Tencent 腾讯 | CSIG 云与智慧产业事业群

Retroformer:

Pushing the Limits of End-to-end Retrosynthesis Transformer

Yue Wan, Chang-Yu Hsieh, Benben Liao, Shengyu Zhang

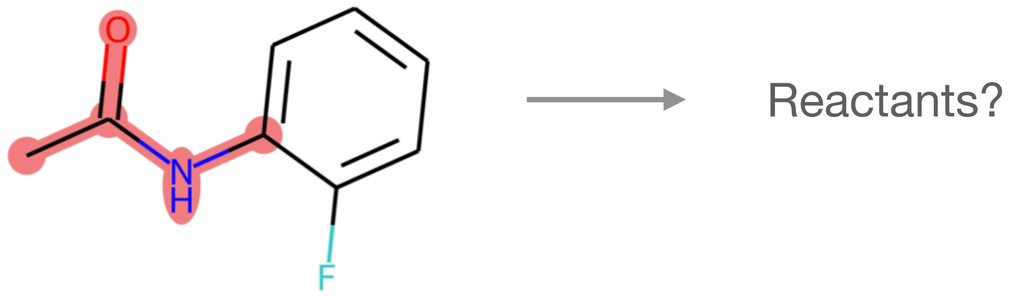
06/22/2022

Introduction

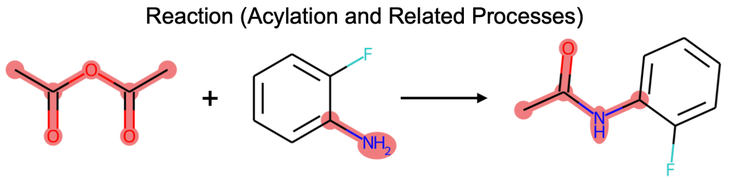
What is retrosynthesis? Why is it important?

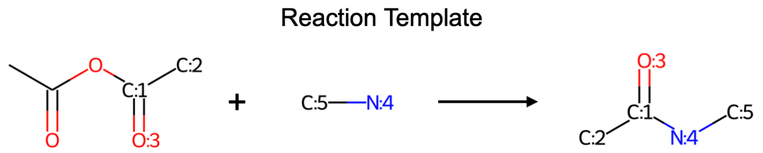
- Major building blocks in organic synthesis
- Aims to discover valid and efficient synthetic routes for a target molecule (e.g., drug)
- Computer-aided synthesis planning to save time and efforts
- Single-step retrosynthesis prediction + Multi-step retrosynthesis planning

  (Predict the reactants given the product) (Recursively expand the reaction tree with search algorithm)



