RetCL: A Selection-based Approach for Retrosynthesis via Contrastive Learning

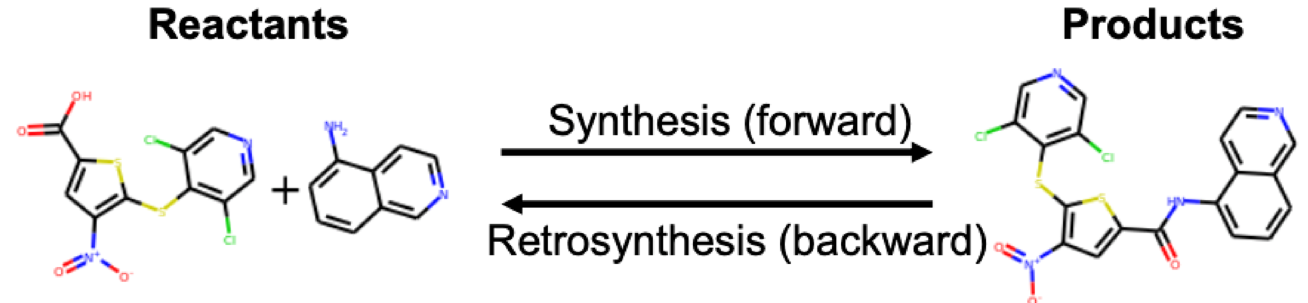
Hankook Lee[1*], Sungsoo Ahn[2], Seung-Woo Seo[3*], You Young Song[4*], Eunho Yang[15], Sung Ju Hwang[15], Jinwoo Shin[1]

[1]KAIST, [2]Mohamed bin Zaeyed University of Artificial Intelligence, [3]Standigm, [4]Samsung Electronics, [5]AITRICS, *this work was partially done while the first author visited Samsung Advanced Institute of Technology

TL; DR. We propose a framework to consider the **commercial availability of reactants** for retrosynthesis

Background: Retrosynthesis

Retrosynthesis aims at finding a synthetic route starting from commercially available reactants to synthesize a target product



Template-based approaches first enumerate known reaction templates and then apply a well-matched template into the target product

- Pros: They can provide chemically interpretable predictions
- Cons: They limit the search space to known reaction templates

Template-free approaches generate the reactants from scratch using deep generative models

- Pros: They can avoid relying on the reaction templates
- Cons: Their predictions could be either unstable or unavailable

Motivation: Retrosynthesis methods are required to consider the availability of reactants and generalize to unseen templates

Contribution

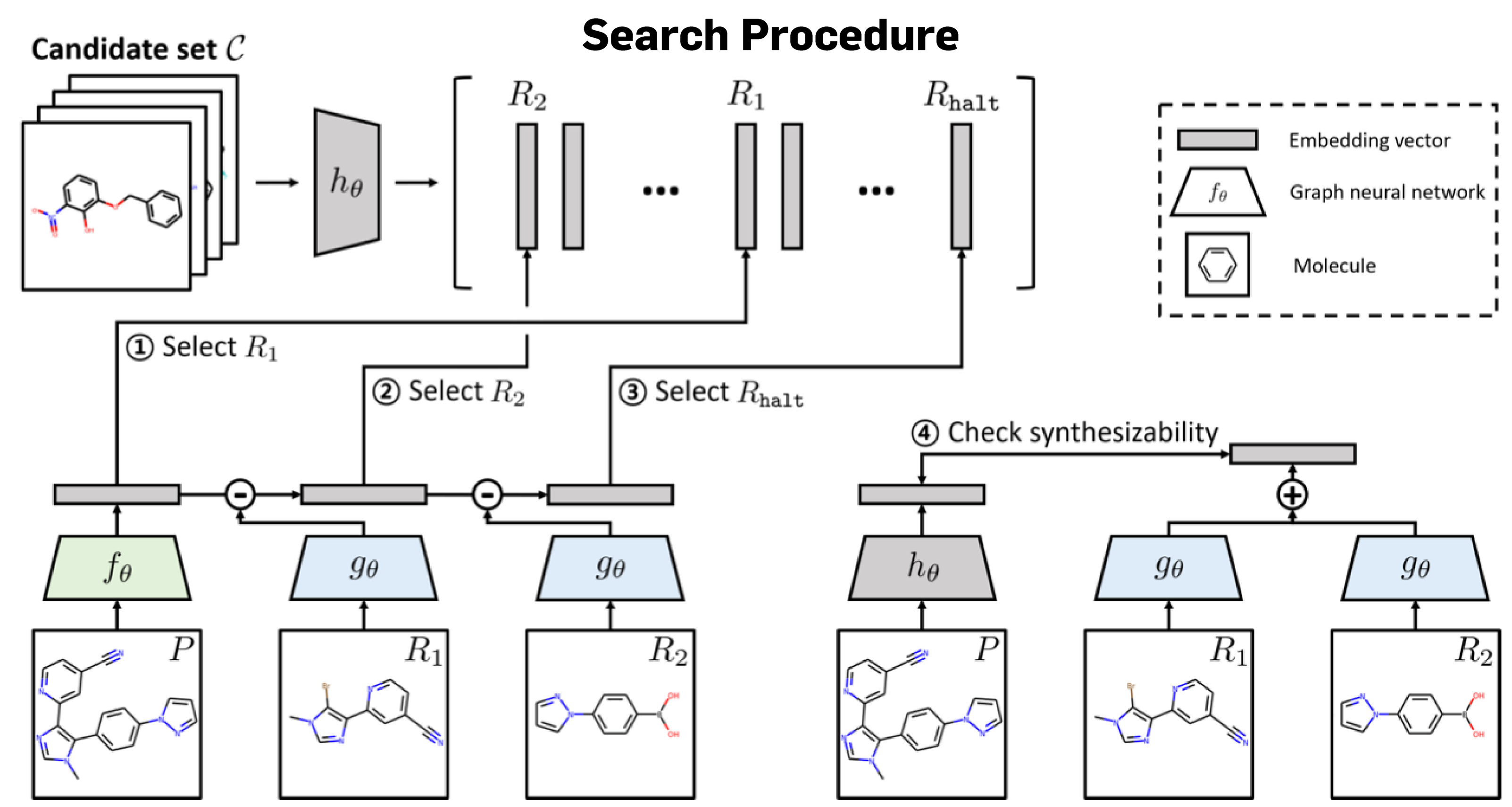
We propose a new **selection-based** approach which allows considering the commercial availability of reactants

