FusionRetro: Molecule Representation Fusion via In-Context Learning for Retrosynthetic Planning

ICML 2023

Songtao Liu[1], Zhengkai Tu[2], Minkai Xu[3], Zuobai Zhang[4], Lu Lin[1], Rex Ying[5], Jian Tang[4], Peilin Zhao[6], Dinghao Wu[1]

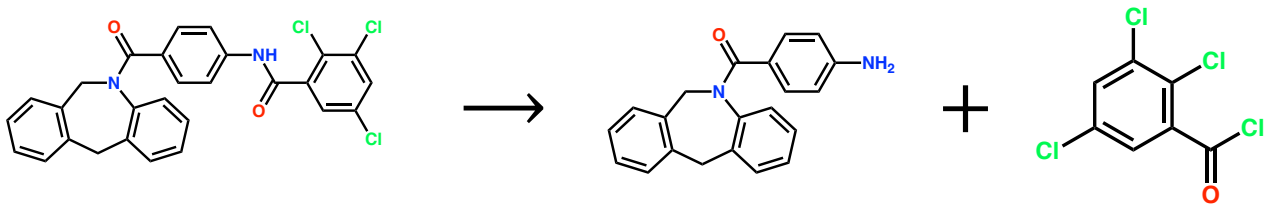
[1]Penn State, [2]MIT, [3]Stanford University, [4]Mila, [5]Yale University, [6]Tencent AI Lab

Contact: Songtao Liu (skl5761@psu.edu)

Background: Retrosynthesis prediction

Retrosynthesis Prediction:

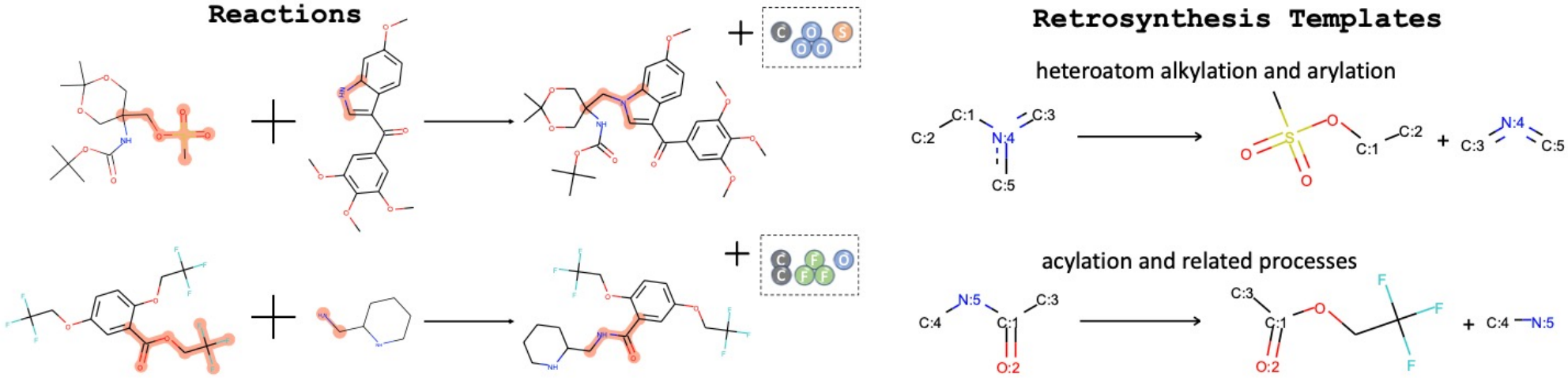
Given a target product molecule $T \in \mathcal{M}$(the set of chemical molecules), the goal of one-step retrosynthesis is to predict a set of reactants $\mathcal{R} = \{r_1, r_2, \cdots, r_n\} \subseteq \mathcal{M}$ that can react to synthesize this product.



Background: Single-step retrosynthesis model

One-step retrosynthesis model: 1. Template-based

Template-based algorithms first extract template rules from the training data, and then formulate the retrosynthesis task as template classification or template retrieval.



