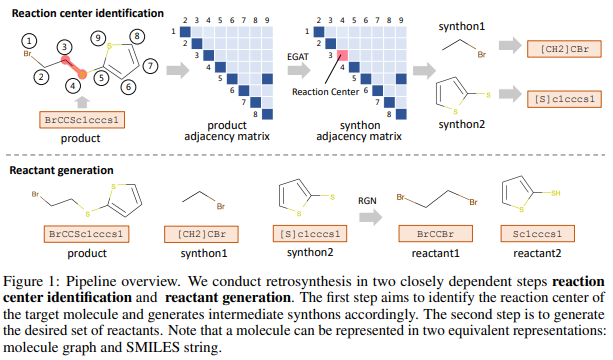
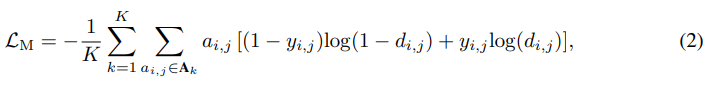
# RetroXpert: Decompose Retrosynthesis Prediction Like A Chemist

* Abstract: - retrosynthesis is the process of recursively decomposing target molecules into available building blocks.
* Most of the retrosynthesis algorithms lack interpretability about their predictions. In this paper the researchers try to devise a novel template-free algorithm for automatic retrosynthesis expansion.
* Their method disassembles retrosynthesis into two steps: (i) identity the potential reaction center of the target molecule through a novel graph neural network and generate intermediate synthons, and (ii) generate the reactants associated with synthons via a robust reactant generation model.
* Introduction: - retrosynthesis mode was formalized in the pioneering work, and now have become one of the fundamental paradigms in the modern chemical society.
* The organic knowledge consists of in the order of 107 reactions and compounds, also the incomplete understanding of the reaction mechanism also increase the difficulty of retrosynthesis, which is typically undertaken by human expert.
* The feasibility of a route is often determined by multiple factors, such as the availability of potential reactants, reaction conditions, reaction yield, and potential toxic byproducts.
* The researcher’s method can be decomposed into two sub-tasks: (i) *Breaking down* the given target molecule into a set of synthons which are hypothetical units representing potential starting reactants in the retrosynthesis of the target, and (ii) *Calibrating* the obtained synthons into a set of reactants, each of which corresponds to an available molecule.
* Template-based methods plan retrosynthesis based on hand-encoded rules or reaction templates.
* Reaction templates are often automatically extracted from the reaction databases and appropriate templates are selected to apply to the target. The key process of these approaches is to select relevant templates for the given target.
* A common limitation is that these methods can only infer reactions within the chemical space covered by the template database, preventing them from discovering novel reactions.
* The template-free methods treat the retrosynthesis as a neural machine translation problem, since molecules can be represented as SMILES strings.
* These models do not fit into the chemists’ analytical process and lack interpretability behind their predictions. Also, such approaches fail to consider rich chemistry knowledge within the chemical reactions.
* the G2Gs proposes to incrementally generate reactants from the associated synthons with a variational graph translation model.
* G2Gs can predict at most one bond disconnection which is not universal. Besides, G2Gs independently generate multiple reactants, which ignores the relationship between multiple reactants.
* To overcome these challenges, the researchers devise a two-step framework named as RetroXpert (**Retro**synthesis e**Xpert**) to automate the retrosynthesis prediction.
* Their model tackles it in two steps firstly – they propose to identify the potential reaction center within the target molecule using a novel Edge-enhanced Graph Attention Network (EGAT).
* Secondly, the Reactant Generation Network (RGN) predicts associated reactants given the target molecule and synthons.
* Different from previous methods, the reactant generation order can be uniquely decided in their method, thanks to the intermediate synthons.
* The researchers also notice that the robustness of the RGN plays an important role.
* To robustify the RGN, they propose to augment the training data of RGN by incorporating unsuccessful predicted synthons.
* The researchers’ main contribution can be summarized as follows: -
  + They propose to identify the potential reaction center with a novel Edge-enhanced Graph Attention Network (EGAT) which is strengthened with chemical knowledge.
  + By splitting the target molecule into synthons, the RGN is able to determine the generation order of reactants. They further propose to augment training data by introducing unsuccessfully predicted synthons, which makes RGN robust and achieves significant improvement.
  + On the standard USPTO-50K dataset, the researchers’ method achieves 70.4% and 65.5% Top-1 accuracy when w/ and wo/ reaction type, respectively, which outperforms SOTA accuracy 63.2% (w/) and 52.6% (wo/) reported in by a large margin.
* Methodology: - given a molecule graph **G** with *N* nodes (atoms), they denote the matrix representation of node features as *X* , the tensor representation of edge features as E , and the adjacency matrix as *A* .



