Project Report
Retrosynthesis reaction prediction architecture inspired by GraphGPS

Doyeon Kim
University of Wisconsin-Madison
dkim676@wisc.edu

1 Introduction

Retrosynthesis is finding organic syntheses with available and simple precursors in reaction path of product molecules and understanding breaking bond and functional group of molecule is important.Sridharan et al. [11] Conventionally, retrosynthesis planing is driven by experience chemists intuition or programs, like Chematica Grzybowski et al. [7], that have thousands of rules that are hand coded by experts. As the computational simulation techniques have developed and advance of computational power, lots of research has been done to understand physic and chemistry in several possible reaction paths between molecules. To be specific, in simulating dynamic of chemical reaction, Molecular Dynamic simulation with potential energy potential (PES) is essential.Schlegel [9] Recently, it is possible to construct PES more efficiently using quantum mechanic simulation data and deep neural network.Zeng et al. [15] Although these two methods are powerful, fast and accurate searching for possible organic molecule reaction pathways is challenging because of overwhelming space of molecules.Chanussot et al. [1] In experiment, research on robotics laboratory to has accelerated new molecules discovery and retrosynthesis planning Granda et al. [6] but technical issues with experiment robots and complex experiment conditions limit studying various possible reaction path. However, as a result of prosperity in both simulation and experiment fields, researchers can access to lots of open source chemical databases. With the development of machine learning techniques, open source chemical and materials databases have been used to search proper reactants of target molecules.Dai et al. [3] Sacha et al. [8] Tetko et al. [12] Moreover, machine learning has helped researches to design particular properties molecules or materials Gómez-Bombarelli et al. [5] Xie et al. [14] but synthesizing new molecules or materials is another challenge to solve. For this reason, in this project, the architecture of LocalRetro Chen and Jung [2], which is state of the art method in retrosynthesis planning, and GraphGPS Thabet et al. [13], which is known for powerful tool describing molecule properties, are used to predict reactants of product molecule. The hybrid architecture is evaluated with USPTO_50K dataset which contains 50,016 reactions from US patents. Schneider et al. [10]

2 Related work

2.1 Retrosysthesis with conditional graph network

Dai et al. [3] uses conditional graph logic network to predict reactants of target molecules. The architecture consists of two steps which are match template step and match reactants step. To be specific, this method uses subgraph and calculates score that scales whether subgraphs are in templates and the product molecule. The first step's template score function calculates probability of whether subgraphs from template matches product's subgraph and the second step's reactant score function obtains probabilty of whether set of subgraphs from template matches set of reactants. This two step probabilistic reasoning gives , the joint probability of retrosythetic proposal.

Preprint. Under review.

2.2 Retrosysthesis with graph attention network

This architecture Sacha et al. [8] uses molecular graph node and adjacency matrix as an input and finds reactants by sequentially editing product molecule. To be specific, the sequential steps are consist of editing atom properties, finding bond between two atoms, adding new atom to the graph, adding new benzene ring to the graph and final stop generation. Each step has encoder and decoder with graph convolution network and atoms and bonds are embedded as one-hot encoding feature. The model finds reactants of product molecule by predicting each step during reaction generation.

2.3 LocalRetro

This approachGilmer et al. [4] first derives a set of local reaction templates from USPTO_50K datasets and this local templates contain before and after the reaction information of changes in the atom and bond. Next step is reaction center is specified after being compared with product molecule and reactants. This architecture first sends featurized atoms and bonds to message passing neural network and updated atoms and bonds features are concatenated into one feature. Concatenated features go through global reactivity attention layer and score that determines changing atoms and bonds is calculated using softmax. This model is trained to predict correct local reaction template at each atom and bond that engages in synthesis by learning product molecules local environments.

2.4 GraphGPS

The characteristic part of this architectureThabet et al. [13] is that it divides molecule features into two parts; positional and structural encodings. Each encoding contains local and global features that represent nodes and relative features for edges and these features go through hybrid message passing neural network (MPNN). In this hybrid step, only the node features passes global attention layers and updated node features with MPNN and updated node feature with transformer are added. Only updating atom (node) feature with transformer does not negatively impact the performance and shows state-of-the-art result in predicting molecules properties also this method improves computational efficiency.

3 Approach

The motivation of this experiment is combining LocalRetro Gilmer et al. [4] and GraphGPS architecure to inspect the impact of applying global attention transformer only in atoms features. The hybrid architecture follows schematic diagram of 1. LocalRetro featurizes molecules into graph. Experiment is done on UPSTO_50K dataset with the hybird model with 1,2 and 3 GPS layer. GPS layer is stacked so updated atom features are fed into message passing neural network again and during GPS loop bond features do not get updated. After GPS loop, bond features are updated with atom features dense layer. Finally updated both features go through dense layer and Softmax layer to get score for each atom and bond in the product molecule. The high score stands for high possibility of change during chemical reaction. Evaluation is comparing predicted reactants from test set ground truth and measuring percentage of molecules that exactly predict reactants. This is called Exact accuracy. Training/Validation/Test set are devided as 40,000/5,000/5,000. Since score is calculated for each bond and atoms, we can find exact reactant by looking through top 1,3,5,10 and 50 candidates. Moreover, after training, the models predict reactants of Lenalidomide, Salmeterol, 5-HT6 receptor ligand, DDR1_037 and DDR1_032 and top score reactants are compared with LocalRetro architecture and hybrid architecture with 1, 2 and 3 laeyers of GPS.