[1] Dai et al., Retrosynthesis prediction with conditional graph logical network. In *NeurIPS*, 2019.

Background: Single-step retrosynthesis model

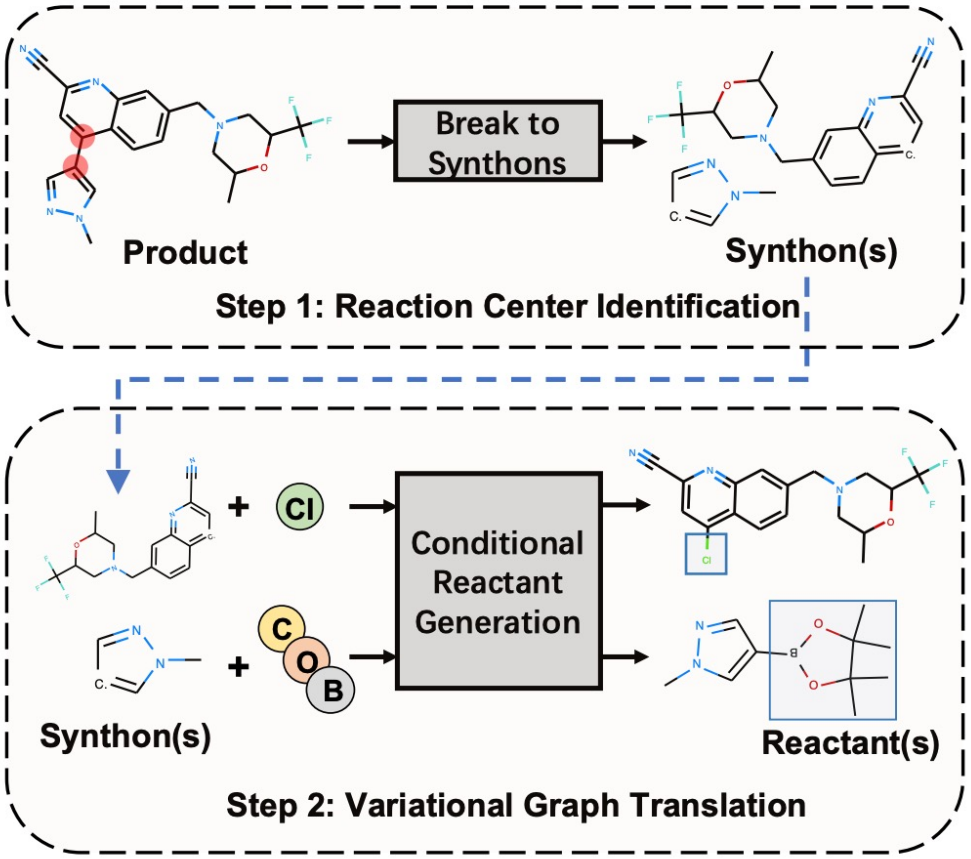
One-step retrosynthesis model: 2. Semi-template-based

Semi-template-based methods first identify the reaction center(s), break the target into several disconnected subgraphs, and recover the full molecule structures of reactants by attaching the leaving groups[2] or generative modeling[3,4].

[2] Somnath et al., Learning graph models for retrosynthesis prediction. In *NeurIPS*, 2021.
[3] Shi et al., G2Gs: A graph to graphs framework for retrosynthesis prediction. In *ICML*, 2020.
[4] Yan et al., RetroXpert: decompose retrosynthesis prediction like a chemist. In *NeurIPS*, 2020.

