Generating Molecules and Reaction Templates with a Transformer Decoder Model

Master Thesis Seminar

Supervisor
Philipp Seidl

Stephan Holzgruber

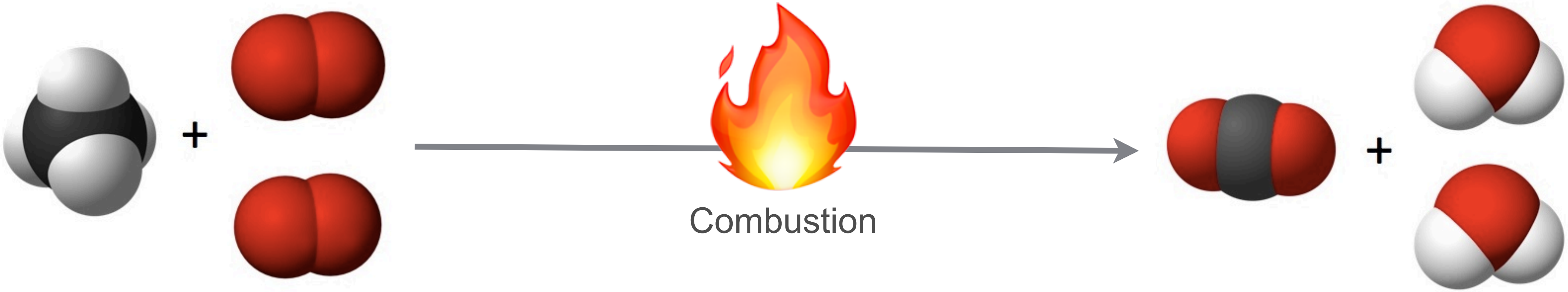
June 12th, 2023

Github	hogru/MolReactGen
Hugging Face	hogru/MolReactGen-GuacaMol-Molecules
	hogru/MolReactGen-USPTO50K-Reaction-Templates

