

DDP End-Term Review

Conditional De Novo Molecular Generation

Saicharan Ganapathy
ED19B065

Guided by Prof. Dr. Nirav Pravinbhai Bhatt

June 04, 2024



Introduction/Motivation

Challenges in conventional approaches to drug discovery

- High costs and time, with low success rates.
- Resource-intensive compound screening in huge libraries.
- The impracticality of manual exploration in vast and unexplored chemical space.
- Increasing demand for more effective treatments and faster development

Opportunities

- Shift towards proactive molecule creation for specific targets.
- Adapting NLP techniques for molecular design using SMILES notation.



Literature Survey

- **MolGPT: Molecular Generation Using a Transformer-Decoder Model:** Introduces a transformer-decoder based model for processing SMILES strings in molecular structures, focusing on its ability to generate valid, diverse molecules with specific properties. Key experiments demonstrate the model's control over molecular properties and use saliency maps for interpretability, highlighting its potential in drug discovery and material science.
- **Molecular Sets (MOSES):** A Benchmarking Platform for Molecular Generation Models, introduces a dataset from the ZINC Clean Leads collection for molecule generation, evaluating several models including neural networks and variational autoencoders against metrics like validity and novelty. The study establishes a benchmark in generative modeling for molecules, demonstrating neural models' effectiveness over non-neural baselines.
- **Reinforced Self-Training (ReST) for Language Modeling:** technique combining reinforcement learning from human feedback (RLHF) with large language models (LLMs) for machine translation. ReST employs a two-step process: dataset expansion through model output sampling and fine-tuning via offline reinforcement learning algorithms, resulting in significantly enhanced translation quality and alignment with human preferences.
- **Searching for High-Value Molecules Using Reinforcement Learning and Transformers:** introduces ChemRLformer, an RL-based algorithm for molecular design, assessing the impact of text representation and training choices in RL. The research, spanning 25 molecular design tasks including protein docking simulations, reveals that SMILES notation outperforms SELFIES, highlights the importance of pretraining molecule quality, and compares transformer and RNN architectures, leading to a refined approach in molecular design with practical insights for future developments.

