Accelerating the inference of string generation-based chemical reaction models for industrial applications

Mikhail Andronov[1,2], Natalia Andronova[5], Michael Wand[1,3], Jürgen Schmidhuber[1,4], Djork-Arné Clevert[2]

[1]IDSIA, USI, SUPSI, 6900 Lugano, Switzerland.
[2]Machine Learning Research, Pfizer Research and Development, Friedrichstr. 110, Berlin, Germany
[3]Institute for Digital Technologies for Personalized Healthcare, SUPSI, 6900 Lugano, Switzerland
[4]AI Initiative, KAUST, 23955 Thuwal, Saudi Arabia.
[5]Independent researcher

Introduction

- Computer-aided synthesis planning (CASP) is one of the core technologies enabling computer-aided drug discovery.
- Machine learning-based CASP systems consist of a single-step retrosynthesis model and a planning algorithm [1].
- State-of-the-art single-step retrosynthesis models like Chemformer are too slow to be successfully incorporated into CASP systems in production [2].
- Transformers for SMILES-to-SMILES transformations need accelerated inference.
- Besides retrosynthesis, transformer-based AI-assistants for reaction prediction like IBM RXN could also benefit from inference acceleration.

Research question

- How to accelerate the inference of the SMILES-to-SMILES encoder-decoder transformer for reaction modeling without compromising on accuracy?

Results

- We reimplement the Molecular Transformer [3] in Pytorch Lightning and make our implementation available on Github.
- We accelerate greedy decoding from Molecular Transformer by ~3 times without the loss in accuracy.
- We accelerate beam search decoding from Molecular Transformer by ~4 times, albeit with a decrease in accuracy.

Method

Chemical insight

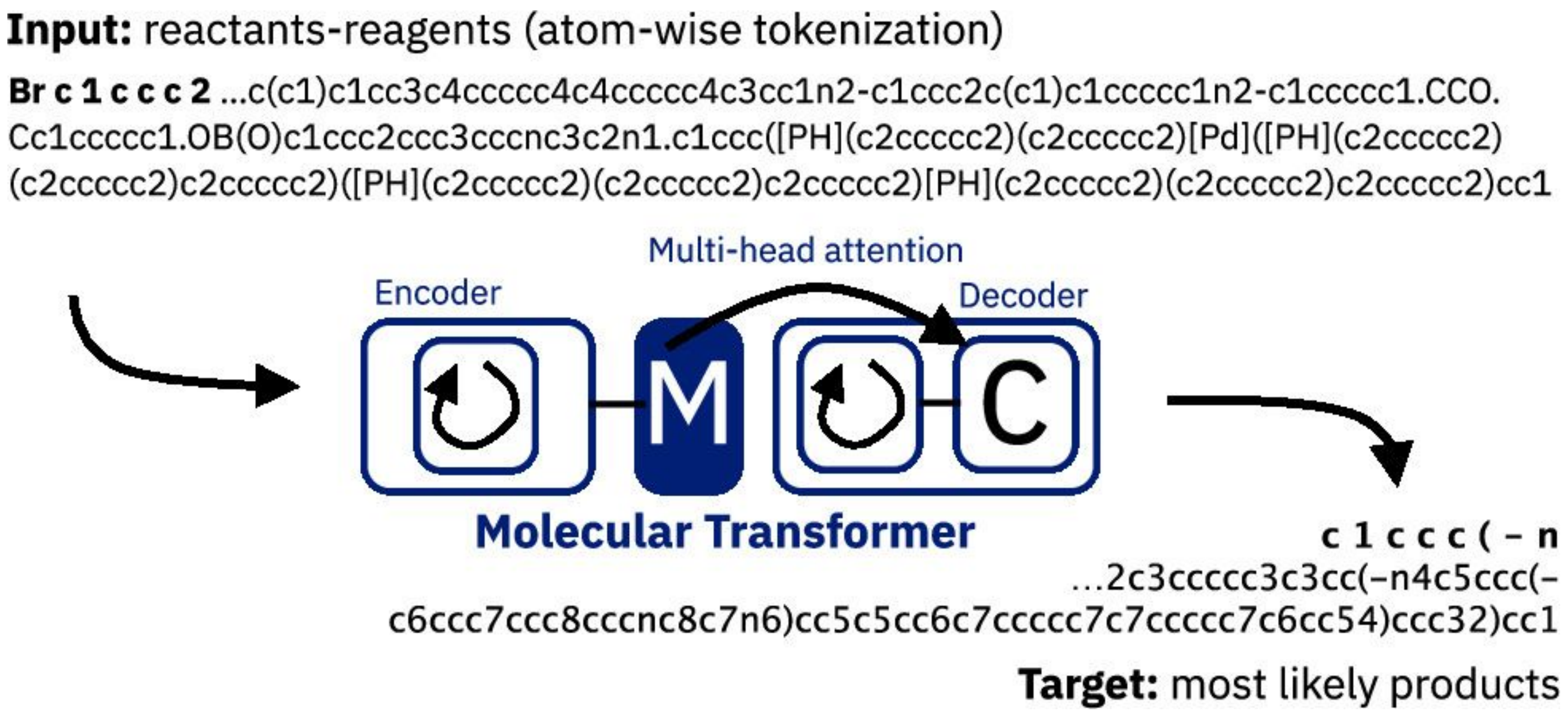
In both reaction product prediction and single-step retrosynthesis (Fig. 1), large fragments of the source molecule remain unchanged. Therefore, in both tasks the target sequence tands to have a lot of common substrings with the source sequence (Fig. 2).