JOHANNES KEPLER UNIVERSITY LINZ
Altenberger Straße 69
4040 Linz, Austria
jku.at

Application Area Chemistry

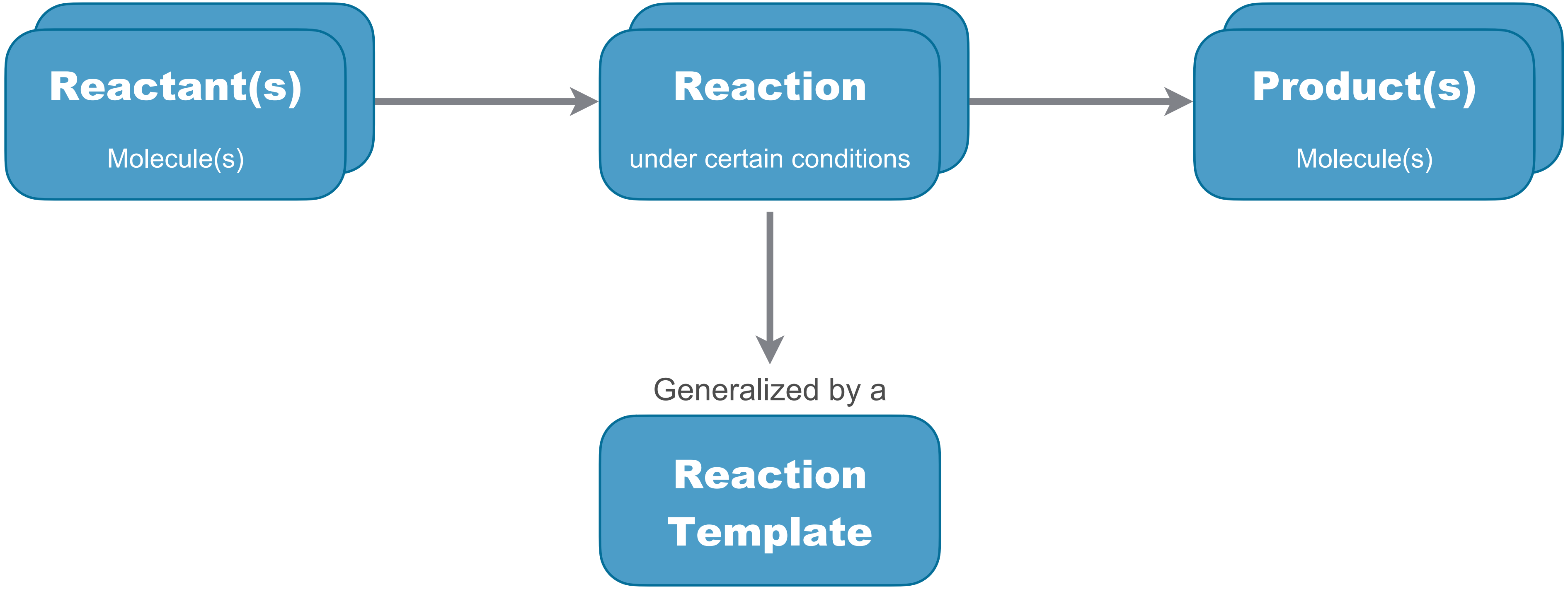
Bunsen Burner



$CH_4 \quad + \quad 2O_2 \qquad\qquad CO_2 \quad + \quad 2H_2O$

Application Area Chemistry

Terms Used



Reactant(s)

Molecule(s)

Reaction

under certain conditions

Product(s)

Molecule(s)

Generalized by a

Reaction Template

Motivation

- Expanding **chemical space**: Uncover novel structures and properties
- **Drug discovery**: Identify new therapeutic agents with improved effectiveness and safety
- **Materials** science: Develop innovative materials with unique properties
- **Environmental sustainability**: Design environmentally friendly processes and materials
- **Fundamental understanding**: Enhance knowledge of chemical principles and reaction mechanisms

Research Questions

- **GuacaMol**[1] considered a reference paper/model for molecule generation
- Research Questions
  - What is the **performance of a transformer decoder** architecture compared to GuacaMol?
  - What is the effect of different **tokenization approaches**?
  - Can we use a model pre-trained on natural language as a basis for **fine-tuning a "molecule language" model**?
  - Can the transformer decoder model also be used to **generate reaction templates**?

[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, doi: 10.1021/acs.jcim.8b00839.

