西南交通大學

Southwest Jiaotong University

Enhanced Template-Free Reaction Prediction with Molecular Graphs and Sequence-based Data Augmentation

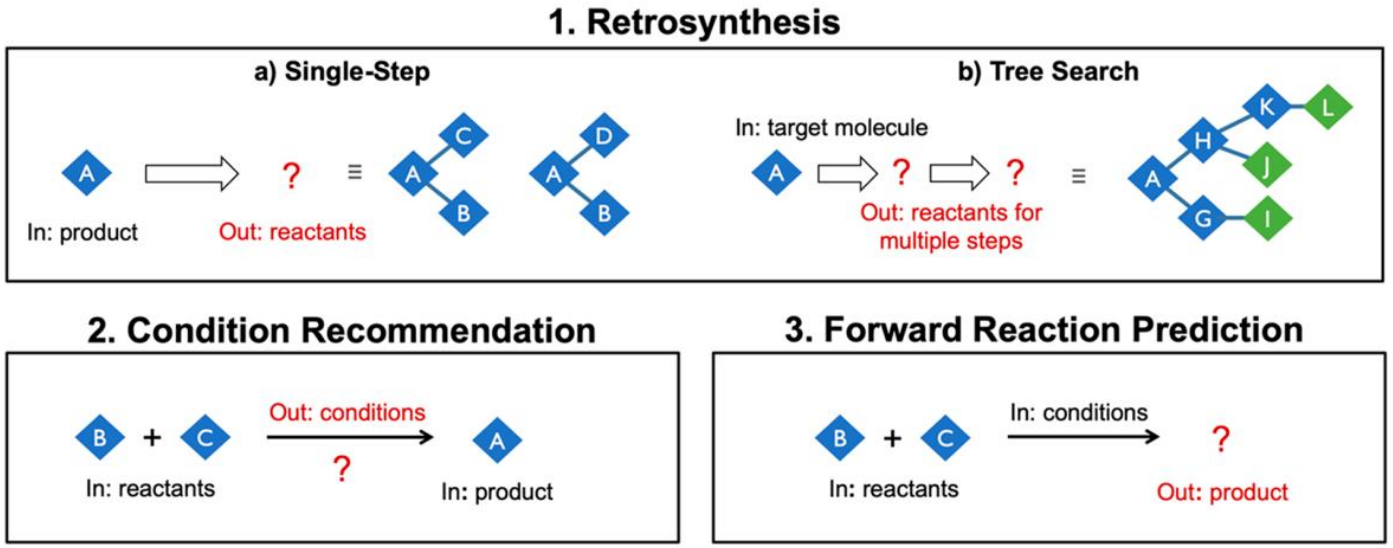
Haozhe Hu[1], Yongquan Jiang*[1], Yan Yang[1], Jim X. Chen[2]