Speculative decoding

Recently, a method of LLM inference acceleration called “speculative decoding” was proposed [4, 5]. It is bases on the draft-and-verify idea:

1. Try to “guess” the continuation of the generated sequence by attaching some draft sequence to the tokens already generated.
2. Accept or discard tokens from the draft sequence in one forward pass.

In our method, substrings of the source sequence serve as drafts which we verify and parallel, selecting the best one.

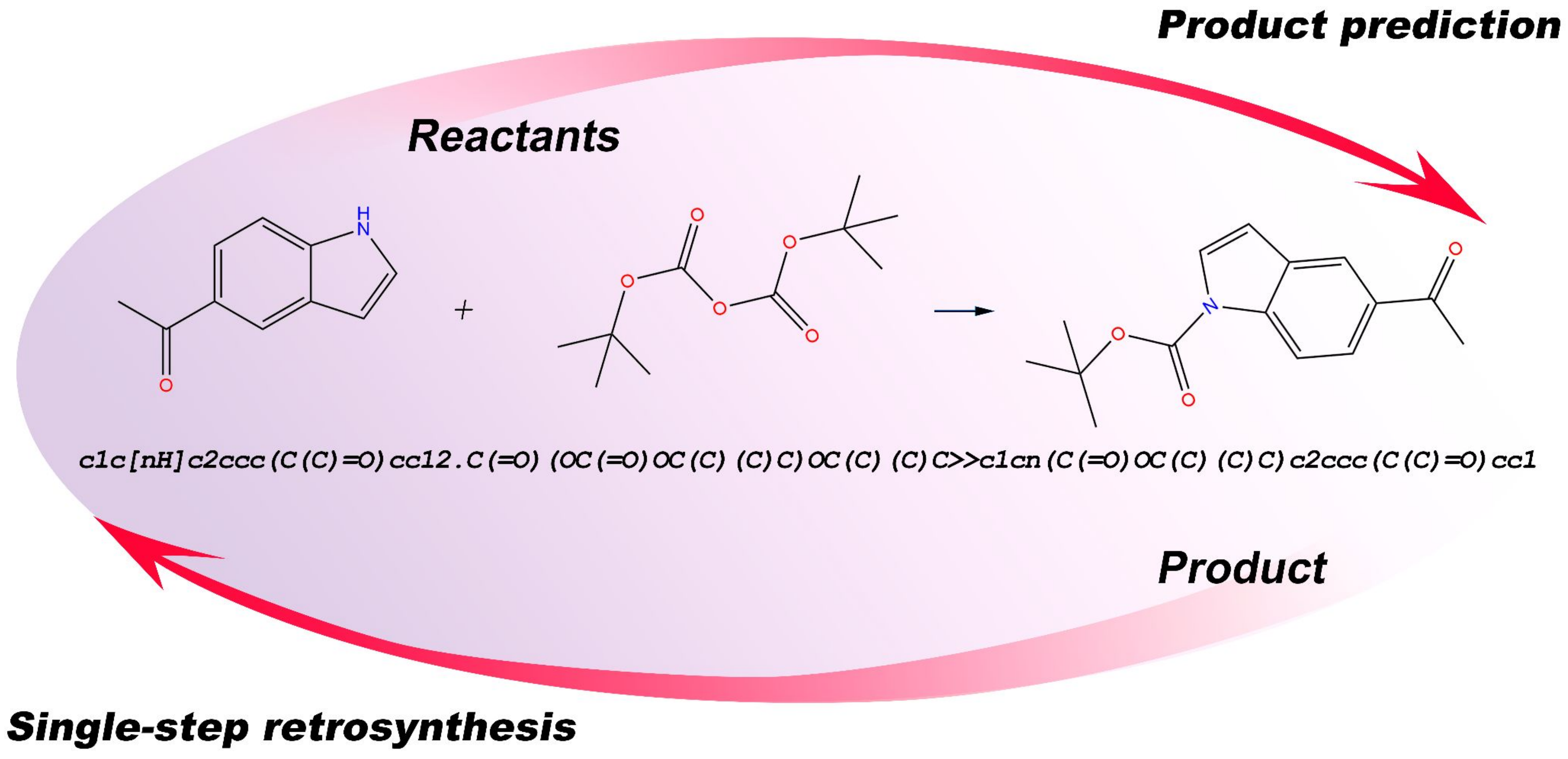

Input: reactants-reagents (atom-wise tokenization)

