GTA: Graph Truncated Attention for Retrosynthesis

Seung-Woo Seo [*†1], **You Young Song**[*1]
June Yong Yang[2], **Seohui Bae**[2], **Hankook Lee**[2], **Jinwoo Shin**[2], **Sung Ju Hwang**[2], **Eunho Yang**[2]
[1] Samsung Advanced Institute of Technology(SAIT), Samsung Electronics
[2] Korea Advanced Institute of Science and Technology (KAIST)
{sw32.seo, yysong02}@gmail.com, {laoconeth, shbae73, hankook.lee, jinwoos, sjhwang82, eunhoy}@kaist.ac.kr

Abstract

Retrosynthesis is the task of predicting reactant molecules from a given product molecule, important in organic chemistry since finding a synthetic path is as demanding as discovering new chemical compounds. Recently, solving the retrosynthesis task automatically without human expertise has become an active topic of research with the aid of powerful deep learning models. Recent deep models are mostly based on Seq2Seq or GNN networks depending on the selection of molecular representation, sequence or graph, respectively. Current state-of-the-art models represent a molecule as a graph, but they require joint training with auxiliary prediction tasks, such as the most probable reaction template or the reaction center prediction. However, they require additional labels by experienced chemist, which is costly. Here, we propose a novel template-free model, *Graph Truncated Attention (GTA)* which leverages both sequence and graph representations by inserting graphical information into a Seq2Seq model. The proposed GTA model masks self-attention layer using adjacency matrix of product molecule in the encoder, and applies a new loss using atom-mappings acquired from an automated algorithm to cross-attention layer in the decoder. Our model achieves new state-of-the-art results such as exact match accuracy of top-1 51.1 % and top-10 81.6 % with USPTO-50k benchmark dataset and top-10 46.0 % and top-10 70.0 % with USPTO-full dataset, both without any reaction class information. GTA model surpasses prior graph-based template-free models over top-1 2 % and top-10 7 % for USPTO-50k and top-1 and top-10 both over 6 % for USPTO-full.

Introduction

In pharmaceuticals and organic chemistry, discovering how to synthesize a certain chemical compound is as important as finding new compounds of desired properties. *Retrosynthesis*, first formulated by (Corey 1988, 1991), is the task of predicting the set of reactant molecules that are synthesized to a given product molecule by finding the inverse reaction pathway. Since its coinage, chemists have tried to adopt computer-assisted methods to retrosynthetic analysis for fast and efficient reactant candidate searches.

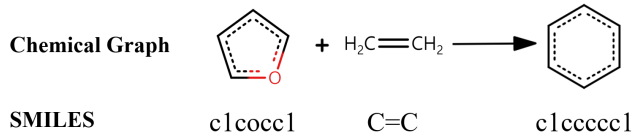

Figure 1: Example of reaction: Synthesis of benzene (right) from furan (left) and ethylene (middle). Each molecule is expressed in SMILES notation.

With the recent success of deep learning in solving various chemical tasks(Feinberg et al. 2018; You et al. 2018; Altae-Tran et al. 2017; Sanchez-Lengeling and Aspuru-Guzik 2018) and with the curation of large datasets (Reaxys 2017; SciFinder 2017; Lowe 2012, 2017; Schneider, Stiefl, and Landrum 2016), works aiming to tackle the retrosynthesis problem in a data-driven fashion with deep learning started to emerge(Liu et al. 2017; Lee et al. 2019; Coley et al. 2017; Dai et al. 2019; Segler and Waller 2017; Shi et al. 2020). These deep learning-based approaches attempt to improve the current retrosynthesis performance and opens up possibilities to automate the task by excluding human intervention and usage of domain knowledge, leading to time and cost-effectiveness.

Recent deep-learning-based approaches for retrosynthesis can be categorized into two groups; 'template-based' and 'template-free'. A template is a set of rules describing how reactants transform into a product using atom-wise mapping information. Blending such information into a model naturally requires well-established domain knowledge handled by a professional. This fact is the prime reason why current state-of-the-art template-based models such as Dai et al. (2019) exhibit higher performance compared to template-free models (Shi et al. 2020). However, reactions not covered by extracted templates are hardly predicted by template-based models (Liu et al. 2017; Chen et al. 2019), so coverage limitation, which hinders generalization, exists. Therefore, template-free models have the strength to generalize beyond extracted templates by learning from data on reactions, reactants, and products.

Template-free retrosynthesis models, the focus of this work, are further dichotomized by molecule representations, i.e., sequence or graph. The dominant representation among recent works of template-free models (Liu et al. 2017; Lee

[*]Equal contribution.
[†]Now at Standigm, Inc. seungwoo.seo@standigm.com

et al. 2019; Karpov, Godin, and Tetko 2019; Chen et al. 2019) is sequence, e.g. simplified molecular-input line-entry system (SMILES) as shown in Figure 1, widely adopted in the field of cheminformatics (Gasteiger and Engel 2006). This form of molecular representation is advantageous as powerful seq2seq models (Vaswani et al. 2017a) can be utilized.

Attempting to utilize the graph nature of molecules, a recent work achieved state-of-the-art performance among template-free models by treating molecules as graphs along with reaction center prediction, splitting molecules into synthons (molecular fragments from products), and translating synthons to reactants (G2Gs)(Shi et al. 2020). However, such an approach requires additional efforts with expert domain knowledge to generate extra labels, and a fully end-to-end graph-to-graph translation model is yet to emerge. Moreover, it is worth noting that the procedures of G2Gs are the same as what the template-based model did. The only difference is that the template is a set of reaction center and functional groups, but G2Gs predict them separately. Thus, G2Gs share the coverage limitation issue with template-based models.

