DrugGPT

Enhancing Efficiency and Quality in AI-Driven Drug Discovery

Molecular generation using machine learning shows great promise for accelerating drug discovery, but faces challenges in efficiency and scalability. We optimize key components of generative models for sequential representations of small molecules:

Tokenization methods, Model architectures, and Decoding strategies

By evaluating various approaches for each component, **we aim to identify optimal combinations that improve both efficiency and generation quality**.

Tokenization

Objectives

Compare Byte Pair Encoding (BPE) and Unigram Language Model (ULM) tokenizers on SAFE and SELFIES molecular representations

Methods

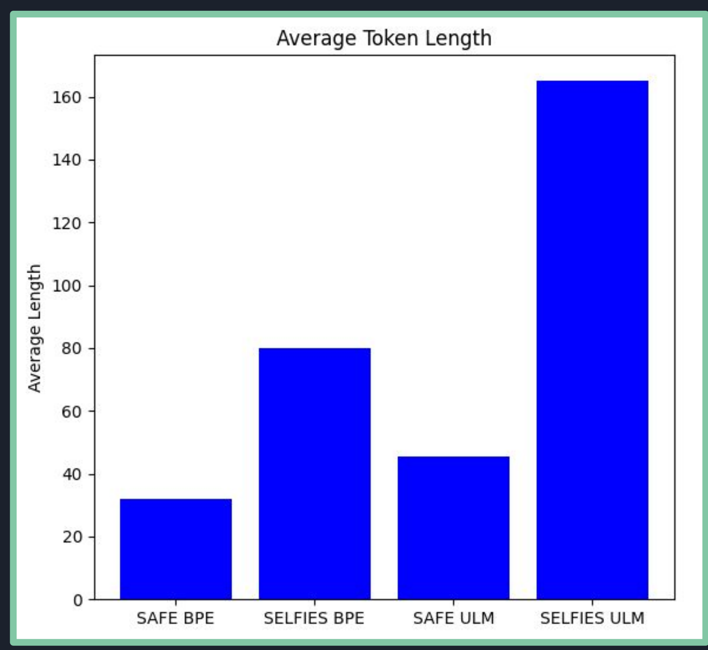
- **Evaluated tokenization efficiency across various vocabulary sizes**
- Tested downstream performance in molecular generation tasks