[1]Southwest Jiaotong University
[2]George Mason University

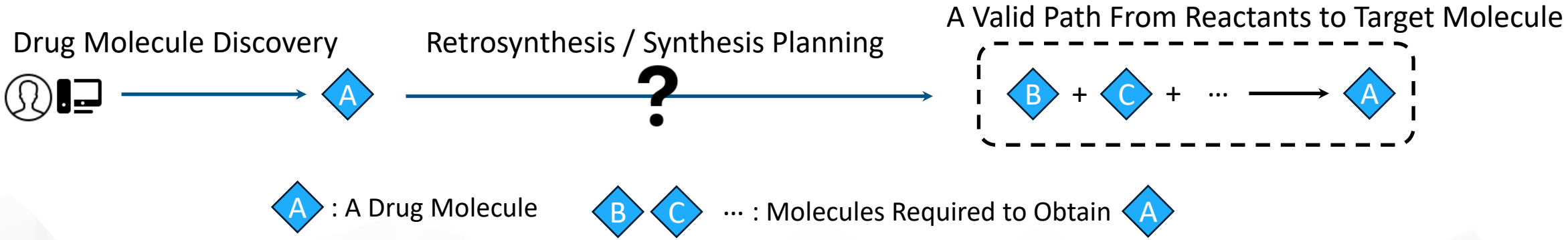
Presenter: Haozhe Hu

2017115054@my.swjtu.edu.cn

Introduction: Retrosynthesis & Forward Synthesis

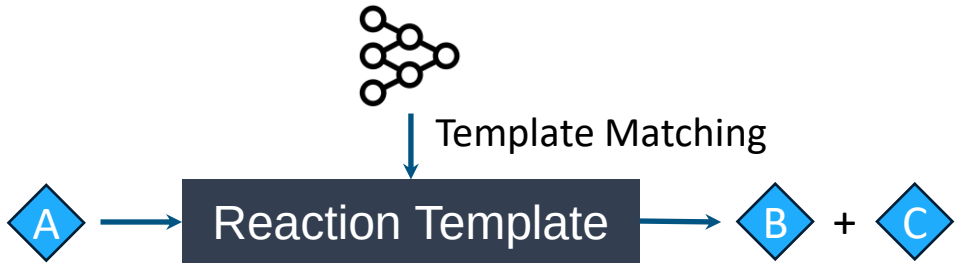


[1]

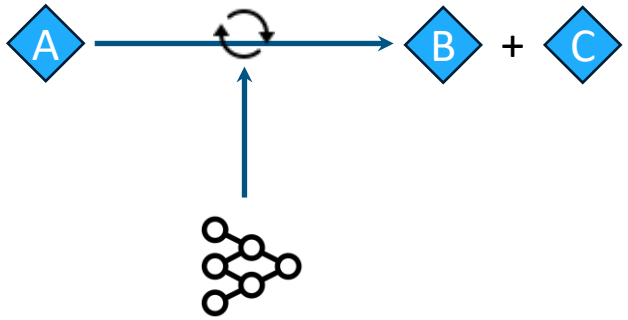


[1] Struble, T. J. et al., Current and future roles of artificial intelligence in medicinal chemistry synthesis, J. Med. Chem., 2020

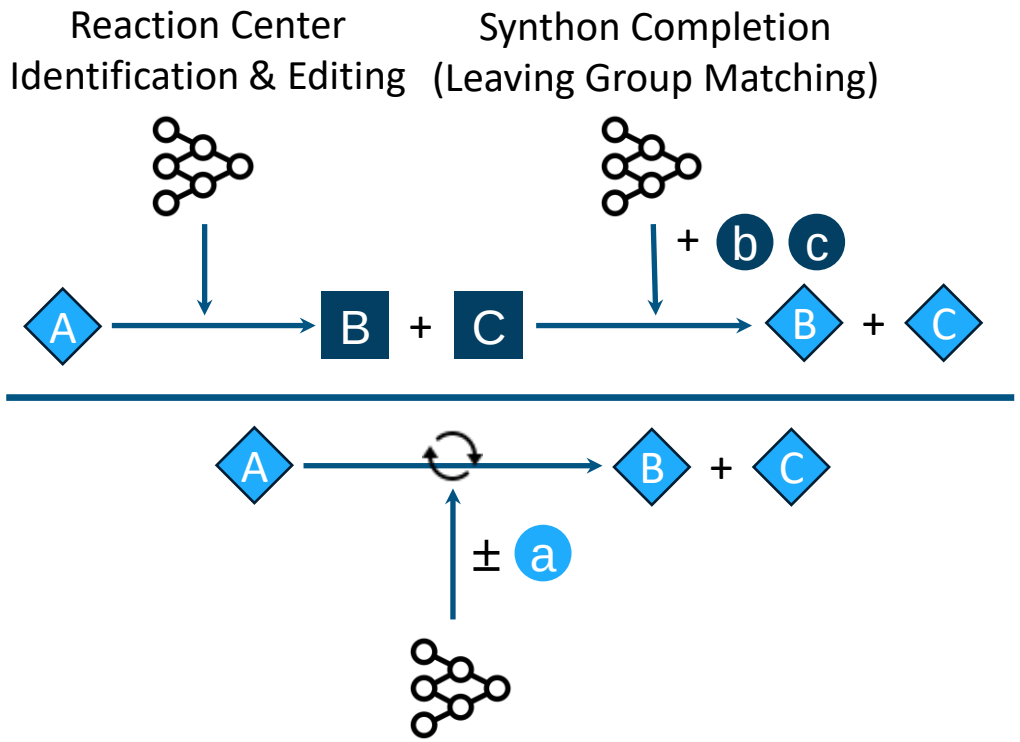
Introduction: Deep Learning in Retrosynthesis



