A Sequence-to-Sequence Transformer Model for Disconnection Aware Retrosynthesis

Andrea Byekwaso
IBM Research Europe, Zürich

Philippe Schwaller
IBM Research Europe, Zürich

Alain C. Vaucher
IBM Research Europe, Zürich

Alessandra Toniato
IBM Research Europe, Zürich

Teodoro Laino
IBM Research Europe, Zürich

Abstract

Retrosynthesis is an approach commonly undertaken when considering the manufacture of novel molecules. During this process, a target molecule is broken down and analyzed by considering the bonds to be changed as well as the functional group interconversion. In modern computer-assisted synthesis planning tools, the predictions of these changes are typically carried out automatically. However there may be some benefit to the decision being guided by those executing the process: typically, chemists have a clear idea where the retrosynthetic change should happen, but not how such a transformation is to be realized. Using a data-driven model, the retrosynthesis task can be further explored by giving chemists the option to explore specific disconnections. In this work, we design an approach to provide this option by adapting a transformer-based model for single-step retrosynthesis. The model takes as input a product SMILES string, in which the atoms where the transformation should occur are tagged accordingly. This model predicts precursors corresponding to a disconnection occurring in the correct location in 88.9% of the test set reactions. The assessment with a forward prediction model shows that 76% of the predictions are chemically correct, with 14.1% perfectly matching the ground truth.

1 Introduction

Society relies on the production of novel molecules and materials. The most common method for developing novel compound synthesis involves breaking existing atom-to-atom bonds and forming new ones. Retrosynthetic analysis (retrosynthesis) is the practice of analyzing a target molecule into commercially available precursors by way of potential disconnections (bond breakages) and functional group interconversion [1]. E. J. Corey, whose primary focus was in studying the structure of organic compounds, was the first one to pioneer this concept [2, 3]. However, despite 60 years of synthetic strategy development, the search space for all potential disconnections is so large that even professionals find it difficult to explore. The fact that compounds can have multiple retrosynthetic routes and that each route is affected by elements such as reaction conditions, availability of reactants, and reaction yields (to mention a few) makes the entire process even more challenging [4].

Several deep-learning-based approaches to single-step retrosynthesis treat the prediction of possible disconnections as a translation task, relying on the use of the Transformer architecture [5] and the simplified molecular-input line-entry system (SMILES) [6, 7] notation [8–11]. Given a target

ELLIS Machine Learning for Molecule Discovery Workshop 2021.

molecule, these approaches suggest the best set of precursors (i.e. reactants, and possibly other reagents) as the translation's outcome, with the possibility to generate multiple such sets.

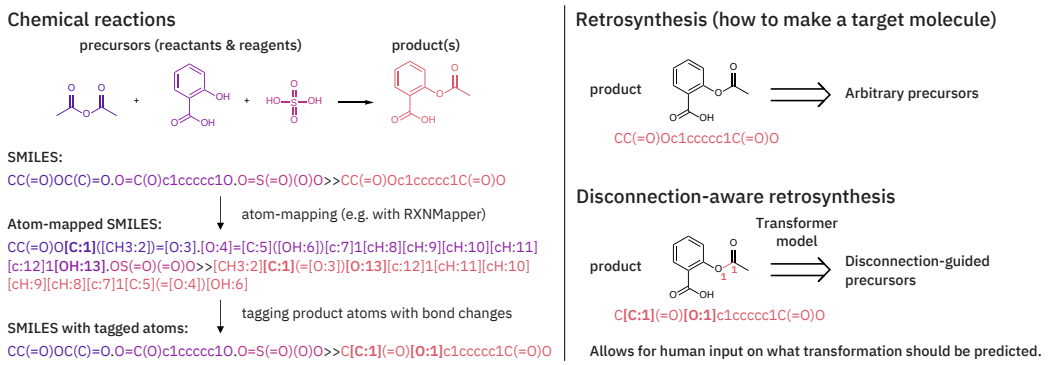

Figure 1: Left: Overview of the processing pipeline of the chemical reactions represented as SMILES to tag the product atoms with changes. Right: Comparison of typical single-step retrosynthesis with disconnection-aware retrosynthesis.