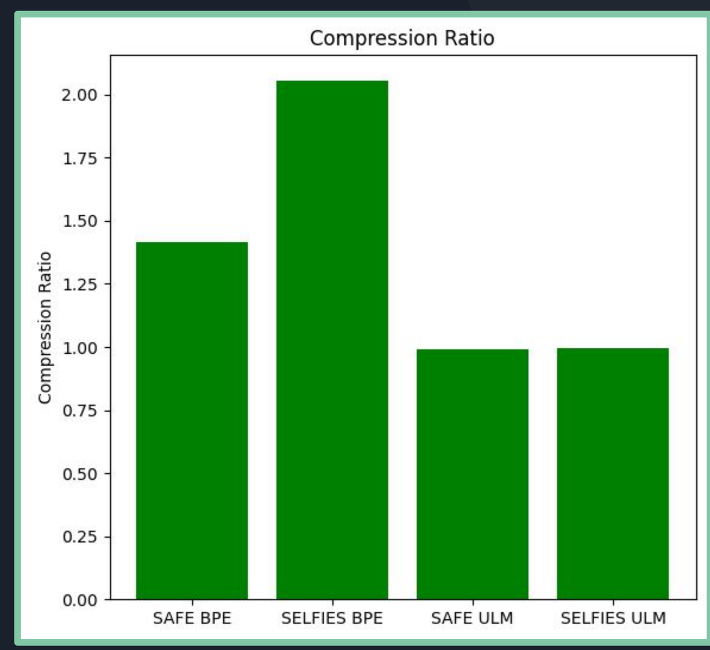
Key Results

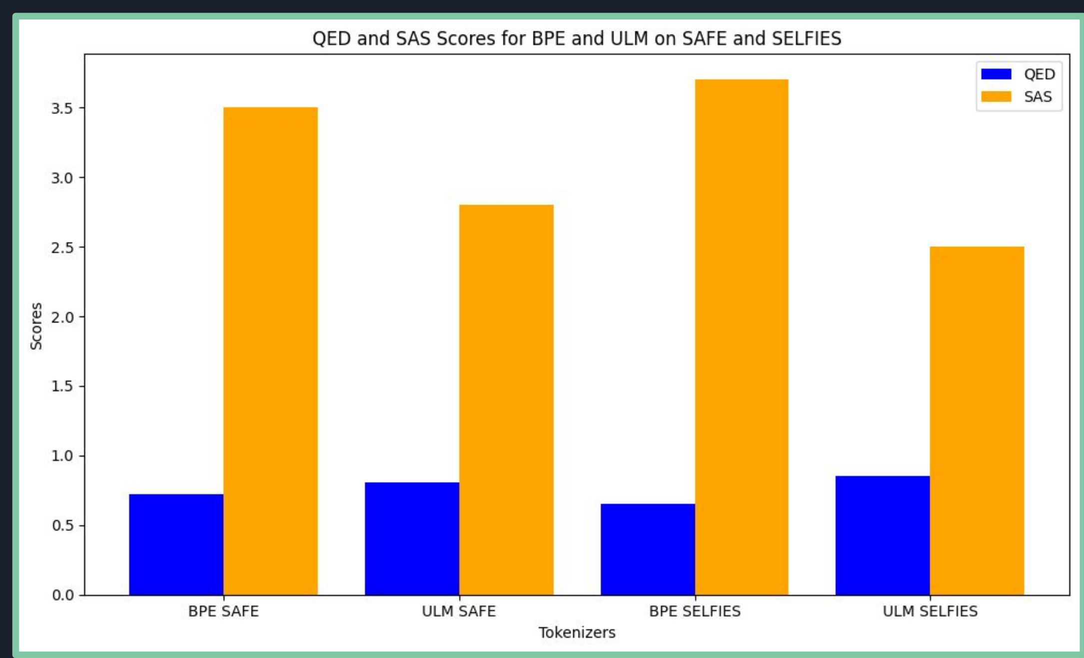
- **BPE achieves more compact representations**
- ULM, especially with SELFIES, produces molecules with better synthetic accessibility
- **Increasing vocabulary size improves efficiency, with diminishing returns** beyond a certain threshold

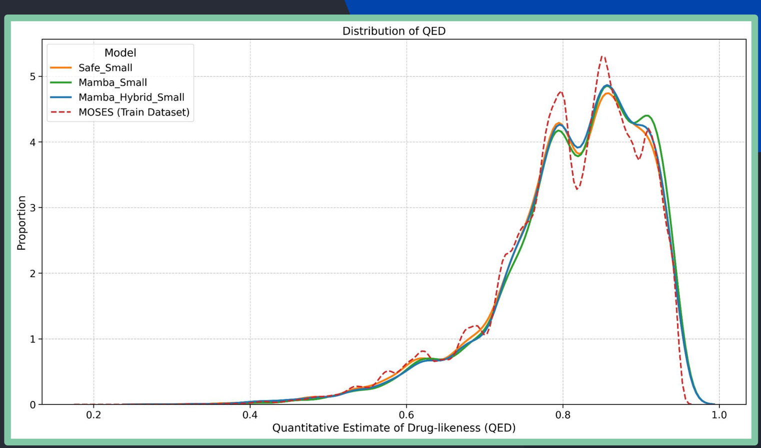
Conclusion

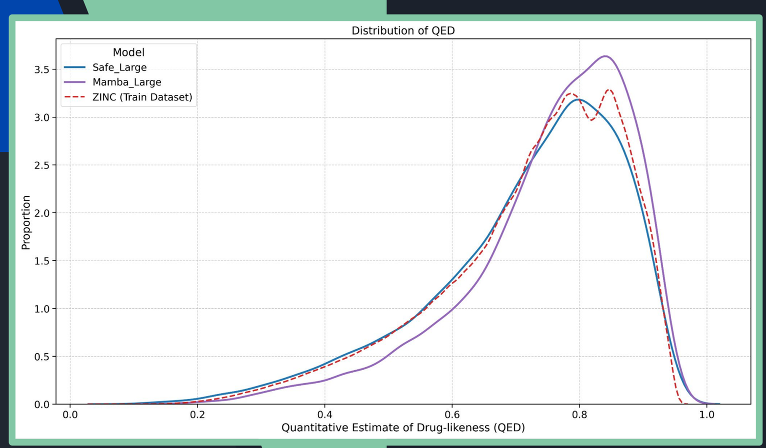
Tokenization choice significantly influences both efficiency and molecular generation performance, highlighting the need to balance these factors in AI-driven molecular design.

