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

Master Thesis Defensio

Author
Stephan Holzgruber

Supervisor
Günter Klambauer

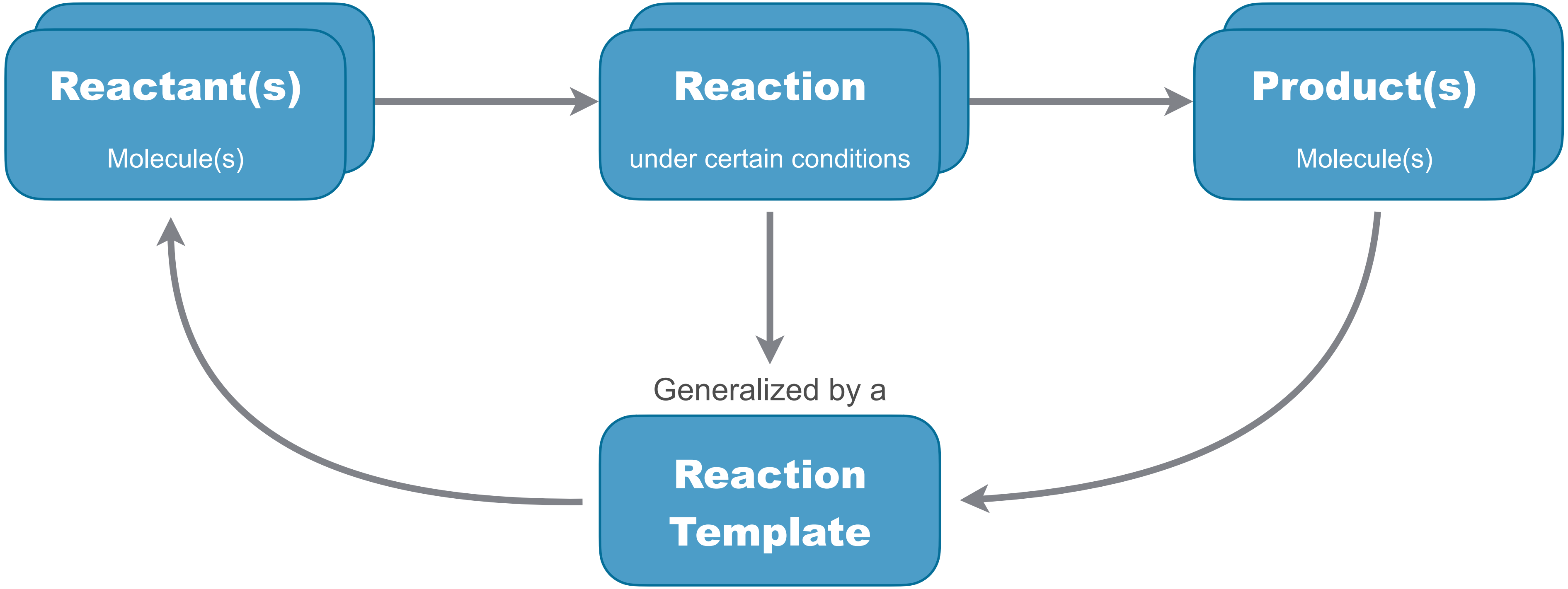
Co-Supervisor
Philipp Seidl

May 29th, 2024

E-Mail	stephan.holzgruber@gmail.com
Master Thesis	JKU ePUB
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: Chemical Reactions