Pipeline 1/3 — Molecules From Scratch

Data

GuacaMol
ChEMBL

Tokenizer

Pre-Tokenizer	Char
	Atom
	SMARTS
Algorithm	Wordlevel
	BPE
	WordPiece
	Unigram

Model / Training

GPT-2	12 Layers
	12 Heads
	144 Hidden Dim

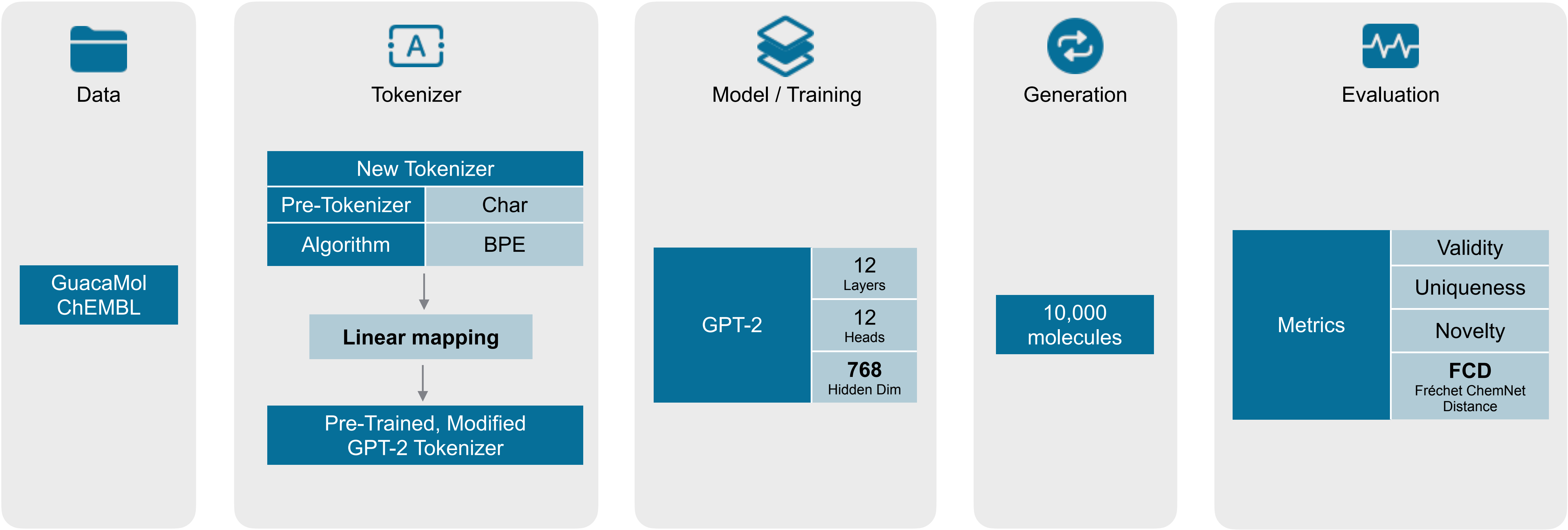
Generation

10,000
molecules

Evaluation

Metrics	Validity
	Uniqueness
	Novelty
	FCD Fréchet ChemNet Distance

Pipeline 2/3 — Molecules from Pre-Trained Model



Data

GuacaMol ChEMBL

Tokenizer

New Tokenizer	
Pre-Tokenizer	Char
Algorithm	BPE

Linear mapping

Pre-Trained, Modified GPT-2 Tokenizer

Model / Training

GPT-2	
	12 Layers
	12 Heads
	768 Hidden Dim

Generation

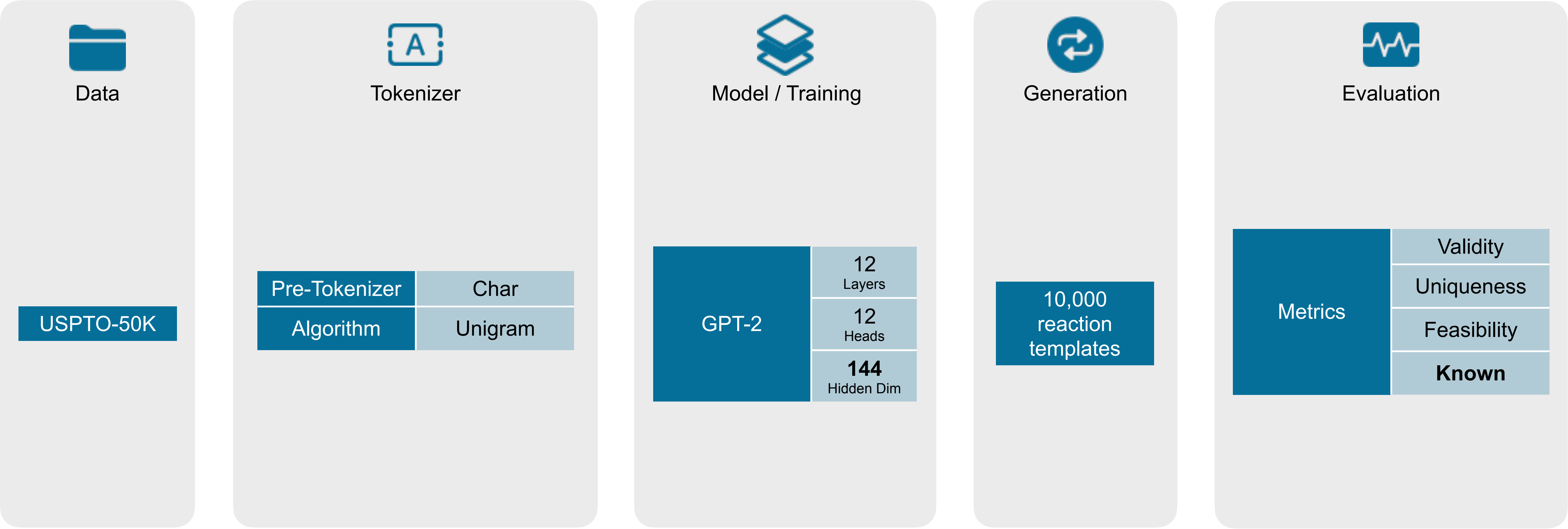
10,000 molecules

Evaluation

Metrics	
