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

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

Abstract

Template-free SMILES-to-SMILES translation models for reaction prediction and single-step retrosynthesis are of interest for industrial applications in computer-aided synthesis planning systems due to their state-of-the-art accuracy. However, they suffer from slow inference speed. We present a method to accelerate inference in autoregressive SMILES generators through speculative decoding by copying query string subsequences into target strings in the right places. We apply our method to the molecular transformer implemented in Pytorch Lightning and achieve over 3X faster inference in reaction prediction and single-step retrosynthesis.

1. Introduction

Automated planning of organic chemical synthesis, first formalized around fifty years ago (Pensak & Corey, 1977), is one of the core technologies enabling computer-aided drug discovery. While first computer-aided synthesis planning (CASP) systems relied on manually encoded rules (Johnson et al., 1989; Gasteiger et al., 2000), researchers now primarily focus on CASP methods powered by artificial intelligence techniques. The design principles of the latter were outlined in the seminal work by Segler et al. (Segler et al., 2018): a machine learning-based single-step retrosynthesis model combined with a planning algorithm. The former proposes several candidate retrosynthetic chemical transformations for a given molecule, and the latter, e.g., Monte-Carlo Tree Search, uses those candidates to construct a synthesis tree. Single-step retrosynthesis models are now commonly developed in two paradigms: template-based models and template-free models. Besides retrosynthesis, one can also build a model to predict the products of chemical reactions (Figure 1).

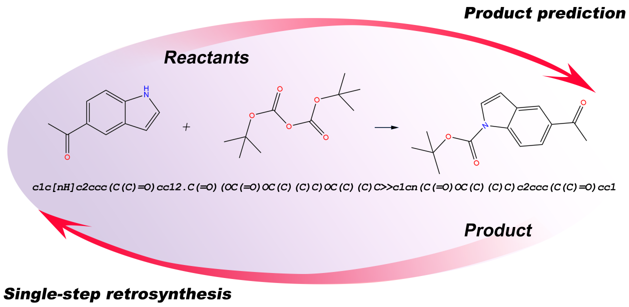
Product prediction
Reactants
N
H
O
O
O
O
O
+
O
O
O
N
c1c[nH]c2ccc(C(C)=O)cc12.C(=O)(OC(=O)OC(C)(C)C)OC(C)(C)C>>c1cn(C(=O)OC(C)(C)C)c2ccc(C(C)=O)cc1
Product
Single-step retrosynthesis

Figure 1. Both reaction product prediction and single-step retrosynthesis can be formulated as SMILES-to-SMILES translation and approached with a model like an encoder-decoder transformer.

The principle of template-based models is to use a set of SMIRKS templates extracted from reaction data and a machine learning model for classification or ranking to select a matching template for a query SMILES that will, upon application, transform the query into the product SMILES (for product prediction), or the SMILES of possible reactants (for single-step retrosynthesis). In contrast, in template-free models, the query transforms into the result without resorting to SMIRKS templates, e.g., with a sequence of predicted graph edits (Sacha et al., 2021; Bradshaw et al., 2018) or through “translation” of the query SMILES into the desired SMILES with a conditioned text generation model (Schwaller et al., 2019; Tetko et al., 2020; Irwin et al., 2022). While CASP systems leveraging template-based single-step models proved to be effective (Genheden et al., 2020), there is an interest in building CASP with template-free models instead, as they demonstrate state-of-the-art accuracy in both single-step retrosynthesis and reaction product prediction. Most accurate template-free models are currently conditional autoregressive SMILES generators based on the transformer architecture (Vaswani et al., 2017), which also serves as the backbone for Large Language Models (LLM) (Brown et al., 2020; Zhao et al., 2023). Unfortunately, the

[1] IDSIA, USI, SUPSI, 6900 Lugano, Switzerland [2] Machine Learning Research, Pfizer Research and Development, 10117 Berlin, Germany [3] Independent researcher [4] Institute for Digital Technologies for Personalized Healthcare, SUPSI, 6900 Lugano, Switzerland [5] AI Initiative, KAUST, 23955 Thuwal, Saudi Arabia. Correspondence to: Mikhail Andronov <mikhail.andronov@idsia.ch>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

high accuracy of autoregressive models like Chemformer (Irwin et al., 2022) comes at the cost of a slow inference speed (Torren-Peraire et al., 2024), which hinders their successful adoption as part of industrial CASP systems. In our work, we propose a method to accelerate inference from SMILES-to-SMILES translation models based on speculative decoding (Leviathan et al., 2023; Xia et al., 2023), a general technique for LLM inference acceleration, combined with insights from the chemical essence of the problem. We reimplement the Molecular Transformer (Schwaller et al., 2019) in Pytorch Lightning and use our method to demonstrate its inference acceleration in single-step retrosynthesis and product prediction by 300% without changing the model architecture, training procedure, or generated SMILES.

