氏　　名　郭 中樑

学位(専攻分野)　博士(統計科学)

学 位 記 番 号　総研大甲第 2236 号

学位授与の日付　2021年3月 24日

学位授与の要件　複合科学研究科　統計科学
学位規則第6条第1項該当

学 位 論 文 題 目　Bayesian inference for chemical synthesis planning

論 文 審 査 委 員　主　査　日野 英逸
統計科学専攻　教授
持橋 大地
統計科学専攻　准教授
Stephen WU
統計科学専攻　准教授
吉田 亮
統計科学専攻　教授
瀧川 一学
理化学研究所 革新知能統合研究センターAIP
iPS 細胞連携医学的リスク回避チーム　研究員

（様式３）

博士論文の要旨

氏　　名　郭　中樑

論文題目　Bayesian inference for chemical synthesis planning

In organic chemistry, predicting the products from the reactants is called reaction prediction, while the design of synthetic routes in the opposite direction from the final products, which are the target molecule, is called chemical synthesis planning. Reaction prediction and chemical synthesis planning have been studied for more than 50 years. In recent years, advances in machine learning have significantly improved the accuracy of reaction prediction and chemical synthesis planning. However, most researches studied the chemical synthesis planning as a separate problem from reaction prediction. Machine-learning models were trained to predict the reactants from a given product directly. As in most reactions, the products consist of target compound and other side products, it is an ill-posed problem that predict the reactants from the target compound backwardly without the information of side products. This ill-posed nature of the backward prediction induced a limited predictive power of backward prediction models. Compared to the prediction accuracy of over 90% for forward prediction models that predict the products from the reactants, previous reported accuracy for the backward prediction models ranged from 37% to 52%. In addition, the majority of candidate reactants simulated from such backward prediction models are rarely contained within a given set of purchasable compounds that span the feasible solution space. For example, if a synthetic target is decomposed into A and B by a backward reaction prediction model, both the reactants will typically be non-purchasable. In such a case, further identification of synthesis routes to both A and B will be necessary.

In this thesis, we redefine the problem of chemical synthesis planning as a combinatorial optimization task with the solution space subject to the combinatorial complexity of all possible pairs of purchasable reactants. We propose a two-stage approach consisting of forward and backward predictions to solve the chemical synthesis planning. A trained forward model with high predictability defines the mapping $Y = f(S)$ from a set of reactants S to their product Y. By solving the inverse mapping $S = f^{-1}(Y^*)$ with a synthetic target Y^* with respect to possible combinations $\mathcal{S}$ of commercially available reactants, we could obtain an algorithm for chemical synthesis planning that has a high synthetic accessibility. As machine learning models are not perfect prediction models, it is possible that all candidate reactants will never reach the target compound with the given forward model. Furthermore, if the model is

incorrect, true reactants are expected to be close to the optimal solution. Considering that the ultimate goal of chemical synthesis planning is to enumerate all possible reaction routes and to facilitate the creativity of the chemists, we addressed the

$$p(S|Y = y^*) \propto p(Y = y^*, S) = p(Y = y^*|S)p(S)$$

problem of reaction mining within the framework of Bayesian inference:

the posterior is a discrete probability distribution that define the probability of reactants S that can synthesis a give target compound y^*. The posterior probability is proportional to the joint distribution, which is formed by the forward prediction model. The support of the posterior consists of all possible combinations of reactants involved in a synthetic route. As exact computation across all candidates is infeasible, the primary objective in the Bayesian computation is to identify a reduced set of reactant combinations with large joint probability, while those with ignorable probability are effectively eliminated. A diverse candidate set can help chemists to find appropriate reaction route to synthesis the target compound.