One possible downside of such models is that they provide chemists little control over the disconnections they want to investigate for a given target molecule. The recommended precursors are not guaranteed to be consistent with the chemist's desired disconnections because the single-step retrosynthetic models suggest the precursors thought to be optimal based on the training dataset. In this work, we extend transformer-based models for single-step retrosynthesis in order to enhance the control by chemists when determining a retrosynthetic route by exploring user-defined disconnections. In a multi-step retrosynthetis settting, this provides a 'human-in-the-loop' component that combines expert knowledge and experience with the power of deep learning. We can therefore use human knowledge and decision-making strategies that statistical and machine learning algorithms can't yet encode due to a lack of relevant training data to provide an enhanced experience in retrosynthetic problems. Figure 1 provides an overview for this work.

2 Method

We base our model on the approach by Schwaller et al. [11]. This model converts a tokenized SMILES string representing a product into a tokenized SMILES string representing a set of precursors (including solvent, reagents, and catalysts), which is achieved with the help of a sequence-to-sequence transformer model. In the current work, we encode the desired disconnection into the input SMILES string, we exploit the atom mapping notation of SMILES: all the atoms involved in the transformation are attributed an atom mapping number of `1`. Note that although we rely on the atom mapping notation to tag the atoms, the produced strings do not carry any information on atom mapping. As a simple example, in order to produce the target molecule butanol, with SMILES string `CCCCO`, from a transformation involving the alcohol bond, the input SMILES string (before tokenization) would be `CCC[CH2:1][OH:1]`.

The dataset to train this model can be generated starting from any set of reaction SMILES. All the reactions are first atom-mapped with RXNMapper [12]. This enables the identification of product atoms with different neighboring atoms or bonds than the precursors (see details in Appendix A). These atoms are assigned an atom map number of `1`, while the atom mapping information is removed from all the other product atoms. The atom mapping information is also removed from all the precursors.

For this study, the dataset was generated from Pistachio (version 2020Q1) [13], resulting, after processing, in training, validation, and test sets of sizes 2,269,560, 10,000, and 126,435, respectively. The model architecture and training are detailed in Appendix B.

Figure 2 (top) illustrates a chemical reaction from the dataset, with the atom mapping obtained from RXNMapper. From the atom mapping, it can be seen that the atoms involved in the chemical transformation are the ones with mappings 2, 3, and 5. Accordingly, those are the atoms that end up being marked during the dataset generation. Figure 2 (bottom) shows the chemical equation after

pre-processing, where the atom mapping information has been removed, and the target molecule contains an indication of what atoms are involved in the chemical transformation.

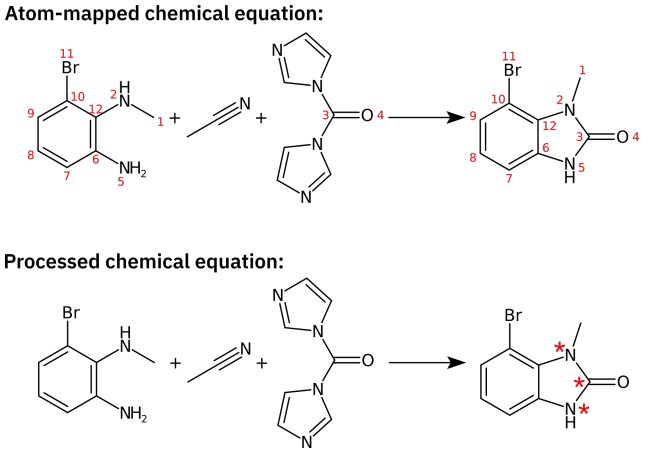

Figure 2: Top: An atom-mapped chemical equation. The red numbers indicate where the precursor atoms end up in the product. It can be seen that the atoms involved in the transformation are the ones with atom mapping numbers 2, 3 and 5. Bottom: The same chemical equation after processing. The atoms involved in the transformation are red indicated using red asterisks. Another example is provided in Figure 1.