	Validity
	Uniqueness
	Novelty
	FCD Fréchet ChemNet Distance

JOHANNES KEPLER UNIVERSITY LINZ

Pipeline 3/3 — Reaction Templates From Scratch



Data

USPTO-50K

Tokenizer

Pre-Tokenizer	Char
Algorithm	Unigram

Model / Training

GPT-2	12 Layers
	12 Heads
	144 Hidden Dim

Generation

10,000 reaction templates

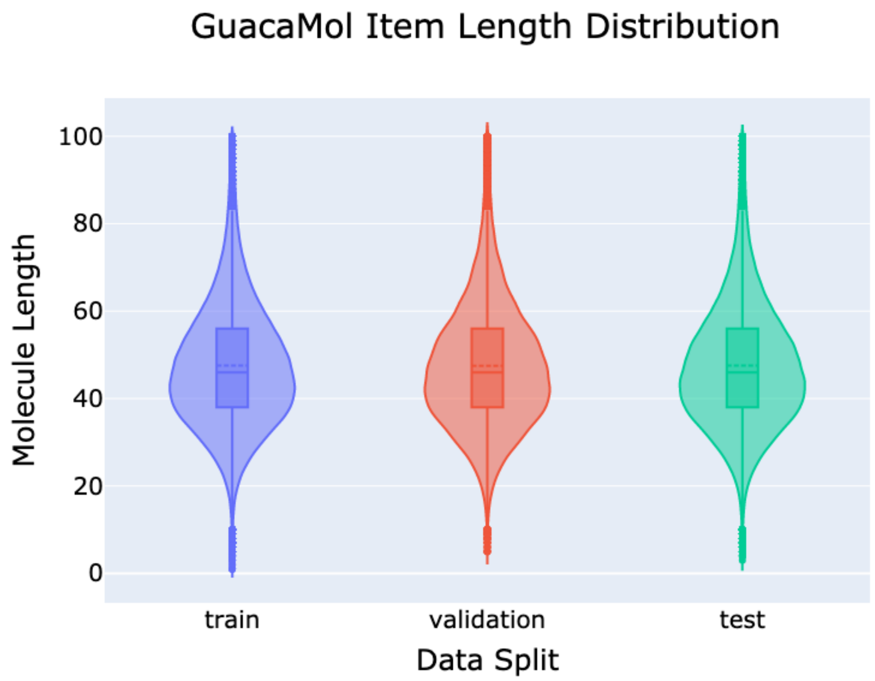
Evaluation

Metrics	Validity
	Uniqueness
	Feasibility
	Known

GuacaMol Dataset

Molecules represented as SMILES[2]

Data Split	Count	Percent
Train	1,273,103	80 %
Validation	79,567	5 %
Test	238,705	15 %
Total	**1,591,375**	**100 %**


GuacaMol Item Length Distribution

Example SMILES	O=C(O)C1CCC(OCC2CC(F)CN2C(=O)Cc2ccc(NC(=O)N3CCc4ccccc43)c(Cl)c2)CC1