Context / Motivation

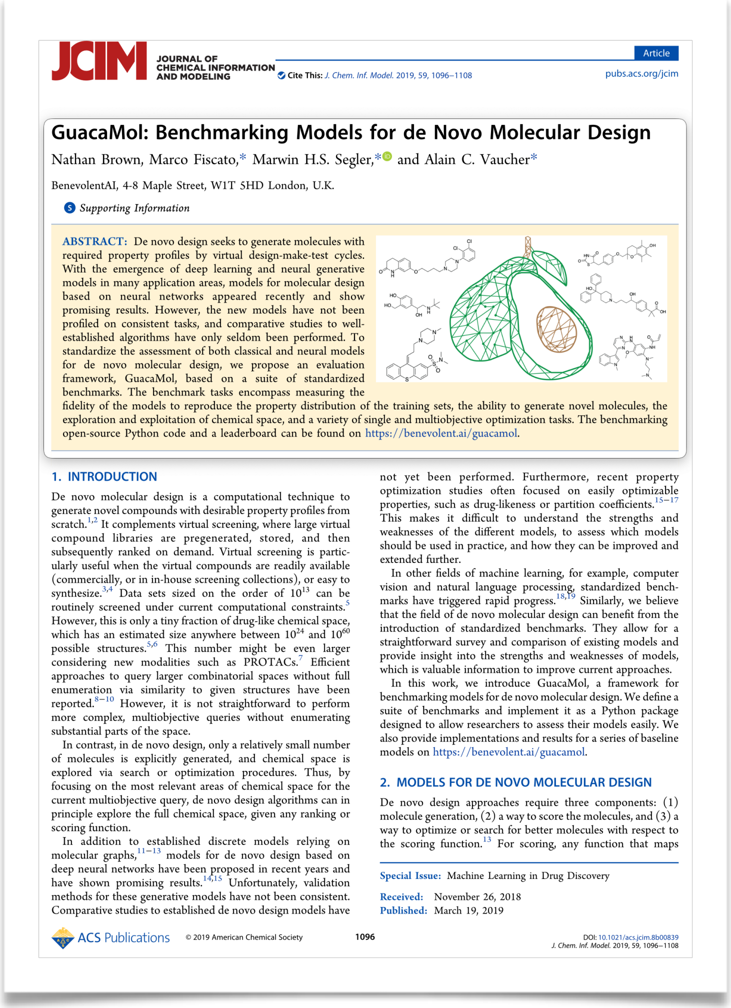
What's it all about?	Design new product molecules
What for?	New drugs to address gaps in disease treatment, ...
Why care?	Multi-drug resistant bacteria, neuro-degenerative diseases, ...
The issue?	Huge search space (~10^{33}[1]), cost, time, skills, equipment, ...
How?	Generative model (LLM) ➔ Similar, but different molecules
How is that useful?	Rapid, cheaper creation and screening of candidate drugs
Fine, and then?	How to synthesize those molecules?
Any ideas?	Start again, with generating reaction templates

[1] P. G. Polishchuk et al., “Estimation of the size of drug-like chemical space based on GDB-17 data,” J Comput Aided Mol Des, vol. 27, no. 8, pp. 675–679, Aug. 2013, doi: 10.1007/s10822-013-9672-4

Research Questions

- **GuacaMol**[2] considered a reference paper/model for de novo 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**?



JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

Cite This: *J. Chem. Inf. Model.* 2019, 59, 1096−1108

pubs.acs.org/jcim

GuacaMol: Benchmarking Models for de Novo Molecular Design

Nathan Brown, Marco Fiscato,* Marwin H.S. Segler,* and Alain C. Vaucher*

BenevolentAI, 4-8 Maple Street, W1T 5HD London, U.K.

Supporting Information

ABSTRACT: De novo design seeks to generate molecules with required property profiles by virtual design-make-test cycles. With the emergence of deep learning and neural generative models in many application areas, models for molecular design based on neural networks appeared recently and show promising results. However, the new models have not been profiled on consistent tasks, and comparative studies to well-established algorithms have only seldom been performed. To standardize the assessment of both classical and neural models for de novo molecular design, we propose an evaluation framework, GuacaMol, based on a suite of standardized benchmarks. The benchmark tasks encompass measuring the fidelity of the models to reproduce the property distribution of the training sets, the ability to generate novel molecules, the exploration and exploitation of chemical space, and a variety of single and multiobjective optimization tasks. The benchmarking open-source Python code and a leaderboard can be found on https://benevolent.ai/guacamol.

1. INTRODUCTION

De novo molecular design is a computational technique to generate novel compounds with desirable property profiles from scratch.[1,2] It complements virtual screening, where large virtual compound libraries are pregenerated, stored, and then subsequently ranked on demand. Virtual screening is particularly useful when the virtual compounds are readily available (commercially, or in in-house screening collections), or easy to synthesize.[3,4] Data sets sized on the order of 10^{13} can be routinely screened under current computational constraints.[5] However, this is only a tiny fraction of drug-like chemical space, which has an estimated size anywhere between 10^{24} and 10^{60} possible structures.[5,6] This number might be even larger considering new modalities such as PROTACs.[7] Efficient approaches to query larger combinatorial spaces without full enumeration via similarity to given structures have been reported.[8−10] However, it is not straightforward to perform more complex, multiobjective queries without enumerating substantial parts of the space.