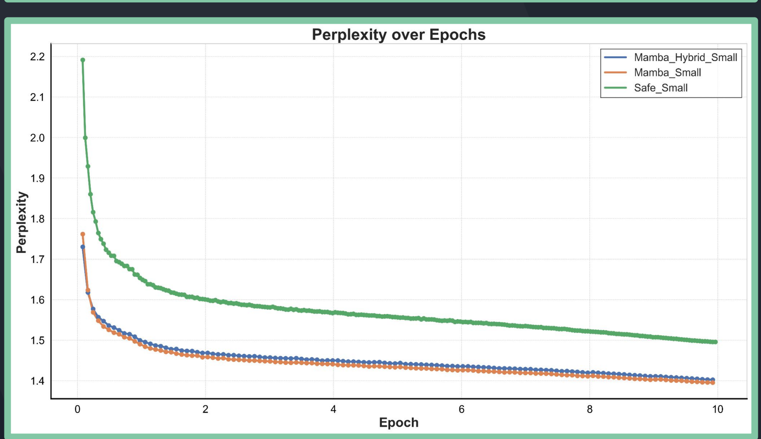


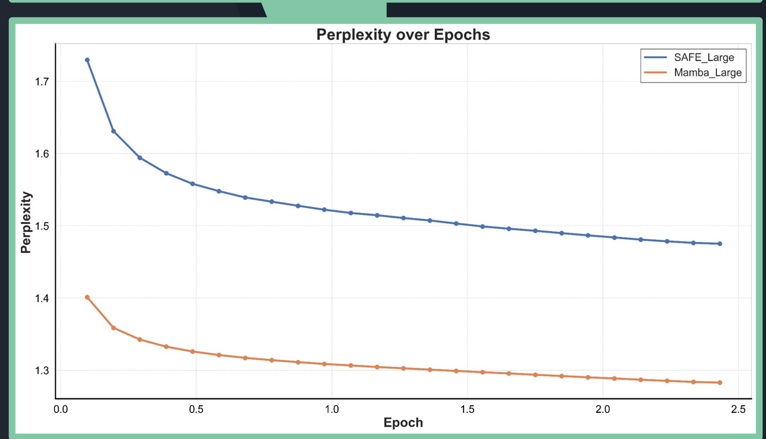












Model	Valid@10K↑	Unique@10K↑	Diversity↑
Safe_Large (87M)	0.98	1	0.880
Mamba_Large (94M)	1	1	0.873
Safe_Small (21M)	1	0.999	0.864
Mamba_Small_Hybrid (20M)	1	0.999	0.862
Mamba_Small (20M)	1	0.999	0.860

Architectures

Objectives

Compare **Transformer-based (SAFE-GPT) and State Space Model (MAMBA) architectures** for molecular generation

Methods

- Evaluated models with **~20M and ~90M parameters**
- Tested on **MOSES and ZINC datasets**
- Focused on generation quality and computational efficiency

Key Results

- Achieved comparable performance:
  - **98-100% valid molecules**
  - **99.9-100% unique molecules**
- Demonstrated lower perplexity
- **Reduced GPU power consumption by up to 30%**

Conclusion

State Space Models offer a computationally efficient alternative for molecular generation tasks, potentially enabling more efficient processing of larger datasets and complex molecular structures.

