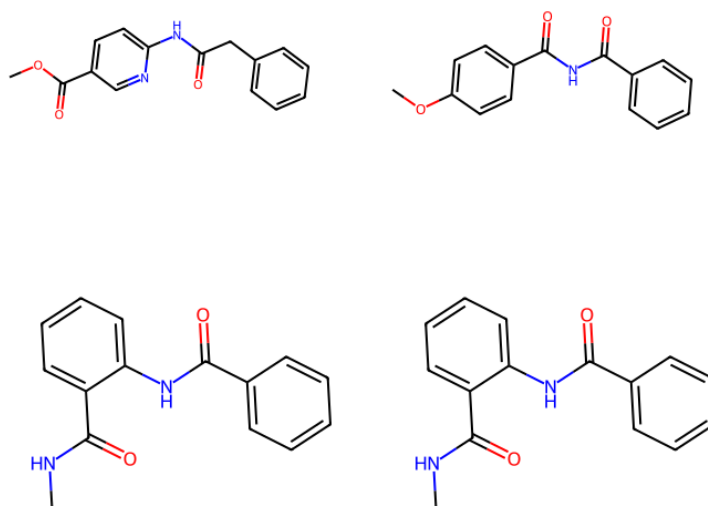


Figure 5.6: *Property Conditioned Generation Samples*

5.3.3 Performance on Evaluation Metrics

Metric	Value	Metric	Value
Valid	0.849	FCD/TestSF	8.185
Unique@1000	0.113	SNN/TestSF	0.568
Unique@10000	0.660	Frag/TestSF	0.920
FCD/Test	7.355	Scaf/TestSF	0.008
SNN/Test	0.613	IntDiv	0.812
Frag/Test	0.919	IntDiv2	0.785
Scaf/Test	0.594	Filters	0.989
logP	0.632	QED	0.031
SA	0.448	Weight	17.584
		Novelty	0.886

Table 5.3: *Performance of Property Conditioned Generation on Evaluation Metrics*

5.4 Conditioned Training - Scaffold, Properties

As a next step, we try to train a model that can control both scaffold and property criteria in the generated molecules. The performance of the generated molecules is again evaluated on several metrics described earlier.

5.4.1 Training Curves



Figure 5.7: Plots for Scaffold + Property Conditioned Training of the Model

In Figure 5.7, the downward trend observed in both the training and validation losses as epochs progress suggests that the model is effectively learning and generalizing well. This is indicated by its stable performance on unfamiliar data, showing no signs of overfitting.

5.4.2 Generation Results

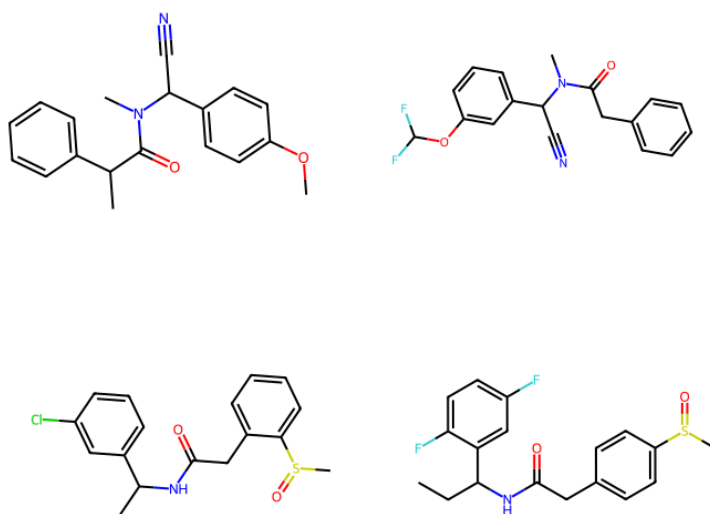


Figure 5.8: Scaffold + Property Conditioned Generation Samples

5.4.3 Performance on Evaluation Metrics

Metric	Value	Metric	Value
Valid	0.220	FCD/TestSF	21.831
Unique@1000	0.591	SNN/TestSF	0.557
Unique@10000	0.530	Frag/TestSF	0.802
FCD/Test	21.247	Scaf/TestSF	0.216
SNN/Test	0.469	IntDiv	0.827
Frag/Test	0.803	IntDiv2	0.806
Scaf/Test	0.001	Filters	0.989
logP	1.039	QED	0.038
SA	0.552	Weight	13.349
		Novelty	1.000

Table 5.4: Performance of Scaffold + Property Conditioned Generation on Evaluation Metrics

The Table 5.4 indicates a relatively poor performance of the model in the evaluation metrics computed (especially the fraction of valid molecules generated), in spite of relatively consistent downward trend in the validation loss. This indicates that the model’s performance may improve if trained for more epochs or with more data.