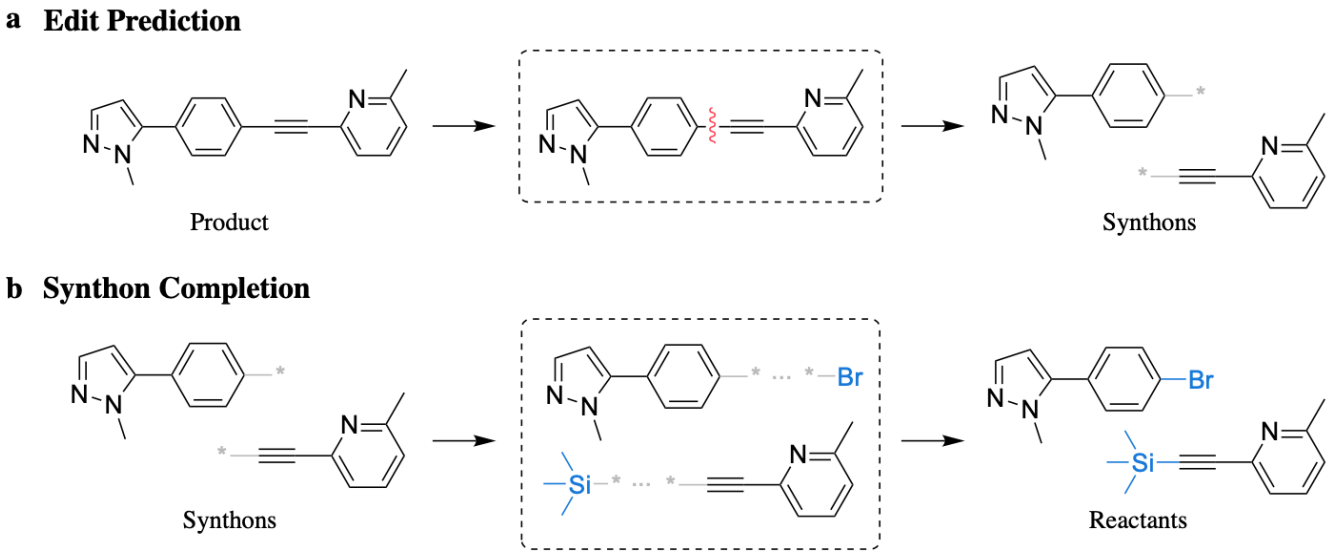
Background: G2Gs

One-step retrosynthesis model: 2. Semi-template-based



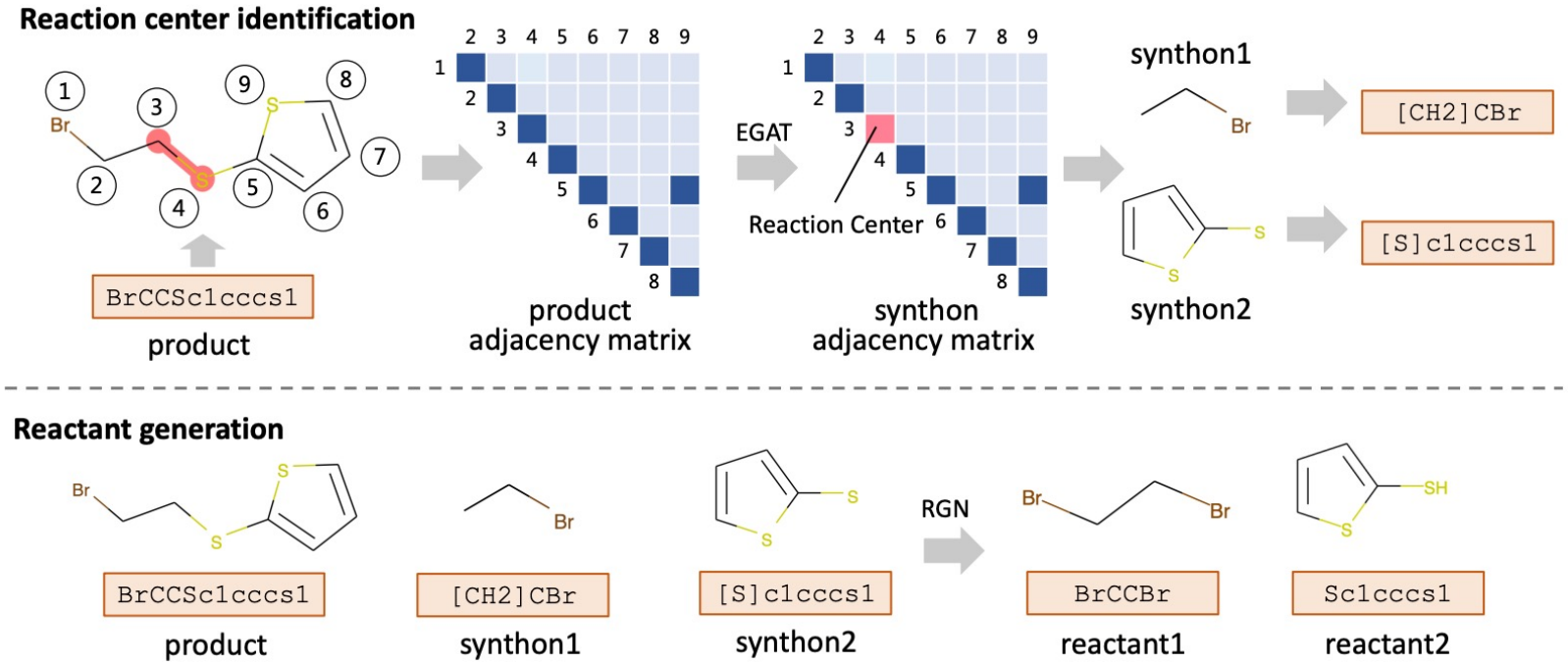
Background: Single-step retrosynthesis model