In this paper, we propose a new model combining SMILES and graphs' strengths, taking the best of both worlds. First, we re-examine the untapped potential of Transformer (Vaswani et al. 2017a)-based seq2seq models. We show that Transformer-based models have been underperforming simply because their various hyperparameters were not fully explored and optimized. If properly tuned, even the vanilla Transformer with simple data augmentation for sequence representation can outperform the state-of-the-art graph model (Shi et al. 2020) by a large margin. Going one step further, we propose a novel method called **Graph-Truncated Attention (GTA)** that takes advantage of both graph and sequence representations by encouraging the graph neural network nature of Transformer architecture. Our method records the new state-of-the-art results both on the USPTO-50k and USPTO-full dataset in every exact match top-k accuracy. Our contribution is as follows.

- We correct the misunderstanding about a recent work that graph-based models outperform Transformer-based models processing sequence representations of molecules. We show that Transformer-based models have been studied without being fully optimized. With our optimized hyperparameters, a vanilla Transformer trained on the augmented USPTO-50k dataset achieve top-1 and top-10 accuracy of $49.0\,\%$ and $79.3\,\%$, respectively.
- We propose a novel graph-truncated attention method that utilizes the graph-sequence duality of molecules for retrosynthesis. We accomplish our purpose using the graph adjacency matrix as a mask in sequence modeling without introducing additional parameters.
- We validate the supremacy of the proposed architecture on the standard USPTO-50k benchmark dataset, and report that the best-achieved accuracy is state-of-the-art top-1 of **51.1 %** and top-10 of **81.6 %**. The best-achieved accuracy on USPTO-full dataset is also state-of-the-art top-1 of **46.0 %** and top-10 of **70.0 %**.

Related Works

Retrosynthesis models Validating templates experimentally requires a lot of time, e.g. 15 years for 70k templates (Bishop, Klajn, and Grzybowski 2006; Grzybowski et al. 2009; Kowalik et al. 2012; Szymkuć et al. 2016; Klucznik et al. 2018; Badowski et al. 2020), and it is hard to follow the growing speed of new reactions added to the database, e.g. about 2 million per year in 2017 to 2019 (Reaxys 2017), although automatic template extraction is available (Coley et al. 2017). Template-free models are free from these coverage problems on templates.

The first template-free model used the seq2seq model to predict the reactant sequence from a given product sequence (Liu et al. 2017). Bidirectional LSTM encoder and LSTM decoder with encoder-decoder attention were used, and the model showed comparable results to a template-based expert system. Then, a multi-head self-attention model or Transformer (Vaswani et al. 2017a) was adopted. (Karpov, Godin, and Tetko 2019) reported that character-wise tokenization and cyclic learning rate scheduling improved the model performance without modification of the Transformer. (Chen et al. 2019) added latent variable on the top of the Transformer to increase the diversity of predictions. Graph-to-graph (G2Gs) (Shi et al. 2020), a recent template-free model that used graph representation, reported state-of-the-art performance for USPTO-50k dataset. However, G2Gs follows similar procedures as the template-based model as it needs additional predictions to find reaction center which relies heavily on atom-mapping labeled by experienced chemists and translation from synthons to equivalent reactants.

Unlike existing models mentioned above, our GTA model focuses on the graph-sequence duality of molecules for the first time. GTA uses both chemical sequence and graph to model retrosynthetic analysis without any additive parameters to vanilla Transformer. GTA guides the self- and cross-attention to more explainable direction only with its graphical nature. Additionally, we re-evaluate the performance of vanilla Transformer for retrosynthesis and figure out that vanilla Transformer can surpass the current state-of-the-art model using a graph in terms of top-1 and top-10 accuracy.

Attention study There are other studies on the nature of attention. Truncated self-attention for reducing latency and computational cost in speech-recognition task was suggested, but the performance was degraded (Yeh et al. 2019). (Raganato, Scherrer, and Tiedemann 2020) reported fixed encoder self-attention is more effective for smaller-sized database. (Maziarka et al. 2020) added softmax of interatomic distance and adjacency matrix of a molecular graph to learned attention for molecule property prediction tasks. (Tay et al. 2020) suggested synthetic self-attention which has wider expressive space than dot-product attention but reducing calculation complexity with dense layer and random attention. However, GTA does not fix or widen the expressive space of self-attention. Instead, GTA limits attention space using graph structure that can be applied both to self- and cross-attention.

Background and Setup

Notation Throughout this paper, let P and R denote product and reactant molecules, respectively. Let also $\mathcal{G}(mol)$ and $\mathcal{S}(mol)$ denote the corresponding molecular graph and SMILES representations for a molecule $mol \in \{P, R\}$. We use T_{mol} and N_{mol} to denote the number of tokens and atoms in $\mathcal{S}(mol)$ respectively. The number of atoms are in fact same as the number of nodes in $\mathcal{G}(mol)$.

Molecule as sequence Simplified molecular-input line-entry system (SMILES) (Weininger 1988) is a typical sequence molecular representation where a molecule is expressed in a sequence of characters. Since a SMILES notation of a molecule is not unique and varies depending on the choice of the center atom or the start of the sequence, canonicalization algorithms are often utilized to generate a unique SMILES string among all valid strings, as in RDkit (Landrum et al. 2006). Thanks to its simplicity, SMILES representations with enough information inside, is widely used as descriptors in cheminformatics, e.g., molecular property prediction (Ramakrishnan et al. 2014; Ghaedi 2015), molecular design (Sanchez-Lengeling and Aspuru-Guzik 2018) and reaction prediction (Schwaller et al. 2019). In this work, we follow the SMILES tokenization of (Chen et al. 2019), which separates each atom (e.g., B, C, N, O), and non-atom tokens such as bonds (e.g., -, =, #), parentheses, and numbers for cyclic structures with whitespace.

