

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

<b>Contents</b>	<b>iii</b>
<b>1 Introduction</b>	<b>2</b>
1.1 Context and Challenges in Drug Discovery . . . . .	2
1.2 Innovative Approach Using Transformer-Decoder Models . . . . .	2
<b>2 Brief review of literature</b>	<b>4</b>
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
<b>3 Datasets</b>	<b>8</b>
3.1 Properties of the Datasets . . . . .	8
3.1.1 Moses . . . . .	8
3.1.2 Guacamol . . . . .	8
3.1.3 ChemBL . . . . .	8
3.1.4 Zinc Datasets(250k, 1M, 10M, 270M, 37B) . . . . .	9
3.2 Dataset Statistics after Processing . . . . .	9
3.3 Inter-Dataset Overlap . . . . .	10
<b>4 Methods</b>	<b>13</b>
4.1 Model . . . . .	13
4.1.1 Motivation for Transformer Architecture . . . . .	13
4.1.2 Key Equations and Components . . . . .	13

---

4.1.3	GPT-2 Architecture . . . . .	14
4.1.4	Unique Qualities and Advantages . . . . .	14
4.1.5	Relevance to Molecular Design . . . . .	14
4.2	Training . . . . .	14
4.3	Evaluation Metrics . . . . .	15
<b>5</b>	<b>Results</b>	<b>18</b>
5.1	Unconditioned Training . . . . .	18
5.1.1	Training Curves . . . . .	18
5.1.2	Generation Results . . . . .	19
5.1.3	Performance on Evaluation Metrics . . . . .	19
5.2	Conditioned Training - Scaffold . . . . .	20
5.2.1	Training Curves . . . . .	20
5.2.2	Generation Results . . . . .	21
5.2.3	Performance on Evaluation Metrics . . . . .	21
5.3	Conditioned Training - Properties . . . . .	22
5.3.1	Training Curves . . . . .	22
5.3.2	Generation Results . . . . .	23
5.3.3	Performance on Evaluation Metrics . . . . .	23
5.4	Conditioned Training - Scaffold, Properties . . . . .	24
5.4.1	Training Curves . . . . .	24
5.4.2	Generation Results . . . . .	25
5.4.3	Performance on Evaluation Metrics . . . . .	25
5.5	Experiment - Conditioned Training - Scaffold, Functional Group, Property	26
5.5.1	Training Curves . . . . .	26
5.5.2	Generation Results . . . . .	27
5.5.3	Performance on Evaluation Metrics . . . . .	27
5.6	Experiment - Conditioned Training - Functional Group . . . . .	28
<b>6</b>	<b>Conclusion and Future Work</b>	<b>29</b>
6.1	Training on Other Larger Datasets . . . . .	29
6.2	Improved Techniques to Present the Data to the Model . . . . .	29
6.3	Fine-tuning for Specialised Downstream Tasks . . . . .	29
6.4	Using RL-based Techniques to Search the Chemical Space . . . . .	30
6.5	Compiling Results and Finishing the Report . . . . .	30
<b>7</b>	<b>Declaration of ai-assisted technologies in the writing process</b>	<b>31</b>

# 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 in-

tricate 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.

## 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.

### 3.2 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](#).

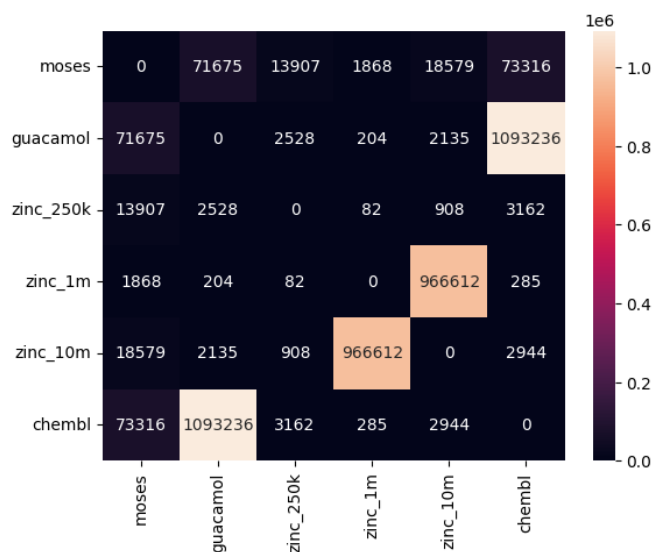
Dataset	Number of Rows	File Size
zinc_250k	249,455	18.18 MB
zinc_1m	999,998	72.13 MB
moses	1,936,962	93.42 MB
guacamol	1,591,011	140.94 MB
chembl	2,066,232	189.25 MB
zinc_10m	9,999,971	722.37 MB
zinc_270m	269,536,671	12.5 GB