One-step retrosynthesis model: 2. Semi-template-based



Background: Single-step retrosynthesis model

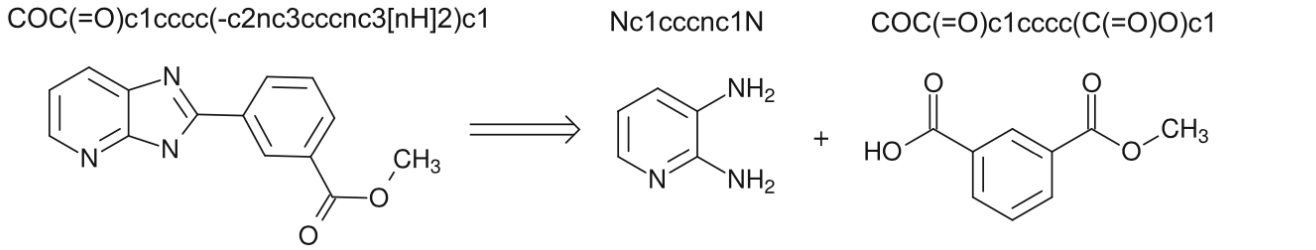
One-step retrosynthesis model: 2. Semi-template-based



Background: Single-step retrosynthesis model

One-step retrosynthesis model: 3. Template-free

Template-free methods use an end-to-end[5] or graph-edit based formulation[6].

