GTA: Graph Truncated Attention for Retrosynthesis

Seung-Woo Seo, You Young Song, June Yong Yang, Seohui Bae,
Hankook Lee, Jinwoo Shin, Sung Ju Hwang, Eunho Yang

2021.7.30

Outlines

Retrosynthesis

- Task of predicting the set of reactant molecules that are synthesized to a given product molecule by finding the inverse reaction pathway.
- Important in discovering new materials and in drug discovery.
- Chemists try to adopt computer-assisted methods to retrosynthetic analysis for fast and efficient reactant candidate searches.

Deep-learning-based approaches for retrosynthesis

- template-based
  - coverage limitation: Reactions not covered by extracted templates are hardly predicted.
- template-free
  - can generalize beyond extracted templates by learning from data on reactions, reactants, and products
  - focus on molecule representations: sequence or graph
  - seq2seq, G2G

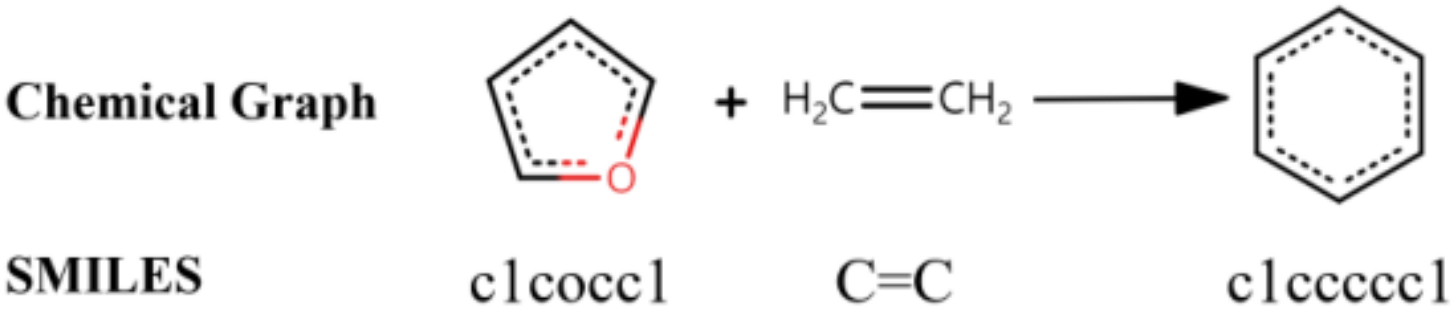
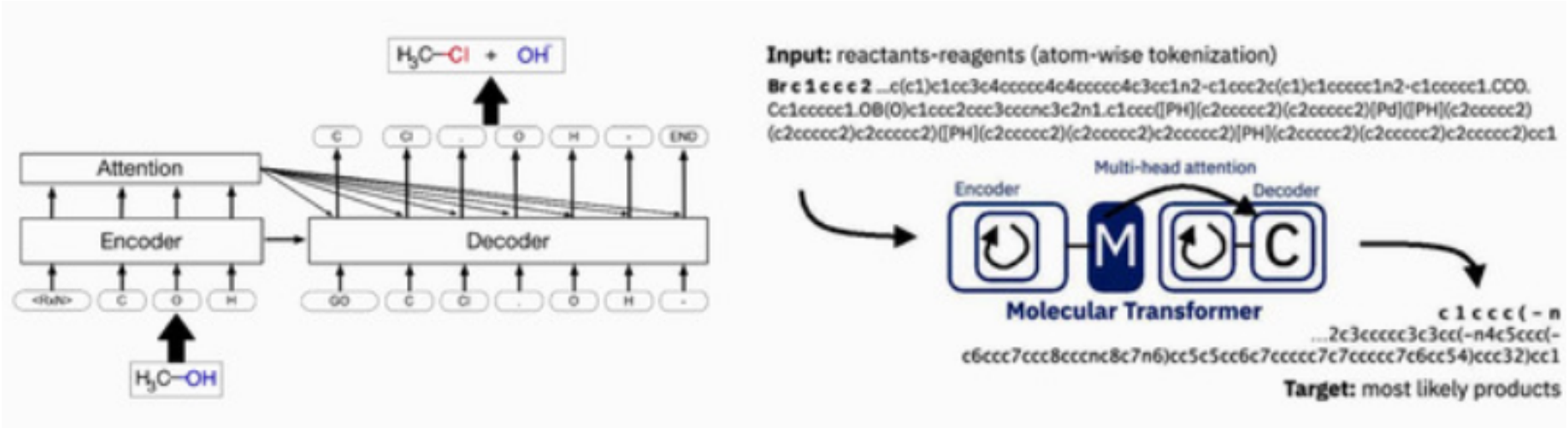
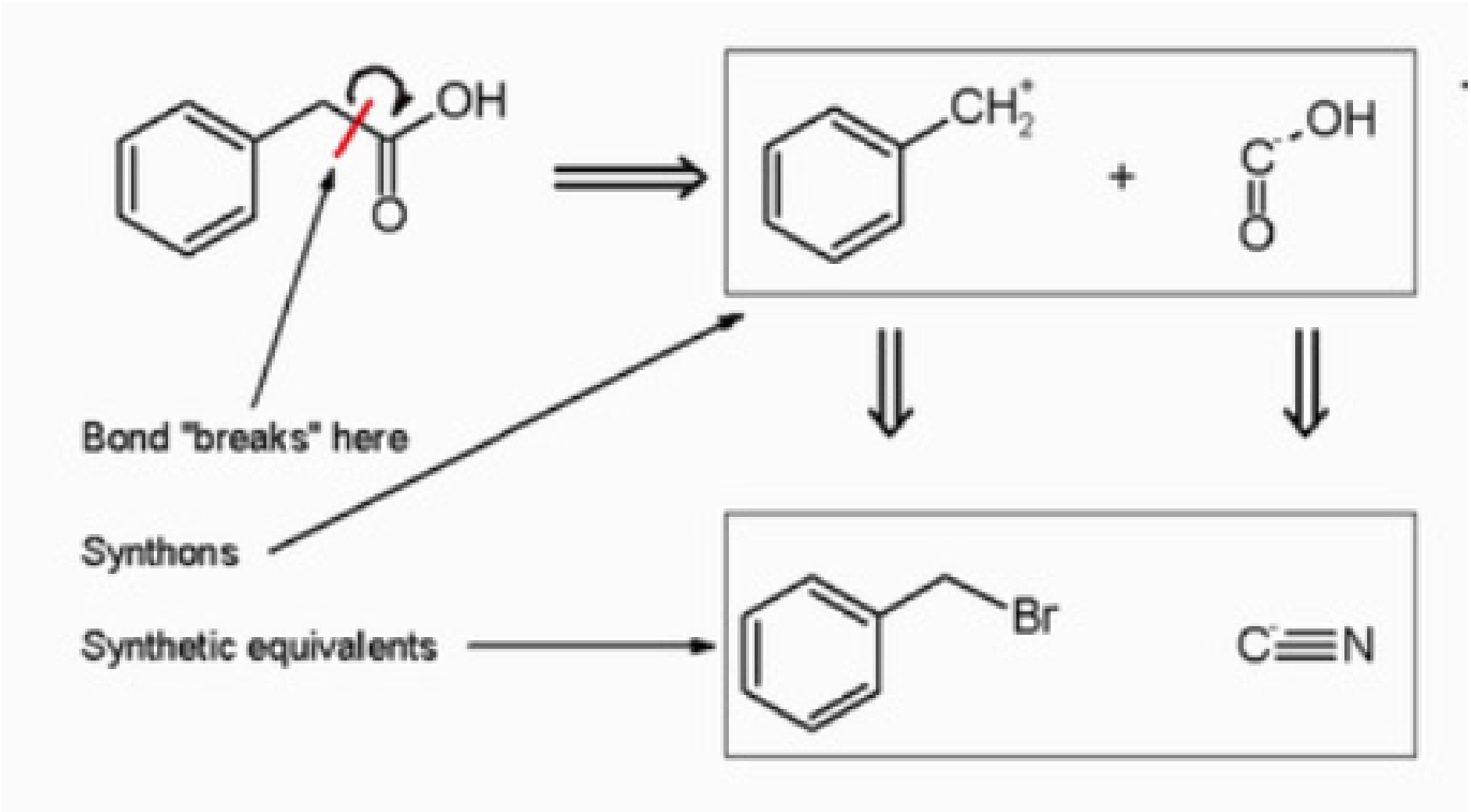