1) Template-Based



2) Template-Free



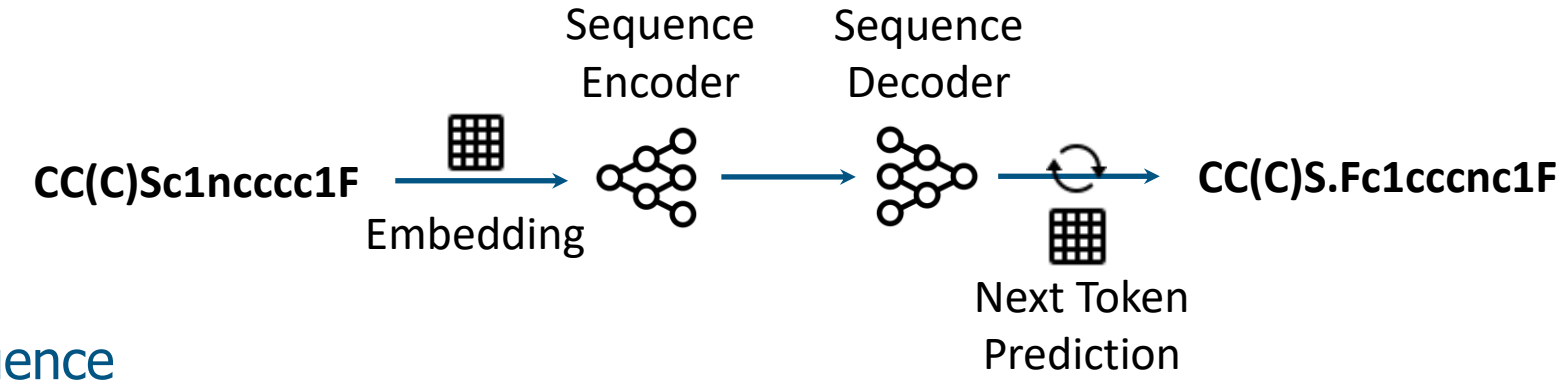
3) Semi-Template

: Neural Networks B C : Synthon b c : Leaving Group a : Atoms / Bonds

Related Work 1: Two Types of Template-Free Methods

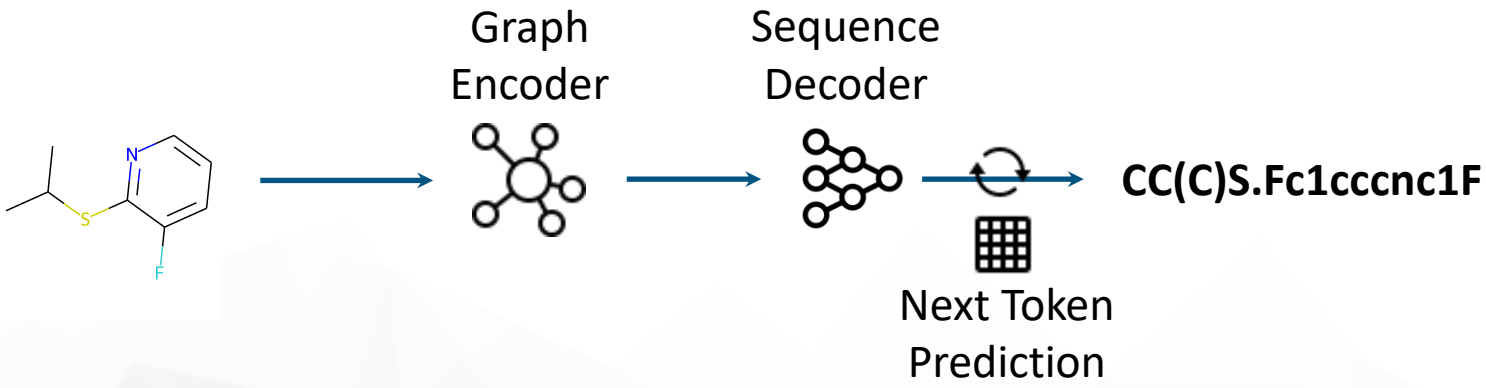
1) Sequence-to-Sequence

- SMILES (or SMILES with Graph Features, e.g., Bonds) as Input, SMILES as Output
- Enable to Use SMILES Data Augmentation (i.e., Sequence-based Data Augmentation)