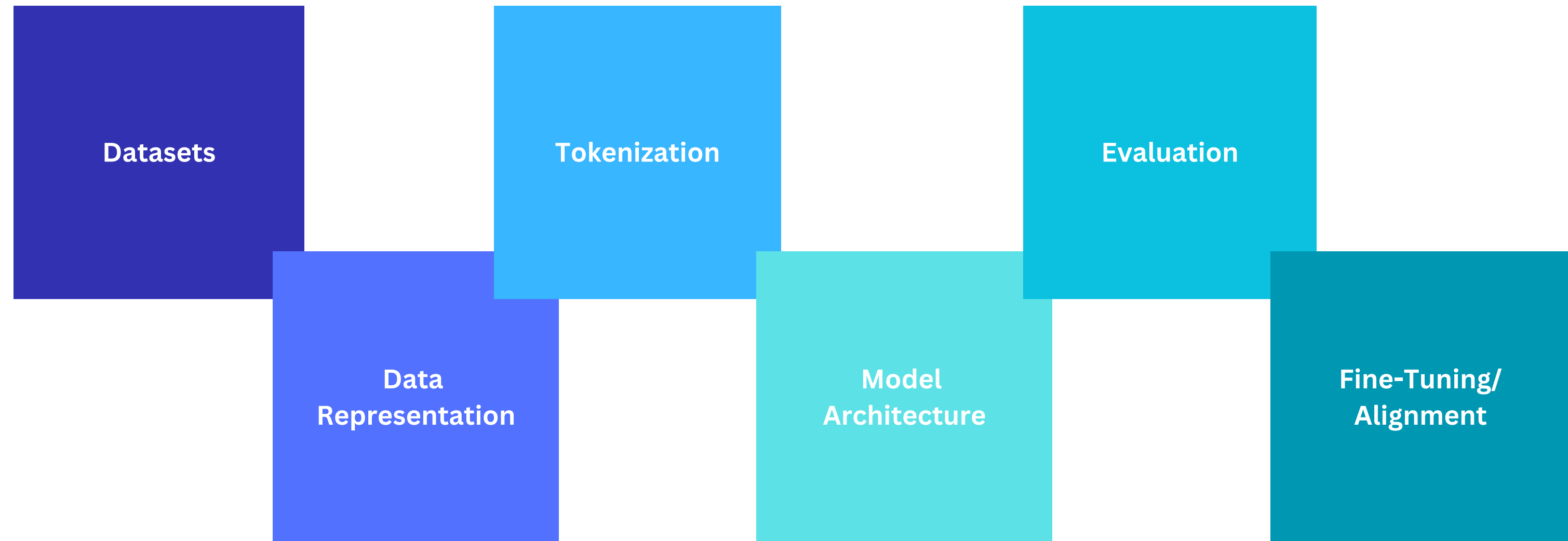


Problem statement or Formulation

Problem statement: Developing a transformer-based model for generating drug- like molecules with the ability to control and generate molecules with desired conditions such as scaffolds, chemical properties and behaviour



Project Workflow



Datasets and representations

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*

Notation	Benzene representation	Benefit
SMILES	<chem>"c1ccccc1"</chem>	Simplest
SELFIES	<chem>[C][=C][C][=C][C][=C][Ring1][=Branch1]</chem>	Ensures syntactic validity
DeepSMILES	<chem>"cccccc6"</chem>	Suitable for ML models
SAFE	Structure-aware encoding	Captures more context