Transformer and masked self-attention The retrosynthesis problem, the goal of this work, is to reversely predict the process of a synthesis reaction in which a number of reactants react to give a single product. The Transformer (Vaswani et al. 2017b) architecture with a standard encoder-decoder structure, is the current *de facto* for solving numerous natural language processing (NLP) tasks such as machine translation as they are capable of learning long-range dependencies in tokens through self-attention. For retrosynthesis tasks, the Molecular Transformer (Schwaller et al. 2019) performs another 'translation' task with SMILES given target product P to produce a set of reactants $\{R\}$.

The key component of the Transformer is the attention layer that allows the tokens to effectively access the information in other tokens. Formally, for query $Q \in \mathbb{R}^{T_{mol} \times d_k}$, 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

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

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

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

where $S = (s_{ij})$ and $M = (m_{ij}) \in \{0, 1\}^{T_{mol} \times T_{mol}}$ are score and mask matrices. The mask matrix M is customized according to the purpose of each attention modules - for example, a lower triangular matrix for decoder self-attention and a matrix of ones for encoder self-attention.

Graph Truncated Attention Framework

In this section, we introduce our graph truncated attention (GTA) framework. The conceptual idea of our model is to inject the knowledge of molecular graphs into the self- and cross-attention layers of the Transformer by truncating their attentions with respect to the graph structure. Based on the fact that the atom tokens in SMILES correspond to the atoms in a molecular graph, the attention modules can focus more on chemically relevant atoms with truncated attention connection.

Although most actively used in NLP fields, Transformer architecture can be reinterpreted as a particular kind of graph neural network (GNN). For example, tokens in source and target sequences can be viewed as nodes, and attentions are edge features connected with every token but its values are unknown initially. Then, the training Transformer model is a step to figure out edge features that well explain training data. More detailed information can be found in Joshi (2020); Ye et al. (2020).

Leveraging this connection, we extract information from both SMILES and graph by enhancing the graph neural network nature of the vanilla Transformer with the connection of given graph and sequence representations. Toward this, we propose a novel attention block of Transformer for SMILES, which we name a graph truncated attention (GTA), that utilizes corresponding graph information in computing attentions using masks, inspired by the recent success of using masks in pre-trained language model (Devlin et al. 2019; Song et al. 2019; Ghazvininejad et al. 2019). GTA can reduce the burden of training while the attention layer learns graph structure and hence perform better than vanilla Transformer. 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) GTA-self constructs a mask $M \in \{0, 1\}^{N_{mol} \times N_{mol}}$ utilizing the graph representation of molecule and truncate the attention using this mask in the attention procedure in (1). More specifically, we set $m_{ij} = 1$ if the geodesic distance between atom i and j on the molecule graph is d (or equivalently, if atom i and j are d-hop neighbors) and allow to attend only between these atoms. Otherwise, $m_{ij} = 0$. Note that d is the tunable hyper-parameter that can be a set if we want to allow to attend to atoms in multiple hops. In this paper, our model uses $d = 1, 2, 3, 4$ for entire experiments.

Through the multi-head attention introduced in the original Transformer (Vaswani et al. 2017b), the above-truncated atom attention according to geodesic distance can be enriched. Let distance matrix, $D = (d_{ij})$ be the geodesic distance between the atoms in $\mathcal{G}(mol)$, then mask matrix for h-th head is set as:

$$m_{ij} = \begin{cases} 1 & \text{if } d_{ij} = d_h \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

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. It is also

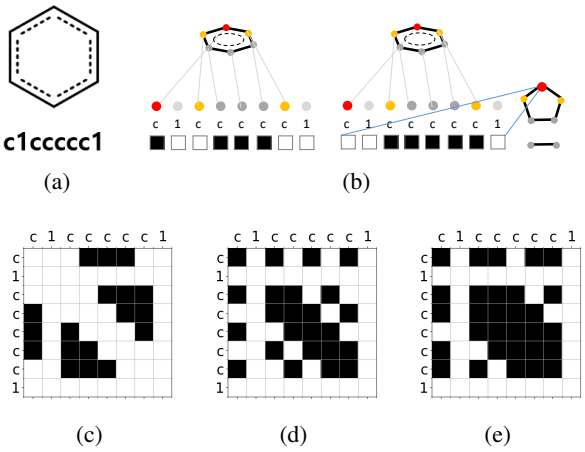

Figure 2: (a) Chemical graph of benzene and its SMILES (b) Graph Truncated Attention (*left*: mask for encoder self-attention (of red), *right*: atom mapping toward which mask for cross-attention is encouraged (of red)) (c), (d), (e) examples of modified adjacency matrix at graph geodesic distance 1, 2 and 3 which applied as mask to self-attention

worth noting that if all heads are using $d_h = 1$, GTA would become similar to Graph Attention Network (Velickovic et al. 2018). In the experiments, each two heads have the same target distance as $d_h = (h \bmod 4) + 1$ where h is indices of heads from 0 to 7.

One caveat here is that not all tokens in SMILES match the nodes of its chemical graph representation, especially, tokens for non-atoms (e.g., =, −, #, ., etc). These tokens are closely related to both atoms and other non-atoms tokens in a wide range. For example, the type of bond tokens such as double, =, or triple, #, can be clarified in the entire context and the digit tokens of cyclic structure mark where the ring opens and closes, which would require a wider range of information. Therefore, GTA-self is designed to allow non-atom tokens to exchange attentions with all other tokens regardless of the molecule graph structure. Overall, when all of these non-atom tokens are considered, the size of the mask matrix becomes larger than the previously considered mask only between atom tokens.

Figure 2 illustrates the examples of mask M with different choices of d for benzene ring. Finally, this mask M is applied to score, S to update the graph-related attentions only (see Figure 3).

Graph-truncated cross-attention (GTA-cross) Since the reaction is not a process that completely breaks down molecules to produce a completely new product, product and reactant molecules usually have quite common structures and hence it is possible to make atom-mappings between product and reactant atoms. From here, we make a simple assumption that ideal cross-attention should catch this atom-mappings because cross-attention reflect the relationship between tokens in product and reactant.