4 Results

The exact accuracy result does not show much difference depends on GPS layer in Table. 1 also computational time of each model is approximately same. However, predicted reactants of five molecules are different 2 and GPS 3 layers have problem of score diminishing. As GPS layers are added, the score decreases and crashes in GPS 3 layers model and score of GPS 3 layers is negligible that the model cannot find DDR1_037 and DDR1_032 reactants. GPS 2 layers has same result with LocalRetro setting but GPS 1 layer separates Salmeterol reactant into two and DDR1_032 reactants have different bond than LocalRetro and GPS 2 layer trained model.

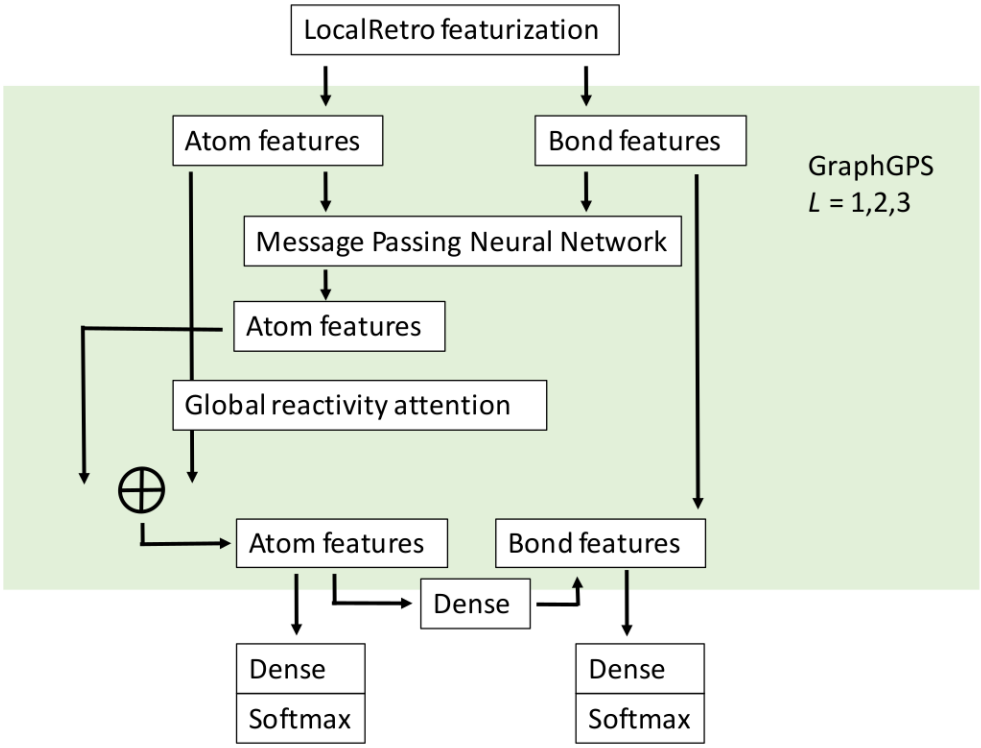


Figure 1: This is hybrid architecture of LocalRetro and GraphGPS. Product molecule is featurized into atoms and bond features and both features go through message passing neural network (MPNN). After MPNN layer atom features are updated and previous atom features pass global reactivity attention (GRA) layer. Atom features updated with MPNN and GRA are added. The green box shows GPS architecture from GraphGPS and GPS layer is stacked from 1, 2 and 3 layers for experiment. Bond features are not updated until atom features pass GPS layers and updated atom features updates bond features by dense layer. Both features obtain score, that estimates possible changes on atoms and bonds, through dense and softmax layer.

	Exact accuracy				
Method	**Top-1**	**Top-3**	**Top-5**	**Top-10**	**Top-50**
LocalRetro	0.641	0.873	0.928	0.970	0.985
GPS 1 layer	0.631	0.867	0.928	0.966	0.985
GPS 2 layers	0.637	0.877	0.931	0.969	0.986
GPS 3 layers	0.643	0.869	0.931	0.967	0.983

Table 1: This is architecture of hybrid architecture of LocalRetro and GraphGPS. Molecule is featurized in to atoms and bond features and both features go through message passing neural network (MPNN). MPNN updates atom features and