5.5 Unconditioned training using GuacaMol Dataset

In the following experiment, we aim to compare these results with those obtained by training on the GuacaMol dataset. As previously discussed, the GuacaMol dataset is a subset of the ChEMBL24 dataset and provides a more realistic representation of drug-like molecules compared to MOSES. This comparison will help us evaluate the impact of the dataset on the model’s performance across several important metrics.

5.5.1 Training Curves

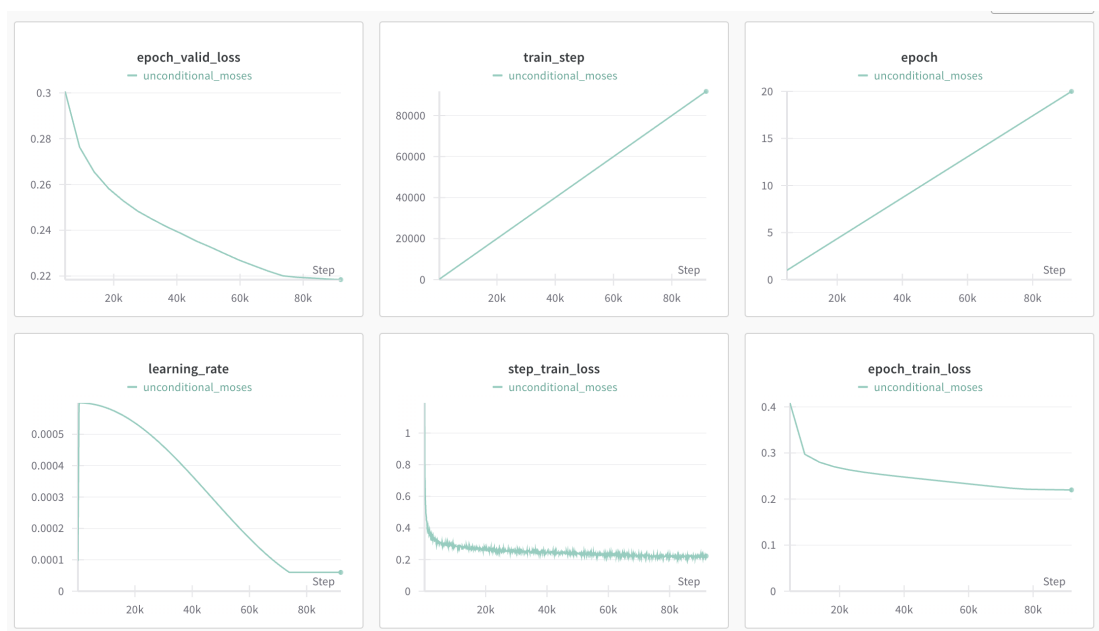


Figure 5.9: Plots for Model Trained using GuacaMol Dataset

The training plots in 5.9 exhibit a similar trend to those in 5.1. The consistent reduction in validation loss indicates that the model is able to learn the chemical grammar of the dataset and generalize well. 5.10 presents a few sample molecules generated using the model.