Unfortunately, how to make atom-mapping is not trivial in general and has become an active research topic in chemistry (Jaworski et al. 2019). For example, many mapping algorithms start from finding maximum common substructure (MCS) which is the largest structure that shared by two molecules. However, finding such MCS is known to be a computationally intractable NP-hard problem (Garey and Johnson 1979). As such, various methods of approximating atom mapping may be controversial depending on performance and computational efficiency, which is out of the scope of our work. As it will become clearer later, unlike other methods (Dai et al. 2019; Shi et al. 2020; Coley, Green, and Jensen 2019) based on atom mapping for retrosynthesis, our GTA-cross does not require exact atom mapping for all nodes but only leverages the information of certain pairs. Therefore, for simplicity, we simply use FMCS algorithm (Dalke and Hastings 2013) implemented in the standard RDkit (Landrum et al. 2006).

Given the (partial) information of atom mapping between product and reactant molecules, the mask for cross-attention $M = (m_{ij}) \in \{0, 1\}^{T_R \times T_P}$ is constructed as follows:

$$m_{ij} = \begin{cases} 1 & \text{if } R_{i'} \overset{mapped}{\longleftrightarrow} P_{j'} \\ 0 & \text{else} \end{cases} \tag{3}$$

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 $P_{j'}$ denote the nodes in $\mathcal{G}(R)$ and $\mathcal{G}(P)$, respectively. That is, the element of mask for cross-attention is 1 when corresponding atoms are matched by the atom mapping and 0 otherwise, as shown in Figure 2(b), and 3.

The way of using mask constructed from (3) in GTA-cross should be a completely different one from that of using a mask in GTA-self following the standard way (1). This is not just because the atom mapping is not perfect as discuss above but because the auto-regressive nature of decoder in cross attention makes incomplete SMILES and unable to find mapping during sequence generation at inference time. To side-step this issue, GTA-cross does not force attention by a *hard* mask but 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:

$$\mathcal{L}_{\text{attn}} = \sum \left[(M_{\text{cross}} - A_{\text{cross}})^2 \odot M_{\text{cross}} \right] \tag{4}$$

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

Finally along with GTA-self component, the overall loss of GTA is $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ce}} + \alpha \mathcal{L}_{\text{attn}}$ where the effect of GTA-self is implicitly represented since the self attention generated by GTA-self contributes to cross-entropy loss $\mathcal{L}_{\text{ce}}$ through model outputs. Here α is the tunable hyper-parameter to balance two loss terms and we set it as 1.0 for all our experiments.

Experiments

In this section, we provide experimental justifications to our statements. First, as stated in our contributions, we show that

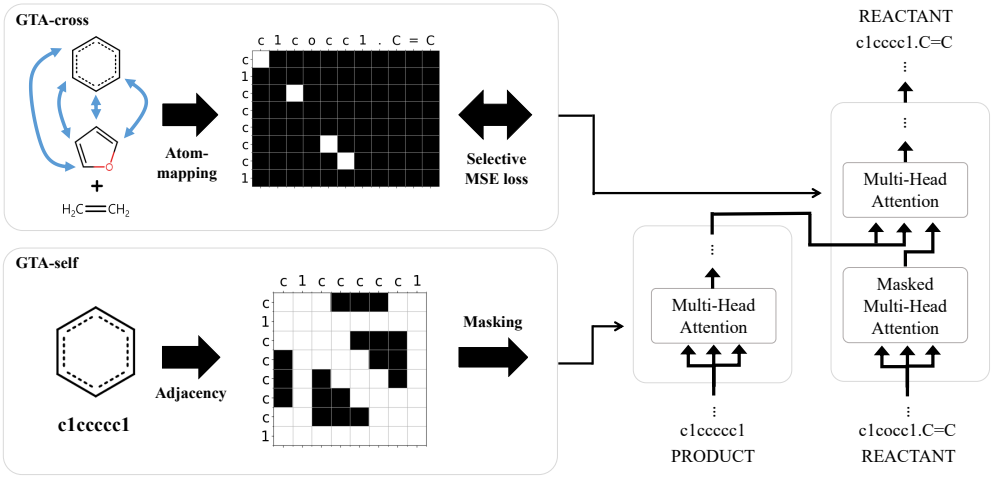

Figure 3: Graph Truncated Attention to Transformer (GTA-self: Self-attention encoder structure, GTA-cross: Selective MSE loss between cross-attention and atom-mapping in decoder)

even the naive Transformer is capable of achieving state-of-the-art performance simply by *tuning hyperparameters*. Second, we demonstrate the even higher performance of the vanilla Transformer equipped with our graph-truncated attention.

Experimental setup

Datasets and augmentation strategy We use the open-source reaction database from U.S. patent, USPTO-full and USPTO-50k as a benchmark in this study which was used in previous studies. Detailed information on each dataset and difference between them is well summarized in (Thakkar et al. 2020). USPTO-full contains the reactions in USPTO patent from 1976 to 2016, curated by (Lowe 2012, 2017) that have approximately 1M reactions. USPTO-50k, a subset of USPTO-full, is refined by randomly choosing 50k out of 630k reactions which are identically and fully atom-mapped reactions using two different mapping algorithms out of whole 1M reactions (Schneider, Stiefl, and Landrum 2016). We follow data splitting strategy of (Dai et al. 2019) which is randomly dividing train/valid/test set to 80 %/10 %/10 % of data. We augment the USPTO-50k dataset by changing the order of reactant molecule(s) as <c1cocc1.C=C> and <C=C.c1cocc1> in SMILES notation, and changing the starting atom of SMILES as in (Tetko et al. 2020). For example, standard or canonical form of a furan molecule in Fig. 1 is <c1cocc1> in SMILES notation, and we can create an alternative SMILES representation <o1cccc1> by changing the starting atom to oxygen. We refer this reactant ordering change as '_s' and altering the starting atom as '2P2R', where 2 denotes one random alternative SMILES of product and reactant SMILES are added to original USPTO-50k. Both augmentation methods are applied to '2P2R_s' dataset. It is worth noting that these kinds of augmentation are incompatible with graph-based approaches.

