Learning Discrete Neural Reaction Class to Improve Retrosynthesis Prediction

Théophile Gaudin[1,2,3], **Yuhuai Wu**[2,3], **Robert Pollice**[3,4],
Animesh Garg[2,3], **Alán Aspuru-Guzik**[2,3,4,5]
[1]IBM Research Zürich, Switzerland
[2]Vector Institute for Artificial Intelligence, Toronto, Ontario, Canada.
[3]Department of Computer Science, University of Toronto, Canada
[4]Department of Chemistry, University of Toronto, Toronto, Ontario, Canada.
[5]Lebovic Fellow, Canadian Institute for Advanced Research (CIFAR), Toronto, Ontario, Canada.
tgaudin@cs.toronto.edu, ywu@cs.toronto.edu
rob.pollice@utoronto.ca, garg@cs.toronto.edu, alan@aspuru.com

Abstract

Computer-aided retrosynthesis accelerate and innovate the process of molecule and material design, allowing the discovery of new pathways and automating part of the overall development process for drugs and materials. Current machine-learning methods applied to retrosynthesis are limited by their lack of control when generating single-step reactions as they rely on sampling or beam search algorithm. In this work, we apply vector quantized representation learning [1] to learn reaction classes along with retrosynthetic predictions. We represent each reaction class with a vector allowing us to condition the retrosynthetic prediction. We show that learning reaction classes increases control as well as generating more diverse predictions than a baseline model. Our results are a significant step forward in the development of multistep retrosynthesis prediction.

1 Introduction

With the emergence of automated labs and material acceleration platforms, there is a need for a reliable automated synthesis planning algorithm to further speed up new material discovery. First formalized by Corey [2], retrosynthesis is a way to approach synthesis backwards and decompose a target molecule sequentially into simpler compounds. This approach was adapted in all computational retrosynthesis methods that use machine learning. Usually, these methods consist of two algorithms: a single-step model predicting the possible precursors given a target molecule and a search algorithm that combines the single-step predictions into a full synthesis.

We can distinguish three types of single-step models: template-based methods [3, 4, 5], graph-based methods [6, 7, 8] and methods based on textual representations of molecules (SMILES) [9, 10]. Template-based methods rely on reaction templates extracted from data. These templates are then matched against a target molecule and the results are typically ranked by a neural network. This type of method suffers from an inability to generalize and does not scale well. The majority of graph-based methods split the problem in two subtasks, first identifying which bond to break and then completing the partial molecule into reactants. While this approach showed promising performance in most cases, it does not cover protection strategies that are crucial for complicated multistep retrosynthesis. Finally, methods using textual representations such as SMILES rely on models developed for neural machine translation, in particular the Transformer network [11, 12, 9]. This approach has the advantage of handling the stereochemistry in a robust way compared to graph-based methods. Notably, reaction

Fourth Workshop on Machine Learning and the Physical Sciences (NeurIPS 2021).

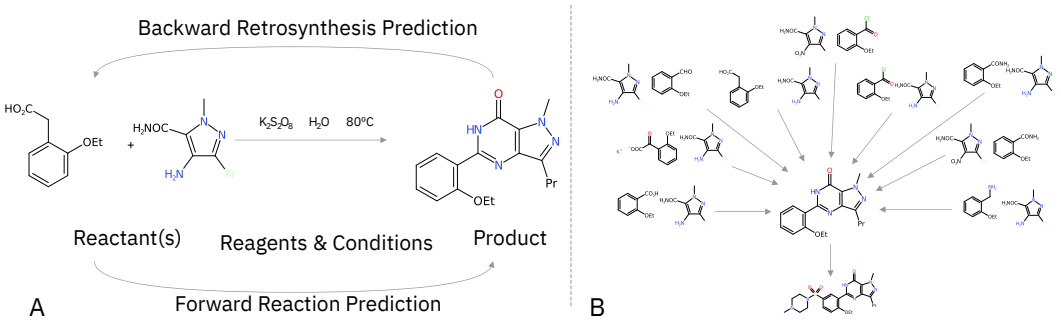

Figure 1: A typical reaction in organic chemistry (a) and the multimodality of retrosynthesis (b)

classes were used in a recent work [10] to succesfully condition the retrosynthesis model to improve its performance.