Br c 1 c c c 2 ...c(c1)c1cc3c4ccccc4c4ccccc4c3cc1n2-c1ccc2c(c1)c1ccccc1n2-c1ccccc1.CCO.Cc1ccccc1.OB(O)c1ccc2ccc3cccnc3c2n1.c1ccc([PH](c2ccccc2)(c2ccccc2)[Pd]([PH](c2ccccc2)(c2ccccc2)c2ccccc2)([PH](c2ccccc2)(c2ccccc2)c2ccccc2)[PH](c2ccccc2)(c2ccccc2)c2ccccc2)cc1

c 1 c c c (- n ...2c3ccccc3c3cc(-n4c5ccc(-c6ccc7ccc8cccnc8c7n6)cc5c5cc6c7ccccc7c7ccccc7c6cc54)ccc32)cc1

Target: most likely products

Fig. 1. We reimplement the Molecular Transformer (image from [3]) and accelerate its inference with speculative decoding.



c1c[nH]c2ccc(C(C)=O)cc12.C(=O)(OC(=O)OC(C)(C)C)OC(C)(C)C>>c1cn(C(=O)OC(C)(C)C)c2ccc(C(C)=O)cc1

Fig. 2. Both reaction prediction and single-step retrosynthesis can be formulated as SMILES-to-SMILES translation and approached with an model like a encoder-decoder transformer.

Example

The product SMILES for the reaction in Figure 3 can be constructed out of subsequences of the source SMILES. With the draft length of four, the product takes 9 runs of the model instead of 39.

Decoding	Time, minutes
Greedy (BS 1)	61.8 ± 5.88
Greedy speculative (BS 1, DL 4)	26.04 ± 2.07
Greedy speculative (BS 1, DL 10)	17.06 ± 0.25
Greedy (BS 32)	4.13 ± 0.06

Table 1. Wall time in product prediction on USPTO MIT. BS is batch size, DL is draft length

Reaction SMILES:

`c1c[nH]c2ccc(C(C)=O)cc12.C(=O)(OC(=O)OC(C)(C)C)OC(C)(C)C>>c1cn(C(=O)OC(C)(C)C)c2ccc(C(C)=O)cc12`

Drafts of length 4 - substrings of the reactants’ SMILES:

c1c[nH]	1c[nH]c	c[nH]c2	[nH]c2c	c2cc	2ccc	ccc(	cc(C	c(C(	(C(C	C(C)
(C)=	C)=O	)=O)	=O)c	O)cc	)cc1	cc12	c12.	12.C	2.C(	.C(=
C(=O	(=O)	=O)(	O)(O	)(OC	(OC(	OC(=	C(=O	(=O)	=O)O	O)OC
)OC(	OC(C	C(C)	(C)(	C)(C	)(C)	(C)C	C)C)	)C)O	C)OC	)OC(
OC(C	C(C)	(C)(	C)(C	)(C)	(C)C					

Fig. 3. Example of product prediction acceleration with speculative decoding

Decoding	Time, minutes
Beam search (5 BW, 5 best)	27.98 ± 0.60
Speculative NS (5 best)	7.06 ± 0.03

Table 2. Wall time in single-step retrosynthesis on USPTO 50k. BW is beam width.

Results

- We test our speculative decoding approach in product prediction on USPTO MIT and single-step retrosynthesis on USPTO 50K.
- The method accelerates greedy decoding by more than 3 times without any loss in accuracy.
- We replace beam search with speculative greedy decoding and accelerate inference by almost 4 times but with some loss in accuracy.
- Accelerating beam search with no loss in accuracy is a part of our ongoing work.

Accuracy	Beam search	Speculative nucleus sampling
TOP-1, %	52.1	51.0
TOP-3, %	75.1	69.8
TOP-5, %	82.0	73.3

Table 2. Wall time in single-step retrosynthesis on USPTO 50k. BW is beam width.

https://github.com/Academich/translation-transformer

References

(1) Segler, M. H., Preuss, M., and Waller, M. P. Planning chem- ical syntheses with deep neural networks and symbolic AI. *Nature*, 555(7698):604–610, 2018.
(2) Torren-Peraire, P., Hassen, A. K., Genheden, S., Verhoeven, J., Clevert, D.-A., Preuss, M., and Tetko, I. V. Models matter: the impact of single-step retrosynthesis on syn- thesis planning. *Digital Discovery*, 3(3):558–572, 2024.
(3) Schwaller, P., Laino, T., Gaudin, T., Bolgar, P., Hunter, C. A., Bekas, C., and Lee, A. A. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS central science*, 5(9):1572–1583, 2019.
(4) Xia, H., Ge, T., Wang, P., Chen, S.-Q., Wei, F., and Sui, Z. Speculative decoding: Exploiting speculative execu- tion for accelerating seq2seq generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 3909–3925, 2023.
(5) Leviathan, Y., Kalman, M., and Matias, Y. Fast inference from transformers via speculative decoding. In *Inter- national Conference on Machine Learning*, pp. 19274– 19286. PMLR, 2023.