Baselines We compare our method with six different baselines from other previous researches and two re-evaluated baselines that we conducted with our optimized hyperparameters and USPTO-50k dataset. **BiLSTM** (Liu et al. 2017) is the first template-free model with seq2seq LSTM layers. **Transformer** is a baseline using self-attention based seq2seq model reported by (Vaswani et al. 2017a). **Latent model** (Chen et al. 2019) implements discrete latent variable for diverse prediction. We refer to the results of latent size equal to five with the plain USPTO-50k dataset and their best result with data augmentation and pre-training. **Syntax correction** (Zheng et al. 2019) added denoising auto-encoder which corrects predicted reactant SMILES. **G2Gs** (Shi et al. 2020) is the only graph model among template-free researches. In this paper, we compare our USPTO-50k results only to the 'template-free' models except for BiLSTM in Table 1. We think 'template-free' is the hardest but the most practical problem setting in novel material discovery, as which it may need a new unseen template or the user may not have enough experience to guess the right reaction class. Furthermore, to examine the scalability, GTA is trained with USPTO-full and compared to the template-based model **GLN** (Dai et al. 2019) in Table 2.

Evalutation metrics We used top-k exact match accuracy which is most widely used, and also used in the aforementioned baselines. Predicted SMILES was standardized using RDkit package first, and then exact match evaluation was done. We use $k = 1, 3, 5, 10$ for performance comparison using beam search. Beam size of 10 and top-50 predictions are found to be the optimal settings for our GTA model with USPTO-50k dataset, while beam size of 10 and top-10 predictions are used in USPTO-full dataset, hyperparameter optimization and ablation study.

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.

Other details GTA is build-up on the work of (Chen et al. 2019) which is based on Open Neural Machine Translation (ONMT) (Klein et al. 2017, 2018) and Pytorch (Paszke et al. 2017). We also used RDkit (Landrum et al. 2006) for extracting distance matrix, atom-mapping, and SMILES pre- and post-processing. GTA implements Transformer architecture with 6 and 8 layers of both encoder and decoder for USPTO-50k and USPTO-full dataset, respectively. Embedding size is set to 256, the number of heads is fixed to 8, and dropout probability to 0.3. We train our model using early-stopping method, training was stopped without improvement within 40 times in validation loss and accuracy for every 1000 (for USPTO-50k) and 10000 (for USPTO-full) steps with a batch size of maximum 4096 tokens in batch. Relative positional encoding (Shaw, Uszkoreit, and Vaswani 2018) is used with maximum relative distance of 4. Adam (Kingma and Ba 2015) optimization method with noam decay (Vaswani et al. 2017a) and learning rate scheduling for 8000 warm-up steps on a single Nvidia Tesla V100 GPU takes approximately 7, 18 hours, and 15 days of training time for USPTO-50k plain, 2P2R_s, and USPTO-full dataset respectively. All experiments are trained with five seeds 2020 to 2024 and averaged to validate pure model performance. Then, we calculate and report the mean and standard error of mean from the experiments.

To explore the best hyperparameters for our model, we optimized early stopping step, dropout, number of layers and maximum relative distance for both ONMT and GTA and optimized GTA-self distance and GTA-cross alpha for GTA. These results can be found in Supplementary. In addition, our implementation, data, and pretrained weight details can be found in Supplementary.

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±0.29	81.6±0.22	**46.0**±0.07	**70.0**±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 ± symbol.

USPTO-50k results

Reproducibility Before reporting GTA result, we found that hyperparameters were not optimized in previous researches with Transformer architecture. They used dropout probability value of 0.1; however, using 0.3 drastically increases the performance of the vanilla Transformer about +2.7 % and +9.9 % point in terms of top-1 and top-10 accuracy, as shown in Table 1 even without data augmentation (compare Transformer and ONMT(Plain)). When augmentation 2P2R_s is applied, the vanilla Transformer breaks previous state-of-the-art result with 49.0 % and 79.3 % in terms of top-1 and top-10 accuracy upon USPTO-50k dataset.

Graph-truncated attention When GTA is applied to plain USPTO-50k dataset, overall top-k performance increases at least +2.6 % point compared to our reproduced vanilla Transformer. As the same in reproduced results, applying 2P2R_s augmentation gives top-1 accuracy above 50 %, surpassing all other template-free models ever reported.

Although the latent model benefited top-10 accuracy, it sacrificed top-1 accuracy a lot, the worst among reported value. In contrast, GTA equally increases all of top-k accuracies without sacrifice. Finally, our result successfully achieves the state-of-the-art template-free result in overall top-k accuracy, which implies GTA is more accurate and diverse than previous retrosynthesis models. GTA records 80.1 % and 81.6 % in top-10 accuracy without and with data augmentation, respectively, and no decreasing performance in top-1 accuracy: 47.3 % and 51.1 % without and with data augmentation, respectively.

USPTO-full results

We also validated GTA with more scalable data USPTO-full, which has a train, validation, and test data with a size of 800k, 100k, and 100k, respectively. As mentioned above, USPTO-50k is a refined dataset out of 630k reactions which relies on atom-mapping consistency between two different mapping algorithm in USPTO-full; USPTO-50k represents exceptional cases of its superset. Although USPTO-50k may not reflect the right character of USPTO-full, none of the previous template-free models showed experimental results with USPTO-full.