In contrast, in de novo design, only a relatively small number of molecules is explicitly generated, and chemical space is explored via search or optimization procedures. Thus, by focusing on the most relevant areas of chemical space for the current multiobjective query, de novo design algorithms can in principle explore the full chemical space, given any ranking or scoring function.

In addition to established discrete models relying on molecular graphs,[11−13] models for de novo design based on deep neural networks have been proposed in recent years and have shown promising results.[14,15] Unfortunately, validation methods for these generative models have not been consistent. Comparative studies to established de novo design models have not yet been performed. Furthermore, recent property optimization studies often focused on easily optimizable properties, such as drug-likeness or partition coefficients.[15−17] This makes it difficult to understand the strengths and weaknesses of the different models, to assess which models should be used in practice, and how they can be improved and extended further.

In other fields of machine learning, for example, computer vision and natural language processing, standardized benchmarks have triggered rapid progress.[18,19] Similarly, we believe that the field of de novo molecular design can benefit from the introduction of standardized benchmarks. They allow for a straightforward survey and comparison of existing models and provide insight into the strengths and weaknesses of models, which is valuable information to improve current approaches.

In this work, we introduce GuacaMol, a framework for benchmarking models for de novo molecular design. We define a suite of benchmarks and implement it as a Python package designed to allow researchers to assess their models easily. We also provide implementations and results for a series of baseline models on https://benevolent.ai/guacamol.

2. MODELS FOR DE NOVO MOLECULAR DESIGN

De novo design approaches require three components: (1) molecule generation, (2) a way to score the molecules, and (3) a way to optimize or search for better molecules with respect to the scoring function.[13] For scoring, any function that maps

Special Issue: Machine Learning in Drug Discovery

Received: November 26, 2018
Published: March 19, 2019

ACS Publications © 2019 American Chemical Society 1096

DOI: 10.1021/acs.jcim.8b00839
J. Chem. Inf. Model. 2019, 59, 1096−1108

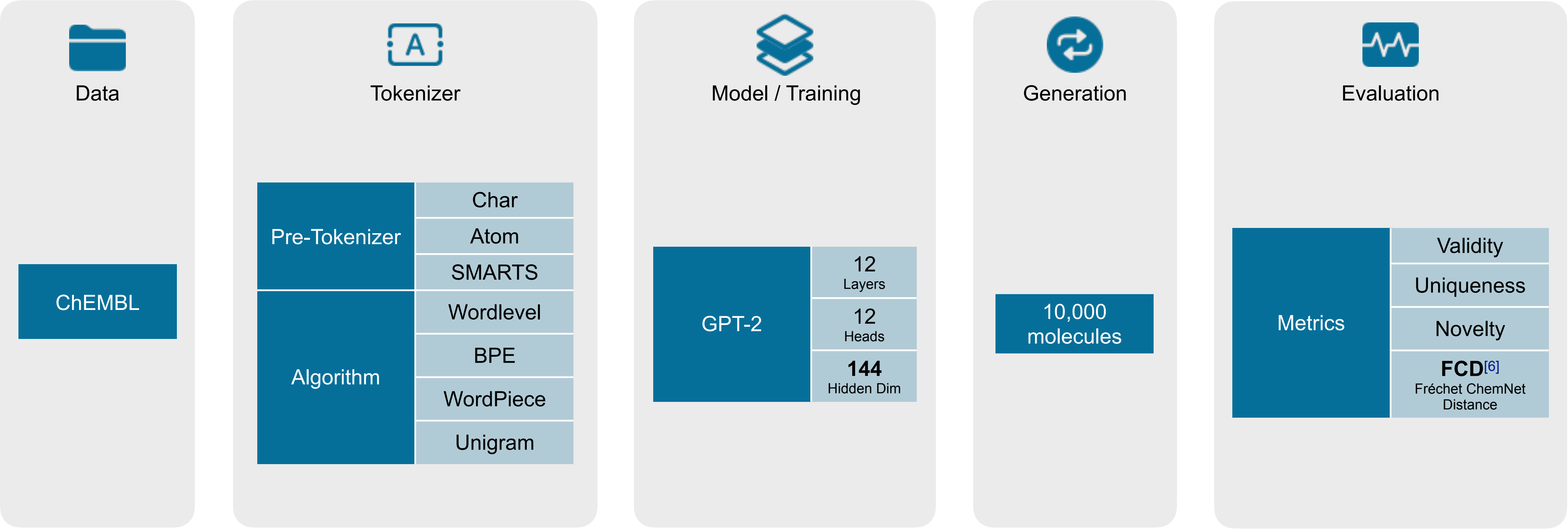
[2] N. Brown et al., "GuacaMol: Benchmarking Models for de Novo Molecular Design," *J. Chem. Inf. Model.*, vol. 59, no. 3, pp. 1096–1108, Mar. 2019, doi: 10.1021/acs.jcim.8b00839.

