Template-free organic retrosynthesis with syntax-directed molecular transformer networks

Stanford CS224N Custom Project
Mentor: Lauren Zhu

David Toomer
Department of Computer Science
Stanford University
djtoomer@stanford.edu

Abstract

Retrosynthesis is one of the fundamental problems in organic chemistry. The advent of generative deep learning models has rapidly improved template-free retrosynthesis planning, where a retrosynthetic step can be modeled as a sequence-to-sequence task between the string representations (SMILES) of the molecules involved in the reaction. However, many existing methods either prune reaction datasets of important stereochemical information, or they output SMILES strings that are often not syntactically correct. We address both of these issues by developing a syntax-directed molecular transformer (SDMT), trained on template- and rule-free reaction data without removal of stereochemical designation. SDMT adds a lightweight modification to the traditional transformer architecture by using the syntactic dependency tree of the input SMILES string to restrict self-attention. SDMT performs competitively in accuracy with the current state-of-the-art text-based and graph-based retrosynthesis models, while outperforming them in invalid SMILES rate. We show that SDMT more consistently outputs syntactically and semantically valid SMILES strings across all top predicted results, and it can be used as an effective way to directly integrate the syntactic structure of SMILES strings into transformer models for reaction prediction.

1 Background

In organic chemistry, retrosynthesis refers to the process of identifying precursors that can be used to synthesize a target product. Since its formulation in 1917, organic retrosynthesis has been a fundamental technique in organic chemistry, and it presents one of the key difficulties associated with drug discovery [1, 2]. A putative drug is only as useful as it is synthesizable, and creating synthesis routes computationally has proven difficult, as the space of possible chemical decompositions is large.

Computer-aided synthesis planning (CASP) has assisted chemists in determining the best reaction pathways to form a target molecule [3]. The most common methods typically use reaction rules and/or reaction templates, which categorize known reactions and predict which reaction type is most probable to form the target molecule [4]. These templates are either meticulously hand-coded or computationally extracted from massive reaction databases, both of which present a large overhead to CASP.

Likewise, template- and rule-free approaches to retrosynthesis have become more prevalent. One such method is modeling chemical reactions as a sequence-to-sequence problem, translating the string representations of one set of molecules into another set of molecules [5]. Molecules can be represented in text using the simplified molecular-input line-entry system (SMILES), where the input molecules are fed into a model, and each token is sequentially predicted by the model [6]. This formulation of the problem works for both synthesis and retrosynthesis prediction.

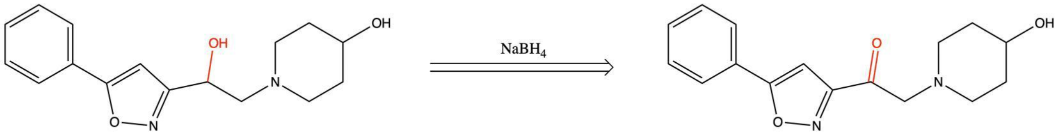

Figure 1: An example of a retrosynthetic step contained in the USPTO-STEREO database. The double-lined arrow signifies that the reaction is happening in reverse.

While this method has proven somewhat effective, it has many common pitfalls. Most notably, in many cases, the generated SMILES strings are syntactically, semantically, or otherwise chemically invalid. In addition, compounds that are chemically similar may have significantly different SMILES representations, making it difficult for traditional models to learn. As such, creating a template-free approach to the synthesis problem that properly ascertains relationships between atoms in SMILES strings is of utmost importance for CASP. Transformers and self-attention have been particularly enticing as atoms that are far apart in SMILES space but proximal in chemical space can be properly associated.

2 Related work

2.1 Synthesis

Transformers have shown promising results for forward synthesis. Schwaller *et al.* develop a transformer model that outperforms resource-intensive quantum chemical calculations in predicting the products of difficult (chemoselective, regioselective, and/or stereoselective) reactions, establishing the tractability of using transformers to model reactions. [7].

2.2 Retrosynthesis

Several sequence-to-sequence models have been applied to the retrosynthesis task. Schwaller *et al.* use a bidirectional LSTM with attention that was shown to compete with state-of-the-art graphical models on chemically diverse datasets. Kim *et al.* improve upon this model through a tied two-way transformer, in which retrosynthesis prediction is coupled with forward synthesis prediction to validate the model's output [8]. However, this model was evaluated without stereochemistry. While removing stereochemistry makes retrosynthesis easier, including the stereochemistry is important because stereochemistry can significantly influence reaction routes and the choice of reagents. Furthermore, in the two models above, there is no deliberate effort to incorporate the SMILES syntax into the model, leading to invalid rates that could be improved.

