Easy-to-use Sequential Model for Retrosynthesis Predition

Suong Tran

Introduction

Synthesis planning involves determining the method to synthesize a chemical compound using available starting materials through a sequence of chemically viable reaction steps. Basically speaking, synthesis planning is to answer the question *how do we make this target molecule?* - which is one of many difficult tasks for chemists. Retrosynthetic analysis, introduced by E. J. Corey in the 1960s, is the process of proposing route to synthesize target molecules from existing starting chemicals.[1] This approach has proven crucial and successful in the chemical industry. Retrosynthesis is difficult task because chemists have to know a lot of possible "moves" starting from a specific state. Therefore, recent years have witnessed a emergence of interest in developing computer-assisted synthetic planning (CASP).[2]

Through the application of advanced machine learning techniques on datasets of organic reactions, such as Reaxys[3] and the USPTO database[4], significant advancement has been made in searching for possible retrosynthetic pathways. More recently, researchers have used deep learning techniques for reaction prediction. This method typically involves a system with neural network (NN) component responsible for ranking candidates. The NN is used for the relevance of each rule in the knowledge base to a given example or evaluates the likelihood of each predicted product generated by applying all the rules in the knowledge base to a given example.[5, 6] However, these deep learning approaches rely heavily on defining highly specific rules tailored to a limited set of reactions with precise reactants and products. Additionally, the previous models are only for one-step retrosynthesis. Notably, these neural models are intricate and challenging for chemistry experts to adjust due to their complexity.

While molecules are commonly depicted as 2D or 3D graphs, it is equally possible to represent them using text sequences in line notation format, such as simplified molecular-input line-entry system (SMILES) strings [7]. SMILES string has been used in recent CASP.[5, 6] However, they are always acquainted with complicated machine learning models. Here, I propose a simple sequential model for multiple-step retrosynthesis working with SMILES string, but instead of using NN to score the candidates, the score will be calculated based on the suggested criteria from experts in organic chemistry. As one molecule has multiple way to make. ***The outcome of the project should be a completed model that given an input SMILES that represents the product, the model directly outputs a SMILES that represents the predicted starting materials. If time allows or if I pursue this project further, I plan to create a user-friendly interface, hopefully a simple website. This is crucial because not many chemists are familiar with coding, and having an accessible interface is essential for their ease of use.***

Data Set

The data set $USPTO\_500\_MT$ derived by Lu *et al* (available on their website) will be used in [8]. This data set is a complete version from Lowe's original USPTO data set.[4] Only columns of $reactants, reagents, products, yield$ in the data set will be used for this project. It was extracted and submitted with this proposal. The data set contains 143,535 reactions from published literature. 80% of the data set will be used for developing the work flow.

The rest will be used for testing.

The data set for commercially-available chemicals is in the process of acquiring. I am asking faculties in Chemistry Department if they have access to this data. If it is not available, I will scrape data, but it will not be sufficient.