Existing works



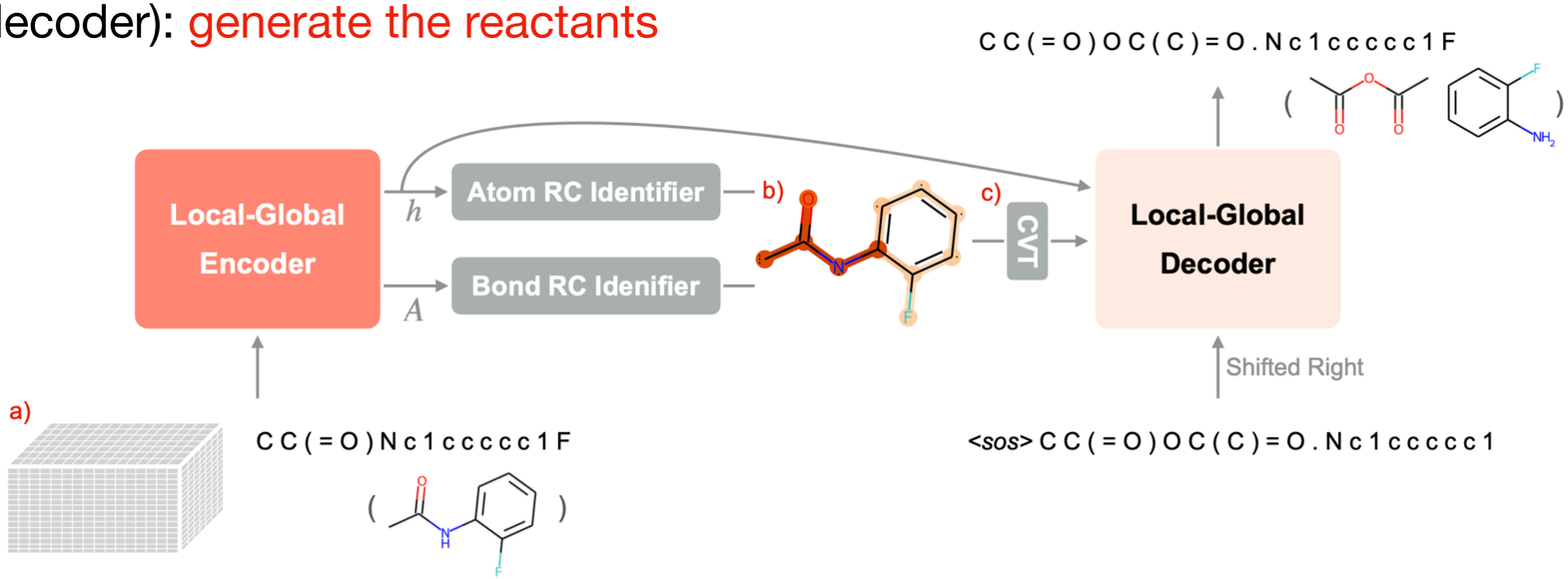


	Template-based	Template-free	Semi-template-based
Framework	product MLP [2], GNN [3], … template database top-*k* templates RDKit [1] top-*k* reactants Retrieval	product (Encoder-decoder) Transformer [4-7] Graph-to-sequence [8] top-*k* reactants Generative	product Trans. [9], GNN [10-12] reaction centers RDKit top-*n* synthons Trans. [9,10], GNN [11,12] top-*k* reactants Retrieval / Generative
Pros	- High validity and accuracy	- Explore larger chemical space (novelty)	- Follow chemists' intuition - Explore relatively larger chemical space
Cons	- Limited chemical space	- Less satisfactory accuracy and validity - Lack interpretability	- Dependency on RDKit for molecule editing - Cannot do end-to-end learning - Complexity can be high

Retroformer

- We tailor a novel Transformer model that mimics the semi-template-based workflow while walking around its disadvantages.

- Local-global encoder (Graph Transformer): jointly encodes the SMILES and topology of the molecule
  - Inputs: topological structure with molecule SMILES S
  - Outputs: updated token representation h and updated edge representation A
- Reaction center detection: predicts the reaction center
  - Inputs: h and A
  - Outputs: reactive probability of each atom and bond, then convert to reactive region S_{rc}
- Local-global decoder (conditional decoder): generate the reactants
  - Inputs: h and reaction center S_{rc}
  - Outputs: reactants SMILES



Retroformer

- Local-global attention head:
  - Encoder: local topological structure of atoms and bonds

$$x_i^{l+1}{}_{global} = \sum_{s_j \in S} \sigma(\frac{q_i {k_j}^T}{\sqrt{d}}) v_j$$

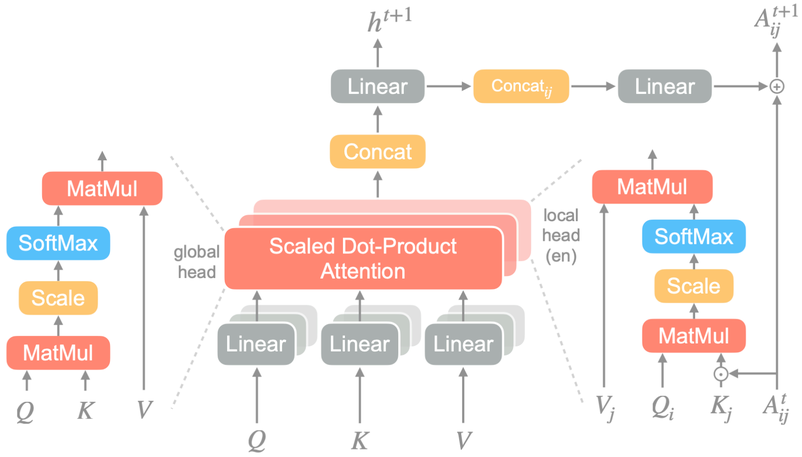
$$[q_i, k_j, v_j] = [h_i^l W^Q, h_j^l W^K, h_j^l W^V]$$

