MolReactGen: Generating Molecules and Reaction Templates with a Transformer Decoder Model

Stephan Holzgruber
Supervisor: Philipp Seidl
Institute for Machine Learning
JOHANNES KEPLER UNIVERSITY LINZ

1 Abstract

The master's thesis focuses on the world of *chemistry*, with the goal of supporting the discovery of drugs to cure diseases or sustainable materials for cleaner energy. The research explores the potential of a *transformer decoder* model in generating chemically feasible *molecules and reaction templates*. We begin with contrasting the performance of *GuacaMol* [1] for molecule generation with a transformer decoder architecture, assessing the influence of various *tokenizers* on performance. The study also involves *fine-tuning* a pre-trained language model and comparing its outcomes with a model trained *from scratch*. It utilizes multiple metrics, including the *Fréchet ChemNet Distance* [2], to evaluate the model's ability to generate new, valid molecules similar to the training data. The research indicates that the transformer decoder model outperforms the GuacaMol model in terms of this metric, and is also successful in generating known reaction templates.

2 Introduction

Application Area Chemistry

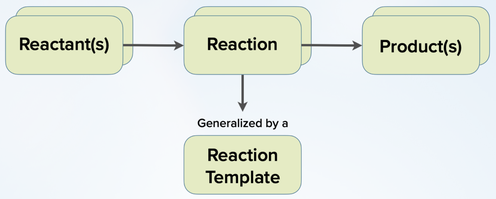

Figure 1: Terms used

Motivation

- Uncover novel structures and properties
- Identify new therapeutics with improved effectiveness and safety
- Develop innovative materials with unique properties
- Design environmentally friendly reactions and materials
- Enhance knowledge of chemical principles

Research Questions

- Performance compared to GuacaMol
- Effect of different tokenization approaches
- Feasibility of fine-tuning a pre-trained NLP model
- Potential of the model to generate reaction templates in addition to molecules