**Table 3.1:** Datasets Information Sorted by File Size

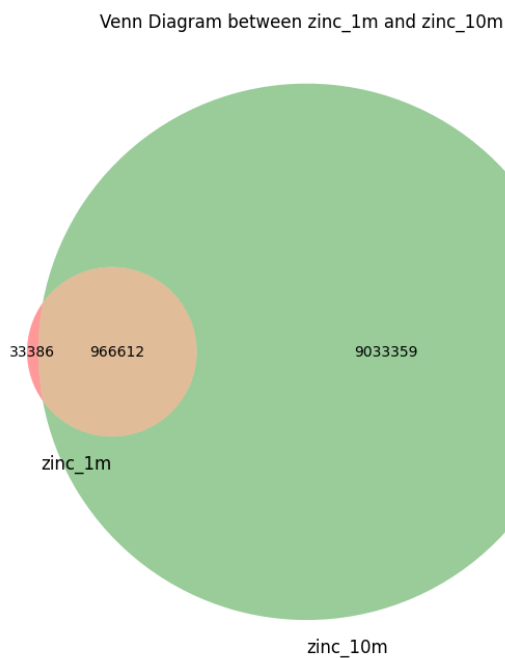
### 3.3 Inter-Dataset Overlap

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*



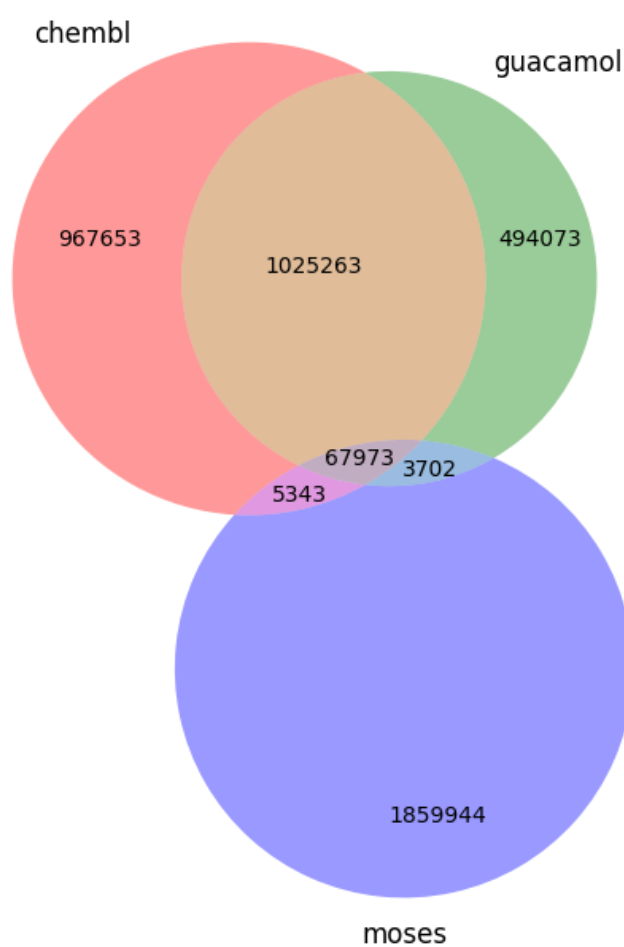
**Figure 3.1:** Heatmap showing the total overlap of SMILES strings between all datasets.



