

DDP End-Term Review

Conditional De Novo Molecular Generation

Saicharan Ganapathy
ED19B065
Guided by Prof. Dr. Nirav Pravinbhai Bhatt

June 04, 2024



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.



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



Algorithm/Solution Strategy

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: Datasets Information Sorted by File Size

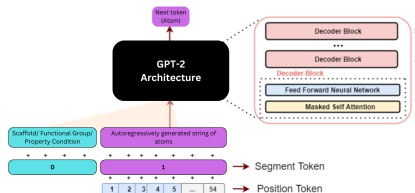


Figure: Model Architecture



Simulation Experiments

Model Conditioning	Performance	Comments
Unconditioned	Good	
Scaffold Conditioned	Good	
Property Conditioned	Good	
Scaffold + Property Conditioned	Adequate	More training, more data
Functional Group Conditioned	Poor	More data, better representation
Scaffold + Property + Functional Group Conditioned	Poor	More data, better representation



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: Performance of Unconditioned Generation on Evaluation Metrics

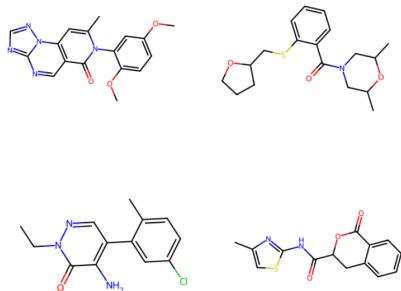
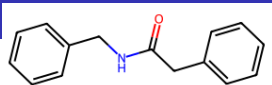


Figure: Unconditioned Generation Samples



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: Performance of Scaffold Conditioned Generation on Evaluation Metrics

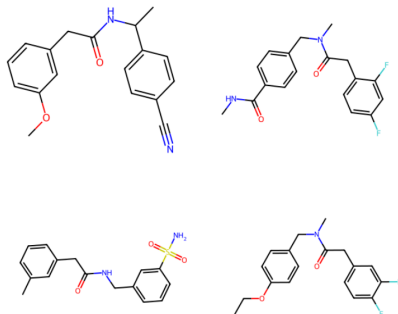
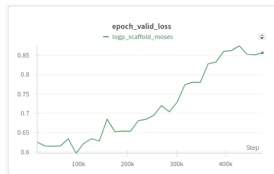


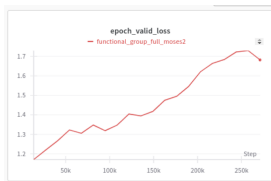
Figure: Scaffold Conditioned Generation Samples



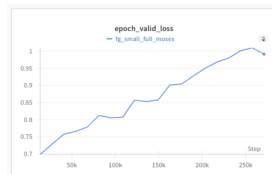
Simulation Experiments



Functional group + Scaffold +
Properties validation loss



Functional group (Full) validation loss



Functional group (Small) validation
loss



Conclusions:

- Great performance in unconditioned as well as scaffold/property conditioned generation.
- Inadequate performance with additional functional group data.

Next Steps

- Training on Other Larger Datasets (*Feb'24 - Mar'24*)
- Improved Techniques to Present the Data to the Model (*Feb'24 - Mar'24*)
- Fine-tuning for Specialised Downstream Tasks (*Mar'24 - Apr'24*)
- Using RL-based Techniques to Search the Chemical Space (*Mar'24 - Apr'24*)