Global (vanilla) self-attention head

$$x_i^{l+1}{}_{local} = \sum_{j \in N(i)} \sigma(\frac{q_i (A_{ij}^l \odot k_j)^T}{\sqrt{d}}) v_j$$

$$[q_i, k_j, v_j] = [h_i^l W^Q, h_j^l W^K, h_j^l W^V]$$

Local self-attention head



$$h^{l+1} = \text{Linear}([x^{l+1}{}_{global}; x^{l+1}{}_{local}])$$

$$A_{ij}^{l+1} = A_{ij}^l + \text{FFN}([h_i^{l+1}; h_j^{l+1}])$$

Edge update

  - Decoder: local reactive region

$$y_i^{l+1}{}_{global} = \sum_{s_j \in S} \sigma(\frac{q_i {k_j}^T}{\sqrt{d}}) v_j$$

$$[q_i, k_j, v_j] = [g_i^l W^Q, g_j^l W^K, g_j^l W^V]$$

Global (vanilla) cross-attention head

$$y_i^{l+1}{}_{local} = \sum_{s_j \in S_{rc}} \sigma(\frac{q_i {k_j}^T}{\sqrt{d}}) v_j$$

$$[q_i, k_j, v_j] = [g_i^l W^Q, h_j W^K, h_j W^V]$$

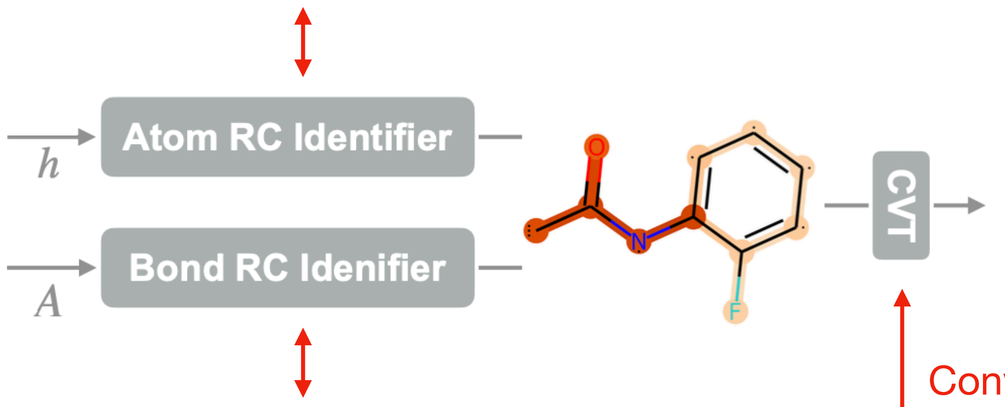
Local cross-attention head

Yellow highlight: local-global difference

Retroformer

- Reaction center detection
  - Probability of atoms and bonds belonging to the reaction center