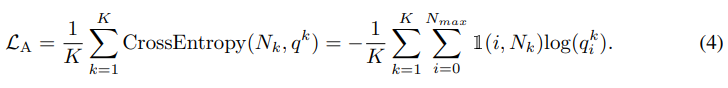
* *M* and *L* are features dimensions of atoms and bonds, respectively. They denote as *P, S, R* the product, synthons, and reactants in the reaction formulation, respectively.
* The single-step retrosynthesis problem can be described as given the desired product *P,* seeking for a set of reactants *R* = {R1, R2, … Rn} that can produce the major product *P* through a valid chemical reaction.
* It is denoted as *PR* which is the reverse process of the forward reaction prediction problem that predicts the outcome products given a set of reactants.
* The researchers’ method decomposes the retrosynthesis task (*PR*) into two closely dependent steps **reaction center identification** (*PS*) and **reactant generation** (*SR*).
* The first step is to identify the potential reaction bonds which will be disconnected during the retrosynthesis, and then the product *P* can be split into a set of intermediate synthons *S* = {*S1*, *S2*, …, *Sn*}.
* The second step is to transform synthons *S* = {*S1*, *S2*, …, *Sn*} into associated reactants *R* = {*R1, R2, … Rn*}.
* EGAT for reaction center identification: - the researchers propose a graph neural network named Edge-enhanced Graph Attention Network (EGAT) which takes the molecule graph **G** as input and predicts disconnection probability for each bond, and this is the main task.
* The current message passing neural networks are shallow and capture only local structure information for each node, and also it is difficult to distinguish multiple reaction centers without global information.
* To overcome or tackle this problem the researchers add a graph-level auxiliary task to predict the total number of disconnection bonds.
* It is distinct from the Graph Attention Network (GAT), which is designed to learn node and graph-level embeddings, the researchers proposed EGAT also learns edge embeddings.
* Given the target **G** = {*A*, *E*, *X*}, the EGAT layers computes node embedding and edge embedding from previous layer’s embeddings and by following equations:
* Here W are trainable parameters, || means concatenation operation, is all neighbors nodes of the node *i*, is the attention weight between the node *i* and its neighbors node j, and as well as are the output node and edge representations, respectively.
* Initial input embeddings , are the input node and edge feature vectors , , respectively, and in this special case the dimensions *F* and *D* equals to the dimensions of associated features, respectively.
* After stacking multiple EGAT, the researchers obtain the final edge representation for the chemical bond between nodes *i* and *j*, as well as the node representation for each node *i*.
* To predict the disconnection probability for a bond, the researchers perform a fully-connected layer parameterized by and a *sigmoid* activation layer to and its disconnection probability is = Sigmoid.
* The optimization goal for bond disconnection predictions is to minimize the negative log-likelihood between prediction and ground-truth through the binary cross entropy loss function:



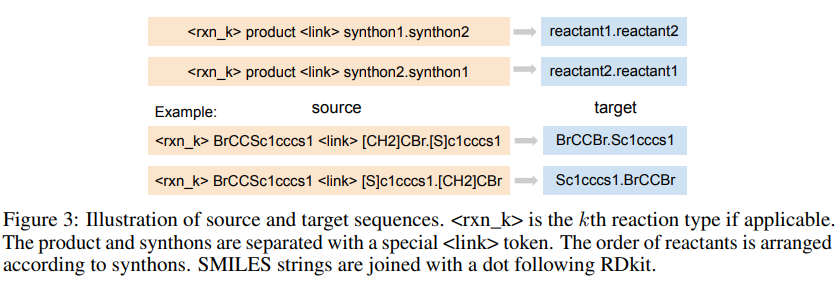
* Here *K* is the total number of training reactions and bond (*i, j*) exists if the associated adjacency element is nonzero. The ground truth = 1 means the bond (*i, j*) is disconnected otherwise remaining the same during the reaction.
* The input of the auxiliary task is the graph-level representation , which is the output of the READOUT operation over all learned node representations. The researchers adopt an arithmetic mean as the READOUT function and it works well in practice.
* Similarly, a fully-connected layer parameterized by and a *Softmax* activation function are applied to to predict the total number of disconnect bonds, which is solved as a classification problem here.
* Here each category represents the exact number of disconnected bonds, so there are 1+*Nmax* classification categories. *Nmax* is the maximum number of possible disconnected bonds in the retrosynthesis.
* The researchers denote the *Softmax* output as *q* = Softmax(. The total number of disconnected bonds for each target molecule is predicted as:



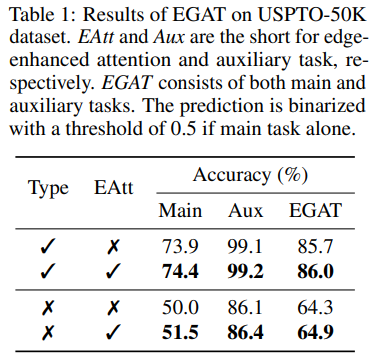
* The ground truth number of disconnections for molecule *k* is denoted as *Nk*, the indicator function is 1 if *i* equals to otherwise it is 0, and the cross-entropy loss for the auxiliary task:



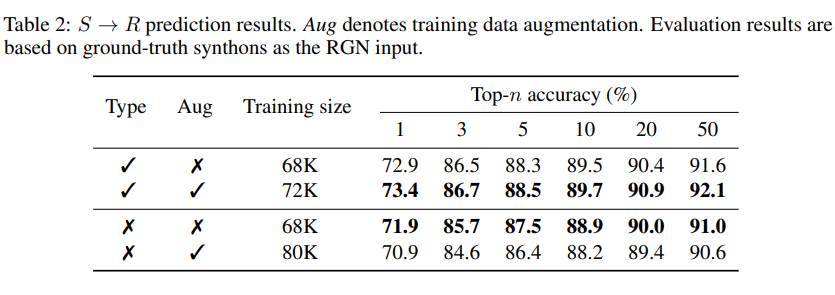
* Finally, the overall loss function for the EGAT is where is fixed to 1 in researchers’ study since they empirically find that is not a sensitive hype-parameter.
* Atom and bond features - the atom feature consists of a series of general atom information such as atom type, hybridization, and formal charge, while the bond feature is composed of chemical bond information like bond type and conjugation.
* To fully utilize the provided rich atom-mapping information of the USPTO datasets, the researchers added the semi-templates indicators to atom feature. For retrosynthesis dataset with given reaction type, a type indicator is also added to the atom feature.
* Semi-templates - there are as many as 11,647 templates for the USPTO-50K train data. Only the product side of templates are kept instead, which the researchers name as semi-template. Because the reaction templates are closely related to the exact reaction, the semi-templates indicator expected to play a significant role in reaction center identification.
* The semi-templates can be considered as subgraph patterns within molecules. The researchers build a database of semi-templates from the training data. For each atom, they mark the indicator bits associated with appeared semi-templates.
* Even reaction templates are introduced and the researchers’ methods is still template-free since (i) only semi-templates are incorporated and their method does not rely on fully templates to plan the retrosynthesis, and (ii) their EGAT still works well in the absence of semi-templates, with only slight performance degradation.
* Reactant generation network: - the reaction center has been identified; synthons can be obtained by applying bond disconnection to decompose the target graph.
* Task *S R* is to generate the set of desired reactants given obtained synthons. The researchers propose that the ideal RGN should meet following three requirements: (R1) be permutation invariant and generate the same set of reactants no matter the order of synthons, (R2) all given information should be considered when generating any reactant, and (R3) the generation of each reactant also depends on those previously generated reactants.
* To match these requirements the researchers’, represent molecules in SMILES and formulate *S R* as a sequence-to-sequence prediction problem.
* They convert synthon graphs to SMILES representations using RDKit, through these synthons may be chemically invalid.