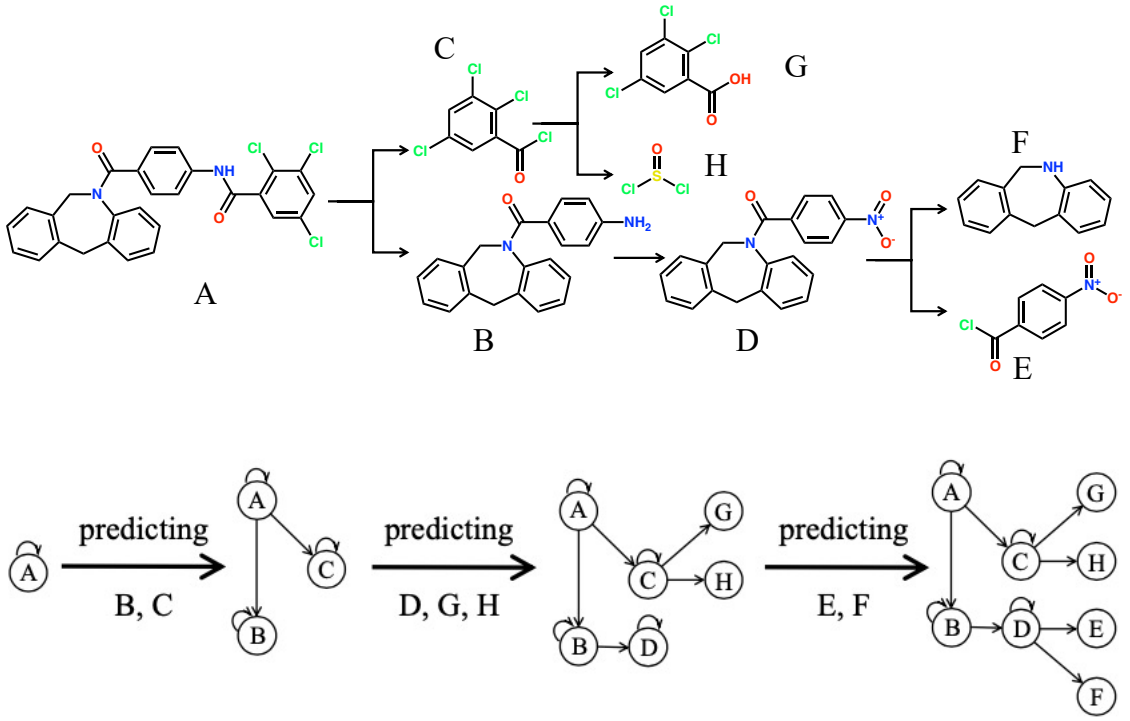


[5] Karpov et al., A transformer model for retrosynthesis. In *ICANN*, 2019.
[6] Sacha et al., Molecule edit graph attention network: modeling chemical reactions as sequences of graph edits. In *ICML workshop*, 2021.

Background: Retrosynthetic planning

Retrosynthetic planning:

Given a target molecule $T \in \mathcal{M}$(the set of molecules), the goal of retrosynthetic planning is to search for the starting materials $\mathcal{R} = \{r_1, r_2, \cdots, r_n\} \subseteq \mathcal{S}$(the set of starting materials) that can synthesize the target molecule through a set of chemical reactions $\tau = \{R_1, R_2, \cdots, R_m\}$



Current solutions for retrosynthetic planning

Current solutions for retrosynthetic planning:

1. Existing strategies model retrosynthetic planning as a search problem.

2. Starting from the target as the root node, these approaches employ some search algorithms to select the most promising node to expand, and then expand it into reaction precursors with a one-step retrosynthesis model.

3. Until a viable route is found in which all the leaf nodes are commercially available.

Drawback:

Previous work lacks the explicit modeling of the contextual information of in-context reactions along the partial synthetic routes preceding any given node.