**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 moses



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

## METHODS

### 4.1 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.

#### 4.1.1 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.

#### 4.1.2 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  $W^O \in \mathbb{R}^{hd_v \times d_{\text{model}}}$ .

### 4.1.3 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.

### 4.1.4 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.

### 4.1.5 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.

## 4.2 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.

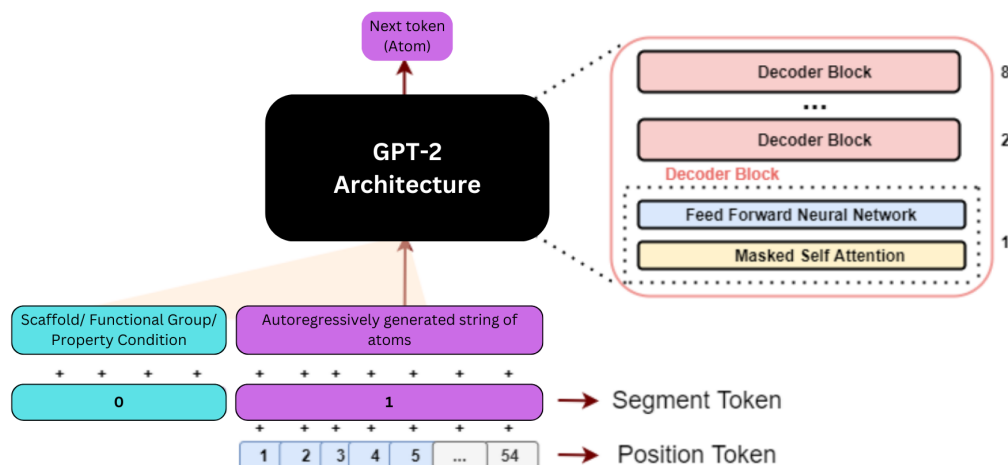


Figure 4.1: Model Architecture (Bagal et al., 2022)

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.

### 4.3 Evaluation Metrics

In evaluating synthesis-aware generative models for molecular generation, it is imperative to employ comprehensive benchmarking methodologies that adequately reflect the complexities inherent in molecular discovery. The MOSES benchmark Polykovskiy et al., 2018 provides a comprehensive set of metrics designed to assess various aspects of the generative models.

- **Validity:** Measures the percentage of generated molecules that are chemically valid.

- **Uniqueness:** Assesses whether the model generates diverse molecules by calculating the proportion of unique molecules in the generated set.
- **Novelty:** Evaluates the model's ability to generate molecules that are not present in the training set, indicating the model's creativity.
- **Internal Diversity:** Quantifies the chemical diversity within the generated set of molecules.
- **External Diversity:** Compares the diversity of the generated set to the diversity of an external set, often the test set.
- **Fréchet ChemNet Distance (FCD):** Uses a deep neural network to capture chemical and biological properties of compounds and measures the distance between the generated and real molecules in this learned feature space.
- **Fragment and Scaffold Similarity:** Measures how closely the distribution of molecular fragments and scaffolds in the generated set matches that of the reference set.
- **Similarity to Nearest Neighbor (SNN):** Calculates the average Tanimoto similarity between the generated molecules and their closest counterparts in the reference dataset.
- **Properties Distribution:** Compares the distribution of certain molecular properties (like molecular weight, logP, etc.) between the generated and reference sets using Wasserstein-1 distance.

While computational benchmarks, such as enrichment factors in virtual screening, standardized tests like Guacamol, and the MOSES (Molecular Sets) benchmark, have significantly propelled the field forward, they often fall short of capturing the entire scope of the discovery process, as highlighted by the implications of "Goodhart's law." These benchmarks serve as proxies and might not entirely encompass the nuances of molecular discovery. To address these limitations and enhance the validity of these models, the following multi-dimensional benchmarking strategies are recommended:

- **Objective Maximization and Ligand Rediscovery:** Assess the capacity of virtual screening or de novo design algorithms to identify molecules that optimize given objectives, along with their ability to rediscover known ligands. Notable benchmarks in this category include Guacamol, various virtual screening benchmarks, and the MOSES benchmark, which is specifically designed for assessing the quality of generative models.
- **Synthesis Prediction and Feasibility:** Employ CASP tools to predict synthetic routes or utilize synthesizability scores for generated molecules to ensure practical feasibility in a laboratory setting.
- **Molecule Quality Assurance:** Implement quality filters akin to those used in Guacamol and MOSES to ascertain the reasonableness of the generated molecules. It is crucial to include visualizations of random, non-cherry-picked molecular samples in machine-learning publications to provide a transparent and accurate representation of the model's output.