Figure 1: Example of reaction: Synthesis of benzene (right) from furan (left) and ethylene (middle). Each molecule is expressed in SMILES notation.

seq2seq


H3C—Cl + OH⁻
Attention
Encoder
Decoder
H3C—OH
Input: reactants-reagents (atom-wise tokenization)
Multi-head attention
Encoder
Decoder
M
C
Molecular Transformer
Target: most likely products

G2G


OH
CH2
OH
Bond "breaks" here
Synthons
Synthetic equivalents
Br
C≡N

SMILES vs Molecular Graph

- SMILES
  - easy to generate, simple augmentation
  - complex, not unique
- Graph
  - straight forward, unique
  - need atom-mapping

We propose a novel graph-truncated attention method called Graph Truncated Attention (GTA) combining SMILES and graphs' strengths.

Notation

- P, R: product and rectant molecules.
- $\mathcal{G}$(mol), $\mathcal{S}$(mol): corresponding molecular graph and SMILES representation for a molecule $\mathbf{mol} \in \{P, R\}$.
- T_{mol}: number of tokens in $\mathcal{S}$(mol).
- N_{mol}: number of atoms in $\mathcal{S}$(mol).

Transformer and Masked Self-attention

For query $Q \in \mathbb{R}^{\mathbb{T}\ \mathrm{mol} \times d_k}$, key $K \in \mathbb{R}^{T_{mol} \times d_k}$ and value $V \in \mathbb{R}^{T_{mol} \times d_v}$ matrices, each of which is linearly transformed by learnable parameters from the input token, we have

$$S = \frac{QK^T}{\sqrt{d_k}}$$

$$[\mathrm{Masking}(S, M)]_{ij} = \begin{cases} s_{ij} & \text{if } m_{ij} = 1 \\ -\infty & \text{if } m_{ij} = 0 \end{cases}$$

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}(\mathrm{Masking}(S, M))\, V$$

where $S = \left(s_{ij}\right)$ and $M = \left(m_{ij}\right) \in \{0, 1\}^{T_{mol} \times T_{mol}}$ are score and mask matrices.

GTA

Transformer architecture can be reinterpreted as a particular kind of graph neural network (GNN).

- Atoms or tokens as nodes
- Attentions as edges

Since self-attention and cross-attention have different shapes, we devise two different truncation strategy for each of them.