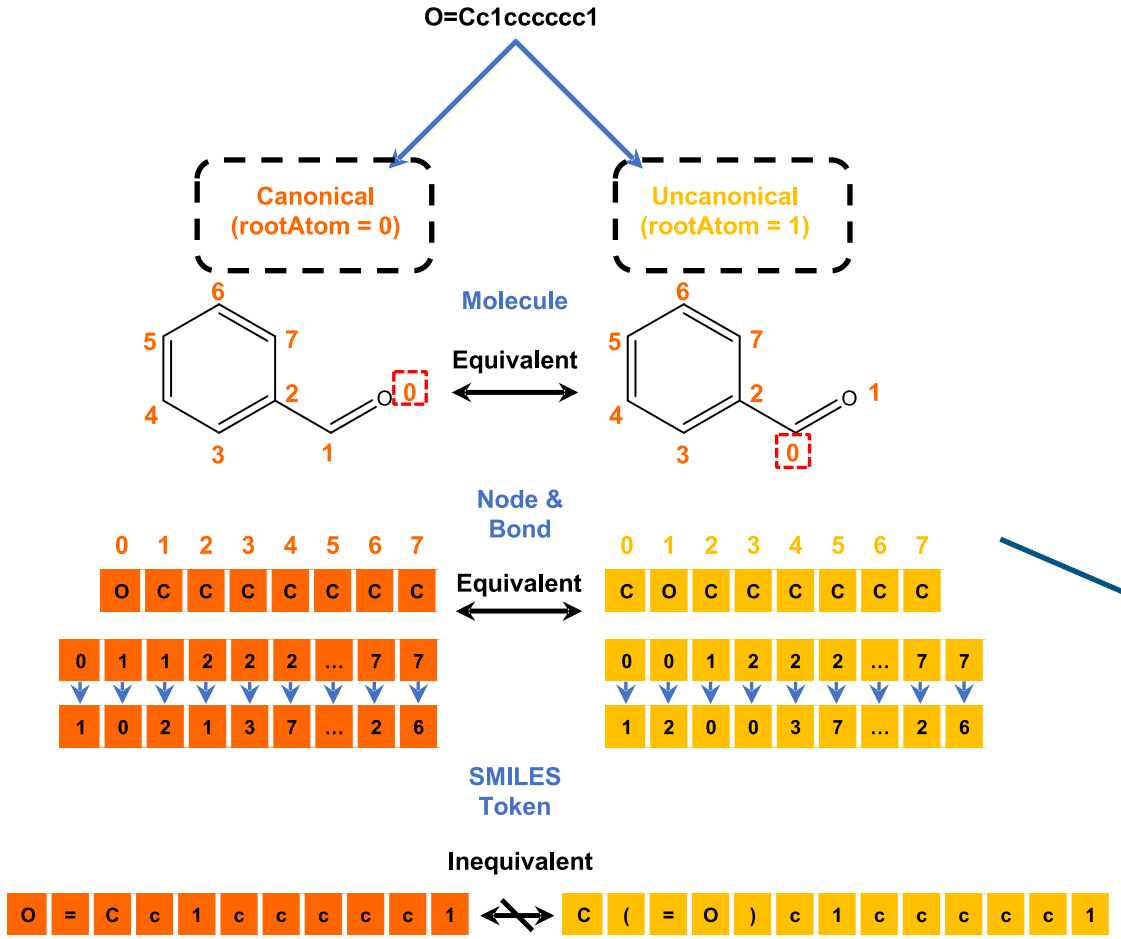


2) Graph-to-Sequence

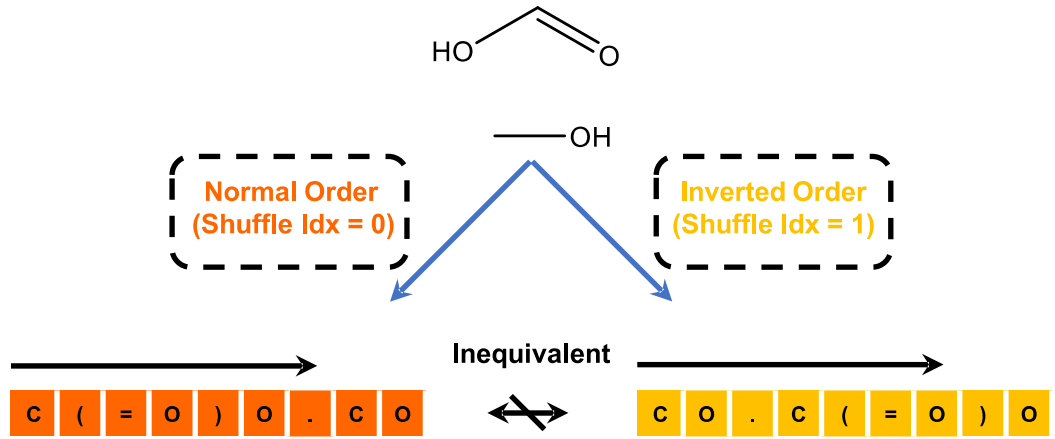
- Graph as Input, SMILES as Output
- Unable to Use SMILES Data Augmentation (Due to Permutation Invariance of Graph)