Tokenization and Model Architecture

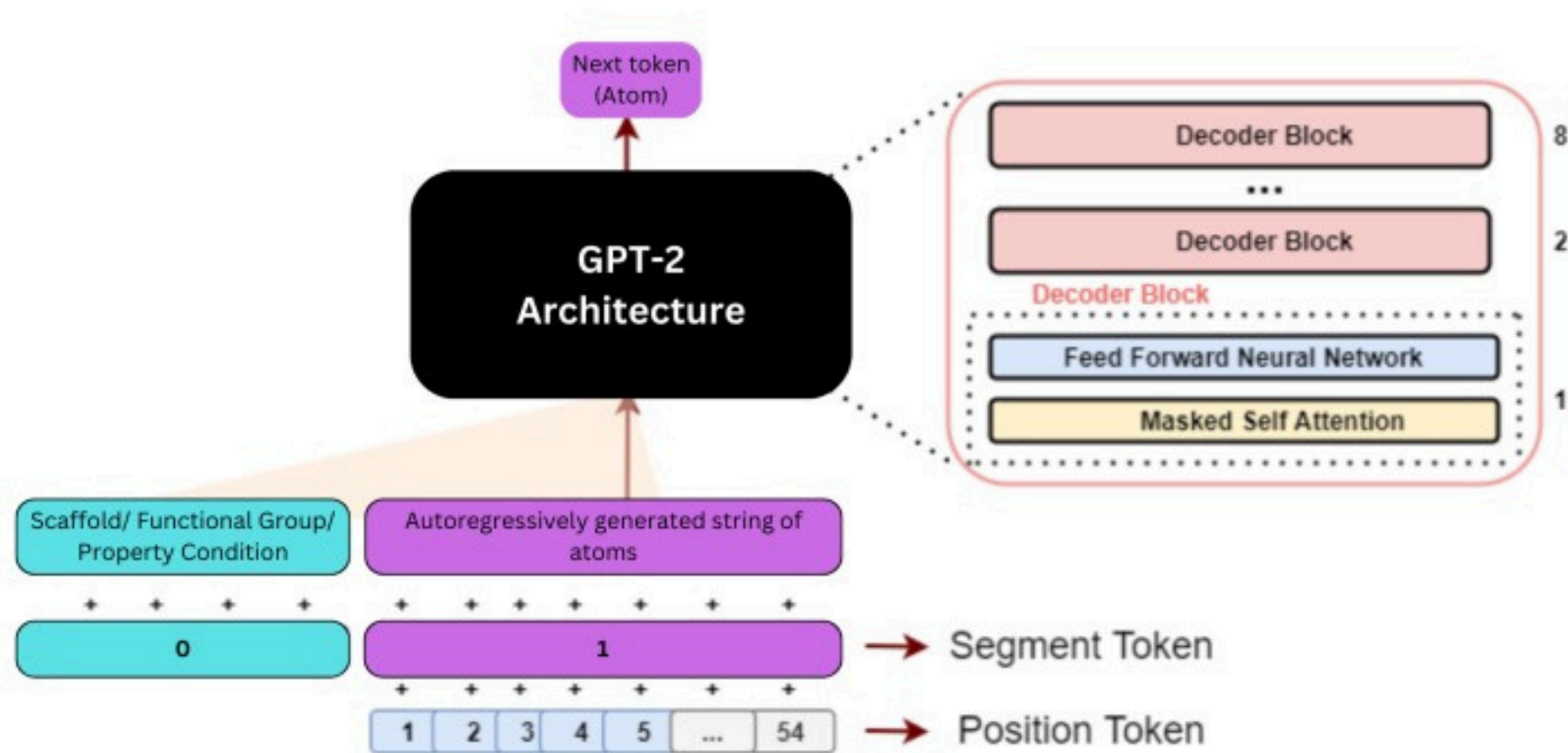


Figure 4.1: Model Architecture (Bagal et al., 2022)

Tokenizer	“c1cccc1” tokenization
Atomwise	[‘c’, ‘1’, ‘c’, ‘c’, ‘c’, ‘c’, ‘c’, ‘1’]
Kmer	[‘c1c’, ‘1cc’, ‘ccc’, ‘ccc’, ‘cc1’]
BPE/SMILESPE	[‘c’, ‘1’, ‘cccc’, ‘1’] - Dataset Dependant



Evaluation and Fine-Tuning

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

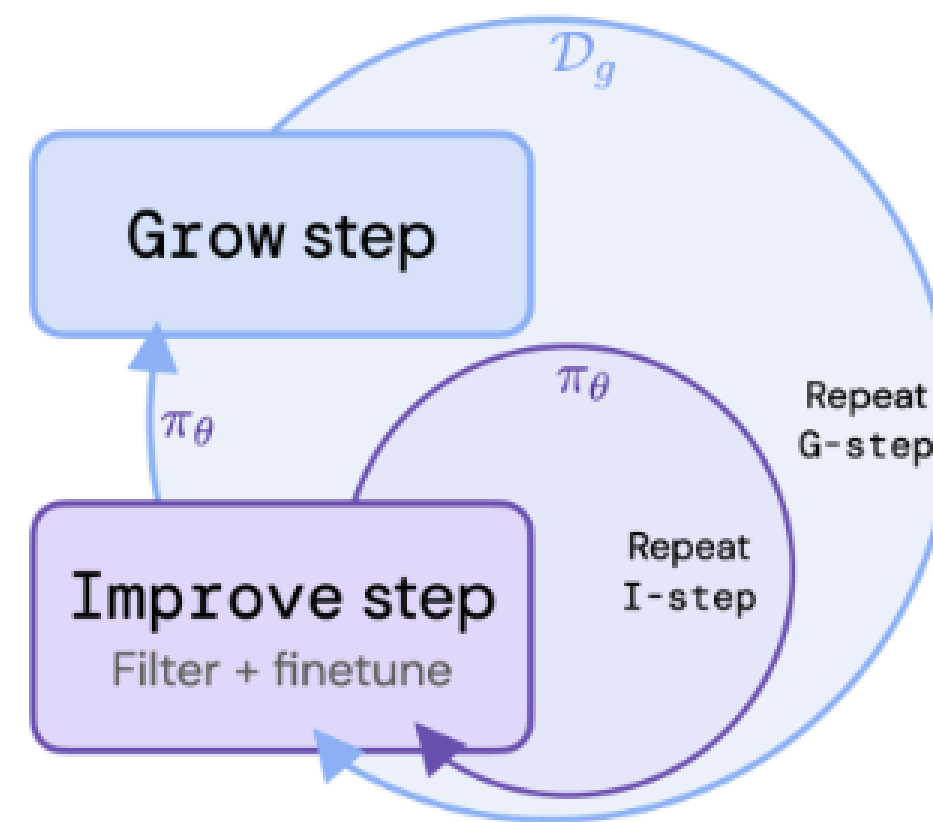


Figure 4.2: ReST method (*Gulcehre et al., 2023*)



Experiments

Over the course of the research, the following aspects of de novo molecular generation were studied:

- Model's ability to control structures and realize set properties in generated samples
- Effect of training dataset on quality of molecules generated
- Potential of RL based frameworks in aligning the generated samples to optimize for downstream tasks of commercial interest



