

## 敵対的生成ネットワークを用いた逆合成の経路探索

<sup>1</sup>株式会社トヨタIT開発センター, <sup>2</sup>東京大学物性研究所

○福島真太郎<sup>1</sup>, 本山裕一<sup>2</sup>, 吉見一慶<sup>2</sup>

## Retrosynthetic Planning with Generative Adversarial Network

○Shintaro Fukushima<sup>1</sup>, Yuichi Motoyama<sup>2</sup>, Kazuyoshi Yoshimi<sup>2</sup>

<sup>1</sup>TOYOTA InfoTechnology Center Co., Ltd., Japan

<sup>2</sup>Institute for Solid State Physics, The University of Tokyo, Japan

**【Abstract】** Recently, retrosynthetic planning with machine learning and deep learning has been proposed actively. We focus on the method proposed by Coley et al. In this method, we first calculate similarities between the target product and products in the reaction database to find similar products. Next, we generate candidate reactions by modifying reactions of the similar targets. The method by Coley et al. is more accurate than other methods. However, its search space is limited because it is based on the matching with the existing reactions.

In this presentation, we propose a method with GAN(Generative Adversarial Network) in order to expand the search space. The idea of the proposed method is to learn a generative model with GAN, generate reactants with the generative model, and then a reaction with reaction prediction. We got new reactants with the proposed method. We continue the detailed calculation of reaction prediction and retrosynthetic planning.

**【序】**近年, 機械学習や深層学習を用いた逆合成の経路探索が活発になっている[1-3]. 本研究では比較的簡便な Coley et al.[1]の方法に着目する. この方法は目標とする生成物と反応データベースのマッチングを行い, 生成物と反応物の類似度をそれぞれ計算する. そして, その類似度に基づき一段階前の反応候補のランキングを出力する. Coley et al.の方法は既存の方法に比べて比較的高精度である. 一方で, 既存の反応とのマッチングに基づくため, 探索空間は過去に生じた反応に限定される.

本発表では以上の問題点に着目し, 探索空間を広げるために深層学習の生成モデルである敵対的生成ネットワーク(Generative Adversarial Network, GAN)[4]を用いた手法を提案する. この手法のアイデアは, 探索空間を広げるために GAN に基づく反応物の生成モデルを学習し, 反応物を生成した後に順方向の反応予測を行って反応式を生成することにある. その結果, 既存の反応データベースには存在しない反応物が生成された. 現在, 反応予測, および逆合成の探索の詳細な計算を進めており, 当日は検証結果も合わせて報告する.

**【方法】**提案手法では, Fig.1 に示すように 4 ステップにより逆合成の経路探索を行う. Step1~3 は反応データベースに含まれる反応式を増加させるために行い, 最後の Step4 は反応式の候補をランキングし逆合成の経路を探索するために行う.

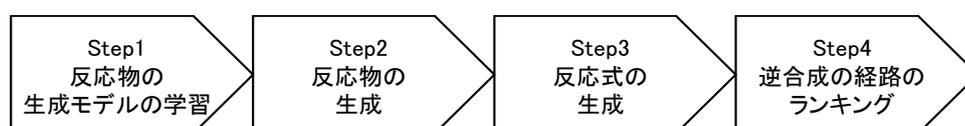


Fig. 1. Calculation flow of the proposed method

### Step1 反応物の生成モデルの学習

与えられた反応データベースに存在する反応物を用いて、反応物の生成モデル (generative model) を学習する。この生成モデルの学習には、反応物の SMILES 表現[5] を入力とし、SeqGAN(Sequential GAN)[6]等、系列データに対して適用可能な GAN を用いる。

### Step2 反応物の生成

Step1 で学習した反応物の生成モデルを用いて、反応物 A, B を生成する。

### Step3 反応式の生成

Step2 で生成した反応物 A と B を反応させて得られる生成物 C を予測する。その結果、反応式「A + B → C」を得る。反応の予測には、Sequence-to-Sequence[7]を用いた順方向の反応予測の方法[8]等を用いる。また、反応予測の結果、反応物 A, B が反応しないと判断された場合は A, B を棄却する。

Step2, 3 は指定した個数  $N$  の反応式が生成されるまで反復する。

### Step4 逆合成の経路のランキング

与えられた反応データベースに Step1~3 で生成した  $N$  個の反応式を追加する。そして、Coley et al.の方法を用いて目標とする生成物に対する逆合成の経路をランキングする。

Coley et al.の方法の概要は以下のとおりである。目標とする生成物 P と反応データベースに存在する生成物 P' の類似度を  $s_1$  とする。次に、P' を生成する反応式 R' の反応中心に着目し、P を生成する反応式の候補 R を生成する。このとき、R と R' の類似度を  $s_2$  とする。生成物と反応物を合わせた類似度  $s = s_1 s_2$  として、類似度  $s$  に基づいて生成物 P に対する反応式 R のランキングを行う。

なお、生成物や反応式に含まれる反応物は ECFP(Extended Connectivity Fingerprint)[9]等の Fingerprint で表現し、類似度  $s_1$  および  $s_2$  は Dice 類似度、Tanimoto 類似度、Tversky 類似度等を用いて計算する。

**【結果・考察】** Coley et al.で使用されたデータを反応データベースとして、提案手法の有効性を検証した。このデータは米国の特許の大規模なコーパス[10]から 50,000 件の反応式を抽出したものであり、各反応式は 10 個のカテゴリのいずれかに分類されている。

Step1, 2 を実行した結果、元々の反応データベースには存在しない反応物が生成された。これらの反応物を用いると、Step3 で反応データベースには存在しない反応式の生成が可能となり、結果として Step4 で逆合成の経路をランキングする際に既存の反応データベースには存在しない経路を列挙できる可能性がある。現在詳細な計算を進め、本手法に関する検証を進めている。

### 【参考文献】

- [1] C.W. Coley et al. *ACS. Cent. Sci.* **3**(12), 1237 (2017).
- [2] M. Segler et al. *Nature* **555**(7678), 604 (2018).
- [3] B. Liu et al. *ACS. Cent. Sci.* **3**(10), 1103 (2017).
- [4] I. Goodfellow et al. *NIPS2014*, 2672 (2014).
- [5] D. Weinberger et al. *J. Chem. Inf. Model.*, **28**, 31 (1988).
- [6] L. Yu et al. *AAAI2017*, 2852 (2017).
- [7] I. Sutskever et al. *NIPS2014*, 3104 (2014).
- [8] P. Schwaller et al. *Chem. Sci.*, **9**, 6091 (2018).
- [9] D. Rogers and M. Hahn, *J. Chem. Inf. Model.*, **50**, 742 (2010).
- [10] D. Lowe et al. *PhD Thesis, University of Cambridge* (2012).