2. Methods

2.1. Algorithm

Autoregressive models, such as Transformer variants (Vaswani et al., 2017; Brown et al., 2020; Schmidhuber, 1992), generate sequences token by token, and every prediction of the next token requires a forward pass of the model. Such a process may be computationally expensive, especially for models with billions of parameters. Therefore, an intriguing question arises whether could one generate several tokens in one forward pass of the model, thus completing the output faster. Speculative decoding (Xia et al., 2023; Leviathan et al., 2023) answers positively. Recently proposed as a method of inference acceleration for Large Language Models, it is based on the draft-and-verify idea. At every generation step, one can append some draft sequence to the sequence generated by the model so far and see if the model ”accepts” the draft tokens.

If the draft sequence has length N, in the best case, the model adds $N + 1$ token to the generated sequence in one forward pass, and in the worst case, it adds one token as in standard autoregressive generation. The acceptance rate for one generated sequence is the number of accepted draft tokens divided by the total number of tokens in the generated sequence. One can also test multiple draft sequences in one forward pass taking advantage of parallelization, and choose the best one. Speculative decoding does not affect the content of the predicted sequence compared to the one-by-one decoding in any way.

One can freely choose a way of generating draft sequences. For LLMs, one would usually use another smaller language model that performs its forward pass faster than the main LLM (Leviathan et al., 2023) or additional generation heads on top of the LLM's backbone (Cai et al., 2024). However, one can also construct draft sequences without calling any learned functions. For example, generate random draft sequences, even though their acceptance rate will be minimal, or assemble draft sequences out of tokens in the query sequence — a prompt for decoder-only language models or a source sequence for translation models. The latter option is perfect for retrosynthesis or reaction prediction as SMILES-to-SMILES translation. In a chemical reaction, large fragments of reactants typically remain unchanged, which means that the SMILES of products and reactants have many common substrings. It is especially true if reactant and product SMILES are aligned to minimize the edit distance between them (Zhong et al., 2022). Therefore, we can extract multiple substrings of a chosen length N from the query SMILES and use them as draft sequences with a high acceptance rate. Figure 2 demonstrates this method in product prediction. Before generating the target string, we assemble a list of token subsequences from the source sequence (reactant tokens) with a sliding window of a desired length (4 in this case) and stride 1. Then, at every generation step, we can try all draft sequences in one forward pass of the model to see if the model can copy up to 4 tokens from one of them. The draft token acceptance rate in this example reaches 78%.

Speculative decoding does not require any changes to the model architecture or training of additional models. The cost of generating draft sequences in this way is negligible compared to that of the forward pass of the reaction model, and the generation acceleration with this method comes practically ”for free”.

2.2. Model

We demonstrate the application of our method to the Molecular Transformer (Schwaller et al., 2019). It is an encoder-decoder transformer model suitable for SMILES-to-SMILES translation. We conduct our experiments on one H100 GPU with 80 GB memory.

The original Molecular Transformer (Schwaller et al., 2019) adopts OpenNMT (Klein et al., 2018), a general framework for neural machine translation, for SMILES-to-SMILES translation. Since the code in this framework is complex and intractable to customize, we decided to re-implement the model in PyTorch Lightning to keep only the necessary code and have more design freedom in the model's inference procedure implementation.

2.3. Data

We used the open reaction data from US patents (Lowe, 2012) for training all models. We trained the model for reaction product prediction as in the original paper (Schwaller et al., 2019) on the USPTO MIT dataset, a standard benchmark for product prediction, without reactant-reagent separation. We trained the model for single-step retrosynthesis on USPTO 50K, a standard dataset for the task. In this dataset, we augmented every reaction in the training set 20