- **Evaluation of Synthesis Planning Algorithms:** Conduct both quantitative and qualitative assessments of synthesis planning algorithms to ensure they are efficient, practical, and innovative.

The discussion further delves into the translational impact of these improvements, questioning the real-world applicability of marginal gains observed in computational benchmarks. Given the often sparse and diverse nature of data in drug discovery, along with the occurrence of distribution shifts, the need for robustness in models is paramount. The authors advocate for a balanced approach towards benchmarking, one that encourages ongoing refinement and innovation in benchmarking practices without necessarily mandating experimental validation due to the varying resource capabilities of computational groups.

Recent prospective validations of virtual screening and de novo design offer promising examples of the field's progress. These include the application of large enumerated on-demand libraries in virtual screening and the integration of synthesis planning with computational algorithms, showcasing their utility, particularly in the early stages of discovery. However, the visibility and publication of such innovations are often delayed in the industrial context due to proprietary concerns or lack of incentives, highlighting an additional layer of complexity in benchmarking and validating these models.

In conclusion, while current benchmarks such as Guacamol, virtual screening benchmarks, and the MOSES benchmark are instrumental in driving advancements in molecular generation models, there is a clear and ongoing need for developing more nuanced, robust, and comprehensive methodologies. These should not only reflect the theoretical and computational excellence but also align closely with practical, real-world utility in the ever-evolving landscape of drug discovery.