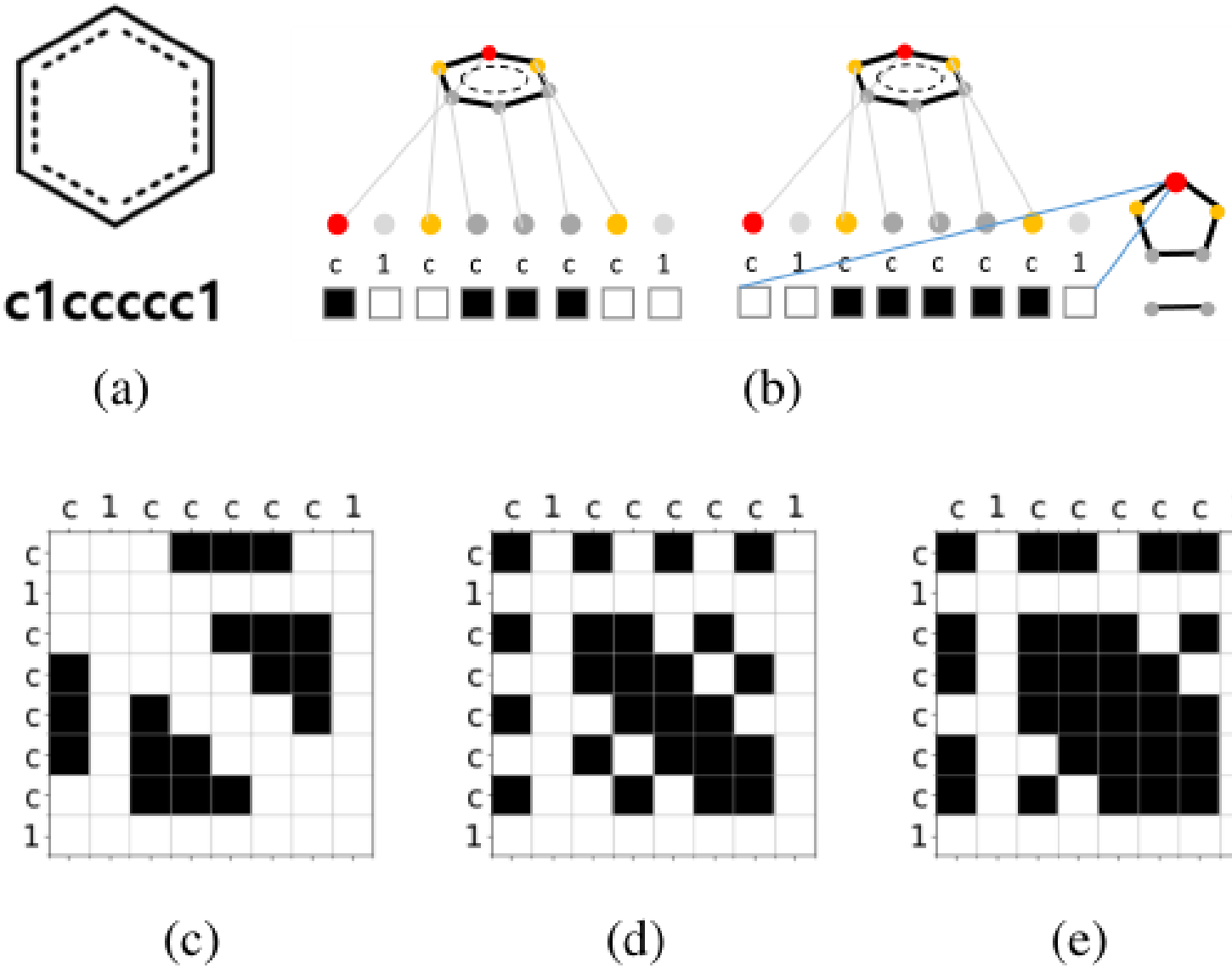
Graph-truncated self-attention (GTA-self)

Given mask matrix $M \in \{0,1\}^{N_{mol} \times N_{mol}}$ and distance matrix $D = \left(d_{ij}\right)$.

- Original transformer: $m_{ij} = \begin{cases} 1 & \text{if } d_{ij} = d \\ 0 & \text{otherwise} \end{cases}$.
- Introduce multi-head attention: $m_{ij} = \begin{cases} 1 & \text{if } d_{ij} = d_h \\ 0 & \text{otherwise} \end{cases}$. where d_h is the target geodesic distance to attend for head h. GTA can learn a richer representation by different heads paying attention to atoms at a different distance.