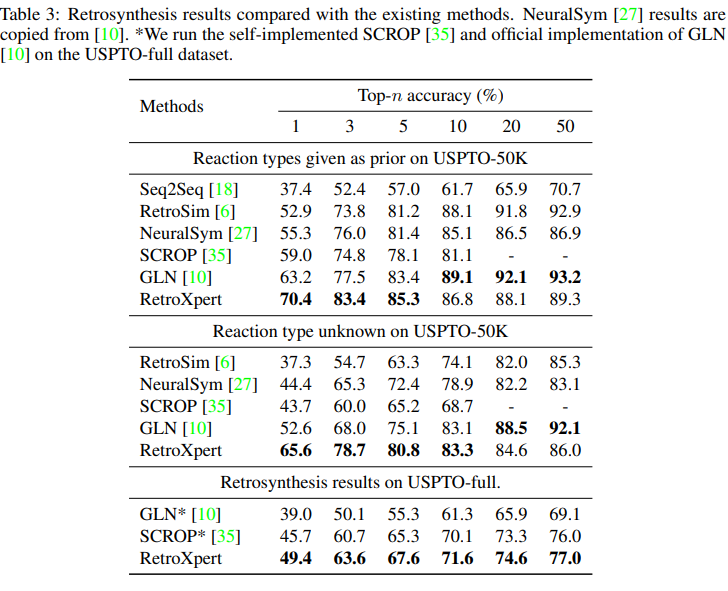
* The researchers approximate the requirement R1 by augmenting train samples with reversely arranged synthons and reactants as shown in above figure.
* The researchers’ studies shows that such approximation works pretty well in practice. To satisfy the R2 requirement, the encoder-decoder attention mechanism is employed, that allow each position in the target sequence attends to all positions in the source sequence.
* A similar masked self-attention mechanism, which masks future positions in the decoder, is adopted to make the RGN meet the requirement R3.
* The researchers build the RGN based on the Transformer module, transformer is a sequence-to-sequence model equipped with two types of attention mechanism; self-attention and encoder-decoder attention.
* Determine the generation order of reactants – the generation order of reactants can be determined by aligning reactants in the target with synthons in the source, thanks to intermediate synthons which are associated with reactants uniquely.
* Robustify the RGN – they find that the EGAT suffers from distinguishing multiple coexisting reaction centers, which is the major bottleneck of their method.
* To make their RGN robust enough and able to predict the desired reactants even if the EGAT fails to recognize the reaction center, they further augment RGN training data by including those unsuccessfully predicted synthons on training data.
* They do not reverse the order of synthons for these augmentation samples, the intuition is that EGAT tends to make similar mistakes on training and test datasets since both datasets follow the same distribution.
* This method can make their RGN able to correct reaction center prediction error and generate the desired set of reactants.
* Experiments: - Dataset and preprocessing – the researchers evaluate their method on USPTO-50K and USPTO-full to verify its effectiveness and scalability. They adopt the same training/validation/test splits in 8:1:1 as.
* For the RGN training data, they add an extra 28K samples of which synthons are reversed as shown above figure if there are at least two synthons.
* The USPTO-full consists of 950K cleaned reactions from: the USPTO 1976-2016, which is randomly partitioned into training/validation/test sets in 8:1:1.
* For the EGAT, they build molecule graphs using DGL and extract atom and bond features with RDKit.
* By comparing molecule graphs of product and reactants, they can identify disconnection bonds within the product graph and obtain training labels for both main and auxiliary tasks.
* For reactions without atom-mapping, a substructure matching algorithm in RDKit can be utilized to accomplish the comparison. They use RDChiral to extract super-templates and obtain 1859 semi-templates for USPTO-50K training data, semi-template that appear less than twice are filtered and finally 654 semi-templates are obtained.
* the product molecule graph is divided into synthon graphs according to the ground truth reaction center, then are converted into SMILES strings.
* Implementation – they use DGL and OpenNMT to implement their EGAT and RGN models, respectively. For the EGAT, they stack three identical four-head attentive layers of which the hidden dimension is 128. They train the EGAT on USPTO-50K for 80 epochs.
* EGAT parameters are optimized with Adam with default settings, and the initial learning rate is 0.0005 and it is scheduled to multiply 0.2 every 20 epochs.
* They train the RGN for 300,000-time steps, they save a checkpoint of RGN parameters every 10,000 steps and average the last 10 checkpoints as the final model.
* Evaluation metric: - the Top-*N* accuracy is used as the evaluation metric for retrosynthesis. Beam search strategy is adopted to keep top K predictions throughout the reactant generation process. A correct predicted set of reactants must be exactly the same as the ground truth reactants.
* Reaction center identification results: - to verify the effectiveness of edge-enhanced attention mechanism, they also include the ablation study by removing edge embedding when computing the coefficient = LeakyReLU.
* The auxiliary task (**Aux**) can successfully predict the number of disconnection bonds for 99.2% test molecules given the reaction type (**Type**) while 86.4% if not given.
* For the main task (**Main**) alone, its prediction accuracy is 74.4% w/reaction type and 51.5% wo/reaction type.
* If the researchers adopt the prediction from the auxiliary task as the prior of the prediction from the auxiliary task as the prior of the number of disconnection bonds, and select the most probable disconnection bonds (**EGAT**), then the prediction accuracy can be boosted to 86.0% (w/) and 64.9% (wo/), respectively.
* The improvement is more significant when the reaction type is unknown, the researchers EGAT is more practical in real world applications without reaction types.
* It shows that the reaction type information plays an important role in the retrosynthesis.