## 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.

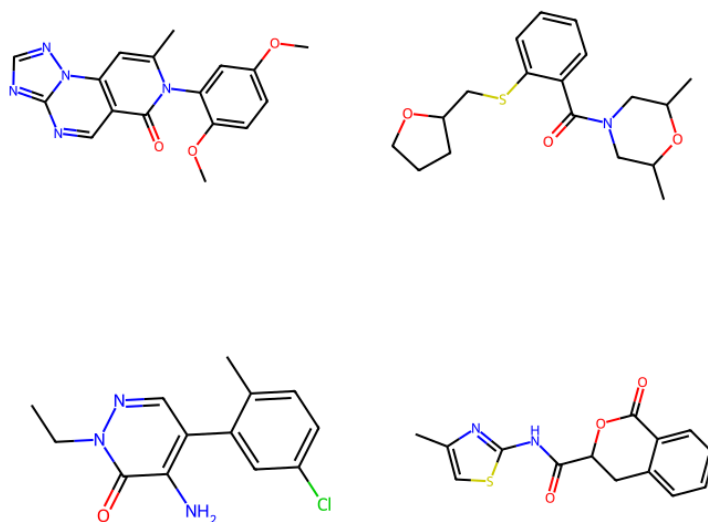
### 5.1.1 Training Curves



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

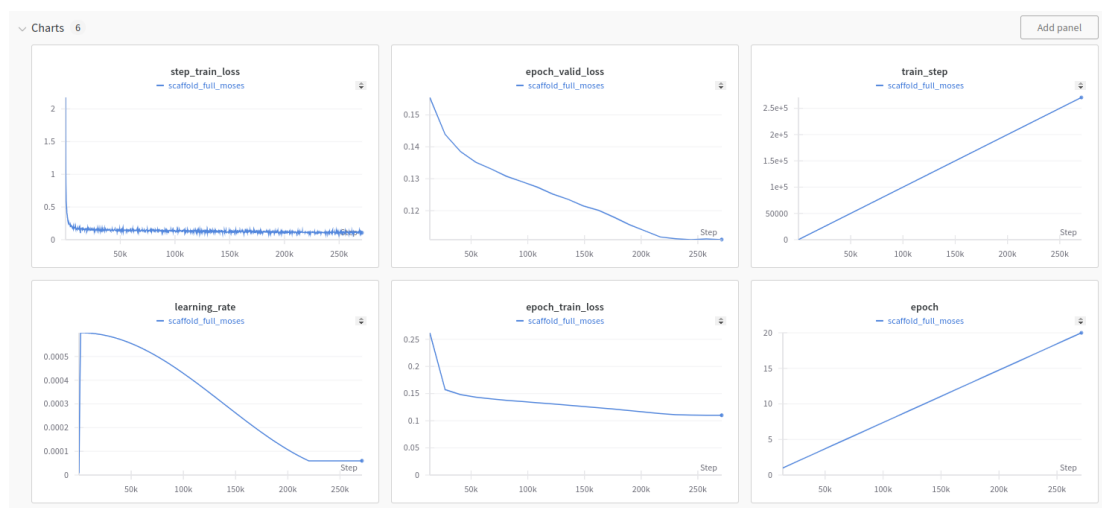
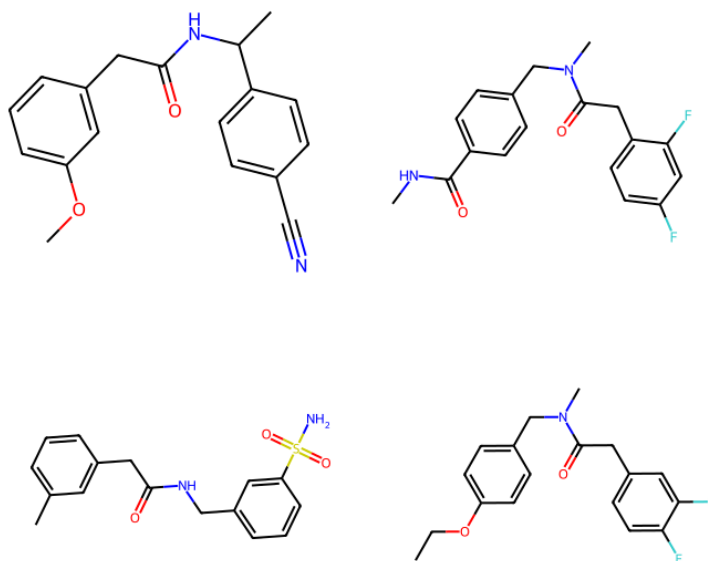


Figure 5.3: Plots for Scaffold Conditioned Training of the Model

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

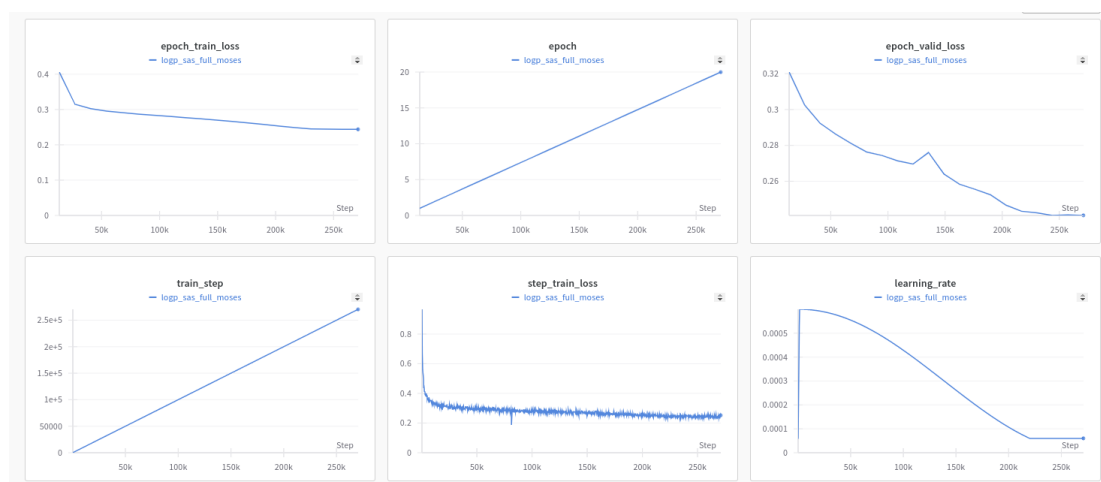
Metric	Value	Metric	Value
Valid	0.985	FCD/TestSF	20.161
Unique@1000	0.894	SNN/TestSF	0.740
Unique@10000	0.702	Frag/TestSF	0.855
FCD/Test	18.911	Scaf/TestSF	0.244
SNN/Test	0.576	IntDiv	0.779
Frag/Test	0.857	IntDiv2	0.763
Scaf/Test	0.000	Filters	0.999
logP	0.194	QED	0.036
SA	0.194	Weight	6.689
		Novelty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