5.5.2 Generation Results

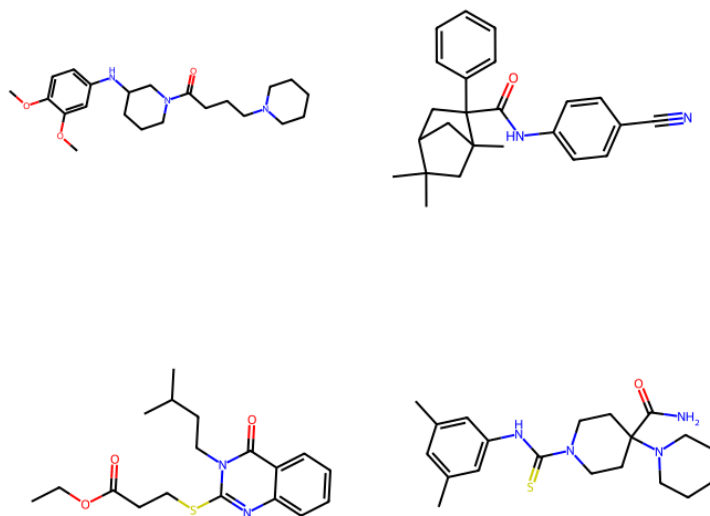


Figure 5.10: Samples of Generation by Model Trained on GuacaMol Dataset

5.5.3 Performance on Evaluation Metrics

Metric	Value	Metric	Value
Valid	1.0	FCD/TestSF	23.383
Unique@1000	1.0	SNN/TestSF	0.267
Unique@10000	0.9995	Frag/TestSF	0.640
FCD/Test	11.113	Scaf/TestSF	0.314
SNN/Test	0.316	IntDiv	0.870
Frag/Test	0.965	IntDiv2	0.865
Scaf/Test	0.314	Filters	0.798
logP	0.939	QED	0.239
SA	0.693	Weight	109.169
		Novelty	1.0

Table 5.5: Performance Metrics for Model Trained on GuacaMol Dataset

From 5.5, we can observe that the model’s performance on downstream metrics is highly dependent on the dataset used for pre-training. Overall, the metrics demonstrate values within the desired range, as expected given the higher quality of molecules in the training dataset.

5.5.4 Comparison - Effect of Dataset on Model Performance

Metric	Moses Value	Guacamol Value
Valid	0.994	1.0
Unique@1000	1.000	1.0
Unique@10000	0.998	0.9995
IntDiv	0.849	0.870
IntDiv2	0.843	0.865
logP	0.017	0.939
QED	0.003	0.239
SA	0.010	0.693
Weight	1.423	109.169
Novelty	0.749	1.0

Table 5.6: *Comparison of Models Trained on MOSES and GuacaMol Datasets*

As illustrated in 5.6, the model trained on the GuacaMol dataset exhibits significantly improved performance on drug-likeness metrics such as QED and LogP. This indicates that the GuacaMol dataset, which is derived from the ChEMBL24 dataset, provides a more realistic and higher-quality set of drug-like molecules compared to the MOSES dataset. The superior performance on these key metrics underscores the importance of using a representative and high-quality dataset for pre-training in order to enhance the model’s effectiveness in generating drug-like molecules.

5.6 Downstream Tasks - Alignment to QED

We try to explore if we can improve the drug-like behaviour of the generated molecules. This is so that the generated molecules can be more useful and fit for human consumption. Having pre-trained a model capable of generating valid, unique and conditioned molecules, we now try to align the generated molecules to certain properties of interest in the downstream tasks, such as the QED score.