To the best of our knowledge, the only transformers applied to retrosynthesis in literature that account for syntax append a correction layer to the output of the transformer decoder. For example, self-corrected retrosynthesis predictor (SCROP) incorporated a full transformer to correct the syntax of their sequence-to-sequence model's results [9]. Adding an additional model for correction is resource-intensive, and would be better incorporated into the original network architecture.

3 Approach

3.1 Problem formulation

We formalize the problem as a sequence-to-sequence task from a single product to potentially multiple reactants and reagents. An example of inputs and outputs are given in Table 1.

Input	N#Cc1ccc(C(=O)CCC(=O)O)cc1N
Output	N#Cc1ccc(C(=O)CCC(=O)O)cc1[N+](=O)[O-]>[NH4+].[OH-]>

Table 1: Example input and output for the retrosynthesis task.

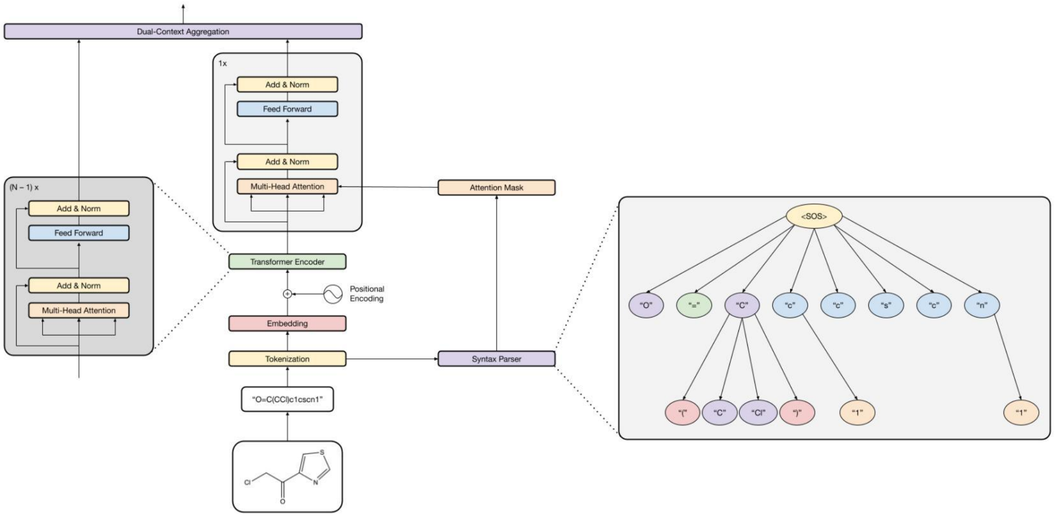

Figure 2: Overview of the syntax-directed molecular transformer encoder.

3.2 Model architecture

We adapt the syntax-guided transformer model, SG-Net, described by Zhang *et al.*, using a custom SMILES parse tree as the dependency tree [10]. Given an input SMILES, we first perform an atom-wise tokenization of the SMILES string into sequence $S = \{s_1, s_2, \ldots, s_n\}$ where n denotes the sequence length. We then use a syntactic parser to generate the dependency tree. This syntactic parser is based on the context-free grammar from the OpenSMILES specification, with the modification that atoms in the same chain are neighbors in the parse tree [11]. The modified context-free grammar is detailed in Figure 3.

After tokenization and parsing, we derive the ancestor node set P_i for each token s_i according to the dependency tree. Finally, we compute mask $\mathcal{M} \in \mathbb{R}^{n \times n}$, where

$$\mathcal{M}_{ij} = \begin{cases} 1 & \text{if } i = j \text{ or } j \in P_i \\ 0 & \text{otherwise} \end{cases}.$$

We then feed S into a vanilla transformer encoder to obtain output representation $H = \{h_1, h_2, \ldots, h_n\}$ [12]. This output is projected into keys, queries, and values for the syntax-guided layer, denoted K_i', Q_i', V_i', respectively, for each attention head i. We apply $\mathcal{M}$ to the multi-head key-query attention in the syntax-guided layer in the following manner:

$$W_i = \text{Softmax}\left(\frac{\mathcal{M} \cdot (Q_i' K_i'^{\top})}{\sqrt{d_k}}\right) V_i'.$$

Finally, W_i is concatenated for all attention heads and fed through the rest of the transformer encoder block to obtain representation $H' = \{h_1', h_2', \ldots, h_n'\}$. We then use dual-context aggregation to combine the two representations H and H' into a final representation $\bar{H} = \{\bar{h}_1, \bar{h}_2, \ldots, \bar{h}_n\}$:

$$\bar{h}_i = \alpha h_i + (1 - \alpha) h_i'$$

where α is a learnable parameter.