GTA model trained on USPTO-full achieved state-of-the-art exact match accuracy of 46.0 %, and 70.0 % for top-1 and top-10, respectively, which is over 5.7 % higher than that of

GTA	GTA	Plain				2P2R_s			
-self	-cross	Top-1	Top-3	Top-5	Top-10	Top-1	Top-3	Top-5	Top-10
-	-	45.0 $\pm$0.29	63.6$\pm$0.30	69.2 $\pm$0.41	73.3 $\pm$0.53	49.6 $\pm$0.31	65.9 $\pm$0.22	72.1 $\pm$0.38	77.8 $\pm$0.47
-	✓	45.9 $\pm$0.28	64.8 $\pm$0.28	70.5 $\pm$0.40	74.7 $\pm$0.45	49.7 $\pm$0.46	66.3 $\pm$0.29	72.9 $\pm$0.36	78.6 $\pm$0.37
✓	-	46.8 $\pm$0.40	65.2 $\pm$0.19	70.5 $\pm$0.29	74.9 $\pm$0.32	51.1 $\pm$0.32	65.8 $\pm$0.17	71.9 $\pm$0.07	77.1 $\pm$0.33
✓	✓	47.3 $\pm$0.28	66.7 $\pm$0.49	72.3 $\pm$0.30	76.5 $\pm$0.30	51.1 $\pm$0.29	67.0 $\pm$0.29	73.1 $\pm$0.38	78.4 $\pm$0.25

Table 3: Ablation study of GTA method. The standard error with 95% confidence interval is given after $\pm$ symbol.

'template-based' GLN with USPTO-full. Table 2 shows that the scalability of template-free GTA is better than template-based GLN. Our result strongly supports that USPTO-full is more appropriate for benchmarking retrosynthesis tasks. Moreover, models that heavily depend on atom-mapping, can exploit USPTO-50k because USPTO-50k dataset has atom-mapping consistency but not USPTO-full. In other words, USPTO-50k are more accessible in mapping the reaction than others. Template-based GLN's performance degradation is $13.2\,\%$ between USPTO-50k and USPTO-full in top-1 accuracy while it is only $5.1\,\%$ in template-free GTA, which is less than half of GLN's degradation when dataset is expaneded to USPTO-full. Here again, we emphasize the generalization of template-free model.

Ablation study

Following ablation study is designed to explore the effect of each GTA-self and GTA-cross modules. Results are shown in Table 3 with top-*k* exact match accuracy evaluated using beam search with beam size of 10 and top-10 predictions (We note again that beam size of 10 and top-50 predictions was used for our best performance in Table 1).

Graph-truncated self-attention (GTA-self) When only GTA-self is applied to Transformer, it gives $+1.4\,\%$ point margins at least, $+1.6\%$ point on average. In particular, its effect on top-1 accuracy is greatest among the others. On both of plain and augmented dataset, GTA-self alone results very close to our best model for top-1 accuracy, showing only $-0.5\,\%$ or under point of difference. This result implies that encouraging the atom on its sequence domain to look the atom nearby on its graph domain takes most part in improvement, supporting that the point of entry of graph-sequence duality was indeed effective. This extends to two important points; first, our model is not heavily relying on the atom-mapping, which are known to require more expertise than FMCS algorithm, and second, the performance has a room for improvements with advanced mapping algorithm.

Graph-truncated cross-attention (GTA-cross) GTA-cross alone, likewise, shows marginal but clear gain for all the case from top-1 to -10 accuracy on plain dataset. It shows smaller margin of increment (+0.9 % point) than GTA-self (+1.6 % point) when they are trained solely on this plain dataset. Interestingly, GTA-cross shows superior performance gain (+0.7 % point) than GTA-self (+0.1 % point), except for top-1, when trained on the augmented dataset. The result of showing the high capacity of GTA-cross especially on larger dataset gives us a presumption that the imperfectness of atom-mapping (derived from FMCS algorithm; trading-off for) could be sufficiently compensated by large number of data points. Consequently, we now take benefits of low computing cost for not generating near-perfect atom-mapping while retaining the highest prediction capacity among all. Lastly, unlike GTA-self which shows a gradual decrease in margin of improvement from top-1 to -10 accuracy, GTA-cross behaves exactly reversely, showing its highest margin of improvement in its top-10 accuracy (+1.3%). GTA-cross and GTA-self behaves in mutual complementary manner, watching each other's back in retrosynthesis prediction.

Conclusion

This paper proposed a method to solve retrosynthetic analysis combining the feature of a molecule as both SMILES sequence and graph called graph-truncated attention (GTA). This kind of sequence-graph duality, while dealing with molecules in deep learning, was previously overlooked.

We revisited Transformer architecture as a graph neural network and found the entry points of chemical graph information. Then, we used the distance matrix of a molecular graph and atom-mapping matrix between product and set of reactant as a mask and guide to self- and cross-attention. Also, we re-evaluated the performance of the vanilla Transformer, which was underestimated because of poor optimization. On top of that, GTA showed the best overall top-k accuracy among the reported results, which recorded over $50\,\%$ top-1 accuracy for the first time in the template-free model with UPSTO-50k dataset.

Finally, we moved to a scalable dataset, USPTO-full, and GTA excels template-based GLN with $5.7\,\%$, and $6.3\,\%$ in top-1 and top-10 accuracy, respectively. This result originated from how USPTO-50k had been constructed. USPTO-50k is built upon the reactions which have full atom-mapping without conflict in the algorithms. This condition gave an advantage to models that utilize full atom-mapping, although $47\,\%$ of USPTO-full does not fall into this category.

We look forward to further performance gain with other models because our method can be combined with other reported retrosynthesis research with attention mechanisms without a conflict. Moreover, other data with graph-sequence duality might get benefits from GTA, too.

Potential Ethical Impact

Our model herein is validated with USPTO dataset which includes reactions on organic molecules, and we believe our model is generally applicable to pharmaceutical or other chemical reaction datasets where chemicals are in SMILES or similar sequence-based representations. Deep learning models on reaction and/or retrosynthesis predictions emphasize the time and cost-effectiveness of using well-trained and automated models instead of solely depending on human expertise and experiments as in the past. The marketplace is positive towards automating the synthesis process and building up autonomous environments dreaming of a one-click system from discovering target product with given properties to finding reactant candidates, optimizing synthetic paths, and testing stability without or minimum human work.