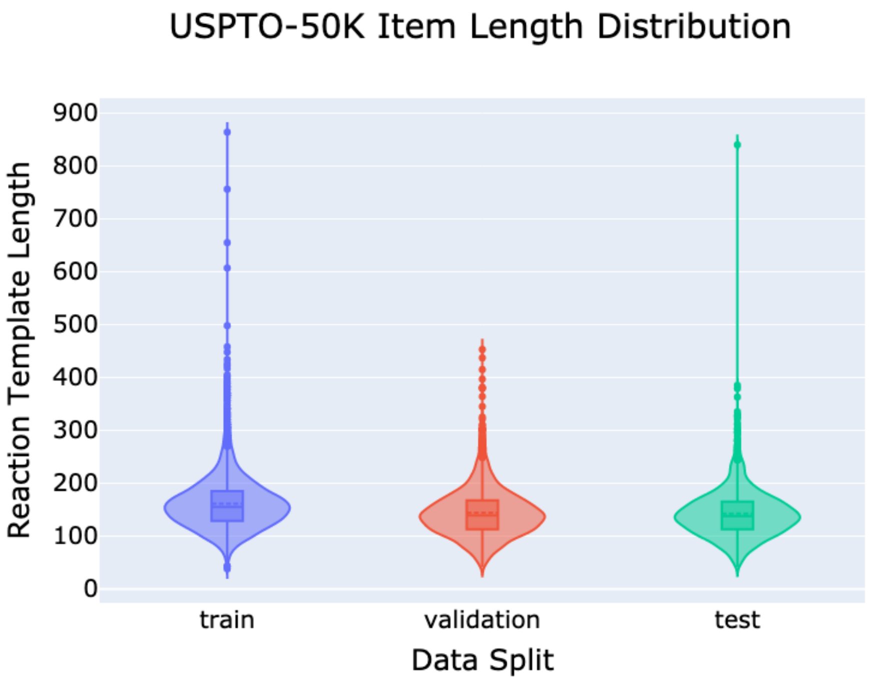
[2] D. Weininger, “SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules,” *J. Chem. Inf. Model.*, vol. 28, no. 1, pp. 31–36, Feb. 1988, doi: 10.1021/ci00057a005.

USPTO-50K Dataset

Reaction Templates represented as SMARTS[3]

Data Split	Count	Percent
Train	7,877	62 %
Validation	2,413	19 %
Test	2,336	19 %
Total	**12,626**	**100 %**

Reaction templates are non-unique and non-disjunct across splits
➔ Remove double entries and make sets disjunct
➔ 25% of data left



Example SMARTS	[#7;a:4]:[c:3]:[c;H0;D3;+0:1](:[#7;a:2])-[n;H0;D3;+0:9]1:[#7;a:5]:[c:6]:[#7;a:7]:[c:8]:1>> Cl-[c;H0;D3;+0:1](:[#7;a:2]):[c:3]:[#7;a:4]. [#7;a:5]1:[c:6]:[#7;a:7]:[c:8]:[nH;D2;+0:9]:1

[3] 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 (accessed Apr. 11, 2022).

Tokenization Approaches

Component	Options	Comment, Example
Normalizer	—	Not needed/used
Pre-Tokenizer	Char	O [C l +] O
	Atom	O [C l +] O
	SMARTS[4]	O [C l +] O
Subword Tokenization Algorithm	WordLevel	A simple lookup table
	BPE[5]	Used by e.g. GPT-2 as *byte-level* BPE
	WordPiece[6]	Used by e.g. BERT
	Unigram[7]	Algorithm for SentencePiece[8], used by e.g. T5
Post-Processor	for WordPiece only	
Decoder	Add BOS and EOS	Did not use GPT-2 default “<|endoftext|>”