3 Methods

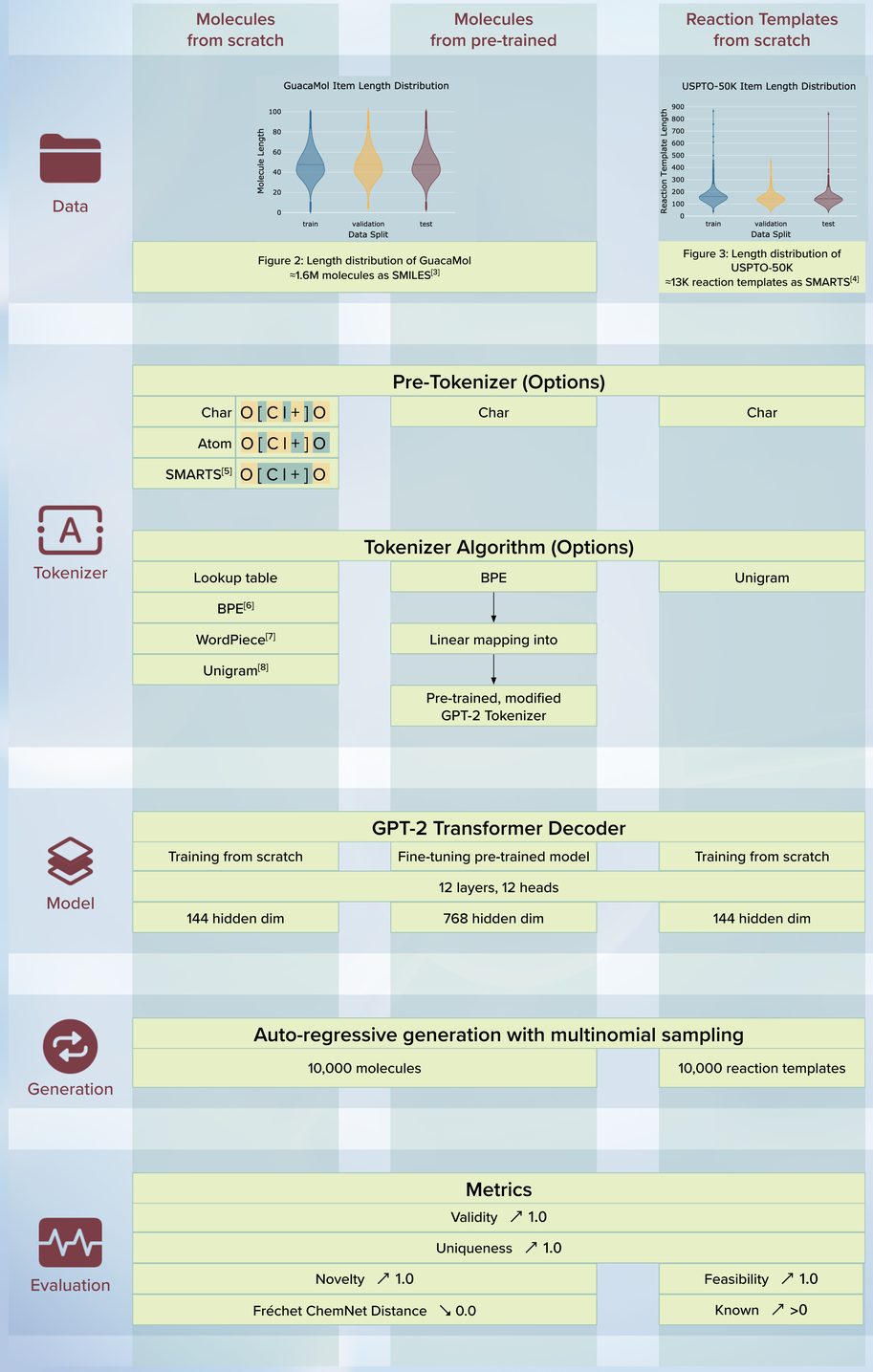

Figure 2: Length distribution of GuacaMol ≈1.6M molecules as SMILES[3]

Figure 3: Length distribution of USPTO-50K ≈13K reaction templates as SMARTS[4]

4 Results

Tokenizers

Pre-Tokenizer	Algorithm	Molecules		Reaction Templates	
		Vocab Size	FCD	Vocab Size	Known
Char	Lookup table	38	0.257	47	677
	BPE	44	0.252		
		88	0.247	88	650
				176	583
	WordPiece	88	0.256		
	Unigram	44	0.258		
		88	**0.239**	88	643
		176	0.246	176	625
Atom	Lookup table	50	0.293	86	678
SMARTS	Lookup table	106	0.277	947	**690**

Figure 4: Effect of tokenizers on main performance metrics

Molecules

Model	Metrics			
	Validity	Uniqueness	Novelty	FCD
GuacaMol	0.959	**1.000**	**0.994**	0.455
MolReactGen *from scratch*	0.976 ± 0.001	0.999 ± 0.000	0.939 ± 0.002	0.223 ± 0.005
MolReactGen *from pre-trained*	**0.992** ± 0.001	0.999 ± 0.000	0.793 ± 0.004	**0.203** ± 0.004

Figure 5: Generation results for molecules
Brown border represent the metric (FCD) our model was optimized for
Other models did improve different metrics
Numbers represent the mean and standard deviation (superscript) across five runs
FCD metric in GuacaMol paper calculated as exp(-0.2 FCD), back-calculated here as -5 ln $FCD_{GuacaMol}$

Reaction Templates

Model	Metrics			
	Validity	Uniqueness	Feasibility	Known
MolReactGen *from scratch*	0.745 ± 0.002	0.841 ± 0.004	0.101 ± 0.003	696 ± 10

Figure 6: Generation results for reaction templates

References

[1] N. Brown, M. Fiscato, M. H. S. Segler, and A. C. Vaucher, "GuacaMol: Benchmarking Models for de Novo Molecular Design," J. Chem. Inf. Model., vol. 59, no. 3, pp. 1096–1108, Mar. 2019
[2] K. Preuer, P. Renz, T. Unterthiner, S. Hochreiter, and G. Klambauer, "Fréchet ChemNet Distance: A Metric for Generative Models for Molecules in Drug Discovery," J. Chem. Inf. Model., vol. 58, no. 9, pp. 1736–1741, Sep. 2018
[3] D. Weininger, "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules," J. Chem. Inf. Model., vol. 28, no. 1, pp. 31–36, Feb. 1988
[4] Daylight Chemical Information Systems, Inc., "Daylight Theory: SMARTS - A Language for Describing Molecular Patterns," SMARTS - A Language for Describing Molecular Patterns. https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html
[5] Bespoke RegEx, inspired by P. Schwaller et al., "Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction," ACS Cent. Sci., vol. 5, no. 9, pp. 1572–1583, Sep. 2019
[6] Haddow, and A. Birch, "Neural Machine Translation of Rare Words with Subword Units." arXiv, Jun. 10, 2016. Accessed: Dec. 12, 2022
[7] M. Schuster and K. Nakajima, "Japanese and Korean voice search," in 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan: IEEE, Mar. 2012, pp. 5149–5152
[8] T. Kudo, "Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates." arXiv, Apr. 29, 2018

stephan.holzgruber@gmail.com
github.com/hogru/molreactgen
huggingface.co/hogru