For the decoder, we make no modifications to the traditional transformer decoder.

4 Experiments

4.1 Data

The data comes from the open-source chemical reactions from granted US Patents (USPTO) curated by Lowe [13]. We use USPTO-STEREO, a subset of USPTO that removes atom mappings, prunes

```
<START>: <chain>
<chain>: (<bond>? <branched_atom>)*
<branched_atom>: <atom> <ringbond>* <branch>*
<atom>: <bracket_atom> | <symbol>
<bond>: "-" | "=" | "#" | "$" | ":" | "/" | "\"
<dot>: "."
<ringbond>: <bond>? <digit> | <bond>? "%" <digit> <digit>
<branch>: "(" <chain> ")" | "(" <bond> <chain> ")"
<bracket_atom>: "[" <isotope>? <symbol> <chiral>? <hcount>? <charge>? <class>? "]"
<symbol>: <element_symbol> | <aromatic_symbol> | "*"
<isotope>: <number>
<element_symbol>: "H" | "He" | . . . | "No" | "Lr"
<aromatic_symbol>: "b" | "c" | "n" | "o" | "p" | "s" | "se" | "as"
<chiral>: "@" | "@@"
<hcount>: "H" | "H" <digit>
<charge>: "-" | "-" <digit> | "+" | "+" <digit> | "–" | "++"
<class>: ":" <number>
<digit>: "0" | . . . |"9"
<number>: <digit>+
```

Figure 3: The modified SMILES context-free grammar for dependency parsing.

duplicate reactions, and only considers reactions leading to the formation of a single product [14]. Unlike the popular subset of the data USPTO-50K, USPTO-STEREO maintains the stereochemical configuration of the atoms. The 1,002,970 reactions are split into 902,581 / 50,131 / 50,258 reactions for train / validation / test, respectively. Due to space constraints, during preprocessing, we further remove reactions whose tokenizations exceed 200 tokens, which removes 60 reactions for a total of 1,002,610 reactions split into 902,261 / 50,113 / 50,236 reactions for train / validation / test, respectively.

The data consists of reaction SMILES, which have the form reactants>reagents>products. Reagents are formally defined by molecules participating in the reaction whose atom mappings are not found in the resulting product. We leverage an atom-wise tokenization of the reactants and products, while we use a molecule-wise tokenization of the reagents [14].

4.2 Baselines

We use two standard sequence-to-sequence models as our baselines. The first is the recurrent neural network described by Schwaller *et al.*, which is comprised of a bidirectional LSTM with learnable attention weights between decoder and encoder outputs, denoted BiLSTM + Attn [14]. Our second sequence-to-sequence baseline is a vanilla transformer model, denoted MT. We also compare our model to highly expressive graph neural networks; our third baseline is a Weisfeiler-Lehman network (WLN), which has reported higher top-3^{+} accuracies than BiLSTM + Attn [15], and currently represents the state of the art for diverse, accurate retrosynthesis prediction.

To the best of our knowledge, no other transformer-based network architecture has been evaluated in the literature on USPTO-STEREO for retrosynthesis. The results for BiLSTM + Attn and WLN were reported from the work of Schwaller *et al.* and Jin *et al.* [14, 15].

4.3 Evaluation method

The models are evaluated on the top-n accuracy of the predictions. The top-n accuracy is the frequency with which the ground truth target appears within the top-n predictions of the model. As a change in one token can result in a completely different molecule, the model's predictions must exactly match the ground truth to be considered correct. However, one molecule can have a large number of valid SMILES representations. It's possible that the model could output the correct molecule with a syntactically correct SMILES string, but in non-canonical form. Since every molecule can only have one canonical SMILES string, we canonicalize the predictions using the RDKit Python API [16]. This method has previously shown to increase calculated accuracies up to 1.5% [14].

The models are also evaluated on the invalid SMILES rate, which is the frequency with which the model outputs syntactically, semantically, or chemically invalid SMILES. This was similarly computed using RDKit.