- We reformulate the task of retrosynthesis as a problem where reactants are selected from a candidate set $\mathcal{C}$ of available molecules
- We design two effective selection scores in synthetic and retrosynthetic manners using graph neural networks
- We propose a novel contrastive learning scheme with hard negative mining to overcome a scalability issue while handling a large-scale candidate set
- We demonstrate the effectiveness of our framework in various single- and multi-step retrosynthesis experiments based on the USPTO database

Method: Selection-based Framework (RetCL)

Notation. $\mathcal{R} \to P$ is a chemical reaction where $\mathcal{R} = \{R_1, \dots, R_n\}$ is a set of reactants and P is a product. $\mathcal{C}$ is a candidate set of commercially-available molecules.

Problem: Find $\mathcal{R} \subset \mathcal{C}$ which can be synthesized to the target product P



①②③ Given P, choose top-T likely reactant-sets $\mathcal{R}_1, \dots, \mathcal{R}_T$ using beam search based on the sequential selection score $\psi(R_i | P, \{R_1, \dots, R_{i-1}\})$

④ For each $\mathcal{R}_i$, evaluate the synthesizability of $\mathcal{R}_i$ based on $\phi(P|\mathcal{R}_i)$

⑤ Decide the rankings of $\mathcal{R}_1, \dots, \mathcal{R}_T$ based on the following overall score:

$$\texttt{score}(P, \mathcal{R}) = \frac{1}{n+2}\left(\max_{\pi\in\Pi}\sum_{i=1}^{n+1}\psi(R_{\pi(i)}|P, \{R_{\pi(1)}, \dots, R_{\pi(i-1)}\}) + \phi(P|\mathcal{R})\right),$$

Score design. We use the cosine similarity using GNNs f_θ, g_θ, h_θ:

$$\psi(R|P, \mathcal{R}_{\text{given}}) = \texttt{CosSim}\left(f_\theta(P) - \sum\nolimits_{S\in\mathcal{R}_{\text{given}}} g_\theta(S),\ h_\theta(R)\right),$$

$$\phi(P|\mathcal{R}) = \texttt{CosSim}\left(\sum\nolimits_{R\in\mathcal{R}} g_\theta(R),\ h_\theta(P)\right),$$

How to learn the score functions ψ and ϕ?

- We use $\psi(R_i|P, \mathcal{R}_{<i})$ and $\phi(P|\mathcal{R})$ as classification scores and learn the classification task of selecting a molecule R_i or P from $\mathcal{C}$
- For efficient learning, we replace $\mathcal{C}$ by the set $\mathcal{C}_\mathcal{B}$ of all molecules in each mini-batch $\mathcal{B}$
- For effective learning, we add hard-negatives in $\mathcal{C}$ into $\mathcal{C}_\mathcal{B}$

Experiment

- RetCL significantly outperforms a previous selection-based approach
- RetCL shows the superiority even if incorporating knowledge of candidates (i.e., $\mathcal{C}$) into baselines, especially, generalizability under the limited template coverage
- RetCL improves multi-step retrosynthesis performance (i.e., length and cost of discovered synthetic routes) with an existing template-free method