3 Results

In order to analyze the results produced by the model, we determined how many predictions were identical to the ground truth after canonicalizing with RDKit [14] and sorting the suggested precursors. We found that 14.1% of the reactions produced by the model were identical to the ground truth. This value is in line with the accuracy of transformer-based models suggesting precursors including reagents, solvents and catalysts. However, because there are typically several ways to synthesize the same molecule [11], the usefulness of this statistic, which corresponds to the top-1 accuracy of the single-step retrosynthesis model, is limited.

More significantly, it is necessary to assess how many predictions involved chemically correct disconnections between the given pair of atoms (disconnection accuracy). To do this, the predicted precursors were combined with the target molecules to produce reaction SMILES that were atom-mapped with RXNMapper [12]. This made it possible to determine if the atoms involved in the predicted transformation are the same as were specified in the model input. It was found that 88.9% of the reactions had disconnections between the correct atoms. We also evaluated the chemical correctness of the predictions by determining the round-trip accuracy on the predictions, which is obtained by comparing the product predicted by a forward reaction prediction model with the initial target molecule. This analysis shows that 76% of the predicted disconnections between the given set of atoms are chemically valid disconnections. Figure 3 illustrates some predictions of the model.

Table 1: Test set accuracies with respect to the number of atoms tagged in the input molecules. The round-trip accuracy is calculated on the subset of reactions with correct disconnections.

Number of tagged atoms	Dataset fraction (%)	Disconnection accuracy (%)	Round-trip accuracy (%)	Identical to ground truth (%)
0	1.29	98.3	7.3	9.0
1	12.65	99.1	84.2	13.9
2	51.22	98.5	88.4	15.6
3	10.99	92.2	59.3	15.4
4	5.64	77.2	45.9	10.6
5	5.34	78.5	53.0	13.4
6-10	8.96	56.3	38.2	9.5
>10	3.91	22.4	20.7	9.7
Overall	100.0	88.9	76.0	14.1

Table 1 presents all these metrics as a function of the number of tagged atoms in the target molecules. The reactions with zero tagged atoms correspond to reactions where the reaction product is present already in the precursors, sometimes with a different stereochemistry (which was ignored during the generation of the dataset). Table 1 shows that the disconnection accuracy decreases with the number of tagged atoms. This is consistent with the difficulty of this task for a chemist: it is easy to suggest reactions with one or two transformed atoms, while it is much harder to suggest precursors corresponding to five or more transformed atoms.

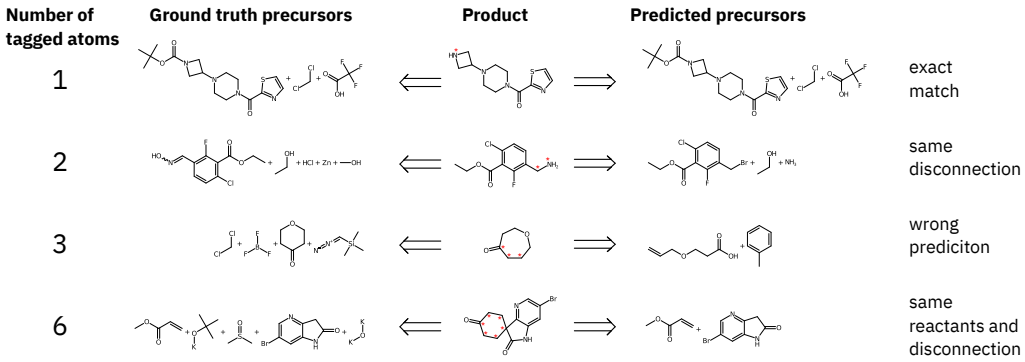

Figure 3: Examples from the test set with different numbers of tagged atoms.