The model

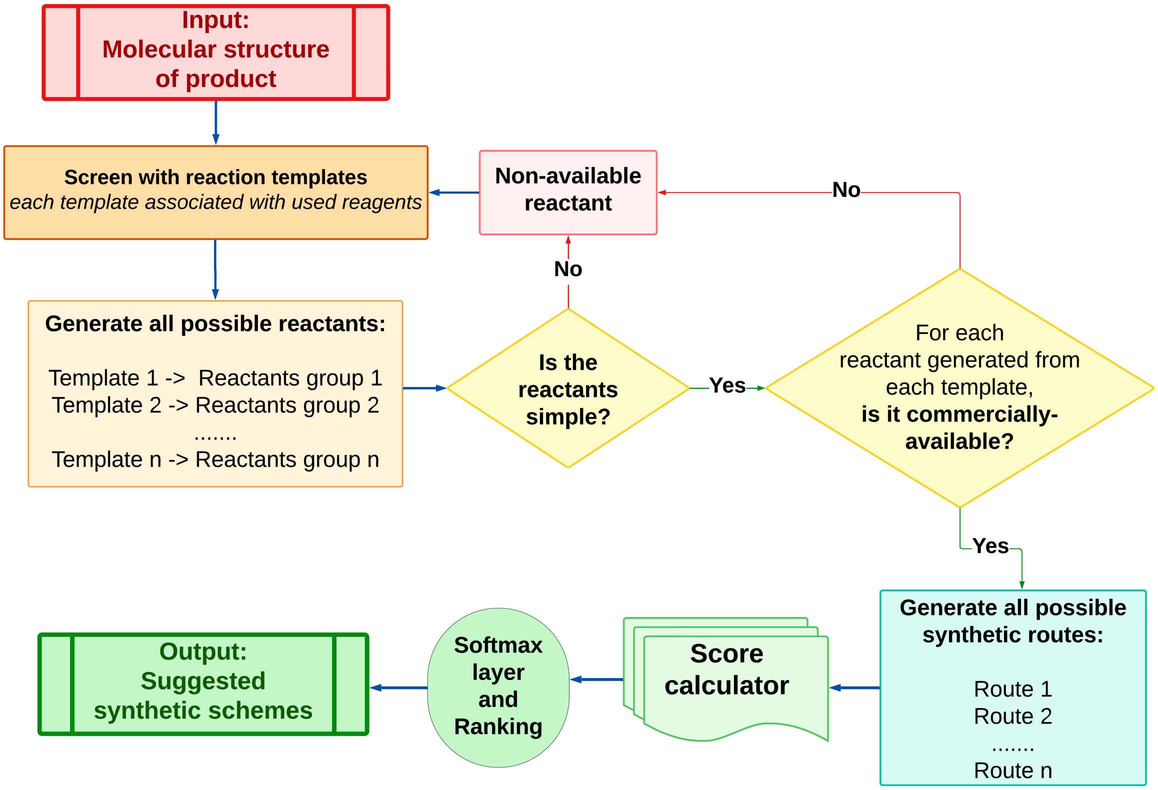
Input:
Molecular structure of product
Screen with reaction templates
each template associated with used reagents
Non-available reactant
No
No
Generate all possible reactants:
Template 1 -> Reactants group 1
Template 2 -> Reactants group 2
....
Template n -> Reactants group n
Is the reactants simple?
Yes
For each reactant generated from each template, is it commercially-available?
Yes
Generate all possible synthetic routes:
Route 1
Route 2
....
Route n
Score calculator
Softmax layer and Ranking
Output:
Suggested synthetic schemes

Figure 1: Flow chart of sequential model.

The primary approach here involves handling SMILES strings, similar to performing string processing in Python. The work flow is depicted in Figure 1.

The templates of reactions can be generated by 80% of the data set with strings manipulation. The score calculator will be developed based on the following questions: How many steps does it take to make the target product?, Does the suggested starting materials have reasonable price?, What is the expected yield of the synthesis?, How toxic it is? These questions might be changed or added after discussed with faculty in Chemistry Department at Dickinson College.

Evaluating model

Accuracy metric

In classification task, accuracy metric is usually used to evaluate the model. Similarly, in this field, top-N accuracy is the percentage that the model predicts the starting materials in top-N candidate score-ranking. It is a good metric to compare between model; however, sometime the top-1-predicted is the material that chemists does not want due to the price or toxicity. Additionally, for retrosynthetis, there are often many ways to break down the target molecule into various starting materials using different reaction types. Each bond in the target molecule can potentially be disconnected in retrosynthetic analysis.

Manually comparing

One other test to do is manually evaluating the model with a chemist. This way is tedious and also cannot be done for the whole test set. However, it can be valuable to enhance the model thanks to the lab experience of chemists.

References

(1) Corey, E. J. General methods for the construction of complex molecules. *Pure and Applied chemistry* **1967**, *14*, 19–38.

(2) Cook, A.; Johnson, A. P.; Law, J.; Mirzazadeh, M.; Ravitz, O.; Simon, A. Computer-aided synthesis design: 40 years on. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2012**, *2*, 79–107.

(3) Reaxys. *https://new.reaxys.com/*.

(4) Lowe, D. M. atent Reaction Extractor. **2014**.

(5) Wei, J.; Duvenaud, D.; Aspuru-Guzik, A. Neural networks for the prediction of organic chemistry reactions. ACS Cent Sci 2: 725–732. **2016**.

(6) Coley, C. W.; Barzilay, R.; Jaakkola, T. S.; Green, W. H.; Jensen, K. F. Prediction of organic reaction outcomes using machine learning. *ACS central science* **2017**, *3*, 434–443.

(7) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences* **1988**, *28*, 31–36.

(8) Lu, J.; Zhang, Y. Unified deep learning model for multitask reaction predictions with explanation. *Journal of Chemical Information and Modeling* **2022**, *62*, 1376–1387.