Retrosynthetic Planning

CS3308 Machine Learning

May 10, 2023

1 Introduction

- Three students at most form a group.
- Deadline for this project is June 16, 2023, 23:59. Each group needs to submit a report pdf on Canvas and the source code is also required by providing the link to your github repo.
- This project will be evaluated from workload (20%), model performance (20%), results analysis (40%) and report writing (20%). It is more crucial to conduct a reasonable analysis of the experimental results.
- Data required for this project is available in `https://jbox.sjtu.edu.cn/l/y1BtE5`.

2 Task1: Single-step retrosynthesis prediction

Given a target molecule and a set of available molecules, the retrosynthetic planning system is to find a feasible route to synthesize the target molecule. The single-step retrosynthesis prediction, which predicts a list of possible direct reactants given product, serves as the foundation for realizing the multi-step retrosynthetic planning. The dataset provided for this project is Schneider50k, which comprises $40,008$, $5,001$, and $5,007$ chemical reactions in the training, validation, and test sets, respectively. Each reaction is in the format of *reactants»products* and each reaction is associated with a corresponding reaction template. In this task, you need to design a model to do template-based single-step retrosynthesis prediction, which is a classification task of predicting reaction templates from products.

Hint: 1. You need to process the data into suitable format. You may need to perform the following steps.:

- Split one reaction into multiple reactions with only one product. For example, reaction $A + B \rightarrow C + D$ is split into $A + B \rightarrow C$ and $A + B \rightarrow D$.
- Extract reaction template from the reaction. rdchiral library is needed and the template can be extracted as follows. If a template cannot be extracted, delete the reaction. The extracted template is the label for the product of the reaction. The corresponding sample is (product, template).

```
from rdchiral.template_extractor import extract_from_reaction
reactants, products = reaction.split('>>')
inputRec = {'_id': None, 'reactants': reactants, 'products': products}
ans = extract_from_reaction(inputRec)
```

```
if 'reaction_smarts' in ans.keys():
return ans['reaction_smarts']
else:
return None
```

- Product is a chemical molecule, which can be transformed into a Morgan FingerPrint vector by the library rdkit as follows.

```
from rdkit import Chem
from rdkit.Chem import AllChem
mol = Chem.MolFromSmiles(product)
fp = AllChem.GetMorganFingerprintAsBitVect(mol, 2, nBits=2048)
onbits = list(fp.GetOnBits())
arr = np.zeros(fp.GetNumBits(), dtype=np.bool)
arr[onbits] = 1
return arr
```

- If you want to recover the reaction from the template, following code is helpful. Sometimes, the reaction can not be recovered, just skip this reaction.

```
from rdchiral.main import rdchiralRunText
out = rdchiralRunText(template, product)
```

- rdchiral is provided with the data file. rdkit library can be installed with

```
pip install rdkit
```

3 Task2: Molecule evaluation

This is a prediction task to predict the synthetic cost of the given molecule. Training and test data are provided in the format of (Packed Morgan FingerPrint, cost). Design your model to predict the cost for each molecule. (unpack the FingerPrint with numpy.unpackbits).

What's more, sometimes we need to predict the synthesis cost of multiple molecules simultaneously. One way to implement this is to predict each molecule separately and then sum them up:

$$cost = f(m_1) + f(m_2) + \cdots + f(m_k) \tag{3.1}$$

Another way is to design a function to predict the cost of multiple molecules:

$$cost = f(m_1, m_2, \cdots, m_k) \tag{3.2}$$

The synthetic cost of one molecule is mainly determined by its composed atoms and structure. Morgan FingerPrint can describe the atoms and structures of given molecule. So, we assume that molecules with similar fingerprints have similar synthetic cost. Discuss which method is better to estimate cost of multiple molecules, Equation 3.1 or Equation 3.2.

Hint: For the situation of Equation 3.2, graph neural network might be helpful. Cosine similarity of the fingerprints can be used as the metric to describe the relationship between molecules. Alternatively, you can try to establish connections between molecules in other ways.

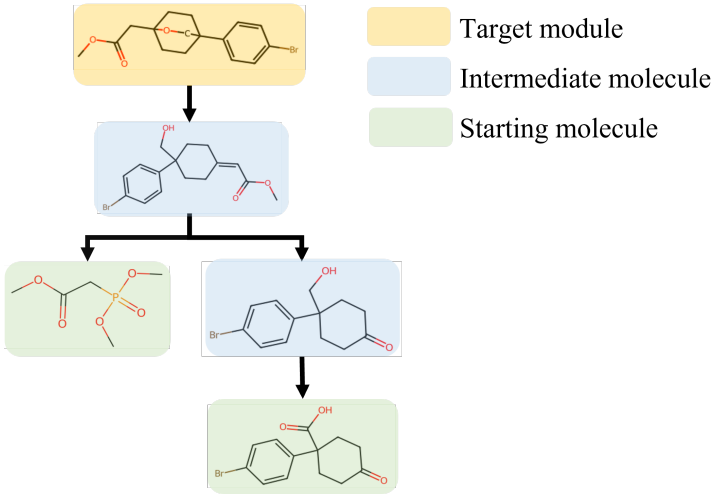

Figure 1: Example of a successful synthetic route.

4 Task3: Multi-step retrosynthesis planning

In general, molecules cannot be synthesized in one step and multiple reactions are required. Therefore, an efficient search algorithm is needed to find such a synthetic route with the help of single step retrosynthesis model and molecule evaluation function. In this task, a set of target molecules to be synthesized and a set of starting molecules are provided. You need to design a search algorithm to find the reactions, which can synthetic target molecules with the starting molecules. Figure 1 provide an example of successful synthetic route.

Hint:

- Available search algorithm includes depth first search, beam search, A* search, Monte Carlo Tree search and so on.
- Retro* is an search algorithm designed for retrosynthesis planning. You can directly test on Retro* and analyze the advantages and disadvantages of Retro*. The real synthetic route for each target molecule is also provided for further analyzation. Code for Retro* is available in `https://github.com/binghong-ml/retro_star`.

5 Project Report

Each group is required to turn in a project report with your main ideas, utilized methods and algorithms, experimental settings, finally experimental results, and your discussion about the results. The project report (.pdf) can be written either in English (encouraged) or in Chinese.

At the end of the report, please attach the contribution of each member as a percentage. And work done by each student is needed to be clarified. For example,

Name	Student ID	Score	Work
A	00000000000	30%	-
B	00000000001	30%	-
C	00000000002	40%	-

You are also required to submit the source code of your classification model by providing the link to your github repo in the report. If you do not know how to use github, please visit its tutorial (`https://guides.github.com/activities/hello-world/`) for some advice

6 Reference material

Segler, Marwin HS, Mike Preuss, and Mark P. Waller. "Planning chemical syntheses with deep neural networks and symbolic AI." Nature 555.7698 (2018): 604-610.

Chen, Binghong, et al. "Retro*: learning retrosynthetic planning with neural guided A* search." International Conference on Machine Learning. PMLR, 2020.

Han, Peng, et al. "Gnn-retro: Retrosynthetic planning with graph neural networks." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 36. No. 4. 2022.