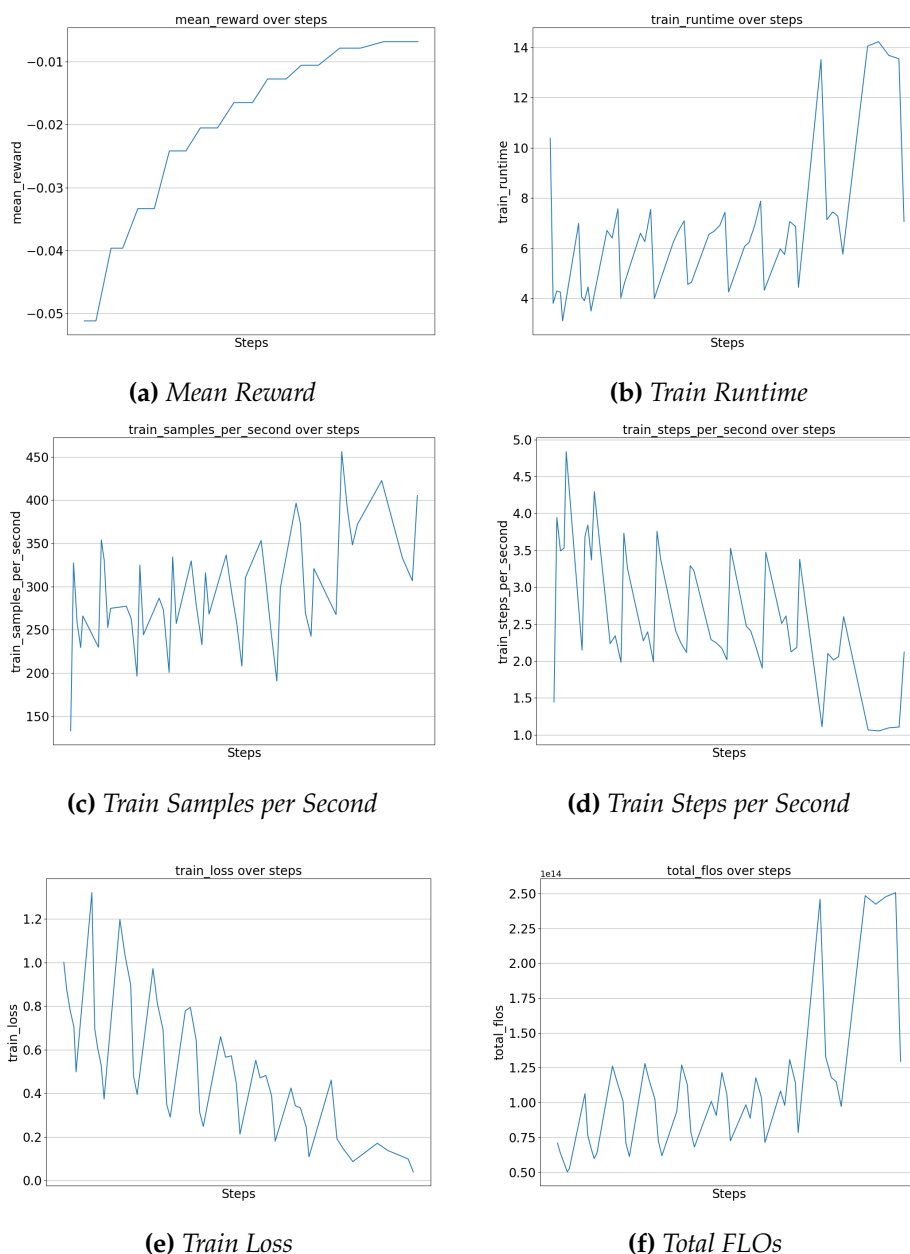


Figure 5.11: QED fine-tuning training plots

The plots in Figure 5.11 show the variation of specific model parameters during the training steps. Notably, the QED score, the property being optimized, increased by over 10% throughout the training. Additionally, Figures 5.12 and 5.13 demonstrate changes in other significant metrics as the model aligns to generate molecules with higher QED values. Analyzing these variations reveals the evolution of model parameters and the impact of these adjustments on the desired output. This analysis provides insights into the effectiveness of the training process and the alignment strategy in optimizing for higher drug-likeness scores, particularly the QED metric.