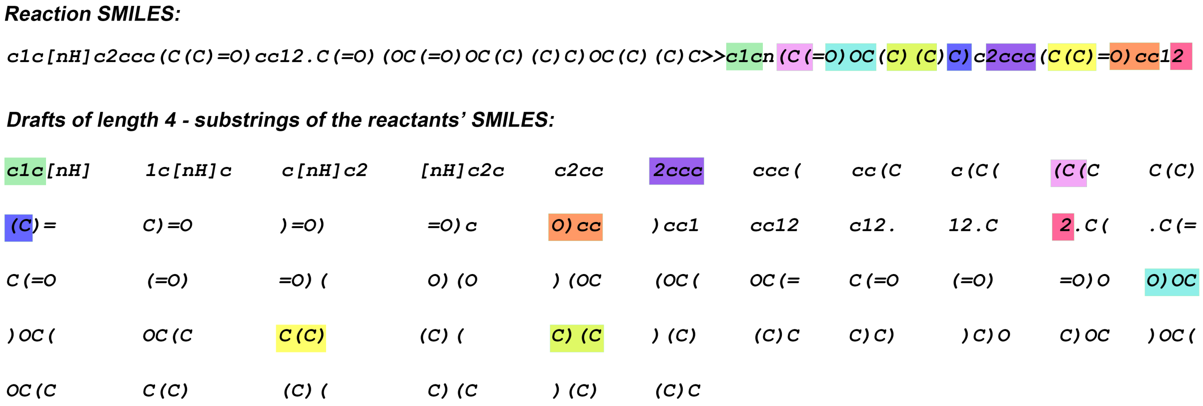
Reaction SMILES:
c1c[nH]c2ccc(C(C)=O)cc12.C(=O)(OC(=O)OC(C)(C)C)OC(C)(C)C>>c1cn(C(=O)OC(C)(C)C)c2ccc(C(C)=O)cc12
Drafts of length 4 - substrings of the reactants' SMILES:
c1c[nH] 1c[nH]c c[nH]c2 [nH]c2c c2cc 2ccc ccc(cc(C c(C((C(C C(C)
(C)= C)=O)=O) =O)c O)cc)cc1 cc12 c12. 12.C 2.C(.C(=
C(=O (=O) =O)(O)(O)(OC (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

Figure 2. Speculative decoding accelerates product prediction with the molecular transformer or a similar autoregressive SMILES generator. Before generating an output sequence, we prepare a list of subsequences of a desired length, e.g., four, of the tokenized query SMILES of reactants. Then, at every generation step, the model can copy up to four tokens from one of the draft sequences to the output, thus generating from one to five tokens in one forward pass.

times using SMILES augmentation (Tetko et al., 2020) with root-aligned SMILES (Zhong et al., 2022). We followed the standard atomwise tokenization procedure (Schwaller et al., 2019) to tokenize SMILES.

3. Results and Discussion

By replacing the standard decoding procedures for the molecular transformer with our method, we achieve a significant speed-up in both product prediction and single-step retrosynthesis. Our implementation of the Molecular Transformer (MT) successfully reproduces the accuracy scores of the original MT (Schwaller et al., 2019) that relies on OpenNMT. Comparing our MT and the original MT, we observe at most 0.2 percentage points discrepancy of top-1 to top-5 accuracy in product prediction with beam search.

3.1. Product prediction

We tested our MT for product prediction on USPTO MIT mixed, i.e., without an explicit separation between reactants and reagents. The test dataset in this benchmark comprises 40 thousand reactions.

When serving a reaction prediction model as an AI assistant for chemists, one could use greedy decoding with a batch size of 1 for inference. Table 1 summarizes our experiments with greedy decoding from MT on the test set of USPTO MIT. The model's inference with standard greedy decoding with a batch size of 1 finishes in around 62 minutes. In contrast, if we use greedy generation enhanced with our speculative decoding, the inference time reduces to 26 minutes with a draft length of 4 and 17 minutes with a draft length of 10, which corresponds to 137 % and 262 % speedup, respectively. The acceptance rate in our drafting method averaged over all test reactions is 79%. Potentially, it can be even higher if one adds more draft sequences to choose from, for example, subsequences of the source sequence dilated by one token. Of course, greedy decoding with a large batch size is much faster and completes in around 4 minutes with 32 reactions in a batch. However, accelerating inference with a batch size of 1 would be sufficient for an improved user experience with reaction prediction assistants. The model's accuracy is 88.3% with both standard and speculative greedy decoding.

Table 1. Wall time of the model's inference on the USPTO MIT test set in reaction product prediction with standard and speculative greedy decoding. BS stands for batch size, and DL stands for draft length. The time is averaged over five attempts.

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

3.2. Single-step retrosynthesis

We carried out single-step retrosynthesis experiments on USPTO 50K, in which the training dataset was augmented 20 times. The augmentation procedure is to construct alternative root-aligned (Zhong et al., 2022) SMILES for every

Table 2. Wall time of the single-step retrosynthesis model's inference USPTO 50K test set without augmentations (5000 reactions) with beam search and speculative nucleus sampling. BW is beam width. The time is averaged across five runs. The batch size is set to 1.