4.4 Experimental details

We train the network on a NVIDIA T4 Tensor Core GPU using Adam optimization with parameters $\alpha = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\varepsilon = 10\text{e}^{-}8$. We additionally include a learning rate decay of 0.1. The input reactions have a maximum reaction length of 200 tokens, and each token has an embedding dimension of 256. The BiLSTM + Attn network has hidden dimension 512 and utilizes the teacher forcing method during training. We use gradient clipping with a max norm of 1.0 to prevent exploding gradients. To train, we use minibatch gradient descent with a batch size of 32. At inference time, we use a beam search with a beam width of 10.

4.5 Results

4.5.1 Top-n accuracy

USPTO-STEREO Top-n Accuracy [%]					
Category	Model	Top-1	Top-2	Top-3	Top-5
sequence-to-sequence	BiLSTM + Attn	80.3	84.7	86.2	87.5
	MT	80.4	84.6	86.7	87.6
	SDMT	**80.8**	**86.1**	87.4	88.5
graph-based	WLN	79.6		**87.7**	**89.2**

Table 2: Top-n accuracies for the evaluated models.

The top-n accuracies evaluated on the USPTO-STEREO test set are reported in Table 2. Notably, SDMT outperforms all of the baselines in top-1 and top-2 accuracy for which the data is known. However, WLN outperforms all other models in the top-3 and top-5 accuracy. This is likely due to a difference in the training procedures: during training, WLN learns to rank multiple potential candidates, whereas each of the sequence-to-sequence models were only trained to predict the top-1 set of reactants. This also supports why SDMT and each of the other sequence-to-sequence models outperform WLN in top-1 accuracy.

SDMT surpasses all other sequence-to-sequence models across all top-n accuracy, suggesting that the syntax-guided layer is improving the model's ability to reason about reactions.

4.5.2 Invalid SMILES rate

The top-n invalid SMILES rate evaluated on the USPTO-STEREO test set are reported in Table 3. There is a large difference between the top-1 invalid rate of the transformer-based models and the BiLSTM + Attn model. Namely, the transformer models perform better at outputting valid SMILES strings than does BiLSTM + Attn. This indicates that transformer networks may be able to

USPTO-STEREO Top-n Invalid SMILES Rate [%]						
Category	Model	Top-1	Top-2	Top-3	Top-5	Top-10
sequence-to-sequence	BiLSTM + Attn	1.3				
	MT	0.2	0.2	0.5	1.2	2.7
	SDMT	**0.1**	**0.1**	**0.1**	**0.3**	**0.5**
graph-based	WLN					

Table 3: Top-n invalid SMILES rate for the evaluated models.

understand the SMILES grammar better than recurrent neural networks, even without the presence of a syntax-guided layer.

Furthermore, the top-2^+ accuracies suggest that SDMT is more consistent than MT in outputting syntactically valid molecules. The top-1, top-2, and top-3 accuracies are the same for SDMT, and they increase marginally for $n \geq 3$, while the top-10 accuracy for MT is over 5 times that for SDMT.

Considering that SDMT and MT have nearly the same number of parameters, yet SDMT has higher top-n accuracies and lower invalid SMILES rates, the syntax-guiding layer is a worthwhile addition to the transformer architecture for modeling chemical reactions.

4.5.3 Alternative retrosynthetic steps

One limitation of the USPTO dataset is that it does not account for retrosynthetic steps that are plausible but not the same as what's been patented. Currently, there are no definitive methods other than human discretion to affirm the feasibility of alternative reactions. We investigated some of the reactions for which SDMT's top-1 prediction was incorrect, and found that there is a nontrivial number of alternate reaction schemes that are plausible within the context of the reaction. An example of this is given in Figure 4, which illustrates that one common mistake is that the model often chooses different leaving groups for nucleophilic substitution reactions. These are marked incorrect by the evaluation metric, but in most cases, they represent plausible reaction schemes.