Sequence-based Data Augmentation in Graph-to-Sequence Methods



1) Root Atom Index



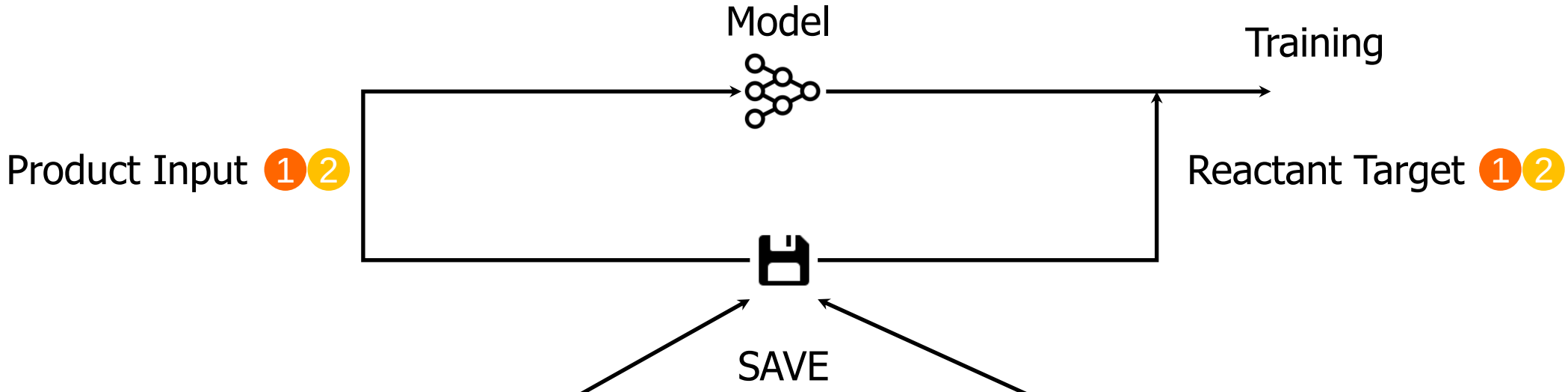
2) Shuffle Index

Chemical Features	1	2	...	...

Node Features for Each Atom

... ... :Dual Task Index & Reaction Class

Implementation: Off-the-shelf and Its Cases



5x Product Augmentation

	SMILES Notation	Root Atom Index	Shuffle Index
1	**Brc1cccc(-c2ccccn2)c1**	**0**	**0**
2	**c1(-c2ccccn2)cccc(Br)c1**	**5**	**0**
	c1ccc(Br)cc1-c1ccccn1	4	0
	c1cc(Br)cc(-c2ccccn2)c1	3	0
	c1(-c2cccc(Br)c2)ccccn1	6	0

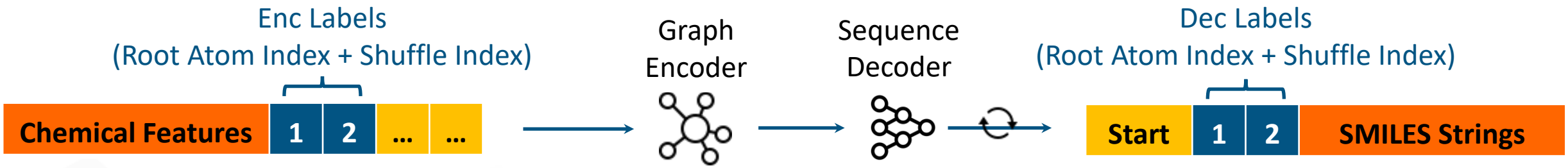
5x Reactant Augmentation

	SMILES Notation	Root Atom Index	Shuffle Index
1	**Brc1ccccn1.OB(O)c1cccc(Br)c1**	**0**	**0**
2	**B(O)(O)c1cccc(Br)c1.Brc1ccccn1**	**1**	**1**
	Brc1ccccn1.c1ccc(Br)cc1B(O)O	4	0
	Brc1ccccn1.c1ccc(B(O)O)cc1Br	6	0
	c1(B(O)O)cccc(Br)c1.Brc1ccccn1	3	1