Graph-truncated self-attention (GTA-self)

Examples of mask M with different choices of d for benzene ring:



Graph-truncated cross-attention(GTA-cross)

Cross-attention reflect the relationship between tokens in product and reactant.
GTA-cross does not require exact atom mapping for all nodes but only leverages the information of certain pairs.
Given the (partial) information of atom mapping between product and reactant molecules, the mask for cross-attention $M = \left(m_{ij}\right) \in \{0,1\}^{T_R \times T_P}$ is constructed as follows:

$$m_{ij} = \begin{cases} 1 & \text{if } R_{i'} \underset{\text{mapped}}{\longrightarrow} P_{j'} \\ 0 & \text{else} \end{cases} \tag{1}$$

where i' and j' are indices of nodes in $\mathcal{G}(R)$ and $\mathcal{G}(P)$ corresponding to i - and j -th tokens in $\mathcal{S}(R)$ and $\mathcal{S}(P)$, $R_{i'}$ and $\tilde{j'}$ denote the nodes in $\mathcal{G}(R)$ and $\mathcal{G}(P)$, respectively.

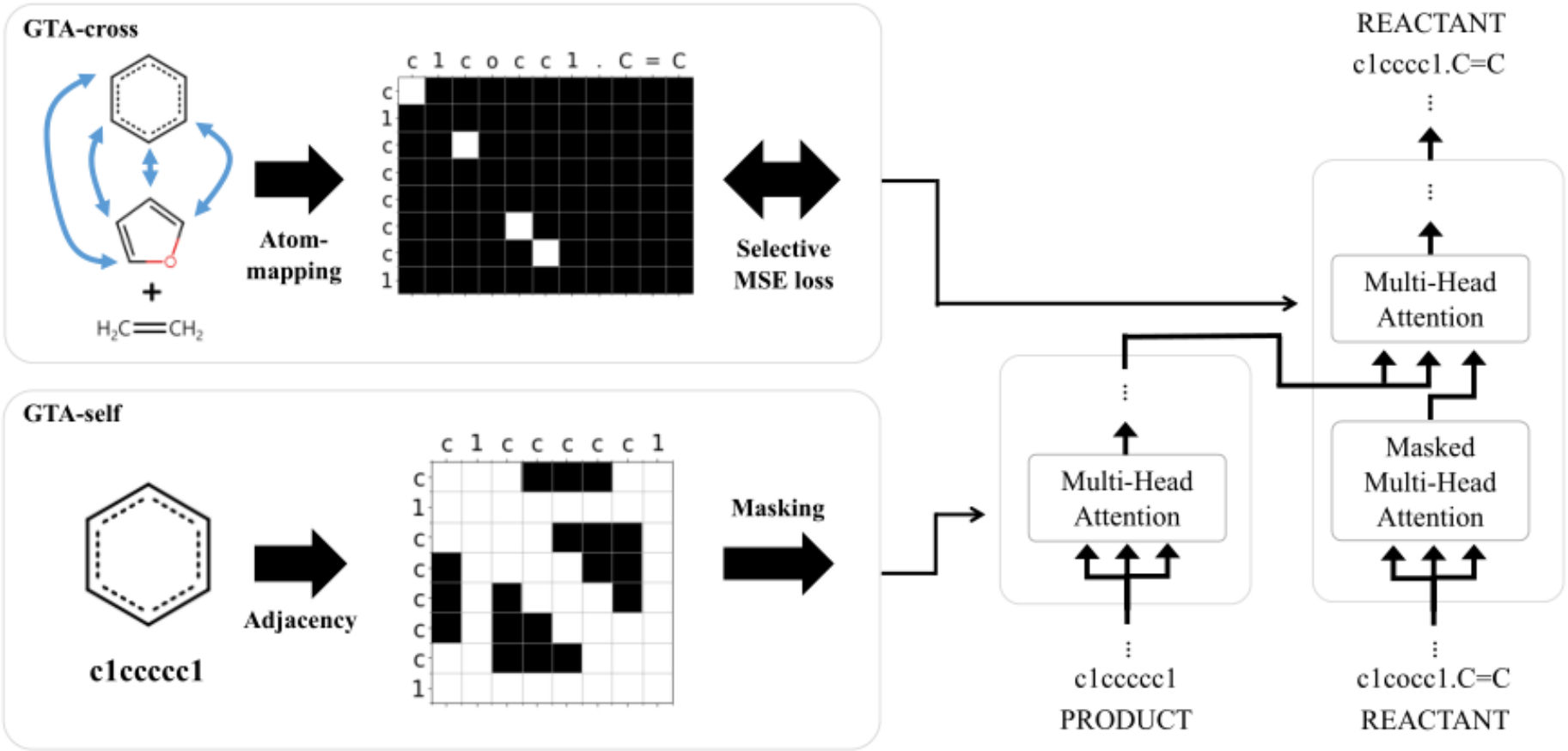
Graph-truncated cross-mapping(GTA-cross)