Data — Intuition

Target	Source / Preprocessing	Format	Count	ø Length (train set)	Example
Molecules	ChEMBL[3] / GuacaMol	SMILES	1.6 M	48	`O=C(O)C1CCC(OCC2CC(F)CN2C(=O)Cc2ccc(NC(=O)N3CCc4ccccc43)c(Cl)c2)CC1`
Reaction Templates	USPTO-50K[4] / MHNReact[5]	SMARTS	12 K	161	`[#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] D. Mendez et al., “ChEMBL: towards direct deposition of bioassay data,” Nucleic Acids Res, vol. 47, no. D1, pp. D930–D940, Jan. 2019, doi: 10.1093/nar/gky1075
[4] D. M. Lowe, “Extraction of chemical structures and reactions from the literature,” Thesis, University of Cambridge, 2012. doi: 10.17863/CAM.16293
[5] P. Seidl et al., “Modern Hopfield Networks for Few- and Zero-Shot Reaction Template Prediction,” arXiv:2104.03279 [cs, q-bio, stat], Jun. 2021, Accessed: Nov. 02, 2021. [Online]. Available: http://arxiv.org/abs/2104.03279

Pipeline 1/3 — Molecules From Scratch



Data

- ChEMBL

Tokenizer

- Pre-Tokenizer: Char, Atom, SMARTS
- Algorithm: Wordlevel, BPE, WordPiece, Unigram