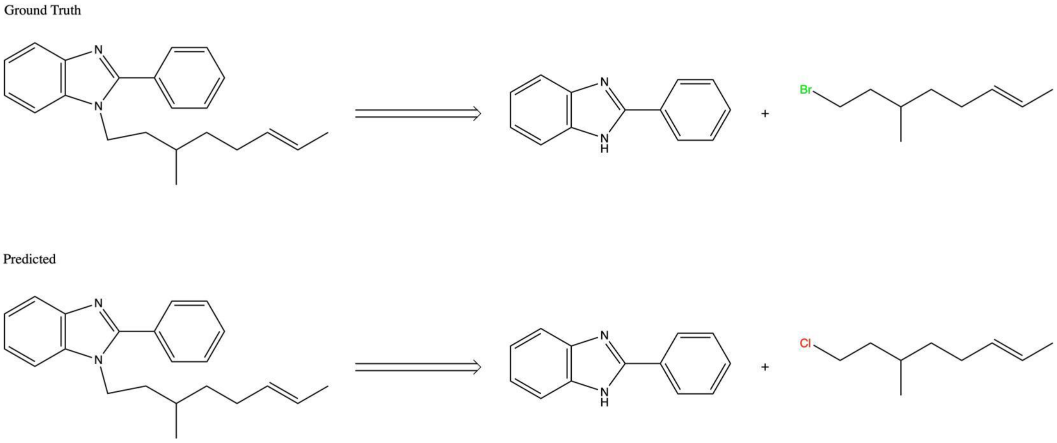

Figure 4: A nucleophilic substitution reaction marked incorrect that differs only in the choice of halide leaving group (ground truth: bromine, predicted: chlorine).

5 Conclusion

In this paper, we develop a transformer-based model that competes with the state-of-the-art for retrosynthesis prediction. Our model, which incorporates a lightweight syntax-guided transformer block, achieves high top-1 and top-2 accuracies. Furthermore, SDMT consistently outputs syntactically, semantically, and chemically valid predictions, suggesting that the syntax-guided layer is improving the transformer model's ability to replicate the SMILES grammar. Our approach is fully data-driven, template- and rule-free, and compatible with stereoselective reactions, allowing it to be applied to a larger and more complex set of molecular decompositions. Hopefully, with the improvements from this model, retrosynthesis is something that can be more reliably automated.

References

[1] Robert Robinson. Lxiii.—a synthesis of tropinone. *J. Chem. Soc., Trans.*, 111:762–768, 1917.

[2] David C. Blakemore, Luis Castro, Ian Churcher, David C. Rees, Andrew W. Thomas, David M. Wilson, and Anthony Wood. Organic synthesis provides opportunities to transform drug discovery. *Nature Chemistry*, 10(4):383–394, Apr 2018.

[3] Matthew H. Todd. Computer-aided organic synthesis. *Chem. Soc. Rev.*, 34(3):247–266, 2005.

[4] Wolf-Dietrich Ihlenfeldt and Johann Gasteiger. Computer-assisted planning of organic syntheses: The second generation of programs. *Angewandte Chemie International Edition in English*, 34(23-24):2613–2633, 1996.

[5] Juno Nam and Jurae Kim. Linking the neural machine translation and the prediction of organic chemistry reactions, 2016.

[6] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, Feb 1988.

[7] Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A. Hunter, Costas Bekas, and Alpha A. Lee. Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction. *ACS Central Science*, 5(9):1572–1583, Sep 2019.

[8] Eunji Kim, Dongseon Lee, Youngchun Kwon, Min Sik Park, and Youn-Suk Choi. Valid, plausible, and diverse retrosynthesis using tied two-way transformers with latent variables. *Journal of Chemical Information and Modeling*, 61(1):123–133, Jan 2021.

[9] Shuangjia Zheng, Jiahua Rao, Zhongyue Zhang, Jun Xu, and Yuedong Yang. Predicting retrosynthetic reactions using self-corrected transformer neural networks. *Journal of Chemical Information and Modeling*, 60(1):47–55, 2020. PMID: 31825611.

[10] Zhuosheng Zhang, Yuwei Wu, Junru Zhou, Sufeng Duan, Hai Zhao, and Rui Wang. Sg-net: Syntax guided transformer for language representation, 2021.

[11] Craig A James. Opensmiles specification. page 29.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[13] Daniel Lowe. Chemical reactions from US patents (1976-Sep2016). 6 2017.

[14] Philippe Schwaller, Théophile Gaudin, Dávid Lányi, Costas Bekas, and Teodoro Laino. "found in translation": predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chem. Sci.*, 9:6091–6098, 2018.

[15] Wengong Jin, Connor W. Coley, Regina Barzilay, and Tommi S. Jaakkola. Predicting organic reaction outcomes with weisfeiler-lehman network. *CoRR*, abs/1709.04555, 2017.

[16] Noel M. O'Boyle. Towards a universal smiles representation - a standard method to generate canonical smiles based on the inchi. *Journal of Cheminformatics*, 4(1):22, Sep 2012.