Results

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

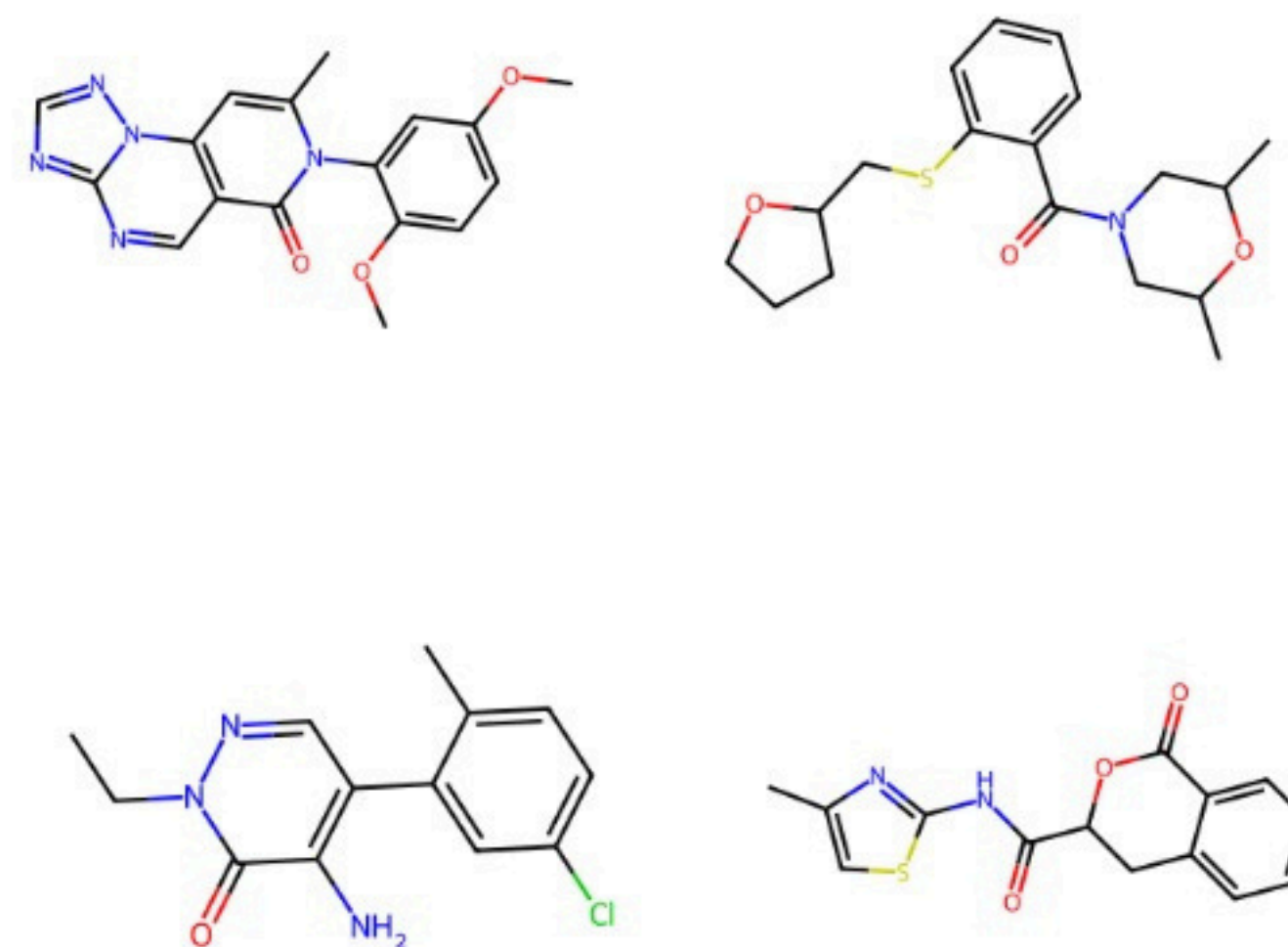
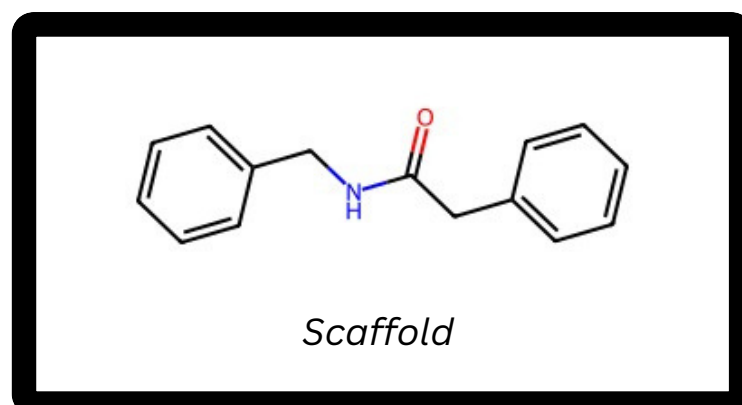