$$P_{rc}(s_i) = \sigma\left(\text{FFN}_{\text{atom}}(h_i)\right), s_i \in V_m$$



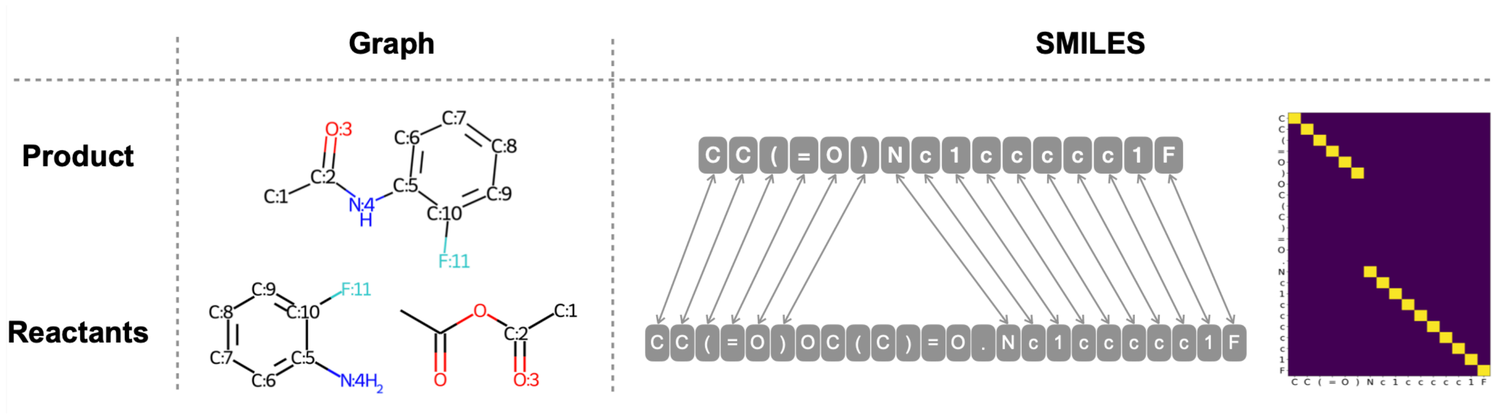
$$P_{rc}(e_{ij}) = \sigma\left(\text{FFN}_{\text{bond}}(A_{ij})\right), e_{ij} \in E_m$$

Convert $P_{rc}(\,.\,)$ into S_{rc}

  - Approach #1: *naively* using a cutoff (e.g., p=0.5)
  - Approach #2: subgraph *search* + rank by total reactive scores (sum of log prob)

  - End-to-end training: generative feedback from the decoder will backpropagate to reaction center detection

Retroformer



- SMILES alignment:
  - Atom alignment => token alignment
  - Additional task: learn guided attention to the context attention module in the decoder

- Data augmentation [6,7,9]:
  - Same molecule can be represented by multiple different SMILES
  - In addition to the canonical form, add permuted SMILES into the training database
  - Our implementation:
    - In each iteration, randomly permute the canonical SMILES by 50% chance

Experiments

Dataset, baselines, & metrics

- Dataset: USPTO50K [13], atom-mapped reactions with labeled reaction type
- Baselines: 2 (template-based) + 6 (template-free) + 5 (semi-template-based) = 13
  - Retroformer$_{base}$: model without data augmentation
  - Retroformer$_{aug}$: model with data augmentation
  - Retroformer$_{aug}$+: model with data augmentation and reaction center subgraph *search*
- Metrics:
  - Top-*k* accuracy
  - Top-*k* SMILES (molecule) validity
  - Top-*k* round-trip accuracy [14] (reaction validity): the percentage of predicted reactants that can lead back to the original product; a proxy for reaction validity

Experiments

Top-k accuracy

- In both settings (reaction class known/unknown), Retroformer surpasses all the template-free baselines.
- Also competitive to the semi-template-based methods.
- Unfortunately, still cannot beat the SOTA template-based methods.