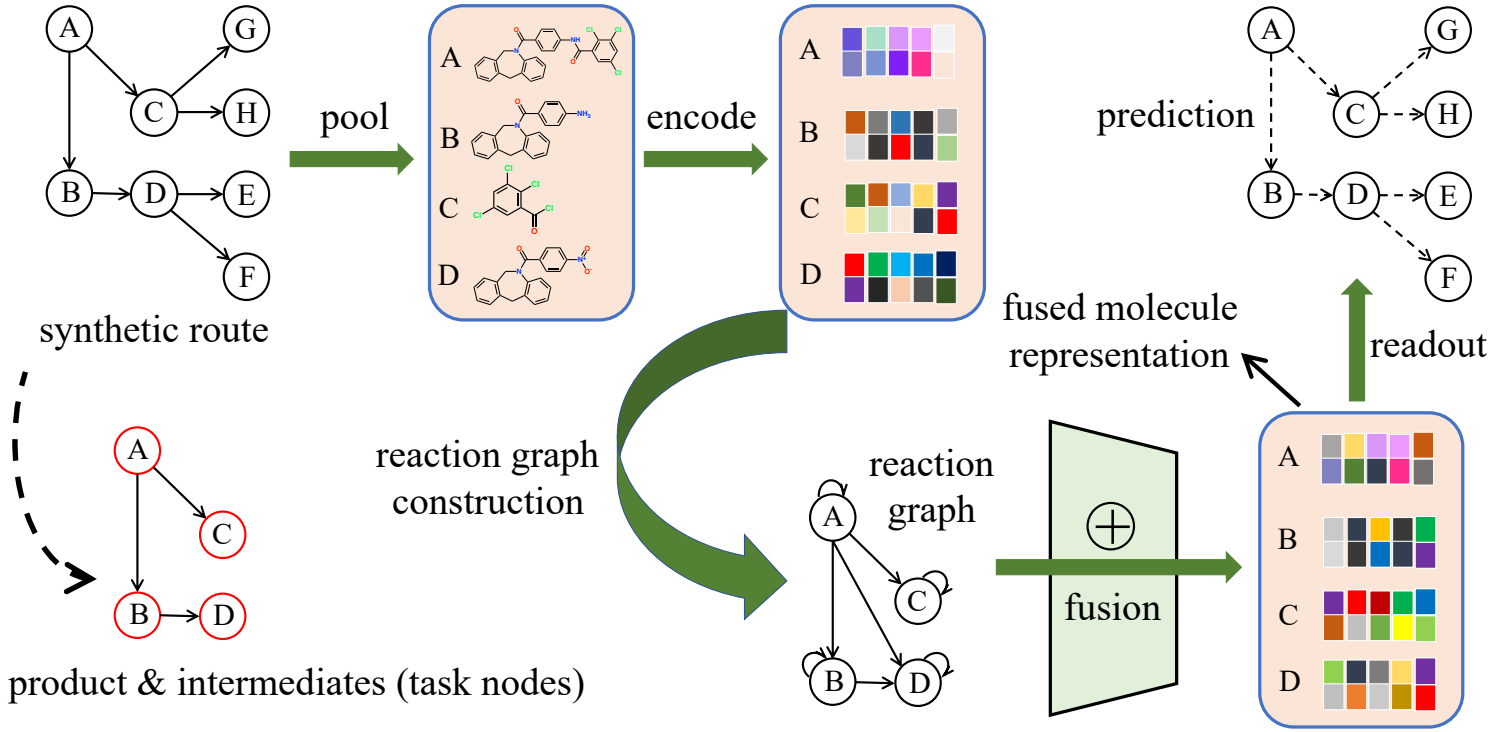
Method: FusionRetro

Motivation:

1. Inspired by the recent advancements in in-context learning within large language models, we utilize in-context examples for retrosynthesis prediction in retrosynthetic planning.

2. Another part of our motivation stems from the discrepancy between machine learning methods prevalent in existing works and the actual thought process of chemists.