This work focuses on the prediction of possible compounds given a target molecule with an emphasis on its application in the context of a multistep retrosynthesis setup. We propose a learnable model where we learn reaction classes along with the retrosynthesis prediction, successfully making the generation process more controllable and thus more reliable. We extend the SMILES based method to learn vector quantized reaction classes and condition the generation of the reactants using these classes. We demonstrate that learning this classes improves the diversity of the relevant generated reactants.

2 Problem Formulation

We can decompose an organic reaction into four pieces of information: reactants, i.e. the molecules to start the reaction with, products, i.e. the molecules that are obtained after the reaction, reagents which enable the reaction such as catalysts or solvents, and reaction conditions, i.e. temperature, reaction time and pressure. An exemple of a reaction is depicted in Figure 1a. Reaction prediction consists of predicting products given sets of reactants, whereas a computational retrosynthesis corresponds to predicting the reactants given a product. Typically, reaction prediction only allows for a very limited number of possible solutions. In contrast, retrosynthesis prediction is a multimodal problem with many equally valid solutions as demonstrated in Figure 1b.

In this work, we represent all molecules using SMILES. Let $\mathbf{X}$ and $\mathbf{Y}$ be the sequences of the product and reactants respectively. Our goal is to predict $\mathbf{Y}$ given $\mathbf{X}$.

3 Methods

Our approach consists of augmenting a single step model by learning reaction classes. We hypothesized that learning to cluster the reactions by type along with the retrosynthesis prediction would enable sampling more diverse retrosynthetic routes. While we used our framework in conjunction with Transformers, it could be applied equally well to graph-based methods without significant changes in architecture.

3.1 Model Architecture

Our model consists of four main components: a product encoder, a reaction class encoder, a reaction class predictor and a reactant decoder. The product encoder and the reaction class predictor are parametrized as a Transformer Encoder, the reactant decoder as a Transformer Decoder and the reaction class predictor is a multilayer perceptron. We use the hidden representation corresponding to the "end-of-string" token as an aggregate representation of the full sequence. The weights of the product encoder and the reaction class encoder are shared. Figure 2 provides a schematic overview the pipeline of our model.

The product encoder $pe(\mathbf{X})$ encodes the product sequence into a hidden representation $\mathbf{h_x}$. The reaction class predictor $rcp(h_x^{eos})$ predicts a distribution over all the classes given the representation

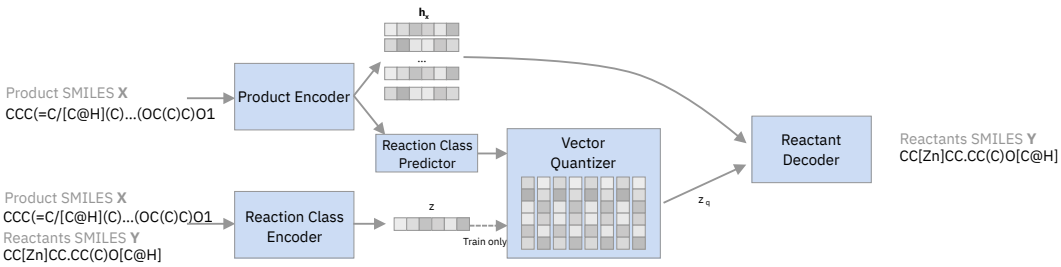

Figure 2: Overview of our model's pipeline

corresponding to the "end-of-string" token. The reaction class encoder $rce([\mathbf{X}, \mathbf{Y}])$ takes a concatenation of the product $\mathbf{X}$ and the reactants $\mathbf{Y}$ and outputs z a single vector representation of the full reaction (which is the hidden representation of the eos token of the reactants). Subsequently, this representation is quantized into z_q, similarly to [1]. We used 128 quantized reaction classes for all our experiments. Finally, the reactant decoder takes $\mathbf{h_x}$ and z_q and defines a distribution over the possible reactants.

At training time, z_q is defined using the reaction class encoder rce, whereas at inference time we use the most likely reaction class according to the reaction class predictor rcp.

3.2 Learning

The training loss of our model has three terms: a reaction prediction term, formalized by the negative log likelihood, a vector quantization loss similar to the one defined in the literature [1] and a reaction negative log likelihood.

$$\mathcal{L} = \log p_\theta(Y|z_q, \mathbf{h_x}) + \log p_\phi(rcp(pe(X))|rce([X,Y])) + \beta||rce([X,Y]) - e||^2 + ||e - rce([X,Y])||^2$$

To avoid the index collapse, where only a couple of classes are used, we pretrained the model using only the retrosynthetic loss for a 5 epochs.