1 : Normal Notation 2 : Random Selected Augmented Notation

Ablation Study for Sequence-based Data Augmentation

Methods	Best T in Softmax	Top-k accuracy(%)↑	
		1	*10*
(1x) Baseline	0.9	52.5	78.1
5x Baseline	0.9	54.6	68.9
5x Baseline + Enc Labels	0.9	**55.0**	79.8
5x Baseline + Enc Labels + Dec Labels	0.8	54.6	**80.3**



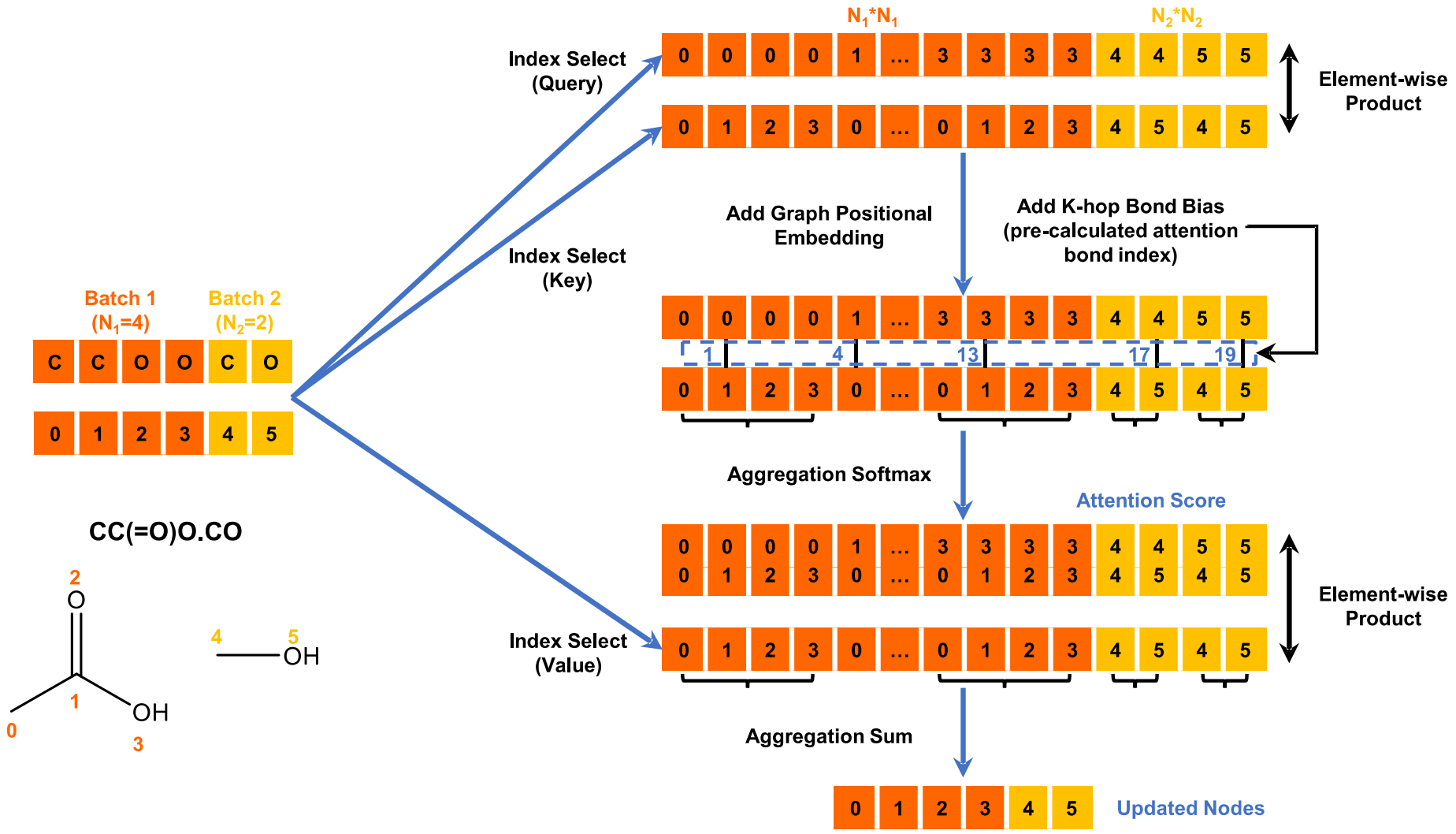
Related Work 2: Graph Transformer with Graph Padding