However, yet these models lack to consider 'chemical stability' and 'overall safety' as related data are not collected enough to be trained or not opened to public. Even though we have up-to-date data, new safety hazards and stability problems can always break out as reaction and/or retrosynthesis models may deal with new chemicals and unseen synthetic paths. Therefore, all concerned researchers and their groups in academies, industries and elsewhere, now need to start to discuss and study how to screen the safety level of the following steps of 'synthesizing predicted reactants' and 'testing stability'.

References

Altae-Tran, H.; Ramsundar, B.; Pappu, A. S.; and Pande, V. 2017. Low data drug discovery with one-shot learning. *ACS central science* 3(4): 283–293.

Badowski, T.; Gajewska, E. P.; Molga, K.; and Grzybowski, B. A. 2020. Synergy Between Expert and Machine-Learning Approaches Allows for Improved Retrosynthetic Planning. *Angewandte Chemie International Edition* 59(2): 725–730.

Bishop, K. J. M.; Klajn, R.; and Grzybowski, B. A. 2006. The Core and Most Useful Molecules in Organic Chemistry. *Angewandte Chemie International Edition* 45(32): 5348–5354.

Chen, B.; Shen, T.; Jaakkola, T. S.; and Barzilay, R. 2019. Learning to Make Generalizable and Diverse Predictions for Retrosynthesis. *arXiv preprint arXiv:1910.09688* .

Coley, C. W.; Barzilay, R.; Jaakkola, T. S.; Green, W. H.; and Jensen, K. F. 2017. Prediction of organic reaction outcomes using machine learning. *ACS central science* 3(5): 434–443.

Coley, C. W.; Green, W. H.; and Jensen, K. F. 2019. RDChiral: An RDKit wrapper for handling stereochemistry in retrosynthetic template extraction and application. *Journal of chemical information and modeling* 59(6): 2529–2537.

Corey, E. J. 1988. Robert Robinson Lecture. Retrosynthetic thinking—essentials and examples. *Chemical Society Reviews* 17(0): 111–133.

Corey, E. J. 1991. The Logic of Chemical Synthesis: Multistep Synthesis of Complex Carbogenic Molecules (Nobel Lecture). *Angewandte Chemie International Edition in English* 30(5): 455–465.

Dai, H.; Li, C.; Coley, C.; Dai, B.; and Song, L. 2019. Retrosynthesis Prediction with Conditional Graph Logic Network. In *Advances in Neural Information Processing Systems 32*, 8872–8882. Curran Associates, Inc.

Dalke, A.; and Hastings, J. 2013. FMCS: a novel algorithm for the multiple MCS problem. *Journal of cheminformatics* 5(1): 1–1.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv* abs/1810.04805.

Feinberg, E. N.; Sur, D.; Wu, Z.; Husic, B. E.; Mai, H.; Li, Y.; Sun, S.; Yang, J.; Ramsundar, B.; and Pande, V. S. 2018. PotentialNet for molecular property prediction. *ACS central science* 4(11): 1520–1530.

Garey, M. R.; and Johnson, D. S. 1979. *Computers and intractability*, volume 174. freeman San Francisco.

Gasteiger, J.; and Engel, T. 2006. *Chemoinformatics: A Textbook*. Wiley. ISBN 9783527606504.

Ghaedi, A. 2015. Predicting the cytotoxicity of ionic liquids using QSAR model based on SMILES optimal descriptors. *Journal of Molecular Liquids* 208: 269–279.

Ghazvininejad, M.; Levy, O.; Liu, Y.; and Zettlemoyer, L. 2019. Mask-predict: Parallel decoding of conditional masked language models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 6114–6123.

Grzybowski, B. A.; Bishop, K. J.; Kowalczyk, B.; and Wilmer, C. E. 2009. The 'wired' universe of organic chemistry. *Nature Chemistry* 1(1): 31–36.

Jaworski, W.; Szymkuć, S.; Mikulak-Klucznik, B.; Piecuch, K.; Klucznik, T.; Kaźmierowski, M.; Rydzewski, J.; Gambin, A.; and Grzybowski, B. A. 2019. Automatic mapping of atoms across both simple and complex chemical reactions. *Nature communications* 10(1): 1–11.

Joshi, C. 2020. Transformers are Graph Neural Networks. URL https://graphdeeplearning.github.io/post/transformers-are-gnns/.

Karpov, P.; Godin, G.; and Tetko, I. V. 2019. A Transformer Model for Retrosynthesis. In Tetko, I. V.; Kůrková, V.; Karpov, P.; and Theis, F., eds., *Artificial Neural Networks and Machine Learning – ICANN 2019: Workshop and Special Sessions*, 817–830. Cham: Springer International Publishing. ISBN 978-3-030-30493-5.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. *CoRR* abs/1412.6980.

Klein, G.; Kim, Y.; Deng, Y.; Nguyen, V.; Senellart, J.; and Rush, A. 2018. OpenNMT: Neural Machine Translation Toolkit. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, 177–184. Boston, MA: Association for Machine Translation in the Americas.

Klein, G.; Kim, Y.; Deng, Y.; Senellart, J.; and Rush, A. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. In *Proceedings of ACL 2017, System Demonstrations*, 67–72. Vancouver, Canada: Association for Computational Linguistics.

Klucznik, T.; Mikulak-Klucznik, B.; McCormack, M. P.; Lima, H.; Szymkuć, S.; Bhowmick, M.; Molga, K.; Zhou, Y.;

Rickershauser, L.; Gajewska, E. P.; Toutchkine, A.; Dittwald, P.; Startek, M. P.; Kirkovits, G. J.; Roszak, R.; Adamski, A.; Sieredzińska, B.; Mrksich, M.; Trice, S. L.; and Grzybowski, B. A. 2018. Efficient Syntheses of Diverse, Medicinally Relevant Targets Planned by Computer and Executed in the Laboratory. *Chem* 4(3): 522 – 532. ISSN 2451-9294.