4 Experimental Setup

4.1 Dataset

For this work, we used the CJHIF dataset [13]. It consists of 3M organic chemistry reactions. We used the SMILES as provided in the dataset and discarded all the information about the reagents as they were provided in various formats. Then, the 3M reactions were filtered according to their yield, keeping only non zero values. The resulting dataset consists of 1.7M reactions and has been split in train, validation and test sets in a 90/5/5 ratio.

4.2 Baseline Model

Our baseline model is a Transformer network and we used the fairseq implementation [14] matching the dimensions of our proposed model. Training of both models was carried out for a similar number of epochs and we used temperature sampling to generate retrosynthetic predictions. Notably, temperature sampling was favored over beam search as the former produces more diverse solutions.

4.3 Evaluation Metrics

The main metric used to compare single-step retrosynthetic models in the literature is the full sequence accuracy on the validation set. However, we argue that this metric is not appropriate for retrosynthesis prediction as inherently it cannot reflect its multimodal nature. Indeed, a model may predict a valid solution that the full sequence accuracy would invalidate as it only accepts a single solution. Furthermore, it may also be prone to overfit to shorter sequences. Hence, we propose the perplexity of the model to be a better metric.

For evaluation, we verify the correctness of a retrosynthetic prediction using a forward synthesis model trained on the same dataset. Similar approaches have already been proposed in previous work [9] and it has also been studied in nlp [15]. To account for the many possible outcomes of a reaction, we used a beam search of width 5. A prediction was considered correct when its product can be correctly backtranslated by the forward model. Importantly, we also verified the uniqueness of each solution using RDKIT to canonicalize the SMILES.

5 Experiments

The correctness and uniqueness of the predictions sampled from both the baseline and our model are reproduced in Table 1. For the baseline model, we used temperature sampling and sampled 128 solutions. For our model, we conditioned the generation on each of the 128 reaction classes and used greedy decoding. As can be seen, our model is better at generating more diverse and correct solutions.

Table 1: Comparison between the performance of the baseline model (Transformer network) and our model. T denotes the temperature used to sample. Using quantized classes improves the number of unique solutions generated.

Model	Correct	Correct and Unique
Baseline (T=1.0)	64.58	11.97
Baseline (T=1.5)	39.49	13.76
Baseline (T=1.75)	26.11	11.07
Transformer + VQ classes	35.04	**22.64**

We performed another experiment to understand the reaction classes better. Hence, we used the reaction class encoder to classify the reactions of the validation set and inspected the outcome. Some reaction classes seem to correspond to the presence of a particular functional group whereas other classes are more difficult to interpret. Figure 3 shows some randomly selected examples classified using the reaction class encoder. While we can see some similarity between the reactions, it is hard to draw any conclusion regarding the interpretability of learned reaction classes.

6 Conclusion

Overall, this work proposes a framework to improve the generation of retrosynthetic predictions using vector quantized reaction classes. Importantly, we have shown an improvement in the diversity of the predictions over a baseline model.

References

[1] Oord, A. v. d., Vinyals, O. & Kavukcuoglu, K. Neural Discrete Representation Learning. *arXiv:1711.00937 [cs]* (2018). URL `http://arxiv.org/abs/1711.00937`. ArXiv: 1711.00937.

[2] Corey, E. J. The Logic of Chemical Synthesis: Multistep Synthesis of Complex Carbogenic Molecules (Nobel Lecture). *Angewandte Chemie International Edition in English* **30**, 455–465 (1991). URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/anie.199104553`. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.199104553.

[3] Segler, M. H. S., Preuss, M. & Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **555**, 604–610 (2018). URL `https://www.nature.com/articles/nature25978`. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 7698 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Automation;Cheminformatics;Computer science;Synthetic chemistry methodology Subject_term_id: automation;cheminformatics;computer-science;methodology.

[4] Coley, C. W., Rogers, L., Green, W. H. & Jensen, K. F. Computer-Assisted Retrosynthesis Based on Molecular Similarity. *ACS central science* **3**, 1237–1245 (2017).

[5] Dai, H., Li, C., Coley, C., Dai, B. & Song, L. Retrosynthesis Prediction with Conditional Graph Logic Network 11 (2019).

[6] Somnath, V. R., Bunne, C., Coley, C. W., Krause, A. & Barzilay, R. Learning Graph Models for Template-Free Retrosynthesis 7 (2020).