DECODING	TIME, MINUTES
BEAM SEARCH (5 BW, 5 BEST)	27.98 ± 0.60
SPECULATIVE NS (5 BEST)	7.06 ± 0.03

Table 3. The top-N accuracy of our single-step retrosynthesis model on USPTO 50K with both beam search and speculative nucleus sampling.

ACCURACY	BEAM SEARCH	SPECULATIVE NS
TOP-1, %	52.1	51.0
TOP-3, %	75.1	69.8
TOP-5, %	82.0	73.3

dataset entry. This augmentation minimizes the edit distance between reactants and products, which simplifies training, pushes the model's accuracy higher, and increases the acceptance rate in our speculative decoding method. Speculative decoding in single-step retrosynthesis accelerates greedy decoding as much as in reaction product prediction. However, this has limited utility. In synthesis planning, one would want a single-step retrosynthesis model to suggest multiple different reactant sets for every query product so that the planning algorithm can choose from them. Usually, one would employ beam search to generate several outputs from the transformer for single-step retrosynthesis. Unfortunately, accelerating beam search with speculative decoding proves to be a challenging task. In beam search, the different beams are ordered by probability and may change order after each generation of a single token. Therefore, one cannot skip individual token generations in beam search and hope to introduce no change to the output compared to the standard beam search. Eventually, if we would like to accelerate the generation of multiple reactant sets with our single-step retrosynthesis model, we would have to consider methods other than beam search. A viable alternative here could be nucleus (top-p) sampling. In nucleus sampling, the model can generate different output sequences for the same query when run several times, and we can accelerate nucleus sampling with speculative decoding just as we can accelerate greedy decoding. In essence, our speculative nucleus sampling algorithm is as follows. First, we generate several diverse complete output sequences for a query using nucleus sampling. Generating a sequence is fast using speculative decoding. Then, we order the obtained sequences by their probabilities. Finally, we keep the N best sequences. We choose nucleus sampling over top-k sampling to ensure the high validity of generated SMILES. We use a high temperature after masking logits not among the positions in the top-p to increase the diversity of generated candidates.

The top-N accuracy of our speculative nucleus sampling falls expectedly behind the accuracy of beam search (Table 3). While top-1 accuracy remains practically unchanged, top-3 accuracy declines by 5.3 percentage points, and top-5 accuracy declines by 8.6 percentage points. The reason for that is an insufficient diversity of candidates generated with nucleus sampling - it may fail to produce N unique predictions and yield identical ones instead. However, this issue does not lie with speculative decoding and can be resolved, which is a part of our ongoing work. Our speculative nucleus sampling works almost four times faster than beam search (Table 2). The wall time our retrosynthesis model takes to process the USPTO 50K test set and generate five predictions for every entry with the batch size of 1 is around 28 minutes for beam search and around 7 minutes for speculative nucleus sampling. Such a speed-up could make the transformer a more attractive single-step model for multi-step synthesis planning.

3.3. Limitations

When generating predictions for a batch of query sequences, the number of accepted speculative tokens is generally different for every sequence. Therefore, formulating speculative decoding batch sizes larger than 1 is a challenging problem, although researchers are seeking the key to solving it (Qian et al., 2024). In any case, the inference speed for a batch would be bottlenecked by the ”least lucky” sequence in terms of the acceptance rate of speculative tokens, and the benefits from speculative decoding would vanish with larger batch sizes. Although this is a serious limitation, it is not critical for industrial application of reaction models like the Molecular Transformer. Chemists would usually enter one query at a time in a user interface for a reaction model like IBM RXN. Similarly, a CASP system calls an underlying single-step model at every step with batch size 1 when building a synthesis tree for a molecule. With that, the latency of the single-step model massively contributes to the overall time of the synthesis tree generation (Torren-Peraire et al., 2024). Therefore, we believe that even the acceleration of the transformer inference with batch size 1 is valuable for industrial applications.

4. Conclusion

We combine speculative decoding and chemical insights to accelerate inference in the molecular transformer, a SMILES-to-SMILES translation model. Our method makes processing the test set more than three times faster in both single-step retrosynthesis on USPTO 50K and reaction product prediction on USPTO MIT compared to the standard decoding procedures. Our method aims at making state-of-the-art template-free SMILES-generation-based models such as the molecular transformer more suitable for industrial applications such as computer-aided synthesis planning systems.