Figure 5.2: Unconditioned Generation Samples



Results



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

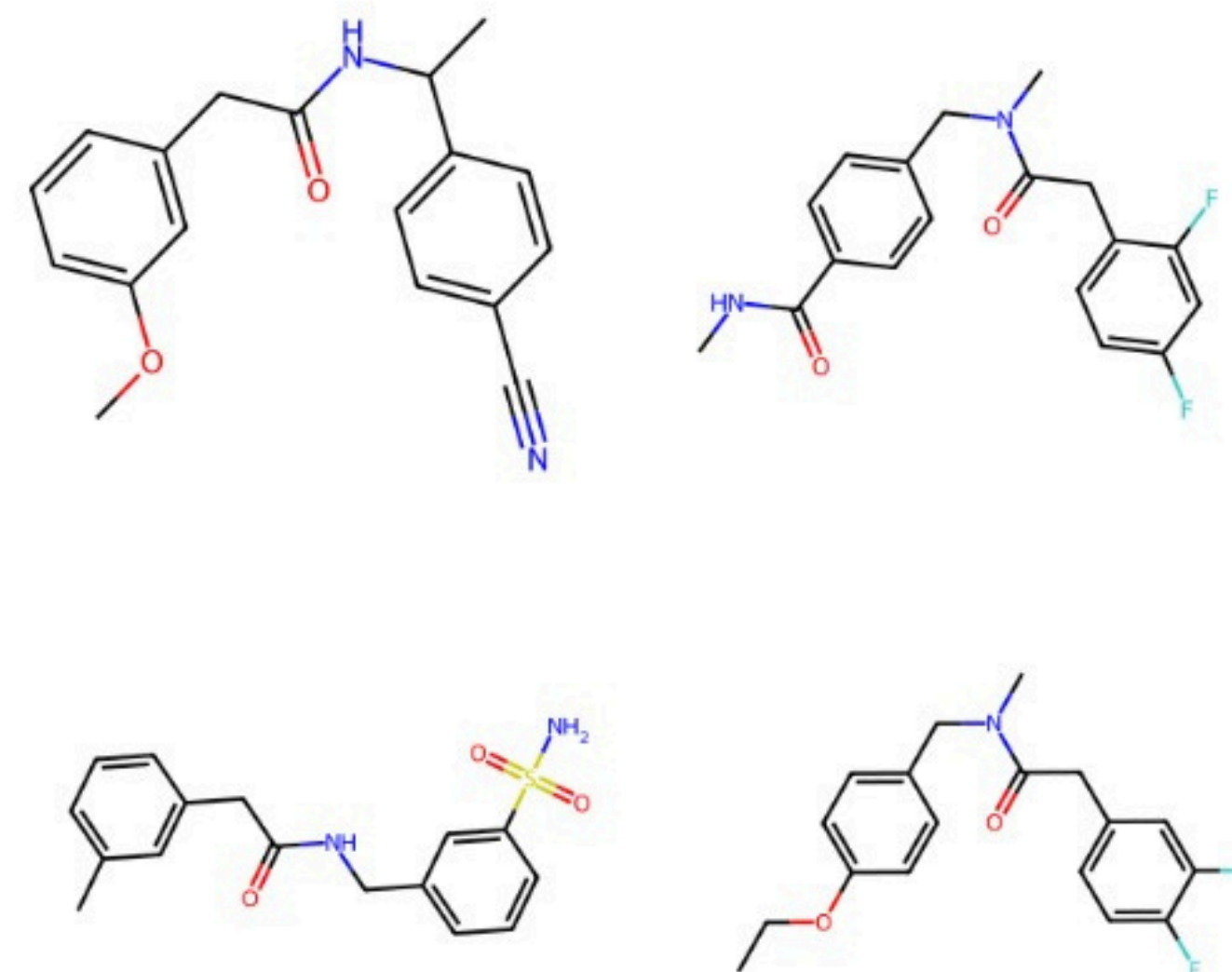


Figure 5.4: Scaffold Conditioned Generation Samples