[7] Shi, C., Xu, M., Guo, H., Zhang, M. & Tang, J. A Graph to Graphs Framework for Retrosynthesis Prediction. *arXiv:2003.12725 [cs, stat]* (2021). URL `http://arxiv.org/abs/2003.12725`. ArXiv: 2003.12725.

[8] Jin, W., Barzilay, R. & Jaakkola, T. Junction Tree Variational Autoencoder for Molecular Graph Generation. In *arXiv:1802.04364 [cs, stat]* (2018). URL `http://arxiv.org/abs/1802.04364`. ArXiv: 1802.04364.

[9] Schwaller, P. *et al.* Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chemical Science* **11**, 3316–3325 (2020). URL `https://pubs.rsc.org/en/content/articlelanding/2020/sc/c9sc05704h`. Publisher: The Royal Society of Chemistry.

[10] Zheng, S., Rao, J., Zhang, Z. & Xu, J. Predicting Retrosynthetic Reaction using Self-Corrected Transformer Neural Networks 19 (2020).

[11] Vaswani, A. *et al.* Attention Is All You Need. *arXiv:1706.03762 [cs]* (2017). URL `http://arxiv.org/abs/1706.03762`. ArXiv: 1706.03762.

[12] Karpov, P., Godin, G. & Tetko, I. V. A Transformer Model for Retrosynthesis. In Tetko, I. V., Kůrková, V., Karpov, P. & Theis, F. (eds.) *Artificial Neural Networks and Machine Learning – ICANN 2019: Workshop and Special Sessions*, Lecture Notes in Computer Science, 817–830 (Springer International Publishing, Cham, 2019).

[13] Jiang, S. *et al.* When SMILES Smiles, Practicality Judgment and Yield Prediction of Chemical Reaction via Deep Chemical Language Processing. *IEEE Access* **9**, 85071–85083 (2021). Conference Name: IEEE Access.

[14] Ott, M. *et al.* fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations* (2019).

[15] Edunov, S., Ott, M., Auli, M. & Grangier, D. Understanding Back-Translation at Scale. *arXiv:1808.09381 [cs]* (2018). URL `http://arxiv.org/abs/1808.09381`. ArXiv: 1808.09381.

A Appendix

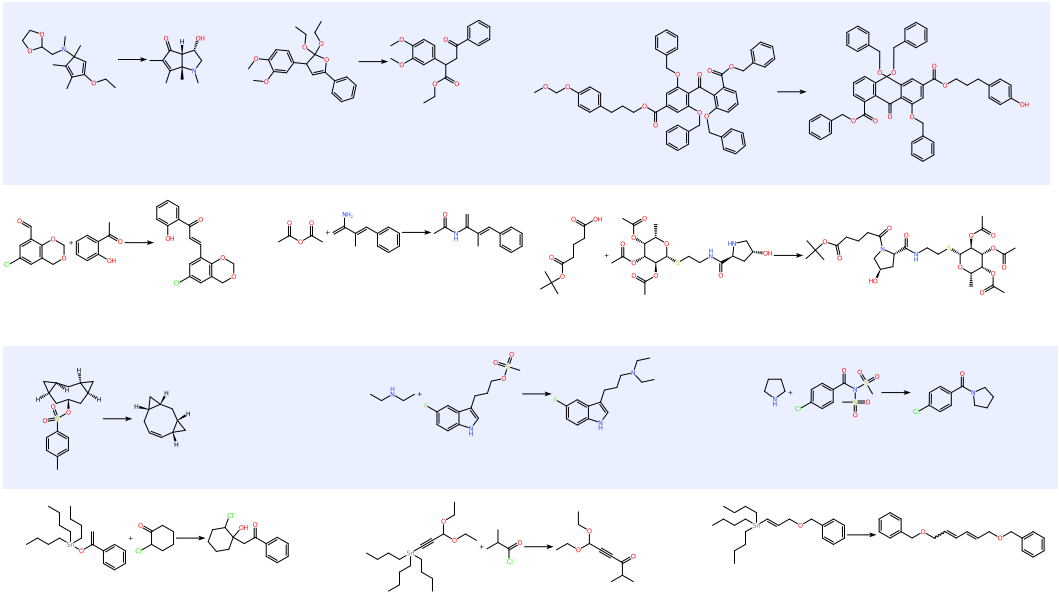

Figure 3: Reactions grouped by learned reaction classes. Each line correspond to a different reaction class. Some similitude can be observed within a reaction class.