- Same as the Vanilla Transformer
- Graph Batch (N, H) → Attention Matrix (B, L, L) → Sequence Batch (B, L, H).

  N = Total Number of Atoms; B = Number of Graphs in the Batch;

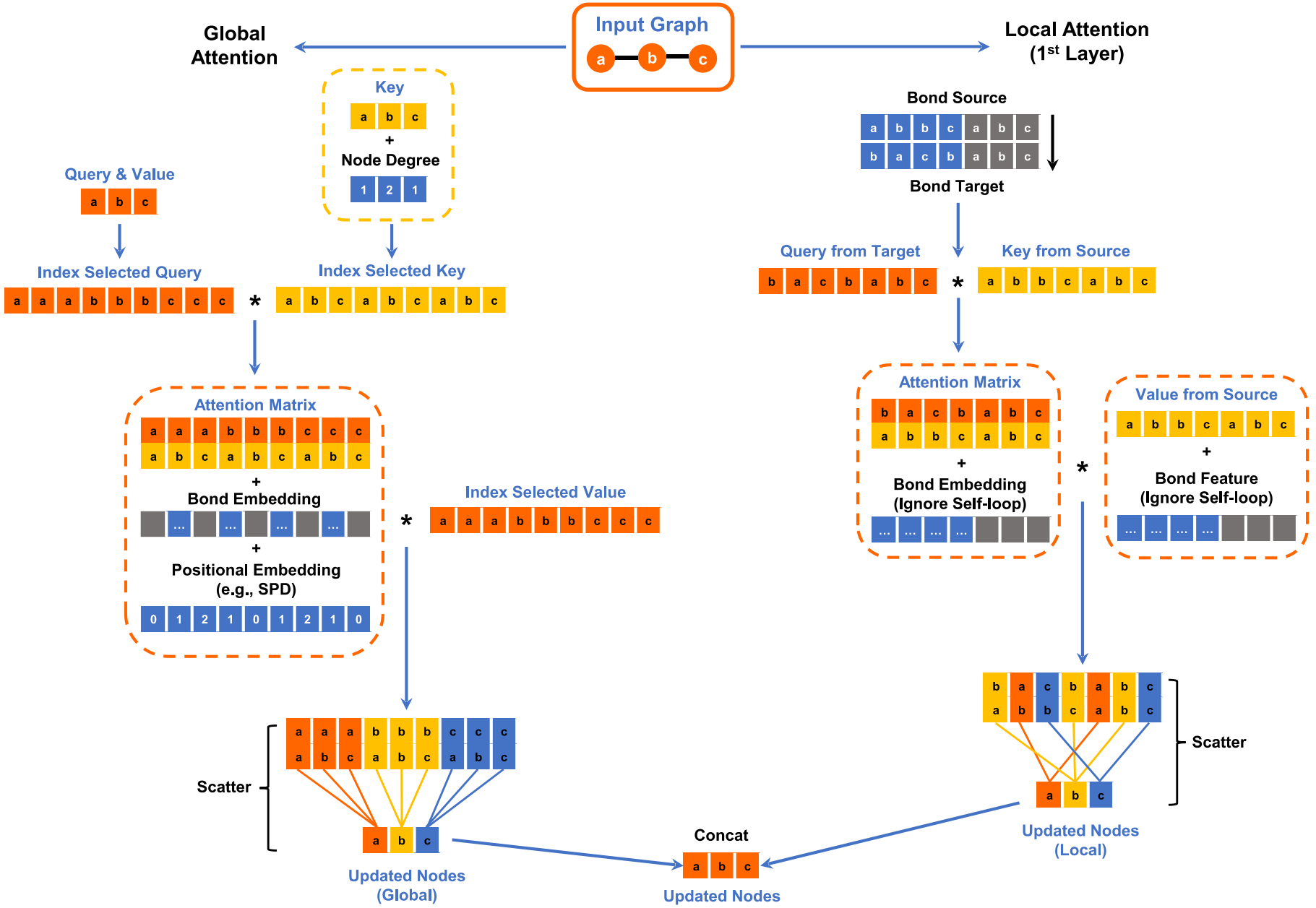
  L = Maximum Number of Nodes in the Graph