Results

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*



Results

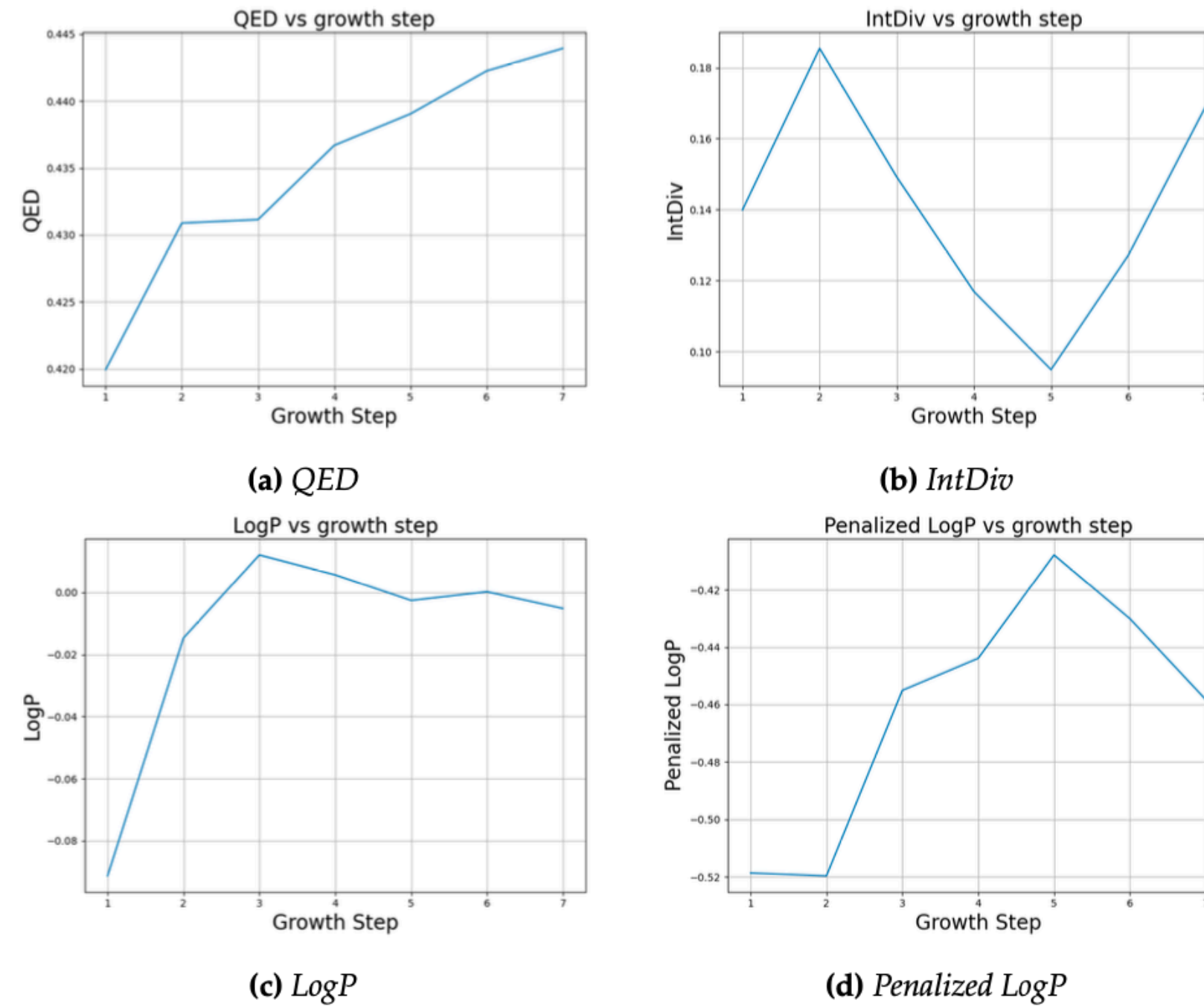


Figure 5.12: QED Generation Plots Over Time (a)