4 Conclusion

In this work, we developed a 'human-in-the-loop' approach for retrosynthesis, where a Transformer-based model for single-step retrosynthesis allows experts to specify where the disconnections should occur. This work gives those performing retrosynthetic analysis the decision of where bonds should be broken. We processed the input training data in such a way that the atoms involved in the chemical transformation during the reaction are marked correspondingly. By using this approach, we were able to train a model that was able to predict 14.1% of reactions identically to the data inputted. 88.9% of the predictions made by the model provided disconnections between the specified atoms, and in 76.0% of these cases, the predictions were chemically correct. This work provides a new way to exploit human knowledge and decision-making strategies through a 'human-in-the-loop' scheme.

Combining additional elements of artificial intelligence in the current development may further increase the accuracy and ability to predict the disconnections in the desired location, improving the overall performance . In fact, the current model requires chemists to specify exactly all the atoms involved in the change. In the future, we plan to train a variant of the model that allows only subsets

of transformed atoms to be tagged. This will be achieved by generating a new data set where target SMILES are represented multiple times, with different numbers of tagged atoms.

References

[1] WT Wipke, H Braun, G Smith, F Choplin, and W Sieber. SECS — simulation and evaluation of chemical synthesis: strategy and planning. ACS Publications, 1977.

[2] E. J. Corey and W. Todd Wipke. Computer-assisted design of complex organic syntheses. *Science*, 166(3902):178–192, 1969. doi: 10.1126/science.166.3902.178. URL `https://www.science.org/doi/abs/10.1126/science.166.3902.178`.

[3] Elias James Corey. *The logic of chemical synthesis*. Wiley, 1991.

[4] Binghong Chen, Chengtao Li, Hanjun Dai, and Le Song. Retro*: learning retrosynthetic planning with neural guided A* search. In *International Conference on Machine Learning*, pages 1608–1616. PMLR, 2020.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[6] David Weininger. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, 28:31–36, 1988.

[7] David Weininger, Arthur Weininger, and Joseph L Weininger. SMILES. 2. Algorithm for generation of unique SMILES notation. *J. Chem. Inf. Comput. Sci.*, 29:97–101, 1989.

[8] Qingyi Yang, Vishnu Sresht, Peter Bolgar, Xinjun Hou, Jacquelyn L Klug-McLeod, Christopher R Butler, et al. Molecular transformer unifies reaction prediction and retrosynthesis across pharma chemical space. *Chem. Commun.*, 55:12152–12155, 2019.

[9] Pavel Karpov, Guillaume Godin, and Igor V Tetko. A transformer model for retrosynthesis. In *International Conference on Artificial Neural Networks*, pages 817–830. Springer, 2019.

[10] Hongliang Duan, Ling Wang, Chengyun Zhang, Lin Guo, and Jianjun Li. Retrosynthesis with attention-based NMT model and chemical analysis of "wrong" predictions. *RSC Adv.*, 10: 1371–1378, 2020.

[11] P Schwaller, R Petraglia, V Zullo, V. H Nair, R. A Haeuselmann, R Pisoni, C Bekas, A Iuliano, and T. Laino. Predicting retrosynthetic pathways using transformer-based models and a hypergraph exploration strategy. *Chem. Sci.*, 2020.

[12] Philippe Schwaller, Benjamin Hoover, Jean-Louis Reymond, Hendrik Strobelt, and Teodoro Laino. Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. *Sci. Adv.*, 7:eabe4166, 2021.

[13] Nextmove Software Pistachio, version 2020q1. URL `http://www.nextmovesoftware.com/pistachio.html`. (Accessed Sep 23, 2021).

[14] Greg Landrum. Rdkit: Open-source cheminformatics. URL `http://www.rdkit.org`.