Model	Top-k accuracy (%)							
	Reaction class known				Reaction class unknown			
	1	3	5	10	1	3	5	10
Template-Based								
GLN (Dai et al., 2019)	**64.2**	79.1	85.2	90.0	52.5	69.0	75.6	83.7
LocalRetro (Chen & Jung, 2021)	63.9	**86.8**	**92.4**	**96.3**	**53.4**	**77.5**	**85.9**	**92.4**
Template-Free								
Transformer	57.1	71.5	75.0	77.7	42.4	58.6	63.8	67.7
SCROP (Zheng et al., 2020)	59.0	74.8	78.1	81.1	43.7	60.0	65.2	68.7
Tied Transformer (Kim et al., 2021)	-	-	-	-	47.1	67.1	73.1	76.3
Aug. Transformer* (Tetko et al., 2020)	-	-	-	-	48.3	-	73.4	77.4
GTA* (Seo et al., 2021)	-	-	-	-	51.1	67.6	74.8	81.6
Graph2SMILES (Tu & Coley, 2021)	-	-	-	-	52.9	66.5	70.0	72.9
$\text{Retroformer}_{\text{base}}$ (Ours)	61.5	78.3	82.0	84.9	47.9	62.9	66.6	70.7
$\text{Retroformer}_{\text{aug}}$* (Ours)	**64.0**	81.8	85.4	88.3	52.9	68.2	72.5	76.4
$\text{Retroformer}_{\text{aug}}$+* (Ours)	**64.0**	**82.5**	**86.7**	**90.2**	**53.2**	**71.1**	**76.6**	**82.1**
Semi-Template-Based								
RetroXpert* (Yan et al., 2020)	62.1	75.8	78.5	80.9	50.4	61.1	62.3	63.4
G2G (Shi et al., 2020)	61.0	81.3	86.0	88.7	48.9	67.6	72.5	75.5
GraphRetro (Somnath et al., 2020)	63.9	81.5	85.2	88.1	**53.7**	68.3	72.2	75.5
RetroPrime* (Wang et al., 2021)	**64.8**	81.6	85.0	86.9	51.4	**70.8**	74.0	76.1
MEGAN (Sacha et al., 2021)	60.7	**82.0**	**87.5**	**91.6**	48.1	70.7	**78.4**	**86.1**

Experiments

Smiles validity & round-trip accuracy

Table 2: Top-k SMILES validity for retrosynthesis prediction on USPTO-50K with reaction class unknown.

Model	Top-k validity (%)			
	1	3	5	10
Transformer	97.2	87.9	82.4	73.1
Graph2SMILES	**99.4**	90.9	84.9	74.9
RetroPrime	98.9	98.2	97.1	92.5
Retroformer$_{aug}$	99.3	98.5	97.2	92.6
Retroformer$_{aug}$+	99.2	**98.5**	**97.4**	**96.7**

Table 3: Top-k round-trip accuracy for retrosynthesis prediction on USPTO-50K with reaction class unknown.

Model	Top-k round-trip acc. (%)			
	1	3	5	10
Transformer	71.9	54.7	46.2	35.6
Graph2SMILES	76.7	56.0	46.4	34.9
RetroPrime	**79.6**	59.6	50.3	40.4
Retroformer$_{aug}$	78.6	71.8	**67.1**	**57.6**
Retroformer$_{aug}$+	78.9	**72.0**	**67.1**	57.2

- Compared with the template-free SMILES generative baselines
- Better molecule validity and “reaction validity”

Experiments

Ablations

- All components are necessary for reaching the best performance

Table 5: Effects of local-global encoder, reaction center search, and data augmentation on reaction center detection performance. Ablation (c) corresponds to $\text{Retroformer}_{\text{base}}$.

Settings	**Top-n accuracy (%)**				
	1		1	2	3
Ablation (d)	55.4	$+search$	71.6	84.1	89.9
Ablation (c)	63.0	$+search$	75.8	88.2	91.3
$\text{Retroformer}_{\text{aug}}$	67.5	$+search$	79.3	90.0	92.9

Table 4: Effects of different components on retrosynthesis performance with reaction class unknown.