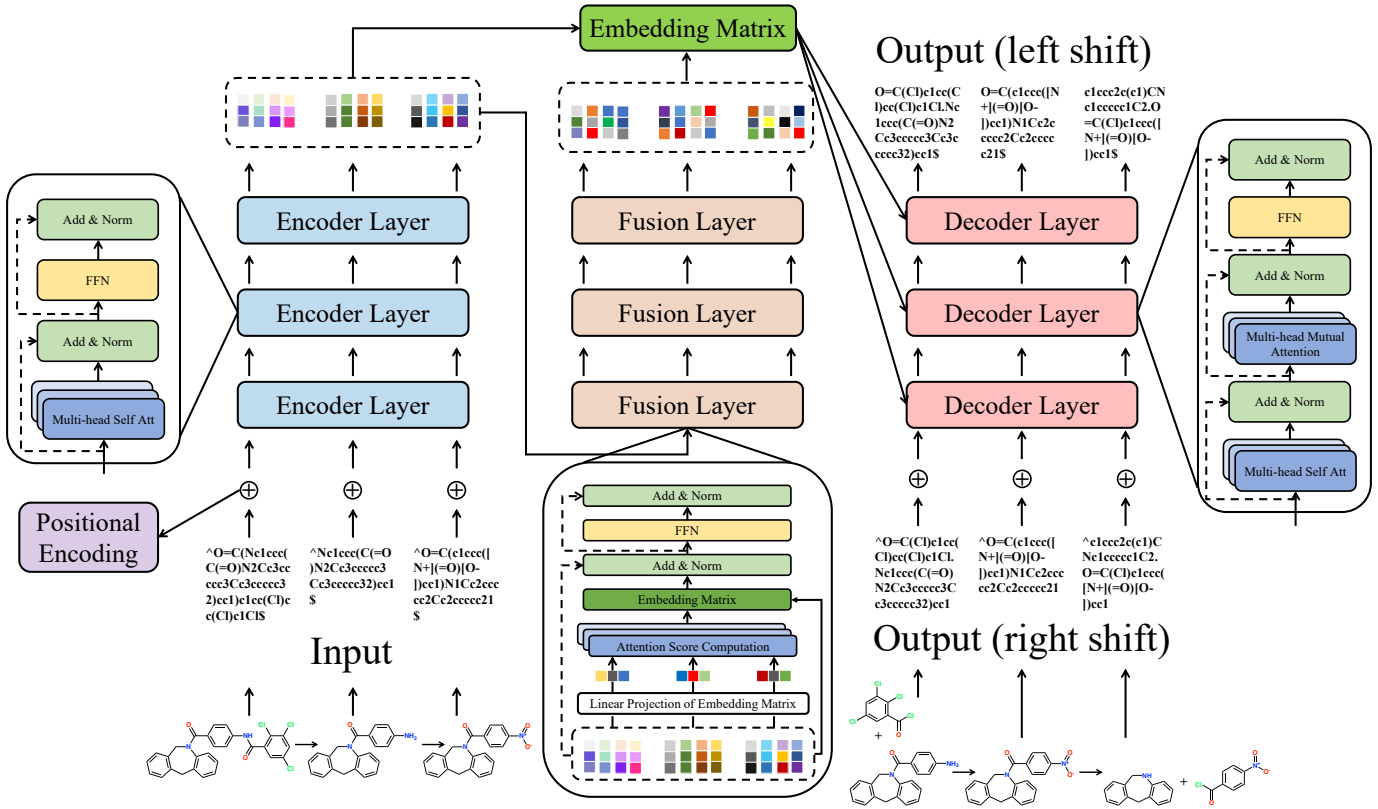
Method: Framework



Our framework is composed of three modules: *encode*, *fusion*, and *readout*.

1. After encoding the molecules,
2. we utilize the fusion module to generate the fused molecule representations (FMR).
3. This FMR is then employed to predict the reactants.

Method: Architecture



Our architecture consists of the encoder, decoder, and fusion modules.

1. We use encoder to transform the embeddings of input SMILES into latent representations.
2. Then we use the fusion module to attain a fused representation,
3. which is fed into decoder to yield the final prediction.

Evaluation: Protocol

Current evaluation metric for retrosynthetic planning:

1. Current evaluation metric of multi-step planning focuses on efficiency or quality.

2. Search efficiency has been measured in the success rate of finding pathways with buyable starting materials, as well as average numbers of iterations and node visits.

Drawback:

Some existing benchmarks[7] based on current metrics do not verify if the searched materials can synthesize the target molecule.

Our solution:

We propose a new evaluation metric: the set-wise exact match between the proposed starting materials and the ground truth.

[7] Chen et al., Retro*: Learning retrosynthetic planning with neural guided a* search. In ICML, 2020.