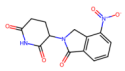
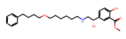
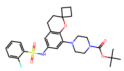
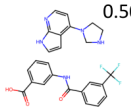
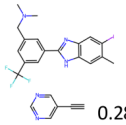
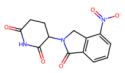
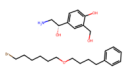
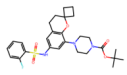
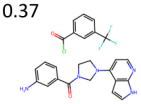
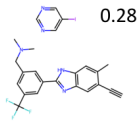
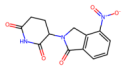
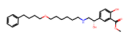
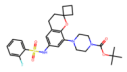
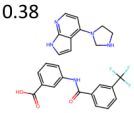
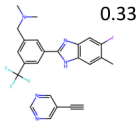
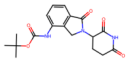
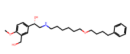
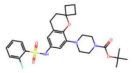
	Lenalidomide	Salmeterol	5-HT6 receptor ligand	DDR1_037	DDR1_032
Local Retro	 0.87	 0.27	 0.57	0.50 	 0.28
GPS 1 layer	 0.89	 0.23	 0.59	0.37 	0.28 
GPS 2 layers	 0.87	 0.26	0.43 	0.38 	0.33 
GPS 3 layers	0.016 	0.023 	0.019 	None	None

Figure 2: Predicted reactants of five molecules from trained models. The number under the molecule is score.

5 Discussion

Since GPS layer added without updating bond feature, the result does not show significant different between LocalRetro and GPS layer added architecture. Also computational time of models to reach certain loss is also similar between models and it is possible that the molecules in UPSTO_50K are not big enough to experience computational advantage of decoupling atom and bond features before applying to attention layer. As stacking GPS layers can cause diminishing of scores of atom and bond, the skip connection to add up the previous score after GPS layer can help properly training score.

6 Future Work

In this work, OI only used GPS layers similar with GraphGPS but the featurizing molecules are still the method of LocalRetro. It would be interesting to find GraphGPS's local and global features affect the result of retrosynthesis plaining. Moreover, computational simulation to find which architecture actually performs better in predicting physically and chemically reliable reactants.

References

[1] Lowik Chanussot, Abhishek Das, Siddharth Goyal, Thibaut Lavril, Muhammed Shuaibi, Morgane Riviere, Kevin Tran, Javier Heras-Domingo, Caleb Ho, Weihua Hu, et al. Open catalyst 2020 (oc20) dataset and community challenges. *ACS Catalysis*, 11(10):6059–6072, 2021.

[2] Shuan Chen and Yousung Jung. Deep retrosynthetic reaction prediction using local reactivity and global attention. *JACS Au*, 1(10):1612–1620, 2021.

[3] Hanjun Dai, Chengtao Li, Connor Coley, Bo Dai, and Le Song. Retrosynthesis prediction with conditional graph logic network. *Advances in Neural Information Processing Systems*, 32, 2019.

[4] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pages 1263–1272. PMLR, 2017.

[5] Rafael Gómez-Bombarelli, Jennifer N Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D Hirzel,

Ryan P Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 4(2):268–276, 2018.

[6] Jarosław M Granda, Liva Donina, Vincenza Dragone, De-Liang Long, and Leroy Cronin. Controlling an organic synthesis robot with machine learning to search for new reactivity. *Nature*, 559(7714):377–381, 2018.

[7] Bartosz A Grzybowski, Sara Szymkuć, Ewa P Gajewska, Karol Molga, Piotr Dittwald, Agnieszka Wołos, and Tomasz Klucznik. Chematica: a story of computer code that started to think like a chemist. *Chem*, 4(3):390–398, 2018.

[8] Mikołaj Sacha, Mikołaj Błaz, Piotr Byrski, Paweł Dabrowski-Tumanski, Mikołaj Chrominski, Rafał Loska, Paweł Włodarczyk-Pruszynski, and Stanisław Jastrzebski. Molecule edit graph attention network: modeling chemical reactions as sequences of graph edits. *Journal of Chemical Information and Modeling*, 61(7):3273–3284, 2021.

[9] H Bernhard Schlegel. Exploring potential energy surfaces for chemical reactions: an overview of some practical methods. *Journal of computational chemistry*, 24(12):1514–1527, 2003.

[10] Nadine Schneider, Nikolaus Stiefl, and Gregory A Landrum. What's what: The (nearly) definitive guide to reaction role assignment. *Journal of chemical information and modeling*, 56 (12):2336–2346, 2016.

[11] Bhuvanesh Sridharan, Manan Goel, and U Deva Priyakumar. Modern machine learning for tackling inverse problems in chemistry: molecular design to realization. *Chemical Communications*, 58(35):5316–5331, 2022.

[12] Igor V Tetko, Pavel Karpov, Ruud Van Deursen, and Guillaume Godin. State-of-the-art augmented nlp transformer models for direct and single-step retrosynthesis. *Nature communications*, 11(1):1–11, 2020.

[13] Slimane Thabet, Romain Fouilland, and Loic Henriet. Extending graph transformers with quantum computed aggregation. *arXiv preprint arXiv:2210.10610*, 2022.

[14] Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi Jaakkola. Crystal diffusion variational autoencoder for periodic material generation. *arXiv preprint arXiv:2110.06197*, 2021.

[15] Jinzhe Zeng, Liqun Cao, Mingyuan Xu, Tong Zhu, and John ZH Zhang. Complex reaction processes in combustion unraveled by neural network-based molecular dynamics simulation. *Nature communications*, 11(1):1–9, 2020.