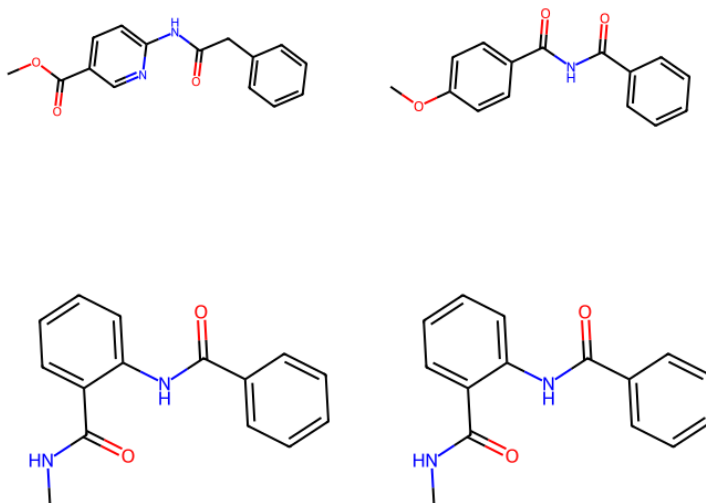
### 5.3.1 Training Curves



**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.

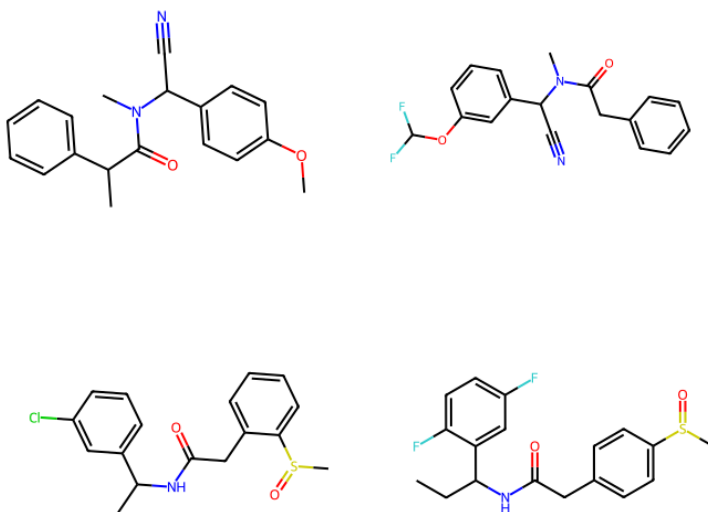
### 5.4.1 Training Curves



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 shows that the 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 Experiment - Conditioned Training - Scaffold, Functional Group, Property

Now we try to further introduce functional group details along with scaffold and chemical properties previously described in the paper. We restrict the maximum number of functional groups information passed along with any molecule to under 35. Further, we also try to expand the chemical property information passed to the model to include 210 other relevant properties that can be computed using the *rdkit* python library.

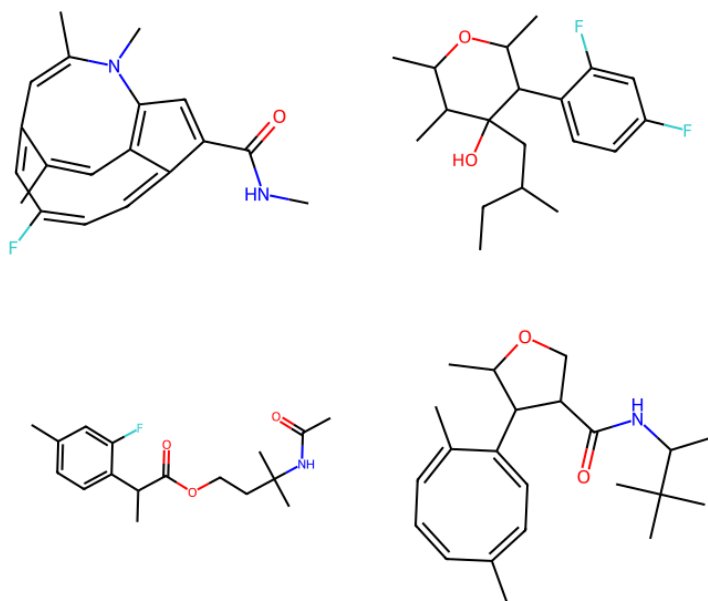
### 5.5.1 Training Curves



**Figure 5.9:** *Plots for Scaffold + Functional Group + Property Conditioned Training of the Model*

Based on the showcased plots in Figure 5.9, we can observe that the validation loss is increasing over epochs. This indicates the model's lack of generalisation and inability to learn meaningful insights from the data.