Evaluation: Results

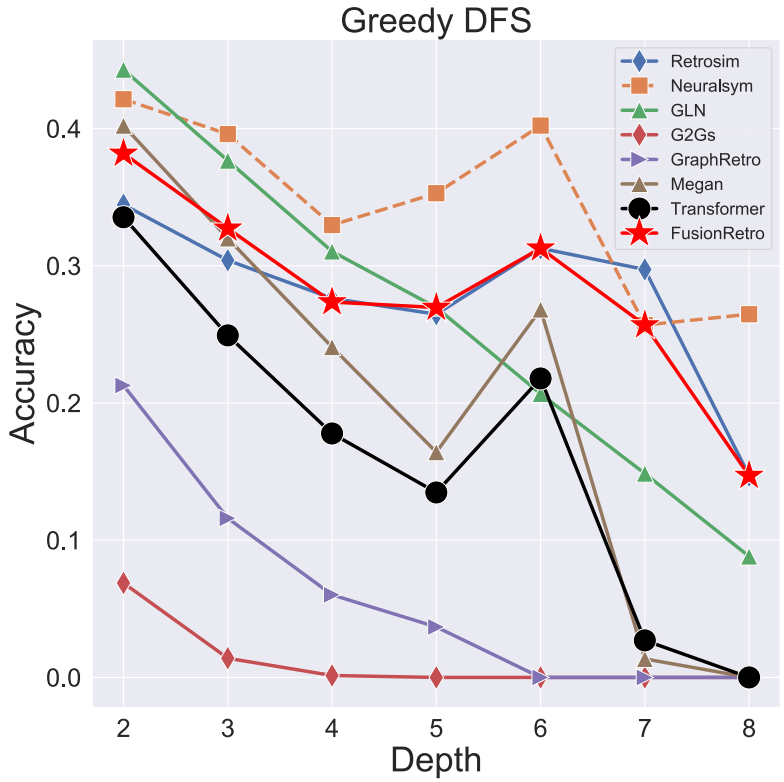
Table 1. Summary of retrosynthetic planning results in terms of exact match accuracy (%).

Search Algorithm	Retro*					Retro*-0					Greedy DFS
	Top-1	Top-2	Top-3	Top-4	Top-5	Top-1	Top-2	Top-3	Top-4	Top-5	Top-1
Template-based											
Retrosim (Coley et al., 2017)	35.1	40.5	42.9	44.0	44.6	35.0	40.5	43.0	44.1	44.6	31.5
Neuralsym (Segler & Waller, 2017)	**41.7**	**49.2**	52.1	53.6	54.4	**42.0**	**49.3**	52.0	53.6	54.3	**39.2**
GLN (Dai et al., 2019)	39.6	48.9	**52.7**	**54.6**	**55.7**	39.5	48.7	**52.6**	**54.5**	**55.6**	38.0
Template-free											
G2Gs (Shi et al., 2020)	5.4	8.3	9.9	10.9	11.7	4.2	6.5	7.6	8.3	8.9	3.8
GraphRetro (Somnath et al., 2021)	15.3	19.5	21.0	21.9	22.4	15.3	19.5	21.0	21.9	22.2	14.4
Megan (Sacha et al., 2021)	18.8	29.7	37.2	42.6	45.9	19.5	28.0	33.2	36.4	38.5	32.9
Transformer (Karpov et al., 2019)	31.3	40.4	44.7	47.2	48.9	31.2	40.5	45.1	47.3	48.7	26.7
FusionRetro	**37.5**	**45.0**	**48.2**	**50.0**	**50.9**	**37.5**	**45.0**	**48.3**	**50.2**	**51.2**	**33.8**

1. FusionRetro outperforms other template-free baseline methods.

2. As the depth of the ground truth synthetic routes increases, the performance gap between the Transformer and FusionRetro generally widens.

Evaluation: Results



2. As the depth of the ground truth synthetic routes increases, the performance gap between the Transformer and FusionRetro generally widens.

Conclusion

1. FusionRetro is the first method in this field that takes context information into account, greatly boosting the performance for realistic multi-step planning.

2. We further introduce new benchmarks for better evaluation of retrosynthesis models, especially for practical multi-step planning settings.

3. We hope our approach can shed light on the research of data-driven retrosynthetic planning, and inspire more studies toward the practical multi-step scenario.

4. Our approach can be viewed as in-context learning and can inspire more works to further explore in-context learning techniques in large language models for scientific problems.

Arxiv&Code

ArXiv: https://arxiv.org/pdf/2209.15315.pdf

Code: https://github.com/SongtaoLiu0823/FusionRetro

Thanks!!!