Single-step Retrosynthesis in USPTO-50k

Category	Method	Top-1	Top-3	Top-5	Top-10	Top-20	Top-50
Reaction type is unknown							
Template-free	Transformer (Karpov et al., 2019)	37.9	57.3	62.7	-	-	-
	SCROP (Zheng et al., 2019)	43.7	60.0	65.2	68.7	-	-
	Transformer (Chen et al., 2019)	44.8	62.6	67.7	71.1	-	-
	G2Gs (Shi et al., 2020)	**48.9**	**67.6**	**72.5**	**75.5**	-	-
Template-based	retrosim (Coley et al., 2017b)	37.3	54.7	63.3	74.1	82.0	85.3
	neuralsym (Segler & Waller, 2017)	44.4	65.3	72.4	78.9	82.2	83.1
	GLN (Dai et al., 2019)	**52.5**	**69.0**	**75.6**	**83.7**	**89.0**	**92.4**
Selection-based	Bayesian-Retro (Guo et al., 2020)	47.5	67.2	77.0	80.3	-	-
	RETCL (Ours)	**71.3**	**86.4**	**92.0**	**94.1**	**95.0**	**96.4**

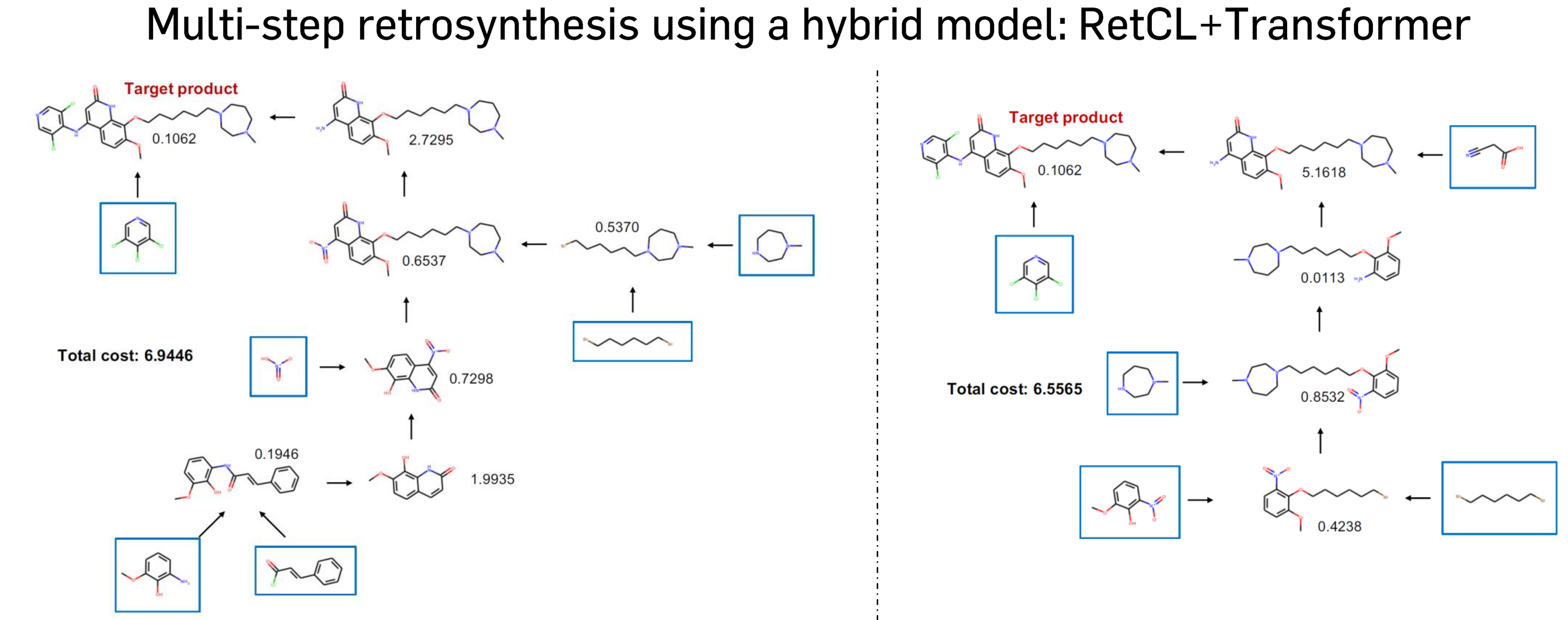
Incorporating knowledge of candidates into baselines

Category	Method	Top-1	Top-5	Top-10	Top-50	Top-100	Top-200
Reaction type is unknown							
Template-free	Transformer (Chen et al., 2019)	59.6	74.3	77.0	79.4	79.5	79.6
	RETCL (Ours)	71.3	**92.0**	**94.1**	**96.4**	**96.7**	**97.1**
Template-based	GLN (Dai et al., 2019)	**77.3**	90.0	92.5	93.3	93.3	93.3

Evaluation of generalizability by training without reaction types from 6 to 10

		Reaction type									
Method	Average	1	2	3	4	5	6	7	8	9	10
GLN (Dai et al., 2019)	39.7	84.3	92.2	70.7	59.3	89.7	0.0	0.0	0.0	0.5	0.0
RETCL (Ours)	**55.6**	**93.9**	**97.6**	**86.4**	**67.0**	**95.6**	**59.1**	**11.9**	**18.3**	**26.1**	0.0

Multi-step retrosynthesis using a hybrid model: RetCL+Transformer



Transformer RetCL+Transformer