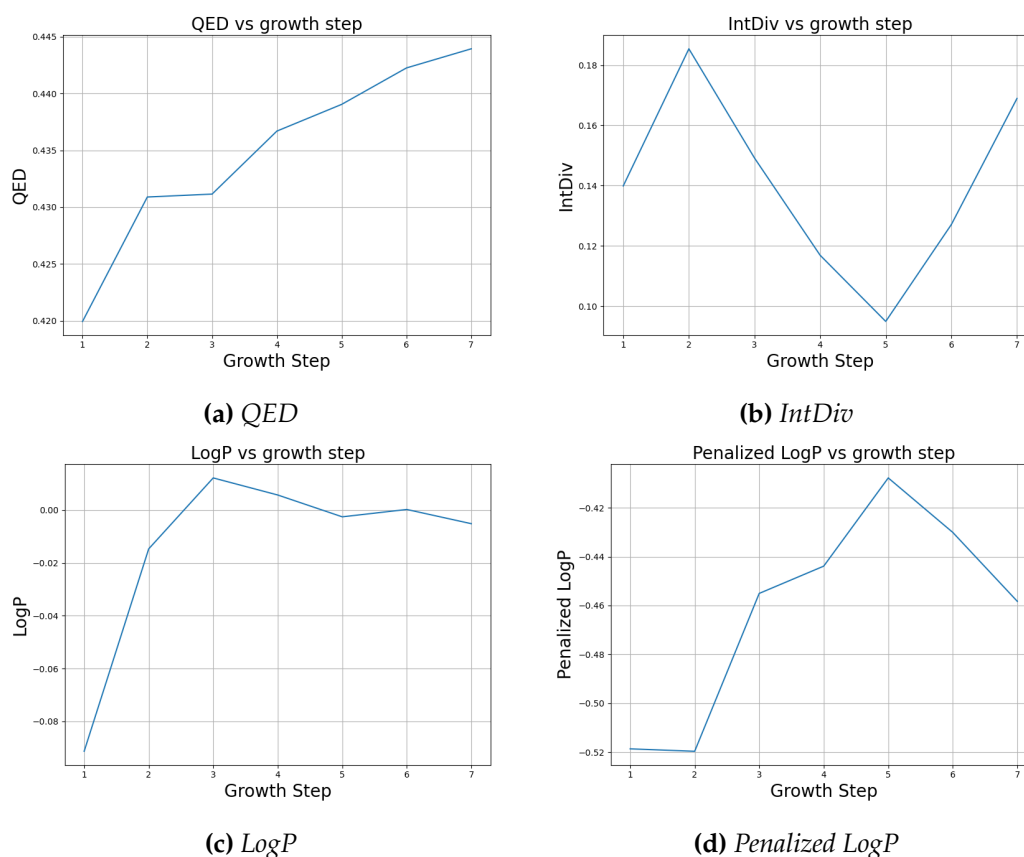


Figure 5.12: QED Generation Plots Over Time (a)

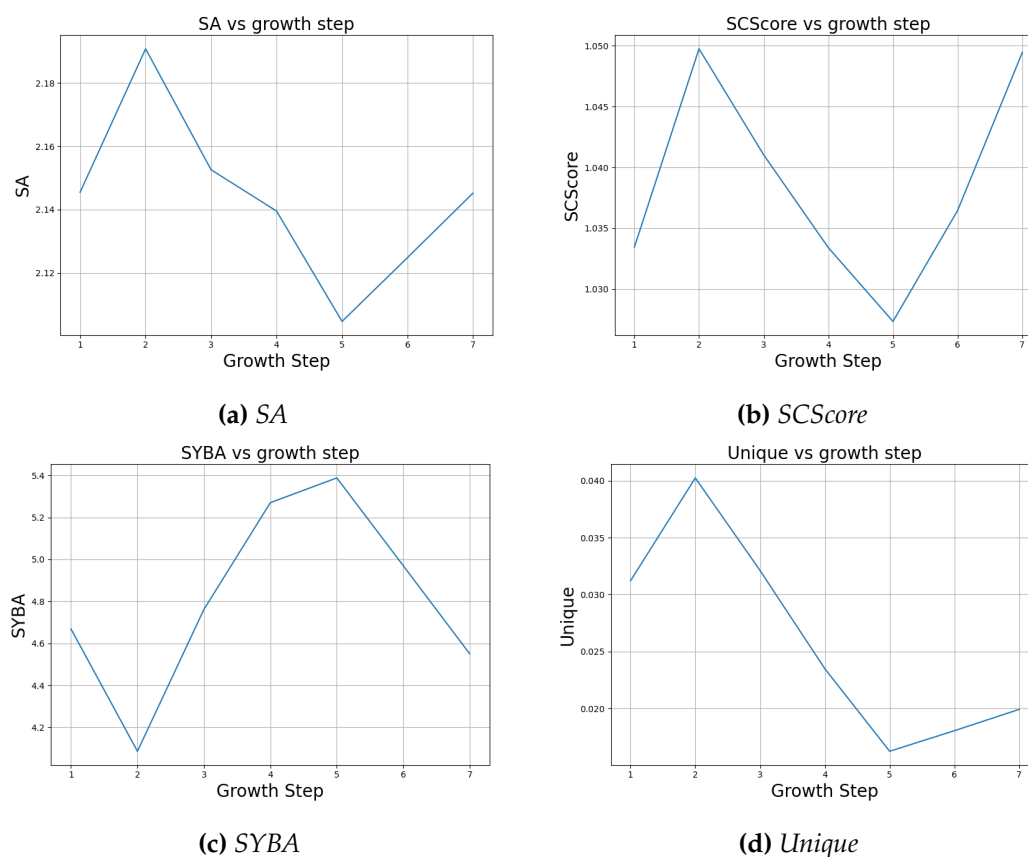


Figure 5.13: QED Generation Plots Over Time (b)

Studying the plots in Figure 5.12 and 5.13, we see that as we maximize the QED value of the generated molecules, we observe concurrent improvements in related metrics such as LogP. Additionally, the synthetic accessibility metrics, computed using different methods (SA, SCScore, and SYBA), exhibit similar trends. This consistency in the observed patterns underscores the robustness of the alignment strategy and its effectiveness in enhancing multiple aspects of drug-likeness simultaneously.