Model / Training

- GPT-2: 12 Layers, 12 Heads, **144** Hidden Dim

Generation

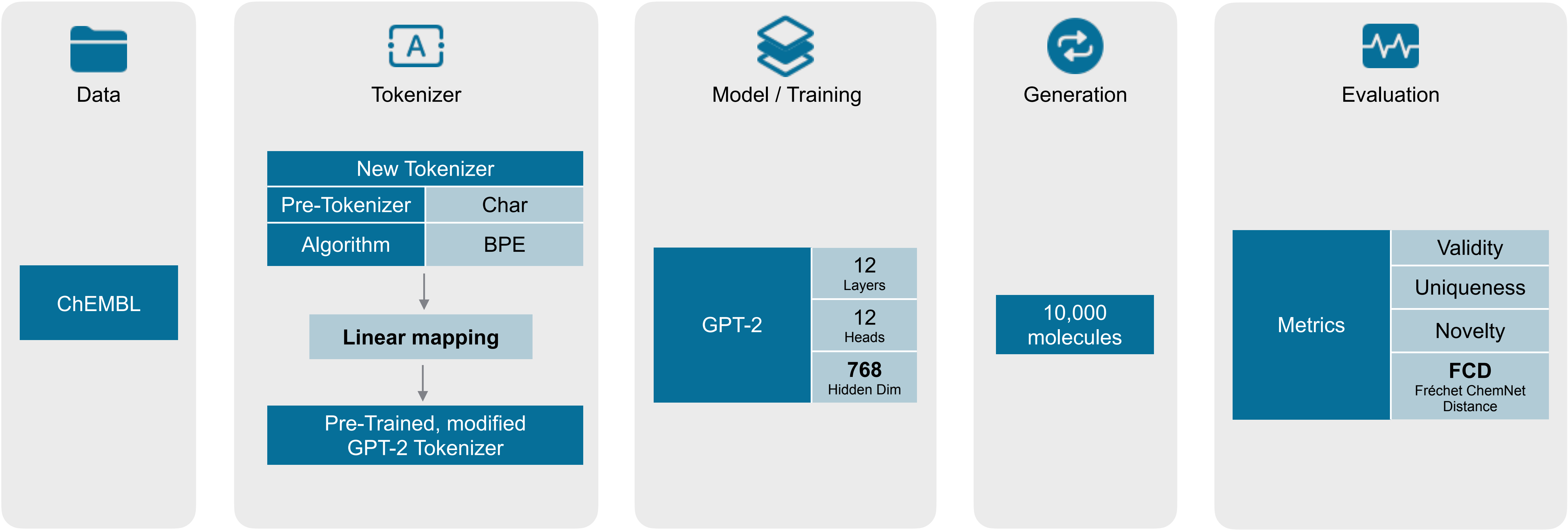
- 10,000 molecules

Evaluation

- Metrics: Validity, Uniqueness, Novelty, **FCD**[6] Fréchet ChemNet Distance

[6] K. Preuer et al., “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

Pipeline 2/3 — Molecules from Pre-Trained Model



Data

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

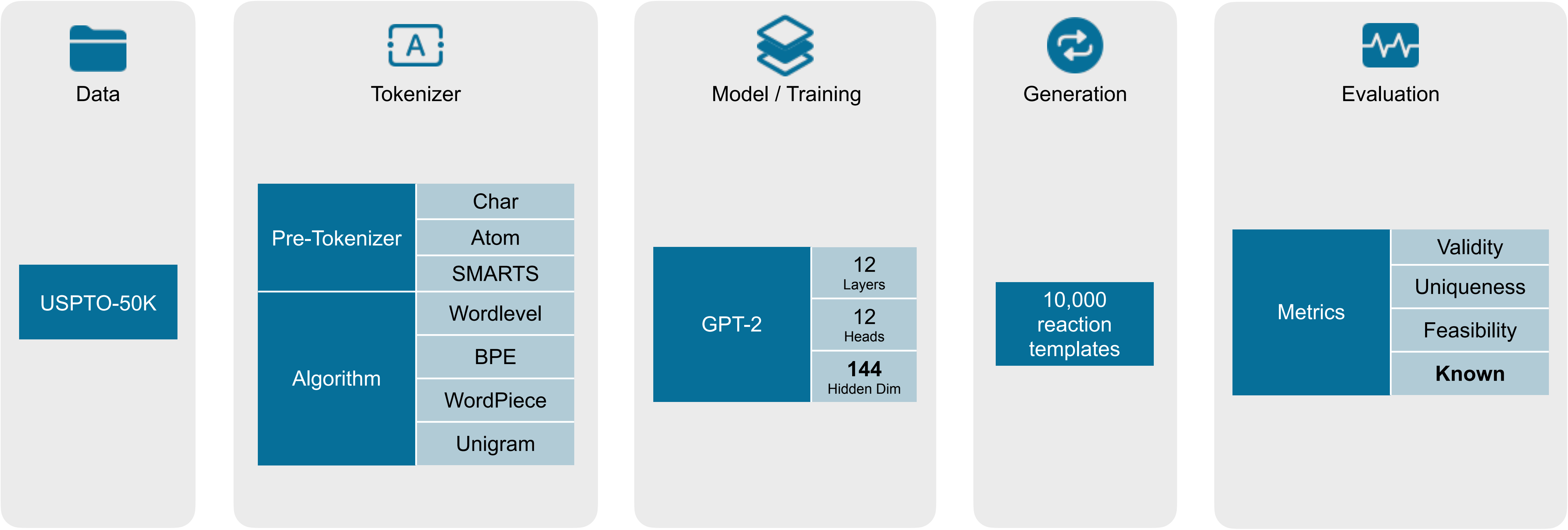
Generation

10,000 molecules

Evaluation

Metrics	
	Validity
	Uniqueness
	Novelty
	FCD Fréchet ChemNet Distance

Pipeline 3/3 — Reaction Templates From Scratch



Data

USPTO-50K

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 reaction templates

Evaluation

Metrics	Validity
	Uniqueness
	Feasibility
	Known

Selected Results
Generation of 10,000 Molecules

Model	**Tokenizer** Pre-Tokenizer | Algorithm | Vocab Size	Metrics			
		Validity ↗	Uniqueness ↗	Novelty ↗	FCD ↘
GuacaMol		0.959	**1.000**	**0.994**	0.455
MolReactGen *from scratch*	Char | Wordpiece | 176	**0.976** $^{\pm 0.001}$	0.999 $^{\pm 0.000}$	0.935 $^{\pm 0.002}$	**0.219** $^{\pm 0.005}$
		0.976 $^{\pm 0.001}$	**0.999** $^{\pm 0.000}$	**0.935** $^{\pm 0.002}$	0.219 $^{\pm 0.005}$
MolReactGen *fine-tuned*	Char | BPE | 50,527	**0.990** $^{\pm 0.001}$	0.999 $^{\pm 0.000}$	0.797 $^{\pm 0.004}$	**0.209** $^{\pm 0.006}$