Acknowledgements

This study was partially funded by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie Innovative Training Network European Industrial Doctorate grant agreement No. 956832 "Advanced machine learning for Innovative Drug Discovery, and also by the Horizon Europe funding programme under the Marie Skłodowska-Curie Actions Doctoral Networks grant agreement "Explainable AI for Molecules - AiChemist" No. 101120466.

Conflict of interests

The authors have no conflicts of interests.

References

Bradshaw, J., Kusner, M. J., Paige, B., Segler, M. H., and Hernández-Lobato, J. M. A generative model for electron paths. *arXiv preprint arXiv:1805.10970*, 2018.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33: 1877–1901, 2020.

Cai, T., Li, Y., Geng, Z., Peng, H., Lee, J. D., Chen, D., and Dao, T. Medusa: Simple LLM inference acceleration framework with multiple decoding heads. *arXiv preprint arXiv:2401.10774*, 2024.

Gasteiger, J., Pförtner, M., Sitzmann, M., Höllering, R., Sacher, O., Kostka, T., and Karg, N. Computer-assisted synthesis and reaction planning in combinatorial chemistry. *Perspectives in Drug Discovery and Design*, 20: 245–264, 2000.

Genheden, S., Thakkar, A., Chadimová, V., Reymond, J.-L., Engkvist, O., and Bjerrum, E. Aizynthfinder: a fast, robust and flexible open-source software for retrosynthetic planning. *Journal of Cheminformatics*, 12(1):70, 2020.

Irwin, R., Dimitriadis, S., He, J., and Bjerrum, E. J. Chemformer: a pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology*, 3(1):015022, 2022.

Johnson, P. Y., Burnstein, I., Crary, J., Evans, M., and Wang, T. Designing an expert system for organic synthesis: the need for strategic planning. ACS Publications, 1989.

Klein, G., Kim, Y., Deng, Y., Nguyen, V., Senellart, J., and Rush, A. M. OpenNMT: Neural Machine Translation Toolkit, 2018.

Leviathan, Y., Kalman, M., and Matias, Y. Fast inference from transformers via speculative decoding. In *International Conference on Machine Learning*, pp. 19274–19286. PMLR, 2023.

Lowe, D. M. Extraction of chemical structures and reactions from the literature. Ph.D. dissertation, University of Cambridge, Cambridge, UK. https://doi.org/10.17863/CAM.16293, 2012.

Pensak, D. A. and Corey, E. J. LHASA—logic and heuristics applied to synthetic analysis. ACS Publications, 1977.

Qian, H., Gonugondla, S. K., Ha, S., Shang, M., Gouda, S. K., Nallapati, R., Sengupta, S., Ma, X., and Deoras, A. BASS: Batched Attention-optimized Speculative Sampling. *arXiv preprint arXiv:2404.15778*, 2024.

Sacha, M., Błaz, M., Byrski, P., Dabrowski-Tumanski, P., Chrominski, M., Loska, R., Włodarczyk-Pruszynski, P., and Jastrzebski, S. Molecule edit graph attention network: modeling chemical reactions as sequences of graph edits. *Journal of Chemical Information and Modeling*, 61(7): 3273–3284, 2021.

Schmidhuber, J. Learning to control fast-weight memories: An alternative to recurrent nets. *Neural Computation*, 4 (1):131–139, 1992.

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.

Segler, M. H., Preuss, M., and Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature*, 555(7698):604–610, 2018.

Tetko, I. V., Karpov, P., Van Deursen, R., and Godin, G. State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis. *Nature Communications*, 11(1):5575, 2020.

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 synthesis planning. *Digital Discovery*, 3(3):558–572, 2024.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

Xia, H., Ge, T., Wang, P., Chen, S.-Q., Wei, F., and Sui, Z. Speculative decoding: Exploiting speculative execution for accelerating seq2seq generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 3909–3925, 2023.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

Zhong, Z., Song, J., Feng, Z., Liu, T., Jia, L., Yao, S., Wu, M., Hou, T., and Song, M. Root-aligned SMILES: a tight representation for chemical reaction prediction. *Chemical Science*, 13(31):9023–9034, 2022.

A. Training details.

For product prediction, we train this model with the same hyperparameters as in Schwaller et al. with four encoder and decoder layers, eight heads, embedding dimensionality of 256, and feedforward dimensionality of 2048, which results in 11,4 million parameters. For single-step retrosynthesis, we set the hyperparameters as in Zhong et al. (six encoder and decoder layers, eight heads, embedding dimensionality of 256, and feedforward dimensionality of 2048), which results in 17,4 million parameters. The dictionary is the same for the encoder and the decoder in both models. We use the Adam optimizer for both models.