CONCLUSION AND FUTURE WORK

We embarked on this project with the goal of leveraging recent advancements in deep learning and natural language processing (NLP) to generate novel molecules that can aid in the drug discovery process. Our initial objective was to explore how state-of-the-art techniques in these fields could be applied to create molecules with specific desired properties, potentially accelerating the development of new therapeutics.

Throughout the course of our research, we delved into the various levers available to control the molecular generation process. This journey took us from understanding the importance of datasets and data representations to experimenting with different tokenizers and models. We also focused on evaluation metrics to assess the quality and relevance of the generated molecules. Additionally, we explored fine-tuning and alignment techniques to ensure that the models could be optimized to generate molecules with desired characteristics for downstream tasks.

Our findings demonstrate that we can effectively control numerous characteristics of the molecules generated by the models. This control extends from the molecular scaffolds to the physicochemical properties and drug-like behaviors. By carefully manipulating the various components of the molecular generation pipeline, we were able to produce molecules that meet specific criteria, showcasing the potential of these advanced techniques in the field of drug discovery.

Looking ahead, the adoption of innovative alignment frameworks like Reinforced Self-Training (ReST) holds great promise for molecular generation. Our results indicate that integrating ReST can significantly enhance the alignment of generated molecules with desired properties. Future work could explore the generation of molecules with specific therapeutic properties, such as anti-malarial, anti-bacterial, and anti-fungal activities. These advancements suggest a bright and exciting future for targeted drug discovery, where sophisticated alignment techniques can pave the way for more effective and efficient therapeutic developments.

DECLARATION OF AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

During the preparation of this work, the author used ChatGPT to improve the organizational flow of the paper and eliminate errors by providing the draft. After using this tool/service, the author reviewed and edited the content as needed and takes full responsibility for the content of the publication.

BIBLIOGRAPHY

- Anstine, Dylan M. and Olexandr Isayev (2023). "Generative Models as an Emerging Paradigm in the Chemical Sciences". In: *Journal of the American Chemical Society* 145.16. PMID: 37052978, pp. 8736–8750. DOI: 10.1021/jacs.2c13467. eprint: <https://doi.org/10.1021/jacs.2c13467>. URL: <https://doi.org/10.1021/jacs.2c13467>.
- Bagal, Viraj et al. (2022). "MolGPT: Molecular Generation Using a Transformer-Decoder Model". In: *Journal of Chemical Information and Modeling* 62.9. PMID: 34694798, pp. 2064–2076. DOI: 10.1021/acs.jcim.1c00600. eprint: <https://doi.org/10.1021/acs.jcim.1c00600>. URL: <https://doi.org/10.1021/acs.jcim.1c00600>.
- Bickerton, G Richard et al. (2012). "Quantifying the chemical beauty of drugs". In: *Nature chemistry* 4.2, pp. 90–98.
- Brown, Nathan et al. (2019). "GuacaMol: Benchmarking Models for de Novo Molecular Design". In: *Journal of Chemical Information and Modeling* 59.3. PMID: 30887799, pp. 1096–1108. DOI: 10.1021/acs.jcim.8b00839. eprint: <https://doi.org/10.1021/acs.jcim.8b00839>. URL: <https://doi.org/10.1021/acs.jcim.8b00839>.
- Coley, Connor W et al. (2018). "SCScore: synthetic complexity learned from a reaction corpus". In: *Journal of chemical information and modeling* 58.2, pp. 252–261.
- Du, Yuanqi et al. (2022). *MolGenSurvey: A Systematic Survey in Machine Learning Models for Molecule Design*. arXiv: 2203.14500 [cs.LG].
- Ertoz, Levent, Michael Steinbach, and Vipin Kumar (2002). "A new shared nearest neighbor clustering algorithm and its applications". In: *Workshop on clustering high dimensional data and its applications at 2nd SIAM international conference on data mining*. Vol. 8.
- Fang, Yin et al. (2024). *Domain-Agnostic Molecular Generation with Chemical Feedback*. arXiv: 2301.11259 [cs.LG].
- Gall , Matthias (2019). "Investigating the effectiveness of BPE: The power of shorter sequences". In: *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 1375–1381.
- Gao, Wenhao et al. (2022). *Sample Efficiency Matters: A Benchmark for Practical Molecular Optimization*. arXiv: 2206.12411 [cs.CE].
- Ghugare, Raj et al. (2023). *Searching for High-Value Molecules Using Reinforcement Learning and Transformers*. arXiv: 2310.02902 [cs.LG].
- Gulcehre, Caglar et al. (2023). *Reinforced Self-Training (ReST) for Language Modeling*. arXiv: 2308.08998 [cs.CL].