Decoders

Objectives

Compare the impact of different decoding strategies on generating molecules using SAFE-GPT models of varying sizes

Methods

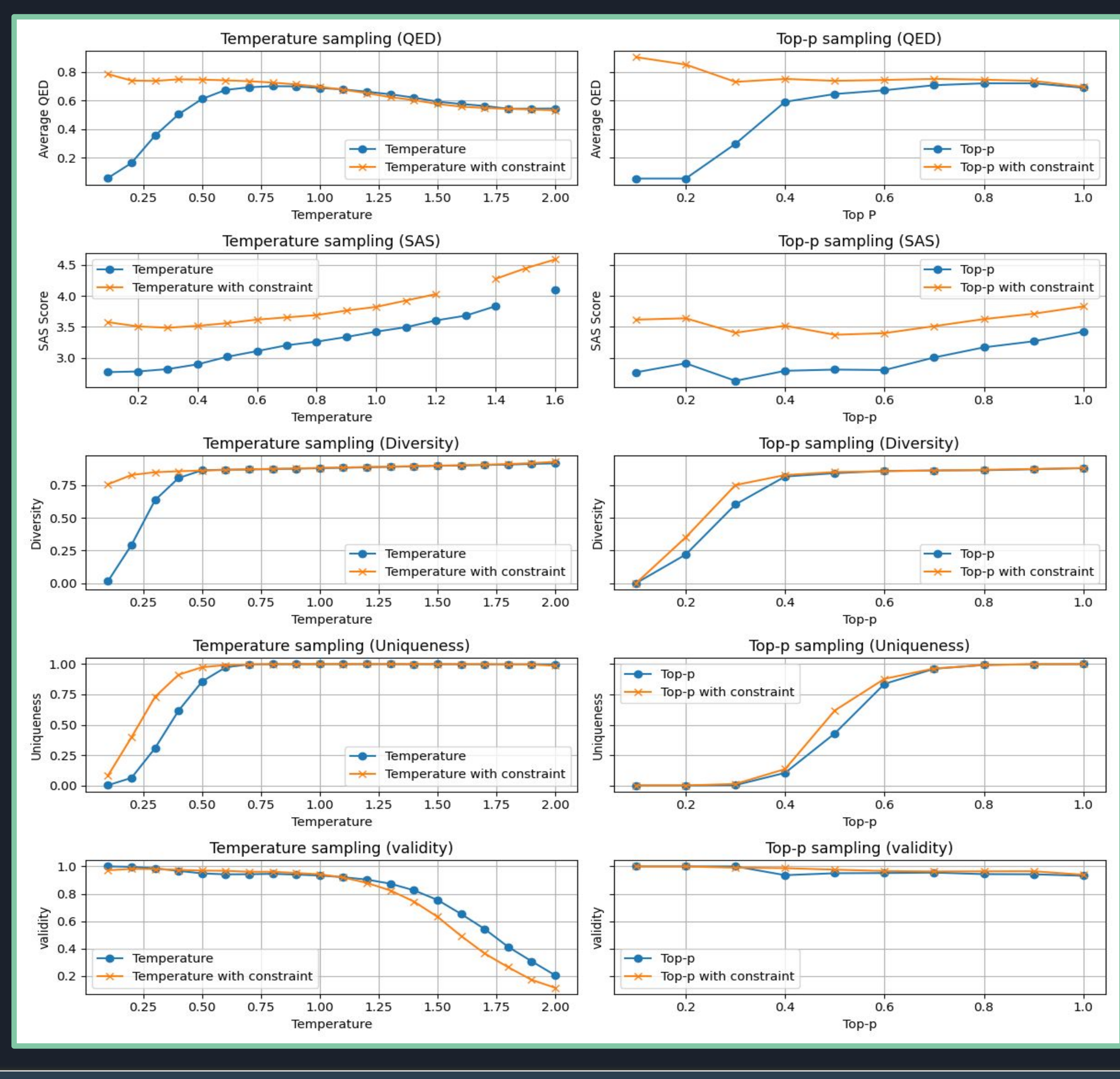
- Examined **consistency across large and small models**
- **Analyzed effect of constraining decoders** on output quality

Key Results

- **Small model:** Top-p sampling without repetition constraint performs best
- **Large model:** Temperature sampling with repetition penalty is most effective
- **Optimal decoding method depends on model size**

Conclusion

Carefully selecting decoders and constraining mechanisms can significantly improve the quality of molecules generated by SAFE-GPT models.



University of Cape Town
Department of Computer Science
www.sit.uct.ac.za

Mahomed Aadil Ally <ALLMAH002@myuct.ac.za>
Anri Lombard <LMBANR001@myuct.ac.za>
Gabriel Marcus <MRCGAB004@myuct.ac.za>
Supervised by **Prof. Jan Buys** <jan.buys@uct.ac.za>