* Reactant prediction results: - the researchers also conduct the *P* *S* prediction on the EGAT training data for USPTO-50K (40K), and the prediction accuracy is 89.0% for the reaction type conditional setting.
* For the unconditional setting, the EGAT accuracy is 70.0% and there are 12K augmentation samples, and the total RGN training size is 80K in this case.
* RGN evaluation – the RGN input consists of the ground truth synthons, the proposed augmentation strategy does not always improve the upper bound.
* Without the given reaction type the RGN commonly performs worse with the augmentation due to the introduced dirty training samples.
* The RGN learns to put more attention on the reaction type and product instead of synthons to generate the reactants.
* Retrosynthesis evaluation – to evaluate the overall retrosynthesis prediction accuracy, the generated synthons from *P* *S* instead of the ground truth are input into the RGN. Through this way, the researchers only need to compare the predicted reactants with the ground truth ones, without considering if the reaction center predictions correct or not.



* retroXpert achieves impressive performance on the test data, specifically, when given reaction types, their proposed method achieves 70.4% Top-1 accuracy, which outperforms the SOTA Top-1 accuracy 63.2% by a large margin.
* The top-1 accuracy of the researchers’ model is quite close to the upper bound 73.4% which indicates the proposed augmentation strategy in Robustify the RGN is considerably effective.
* While the researchers’ model outperforms in Top-1, Top-2, and Top-5 accuracy, template-based methods GLN and RetroSim are better at Top-20 and Top-50 predictions since they enumerate multiple different reaction templates for each product to increase the hit rate.
* To increase the diversity, they design new strategies to enumerate multiple reaction centers for each product.
* They notice that the gap between Top-1 and Top-2 accuracy is around 10%, after analyzing them with experience chemists they find that about 9/10 of these Top-1 predictions are actually reasonable. It indicates that their method can learn general chemical reaction knowledge.
* Large scale experiments: - to demonstrate the scalability of their method, they also experiment on the USPTO-full dataset. They extract 75,129 semi-templates and keep only 3,788 ones that appear at least 10 times. They set as 5 to cover 99.87% training data.
* They obtain 1.35M training data after reversing synthons, the final accuracy of the *P* *S* on training set is 60.5%, and there are 0.3M unsuccessful synthons data and the total RNG training data size is 1.65M.
* They train the RNG for 500,000-time steps on USPTO-full while keeping the other settings the same, they run the official implementation of GLN following their instructions, as well as the self-implemented SCROP on the USPTO-full dataset.
* The researchers’ method outperforms the SCROP and GLN, but both template-free methods SCROP and RetroXpert outperform the GLN significantly, which may indicate the scalability of template-based methods is very limited.
* Prediction visualization: - under the guidance of the auxiliary task, EGAT is able to identify the true reaction center. Even if the predicted synthons are different from the ground truth, the RGN still successfully generates desired reactants.



* Discussion: - one major common limitation of current retrosynthesis work is the lack of reasonable evaluation metrics.
* Broader Impact: - the proposed new retrosynthesis method RetroXpert solves the retrosynthesis prediction in two steps like chemists do, and it achieves impressive performance.
* It is template-free and very scalable to the large real-world dataset.

Chaocaho Yan, Qianggang Ding, Peilin Zhao, Shuangjia Zheng, Jinyu Yang, Yang Yu, Junzhou Huang