- Kim, Sunghwan et al. (2023). "PubChem 2023 update". In: *Nucleic acids research* 51.D1, pp. D1373–D1380.
- Krenn, Mario et al. (2020). "Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation". In: *Machine Learning: Science and Technology* 1.4, p. 045024.
- Li, Xinhao and Denis Fourches (2021). "SMILES pair encoding: a data-driven substructure tokenization algorithm for deep learning". In: *Journal of chemical information and modeling* 61.4, pp. 1560–1569.
- Mannhold, Raimund et al. (2009). "Calculation of molecular lipophilicity: State-of-the-art and comparison of log P methods on more than 96,000 compounds". In: *Journal of pharmaceutical sciences* 98.3, pp. 861–893.
- Murthy, Nitin Madhava (2021). "Synthetic Accessibility Scoring and Its Potential Applications in Chemical Library Generation". PhD thesis. State University of New York at Buffalo.
- Nath, Mriganka and Subhasish Goswami (2021). "Toxicity detection in drug candidates using simplified molecular-input line-entry system". In: *arXiv preprint arXiv:2101.10831*.
- Noutahi, Emmanuel et al. (2024). "Gotta be SAFE: a new framework for molecular design". In: *Digital Discovery* 3.4, pp. 796–804.
- O'Boyle, Noel and Andrew Dalke (2018). "DeepSMILES: an adaptation of SMILES for use in machine-learning of chemical structures". In.
- Polykovskiy, Daniil et al. (2018). "Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models". In: *CoRR* abs/1811.12823. arXiv: 1811 . 12823. URL: <http://arxiv.org/abs/1811.12823>.
- Preuer, Kristina et al. (2019). "Fréchet ChemNet Distance: A metric for generative models for molecules in drug design—Supporting Information—". In: *Deep Learning in Drug Discovery*, p. 59.
- Radford, Alec et al. (2019). "Language Models are Unsupervised Multitask Learners". In.
- Skoraczynski, Grzegorz et al. (Jan. 2023). "Critical assessment of synthetic accessibility scores in computer-assisted synthesis planning". In: *Journal of Cheminformatics* 15.1. ISSN: 1758-2946. DOI: 10.1186/s13321-023-00678-z. URL: <http://dx.doi.org/10.1186/s13321-023-00678-z>.
- Thakkar, Amol et al. (2021). "Retrosynthetic accessibility score (RAscore)—rapid machine learned synthesizability classification from AI driven retrosynthetic planning". In: *Chemical science* 12.9, pp. 3339–3349.
- Tingle, Benjamin I. et al. (2023). "ZINC-22A Free Multi-Billion-Scale Database of Tangible Compounds for Ligand Discovery". In: *Journal of Chemical Information and Modeling* 63.4. PMID: 36790087, pp. 1166–1176. DOI: 10.1021/acs.jcim.2c01253. eprint: <https://doi.org/10.1021/acs.jcim.2c01253>. URL: <https://doi.org/10.1021/acs.jcim.2c01253>.
- Vaswani, Ashish et al. (2017). "Attention Is All You Need". In: *CoRR* abs/1706.03762. arXiv: 1706.03762. URL: <http://arxiv.org/abs/1706.03762>.
- Voršilák, Milan et al. (2020). "SYBA: Bayesian estimation of synthetic accessibility of organic compounds". In: *Journal of cheminformatics* 12, pp. 1–13.

Zdrazil, Barbara et al. (Nov. 2023). "The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods". In: *Nucleic Acids Research* 52.D1, pp. D1180–D1192. ISSN: 0305-1048. DOI: 10.1093/nar/gkad1004. eprint: <https://academic.oup.com/nar/article-pdf/52/D1/D1180/55040046/gkad1004.pdf>. URL: <https://doi.org/10.1093/nar/gkad1004>.

Zhang, Jie et al. (2021). "Comparative Study of Deep Generative Models on Chemical Space Coverage". In: *Journal of Chemical Information and Modeling* 61.6. PMID: 34015916, pp. 2572–2581. DOI: 10.1021/acs.jcim.0c01328. eprint: <https://doi.org/10.1021/acs.jcim.0c01328>. URL: <https://doi.org/10.1021/acs.jcim.0c01328>.