Kowalik, M.; Gothard, C. M.; Drews, A. M.; Gothard, N. A.; Weckiewicz, A.; Fuller, P. E.; Grzybowski, B. A.; and Bishop, K. J. M. 2012. Parallel Optimization of Synthetic Pathways within the Network of Organic Chemistry. *Angewandte Chemie International Edition* 51(32): 7928–7932.

Landrum, G.; et al. 2006. RDKit: Open-source cheminformatics .

Lee, A. A.; Yang, Q.; Sresht, V.; Bolgar, P.; Hou, X.; Klug-McLeode, J. L.; and Butler, C. R. 2019. Molecular Transformer unifies reaction prediction and retrosynthesis across pharma chemical space. *Chemical Communications* 55: 12152–12155.

Liu, B.; Ramsundar, B.; Kawthekar, P.; Shi, J.; Gomes, J.; Luu Nguyen, Q.; Ho, S.; Sloane, J.; Wender, P.; and Pande, V. 2017. Retrosynthetic Reaction Prediction Using Neural Sequence-to-Sequence Models. *ACS Central Science* 3(10): 1103–1113.

Lowe, D. 2017. Chemical reactions from US patents (1976-Sep2016) doi:10.6084/m9.figshare.5104873.v1.

Lowe, D. M. 2012. *Extraction of chemical structures and reactions from the literature*. Ph.D. thesis, Department of Chemistry, University of Cambridge.

Maziarka, L.; Danel, T.; Mucha, S.; Rataj, K.; Tabor, J.; and Jastrzebski, S. 2020. Molecule Attention Transformer. *ArXiv* abs/2002.08264.

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in pytorch .

Raganato, A.; Scherrer, Y.; and Tiedemann, J. 2020. Fixed Encoder Self-Attention Patterns in Transformer-Based Machine Translation. *ArXiv* abs/2002.10260.

Ramakrishnan, R.; Dral, P. O.; Rupp, M.; and Von Lilienfeld, O. A. 2014. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data* 1: 140022.

Reaxys. 2017. Reaxys. URL https://reaxys.com.

Sanchez-Lengeling, B.; and Aspuru-Guzik, A. 2018. Inverse molecular design using machine learning: Generative models for matter engineering. *Science* 361(6400): 360–365.

Schneider, N.; Stiefl, N.; and Landrum, G. A. 2016. What's What: The (Nearly) Definitive Guide to Reaction Role Assignment. *Journal of Chemical Information and Modeling* 56(12): 2336–2346.

Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; and Lee, A. A. 2019. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Central Science* 5: 1572–1283.

SciFinder. 2017. SciFinder. URL https://scifinder.cas.org.

Segler, M. H.; and Waller, M. P. 2017. Neural-symbolic machine learning for retrosynthesis and reaction prediction. *Chemistry–A European Journal* 23(25): 5966–5971.

Shaw, P.; Uszkoreit, J.; and Vaswani, A. 2018. Self-Attention with Relative Position Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 464–468. New Orleans, Louisiana: Association for Computational Linguistics.

Shi, C.; Xu, M.; Guo, H.; Zhang, M.; and Tang, J. 2020. A Graph to Graphs Framework for Retrosynthesis Prediction. In *International conference on machine learning*.

Song, K.; Tan, X.; Qin, T.; Lu, J.; and Liu, T.-Y. 2019. MASS: Masked Sequence to Sequence Pre-training for Language Generation. *ArXiv* abs/1905.02450.

Szymkuć, S.; Gajewska, E. P.; Klucznik, T.; Molga, K.; Dittwald, P.; Startek, M.; Bajczyk, M.; and Grzybowski, B. A. 2016. Computer-Assisted Synthetic Planning: The End of the Beginning. *Angewandte Chemie International Edition* 55(20): 5904–5937.

Tay, Y.; Bahri, D.; Metzler, D.; Juan, D.-C.; Zhao, Z.; and Zheng, C. 2020. Synthesizer: Rethinking Self-Attention in Transformer Models. *ArXiv* abs/2005.00743.

Tetko, I. V.; Karpov, P.; Deursen, R. V.; and Godin, G. 2020. State-of-the-Art Augmented NLP Transformer models for direct and single-step retrosynthesis.

Thakkar, A.; Kogej, T.; Reymond, J.-L.; Engkvista, O.; and Bjerrum, E. J. 2020. Datasets and their influence on the development of computer assisted synthesis planning tools in the pharmaceutical domain. *Chemical Science* 11: 154–168.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017a. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017b. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Velickovic, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; and Bengio, Y. 2018. Graph Attention Networks. *ArXiv* abs/1710.10903.

Weininger, D. 1988. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences* 28(1): 31–36.

Ye, Z.; Zhou, J.; Guo, Q.; Gan, Q.; and Zhang, Z. 2020. Transformer tutorial.

Yeh, C.-F.; Mahadeokar, J.; Kalgaonkar, K.; Wang, Y.; Le, D.; Jain, M.; Schubert, K.; Fuegen, C.; and Seltzer, M. L. 2019. Transformer-Transducer: End-to-End Speech Recognition with Self-Attention. *ArXiv* abs/1910.12977.

You, J.; Liu, B.; Ying, Z.; Pande, V.; and Leskovec, J. 2018. Graph convolutional policy network for goal-directed molecular graph generation. In *Advances in neural information processing systems*, 6410–6421.

Zheng, S.; Rao, J.; Zhang, Z.; Xu, J.; and Yang, Y. 2019. Predicting Retrosynthetic Reactions using Self-Corrected Transformer Neural Networks. *Journal of Chemical Information and Modeling* .