### 5.5.2 Generation Results



**Figure 5.10:** Scaffold + Property + Functional Group Generation Samples

While the model is able to generate some valid molecules as shown in Figure 5.10, a majority of the samples generated are in fact a random sequence of atoms that do not represent valid molecules. Moreover, even the generated molecules don't necessarily showcase the desired conditions prompted to the model.

### 5.5.3 Performance on Evaluation Metrics

Due to the model's inability to make sense of all the additional information provided, the majority of the molecules generated are an invalid sequence of atoms that don't form realistic molecules. Based on the results, we hypothesize that we will have to either training the model with larger datasets that we have prepared (Table 3.1) or reduce the amount of conditional information passed at each time step. Due to these factors, we refrain from evaluating the outputs on metrics that correspond to other downstream tasks.



## 5.6 Experiment - Conditioned Training - Functional Group

Since the model was unable to learn to control all three - scaffold, chemical property and functional group - information in our previous experiment, we try to observe the model's performance when only the functional group information is presented to it. From the plots in Figure 5.11, we again notice that the validation loss is increasing over



**Figure 5.11:** *Plots for Functional Group Conditioned Training of the Model*

epochs. This indicates the model's inability to learn the functional group representations from the data. To further confirm our hypothesis, we train the model by restricting the total functional group conditions fed to the model to under 5 per molecule. The training curves corresponding to this reduced training are shown in Figure 5.12.



**Figure 5.12:** *Plots for Functional Group (small) Conditioned Training of the Model*

We see that even after significantly reducing the complexity of the input functional group information, the model's validation loss does not converge as desired. This indicates the need to find better representations for the functional group condition, or training with larger quantities of data that can increase the chances of the model's convergence and improved performance.

## CONCLUSION AND FUTURE WORK

This report opens up several avenues for continued research in the field of AI-driven molecular generation. The proposed methods and findings lay a foundation for further exploration and enhancement. As seen in the previous sections, we have been able to demonstrate reasonable performance of the model in unconditional as well as various levels of scaffold + chemical property conditioned generation, but the model is unable to perform adequately when presented with additional functional group information. Some immediate next steps will involve exploring possible solutions to overcome this shortcoming. Some of the key areas of future work include:

### 6.1 Training on Other Larger Datasets

- Feb'24 - Mar'24

As previously mentioned, we have prepared several datasets of varying sizes and chemical characteristics (Table 3.1). By training on varied datasets, we can improve the model's understanding of the chemical space, potentially improving the model's performance in areas where it is currently failing.

### 6.2 Improved Techniques to Present the Data to the Model

- Feb'24 - Mar'24

By exploring some novel ways to present the data to the model, we may be able to help the model better understand previously unexplored molecular characteristics like functional groups.

### 6.3 Fine-tuning for Specialised Downstream Tasks

- Mar'24 - Apr'24

After achieving satisfactory performance on the defined metrics, we can explore the model's performance in generating molecules for more specialised tasks by fine-tuning on suitable datasets.

## 6.4 Using RL-based Techniques to Search the Chemical Space

- Mar'24 - Apr'24

Finally, we can further enhance the selectivity of the model using reinforcement learning-based techniques ([Gulcehre et al., 2023](#), [Ghugare et al., 2023](#)) by defining appropriate reward functions for desired actions.

## 6.5 Compiling Results and Finishing the Report

- Apr'24 - May'24

The last phase of the project will involve putting together all the findings and documenting the entire project.

## 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>.
- 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>.
- Du, Yuanqi et al. (2022). *MolGenSurvey: A Systematic Survey in Machine Learning Models for Molecule Design*. arXiv: 2203.14500 [cs.LG].
- 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].
- 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>.
- Radford, Alec et al. (2019). "Language Models are Unsupervised Multitask Learners". In: 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>.
- 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>.