Apply Graph Transformer Without Graph Padding



- Graph Batch (N, H) → Attention Matrix ($N_1^2 + N_2^2 + \ldots$, 1) → Graph Batch (N, H)

Aggregation-based Graph Transformer: A Simple Example



Forward and Backward Speed: Vanilla & Aggregation

Data Composition in a USPTO-50k Batch

Avg 27.13 Atoms | **Avg 28.67 Paddings**

Methods	Global Attention↑	Normalization & Feed Forward↑	Backward↑
Vanilla (Baseline)	**1.00x**	1.00x	**1.00x**
Naive Aggregation	0.32x	**2.85x**	0.55x
Aggregation	0.41x	**2.85x**	0.94x

Test Settings: Separated Single-layer Encoder & Pytorch Profiler

- Vanilla: Matmul(Q, K^T) → Softmax(QK) → Matmul(QK, V)
- Naive Aggregation: Element-wise(Index Select(Q), Index Select(K)) → Scatter Softmax(QK) → Element-wise(QK, Index Select(V)) → Scatter Sum(QKV)
- Aggregation: **Coalesced** Element-wise(Q, K) → Scatter Softmax(QK) → **Coalesced** Element-wise(QK, V) → Scatter Sum(QKV)

Retrosynthesis Benchmark Results (USPTO-50k, USPTO-full)

: Template-Free

: Semi-Template

Methods	**USPTO-50k** Top-k accuracy(%)↑	
	1	*10*
Graph2SMILES[1]	53.1	74.3
RetroDCVAE[2]	53.2	82.1
G2GT[3]	54.1	77.7
SeqAGraph (***Ours, 5x***)	**55.7**	**82.7**
Graph2Edit[4]	55.1	89.4

Methods	**USPTO-full** Top-k accuracy(%)↑	
	1	*10*
Graph2SMILES[1]	45.7	63.4
G2GT[3]	**49.3**	**72.7**
SeqAGraph (***Ours, 2x***)	44.6	69.9

[1] Tu, Z. et al., Permutation Invariant Graph-to-sequence Model for Template-free Retrosynthesis and Reaction Prediction, J. Chem. Inf. Model, 2022
[2] He, H. et al., Modeling Diverse Chemical Reactions for Single-Step Retrosynthesis via Discrete Latent Variables, CIKM'22
[3] Lin, Z. et al., G2GT: Retrosynthesis Prediction with Graph-to-Graph Attention Neural Network and Self-Training, J. Chem. Inf. Model, 2023
[4] Zhong, W. et al., Retrosynthesis Prediction Using an End-to-end Graph Generative Architecture for Molecular Graph Editing, Nat. Commun., 2023

Summary

1) Graph-to-Sequence Methods Struggle in Applying Existing Sequence-based Data Augmentation

- In order to differentiate between various augmented SMILES notations with the same graph representation, we directly introduced the index of the first atom in SMILES (i.e., root atom index) and the shuffle index as additional input features.
- The existing off-the-shelf implementation of sequence-based data augmentation faces challenges in storage and preprocessing time, a more efficient approach is needed for faster and larger-scale (e.g., 20x) data augmentation.

2) Implementation of Existing Graph Transformer is Not Compatible with MPNN Framework

- Based on the pattern of Query, Key, Value, and the lower-dimensional attention matrix, we equivalently implement the attention layer by indexing, element-wise operations, and aggregation operations without graph padding.
- The advantage of padding-free attention computation is offset by under-optimized indexing and element-wise operations, further optimization of the aggregation-based Graph Transformer is needed.

西南交通大學

Southwest Jiaotong University

Presenter: Haozhe Hu

2017115054@my.swjtu.edu.cn

https://github.com/AILBC/SeqAGraph