[15] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. OpenNMT: Open-Source Toolkit for Neural Machine Translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada, 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-4012.

[16] OpenNMT-py library, version 1.2.0. `https://github.com/OpenNMT/OpenNMT-py` (Accessed Sep 23, 2021).

A Determination of atoms involved in the reaction

Algorithm 1 sketches the pseudo code used to prepare the training data for the model.

Algorithm 1 A function that converts the molecule objects for the precursors and product (typically in RDKit Mol format) to the format required for training the model. *all_atom_map_numbers* is a function returning the list of atom map indices present in the product object. *neighborhood* refers to the neighboring atoms and corresponding bond types.

Input:
precursors: Molecule object for the precursors, including atom mapping information
product: Molecule object for the product, including atom mapping information
Output:
precursors, product: a tuple containing the new molecule objects for precursors and product

```
transformed_atoms ← list()

for index ∈ all_atom_map_numbers(product) do
    precursors_atom ← precursors[index]
    product_atom ← product[index]
    if neighborhood(precursors_atom) ≠ neighborhood(product_atom) then
        transformed_atoms.append(product_atom)
    end if
end for

for atom ∈ precursors do
    atom.mapping = 0
end for

for atom ∈ product do
    if atom ∈ transformed_atoms then
        atom.mapping = 1
    else
        atom.mapping = 0
    end if
end for
```

B Model implementation and training

The model was implemented using the OpenNMT-py library [15, 16] version 1.0.0. It was trained with the following command:

```
onmt_train -data $DATA -save_model $SAVE_MODEL -seed 42 -gpu_ranks 0 \
  -save_checkpoint_steps 5000 -keep_checkpoint 20 -train_steps 260000 \
  -param_init 0 -param_init_glorot -max_generator_batches 32 \
  -batch_size 6144 -batch_type tokens -normalization tokens \
  -max_grad_norm 0 -accum_count 4 -optim adam -adam_beta1 0.9 \
  -adam_beta2 0.998 -decay_method noam -warmup_steps 8000  \
  -learning_rate 2 -label_smoothing 0.0 -report_every 1000  \
  -valid_batch_size 8 -layers 4 -rnn_size 384 -word_vec_size 384 \
  -encoder_type transformer -decoder_type transformer -dropout 0.1 \
  -position_encoding -share_embeddings -global_attention general \
  -global_attention_function softmax -self_attn_type scaled-dot \
  -heads 8 -transformer_ff 2048
```

This defines a transformer-based sequence-to-sequence model and trains it by optimizing the negative log-likelihood. The training curves for the cross entropy are shown in Figure 4.

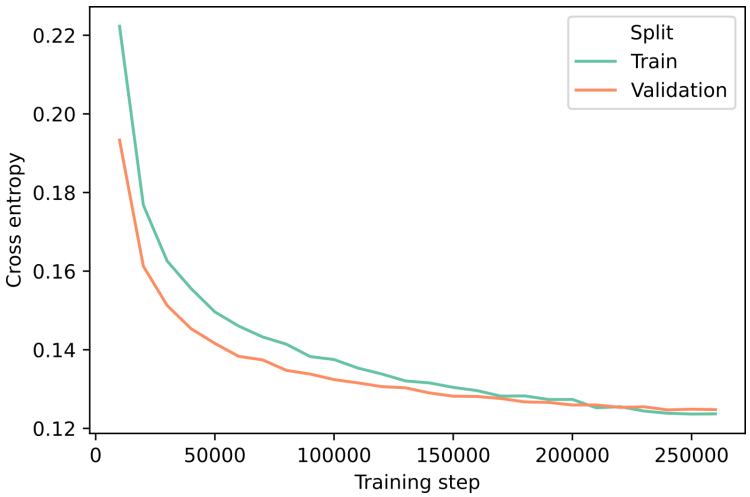

Figure 4: Training curves for the cross entropy.