Settings	**Modules**					**Top-k accuracy (%)**			
	$\text{Guided}_{\text{last}}$	$\text{Guided}_{\text{all}}$	Local-global Encoder	Local-global Decoder	Reaction Center Search	1	3	5	10
(a)			√	√		45.5	60.7	65.4	69.9
(b)		√	√	√		47.0	63.1	66.9	71.1
(c)	√		√	√		47.9	62.9	66.6	70.7
(d)	√			√		44.1	60.1	64.7	70.2
(e)	√		√			46.7	63.7	68.4	73.9
(f)	√		√	√	√	**48.4**	**66.8**	**73.2**	**78.8**

Conclusions

- We propose Retroformer, a novel Transformer-based architecture for retrosynthesis
- Reaches the state-of-the-art performance for template-free retrosynthesis, with better accuracy, validity, interpretability.
- In the future, we plan to:
  - Study the diversity issue of SMILES generative models.
  - Study the multi-step retrosynthesis planning problem using Retroformer as single-step backbone.

Code: https://github.com/yuewan2/Retroformer

References

[1] Rdkit: Open-source cheminformatics. URL http://www. rdkit.org.

[2] Segler, M. H. S. & Waller, M. Neural-symbolic machine learning for retrosynthesis and reaction prediction. Chemistry 23 25, 5966–5971 (2017).

[3] Dai, H., Li, C., Coley, C., Dai, B., and Song, L. Retrosynthesis prediction with conditional graph logic network. In *Advances in Neural Information Processing Systems*, pp. 8870–8880, 2019.

[4] Lin, K., Xu, Y., Pei, J., and Lai, L. Automatic retrosynthetic route planning using template-free models. *Chem. Sci.*, 11:3355–3364, 2020. doi: 10.1039/C9SC03666K.

[5] Kim, E., Lee, D., Kwon, Y., Park, M. S., and Choi, Y.-S. Valid, plausible, and diverse retrosynthesis using tied two-way transformers with latent variables. *Journal of Chemi- cal Information and Modeling*, 61(1):123–133, 2021. doi: 10.1021/acs.jcim.0c01074.

[6]. Tetko, I. V., Karpov, P., Van Deursen, R., and Godin, G. State-of-the-art augmented nlp transformer models for direct and single-step retrosynthesis. *Nature Communications*, 11 2020. doi: 10.1038/ s41467-020-19266-y.

[7] Seo, S.-W., Song, Y. Y., Yang, J. Y., Bae, S., Lee, H., Shin, J., Hwang, S. J., and Yang, E. Gta: Graph truncated attention for retrosynthesis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(1):531– 539, May 2021.

[8] Tu, Z. and Coley, C. W. Permutation invariant graph-to-sequence model for template-free retrosynthesis and reaction prediction, 2021.

[9] Wang, X., Li, Y., Qiu, J., Chen, G., Liu, H., Liao, B., Hsieh, C.-Y., and Yao, X. Retroprime: A diverse, plausible and transformer-based method for single-step retrosynthesis predictions. *Chemical Engineering Journal*, 420:129845, 2021. ISSN 1385- 8947. doi: https://doi.org/10.1016/j.cej.2021.129845.

[10] Yan, C., Ding, Q., Zhao, P., Zheng, S., YANG, J., Yu, Y., and Huang, J. Retroxpert: Decompose retrosynthesis prediction like a chemist. In *Advances in Neural Information Processing Systems (NeurIPS) 33*, pp. 11248–11258. Curran Associates, Inc., 2020.

[11] Shi, C., Xu, M., Guo, H., Zhang, M., and Tang, J. A graph to graphs framework for retrosynthesis prediction. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pp. 8818–8827, 2020.

[12] Somnath, V. R., Bunne, C., Coley, C. W., Krause, A., and Barzilay, R. Learning graph models for template-free retrosynthesis, 2020.

[13] Schneider, N., Stiefl, N., and Landrum, G. A. What's what: The (nearly) definitive guide to reaction role assignment. *Journal of Chemical Information and Mod- eling*, 56(12):2336–2346, 2016. doi: 10.1021/acs.jcim. 6b00564.

[14] Schwaller, P., Petraglia, R., Zullo, V., Nair, V. H., Haeusel-mann, R. A., Pisoni, R., Bekas, C., Iuliano, A., and Laino, T. Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chem. Sci.*, 11:3316–3325, 2020. doi: 10. 1039/C9SC05704H.