[4] 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, doi: 10.1021/acscentsci.9b00576.
[5] Haddow, and A. Birch, “Neural Machine Translation of Rare Words with Subword Units.” arXiv, Jun. 10, 2016. Accessed: Dec. 12, 2022. [Online]. Available: http://arxiv.org/abs/1508.07909
[6] 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. doi: 10.1109/ICASSP.2012.6289079.
[7] T. Kudo, “Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates.” arXiv, Apr. 29, 2018. doi: 10.48550/arXiv.1804.10959.
[8] T. Kudo and J. Richardson, “SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing.” arXiv, Aug. 19, 2018. doi: 10.48550/arXiv.1808.06226.

Metrics

Metric	Applies to	Pseudo Formula	Target	Description
Validity	Molecules	$\frac{items_{valid}}{items_{generated}}$	↗ 1.0	Valid ≙ generated molecule can be parsed by `rdkit`
	Reaction Templates			Valid ≙ generated reactant(s) / product(s) comprise valid molecules
Uniqueness	Molecules	$\frac{items_{unique}}{items_{valid}}$	↗ 1.0	Unique ≙ valid item generated only once
	Reaction Templates			
Novelty	Molecules	$\frac{items_{novel}}{items_{unique}}$	↗ 1.0	Novel ≙ unique molecule not in training set
Fréchet ChemNet Distance (FCD)	Molecules	see paper[9]	↘ 0.0	The similarity between two sets of molecules, in this case the GuacaMol training set and the generated valid molecules
			↗ 1.0	$FCD_{GuacaMol} = e^{-0.2FCD}$
Feasibility	Reaction Templates	$\frac{items_{feasible}}{items_{unique}}$	↗ 1.0	Feasible ≙ ∃ product in validation/test set that the generated reaction template can be applied to Applied to ≙ `rdkit` can compute a reaction
Known	Reaction Templates	—	↗ >0	Known ≙ Generated reaction template *not* in training set, but in validation and/or test set

[9] 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, doi: 10.1021/acs.jcim.8b00234.

Results — Molecules

Dataset	Model	Metrics				
		Validity	Uniqueness	Novelty	FCD	FCD Guacamol
GuacaMol Molecules	GuacaMol	0.959	**1.000**	**0.994**	0.455	0.913
	MolReactGen *from scratch*	0.976 $^{\pm 0.001}$	0.999 $^{\pm 0.000}$	0.939 $^{\pm 0.002}$	0.223 $^{\pm 0.005}$	0.956 $^{\pm 0.001}$
	MolReactGen *fine-tuned*	**0.992** $^{\pm 0.001}$	0.999 $^{\pm 0.000}$	0.793 $^{\pm 0.004}$	**0.203** $^{\pm 0.004}$	**0.960** $^{\pm 0.001}$

Red 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 not stated in GuacaMol paper, calculated as $-5 \ln FCD_{GuacaMol}$

Results — Reaction Templates

Dataset	Model	Metrics			
		Validity	Uniqueness	Feasibility	Known
USPTO-50K **Reaction Templates**	MolReactGen *from scratch*	0.745 $^{\pm 0.002}$	0.841 $^{\pm 0.004}$	0.101 $^{\pm 0.003}$	696 $^{\pm 10}$

Red border represents the metric (Known) our model was optimized for; other models did improve different metrics
Numbers represent the mean and standard deviation (superscript) across five runs

Conclusion

- Used GuacaMol data and metrics as a reference for molecule generation
- Encoded the molecule SMILES with different pre-tokenizers and tokenization algorithms
- Trained a GPT-2 transformer decoder model from scratch
- Compared performance with GuacaMol
- Mapped molecules vocabulary into GPT-2 vocabulary
- Fine-tuned the pre-trained GPT-2 model
- Compared performance of training from scratch with fine-tuning
- Used USPTO-50K to train the model on reaction templates
- Showed that the model can generate reaction templates it has not seen before

JKU

JOHANNES KEPLER
UNIVERSITY LINZ