Results

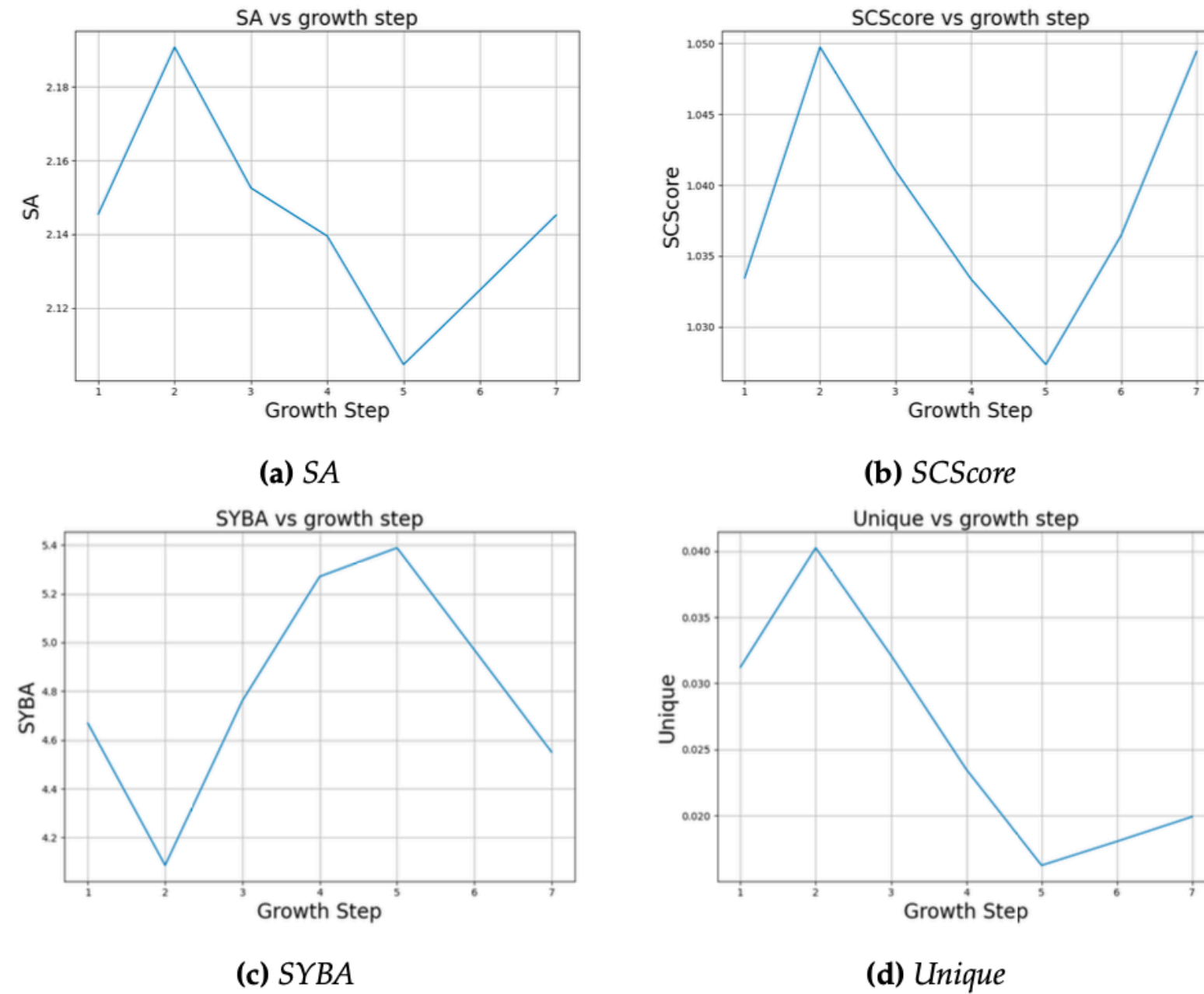


Figure 5.13: QED Generation Plots Over Time (b)



Conclusions and Future work

Conclusions:

- High degree of structural and functional control can be achieved
- The quality of dataset may be more important than the size of the dataset
- RL based fine-tuning methods such as ReST can significantly align the model's behaviour to maximize for desired properties

Next steps:

- Comprehensive study comparing all options available for data representation, tokenization, model architecture, fine-tuning
- Fine-tuning for specific therapeutic properties (anti-malarial, anti-bacterial, anti-fungal)