Molecule generation results
Numbers represent the mean and standard deviation (superscript) across five training and generation runs
FCD metric not stated in GuacaMol paper, calculated here as $-5 \log \mathrm{FCD}_{\mathrm{GuacaMol}}$

Selected Results

Generation of 10,000 Reaction Templates

<table>
<tr><th rowspan="2">Model</th><th rowspan="2">Tokenizer
Pre-Tokenizer | Algorithm | Vocab Size</th><th colspan="4">Metrics</th></tr>
<tr><th>Validity ↗</th><th>Uniqueness ↗</th><th>Feasibility ↗</th><th>Known ↗</th></tr>
<tr><td>MolReactGen <em>from scratch</em></td><td>SMARTS | Wordlevel | 947</td><td>0.804 $^{\pm 0.022}$</td><td>0.795 $^{\pm 0.008}$</td><td>0.110 $^{\pm 0.004}$</td><td>735.2 $^{\pm 27.0}$</td></tr>
</table>

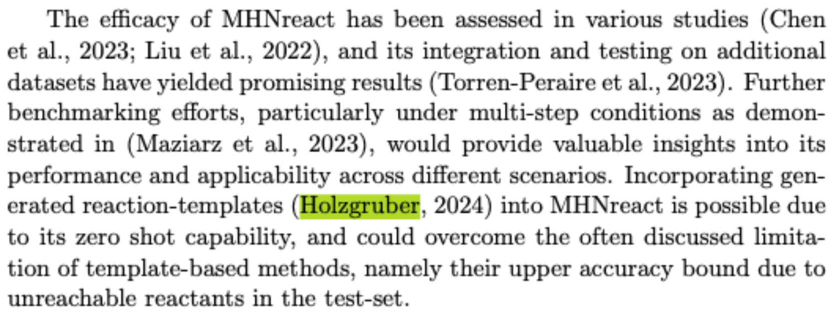
Reaction template generation results
Numbers represent the mean and standard deviation (superscript) across five training and generation runs

Conclusion wrt Research Questions

- Transformer decoder (GPT-2) architecture exhibits a **better FCD** than the GuacaMol baseline
- **Tokenizers** exert “some” influence on FCD (max `0.022`, but statistically significant)
- **Fine-tuning** a pre-trained language model **works**, but at a (computational) cost
- Approach (with different optimal tokenizer) also **works with reaction templates**

"Giving back to the community"

- `PySmilesUtils`: 2 accepted PRs (#1, #2)
- `rdkit`: bug report (fixed)
- `Hugging Face`
  - Implementation of P. Schwaller's feature request
  - 4 issue contributions (#1, #2, #3, #4)
  - Member of HF "Helping Hands"
  - **> 1K pre-trained model downloads** (#1, #2)
- Initial port of GuacaMol evaluation code to current packages
- `FCD`: 2 accepted PRs (#1, #2)
- First citation of master's thesis 😉



The efficacy of MHNreact has been assessed in various studies (Chen et al., 2023; Liu et al., 2022), and its integration and testing on additional datasets have yielded promising results (Torren-Peraire et al., 2023). Further benchmarking efforts, particularly under multi-step conditions as demonstrated in (Maziarz et al., 2023), would provide valuable insights into its performance and applicability across different scenarios. Incorporating generated reaction-templates (Holzgruber, 2024) into MHNreact is possible due to its zero shot capability, and could overcome the often discussed limitation of template-based methods, namely their upper accuracy bound due to unreachable reactants in the test-set.

Details — Links into the Master's Thesis

- Molecules — From Scratch
  - Data and Tokenization
  - Hyper Params
  - Metrics
  - Tokenizer Selection Process
  - Tokenizer Results
  - Results
- Molecules — Pre-trained
  - Tokenizer Mapping
  - Hyper Params
  - Results

- Reaction Templates
  - Data and Tokenization
  - Hyper Params
  - Metrics
  - Reaction Template Feasibility
  - Tokenizer Results
  - Results
- RegEx Patterns

JOHANNES KEPLER
UNIVERSITY LINZ