GTA-cross encourages the attention by selective ℓ_2 loss only with certain information (i.e. where $m_{ij} = 1$) among uncertain and incomplete atom mapping so that the cross attention gradually learns complete atom-mapping

$$\mathcal{L}_{\text{attn}} = \sum \left[(M_{\text{cross}} - A_{\text{cross}})^2 \odot M_{\text{cross}} \right] \tag{2}$$

where M_{cross} is the mask from (3), A_{cross} is a cross-attention matrix and $\odot$ is Hadamard product.

GTA


GTA-cross
Atom-mapping
Selective MSE loss
GTA-self
Adjacency
c1ccccc1
Masking
REACTANT
c1ccccc1.C=C
Multi-Head Attention
Masked Multi-Head Attention
Multi-Head Attention
c1ccccc1
PRODUCT
c1cocc1.C=C
REACTANT

Datasets and Augmentation Strategy

- Datasets: USPTO-full and USTPO-50k
- Augmentation:
  - change the order of reactant molecule(s) as <c1cocc1.C=C> and <C=C.c1cocc1> in SMILES notation
  - change the starting atom of SMILES
- Evaluation metrics: top-k exact match accuracy

Test of Accuracy

Method (Dataset)	Top-1	Top-3	Top-5	Top-10
BiLSTM	37.4	52.4	57.0	61.7
Transformer	42.0	57.0	61.9	65.7
Syntax correction	43.7	60.0	65.2	68.7
Latent model, l=1	44.8	62.6	67.7	71.7
Latent model, l=5	40.5	65.1	72.8	79.4
G2Gs	48.9	67.6	72.5	75.5
ONMT (Plain)	44.7 (±0.29)	63.6 (±0.20)	69.7 (±0.25)	75.6 (±0.04)
ONMT (2P2R_s)	49.0 (±0.30)	65.8 (±0.39)	72.5 (±0.14)	79.3 (±0.14)
GTA (Plain)	47.3 (±0.29)	67.8 (±0.35)	73.8 (±0.20)	80.1 (±0.19)
GTA (2P2R_s)	**51.1** (±0.29)	**67.6** (±0.22)	**74.8** (±0.36)	**81.6** (±0.22)

Table 1: Top-k exact match accuracy (%) of temmplate-free models trained with USPTO-50k dataset. ONMT and GTA accuracies are achieved with our optimized hyperparameters. The standard error with 95% confidence interval is given after ± symbol.

Test of Scalability

GLN: template-based model

Method	USPTO-50k		USPTO-full	
	Top-1	Top-10	Top-1	Top-10
GLN	52.5	83.7	39.3	63.7
GTA	$51.1_{\pm 0.29}$	$81.6_{\pm 0.22}$	$\mathbf{46.0}_{\pm 0.07}$	$\mathbf{70.0}_{\pm 0.19}$

Table 2: Top-k exact match accuracy (%) of template-based GLN and our template-free GTA trained with USPTO-50k and USPTO-full. The standard error with 95% confidence interval is given after $\pm$ symbol.

Conclusions

- GTA utilize graph information in sequence.
- GTA outperforms other results in large scale database and has good scalability.