To enhance the search efficiency and exhaustively enumerate alternative pathways, a sequential Monte Carlo algorithm to sample in the discrete chemical space were developed. A cluster-level resampling was introduced to prevent the particle impoverishment in the resampling step. In addition, a surrogate model was used to save the cost of repeatedly evaluating the computationally expensive forward prediction model. Using a forward model prediction accuracy of approximately 87%, the Bayesian retrosynthesis algorithm successfully rediscovered 81.8 and 33.3% of known synthetic routes of one-step and two-step reactions, respectively, with top-10 accuracy. Remarkably, as the Monte Carlo algorithm is specifically designed to exhaustively explored highly probably reaction sequence ending with a given synthetic target, over 500 synthetic routes on average for each target were identified by the Bayesian retrosynthesis algorithm. In addition, we investigated the potential applicability of such diverse candidates based on expert knowledge of organic chemistry and revealed the influence of the publication bias in reaction datasets.

Form8・Separate Sheet
（様式8・別紙1）

Results of the doctoral thesis screening

博士論文審査結果

Name in Full
氏名　郭　中樑

Title
論文題目　Bayesian inference for chemical synthesis planning

申請論文は全6章70頁からなる．テーマは，有機化合物の合成経路を設計する機械学習の方法論とその概念実証である．合成反応の大量データを用いて，機械翻訳用に開発されたニューラルネットワーク（Transformer）を訓練する．このモデルを用いることで，任意の反応物の組に対し，生成物の化学構造を約90%の精度で予測できる．さらに，この順方向の合成反応モデルの数学的逆写像を求めることで，所望の生成物を合成する反応物の組を同定する．本論文では，このワークフローをベイズ推論の枠組みに帰着させ，逆問題を解くことを提唱している．事後分布の定義域は，購入可能な反応物リストの全ての組み合わせからなる．ここで，反応のステップ数をpとすれば，候補経路の数は10^{6p}のオーダーとなる．この離散確率分布から多様な合成経路を効率的にサンプリングするための逐次モンテカルロ法のアルゴリズムを開発した．

各章の概要は，以下の通りである．

第1章では，50年前に登場したルールベース型AIよる合成経路設計の研究から2018年頃を境に急速に活発化した深層学習に基づく研究の流れを概説したのち，本研究の学術的位置付けを明確にしている．

第2章は，有機合成化学における予測のタスクを順問題と逆問題に分類した上で，先行研究のサーベイを行っている．

第3章では，提案手法であるベイズ推論に基づく合成経路設計の問題の定式化とアイデアの特徴や新規性を説明している．順方向と逆方向の予測からなる2段階アプローチで問題を解くことが提案手法の新規性の一つである．これにより，既存手法の致命的な欠点であった不良設定問題を回避できることを示している．

第4章では，逐次モンテカルロ法の解説，数値実験による性能検証，先行研究との比較実験をまとめている．さらに，発掘された合成経路に対し，有機合成化学の専門家が包括的なレビューを実施している．その結果に基づき，候補経路の内，約35%が化学的に妥当であるという結論を得ている．

第5章は，まとめの章であり，第6章は今後の展望を議論している．

[論文の評価]

合成経路設計の研究の起源は，約50年前にElias James Corey博士が提唱したルールベース型AIに遡る．その後，「逆合成解析」という呼称のもとで技術的な発展を遂げてきたが，専門家の知識の範囲外の予測が難しいということが問題視されてきた．一方，2018年頃を境に大量の合成反応のデータから訓練されたニューラルネットワークを用いることで，ルールベース型の限界の突破を図るという研究が急速に活発化してきた．そのような学術

的潮流の中，本研究は以下の 3 点において重要な貢献を果たした．

(1) 順方向と逆方向の予測からなる 2 段階アプローチで問題を解くことで，既存手法の致命的な欠点である不良設定問題を回避できることを示した．これにより予測精度の大幅な改善を実現した．

(2) 膨大な候補点の上に定義される離散確率分布の近似計算アルゴリズムを開発した．これにより，既存手法では発見できなかった多様な合成経路を検出できるようになった．

(3) 有機合成の専門家による包括的な評価実験を実施し，予測された合成経路の内，約 35%の候補が化学的に妥当であることを実証した．

有機化合物の新規合成手法の開発は，有機化学や化学産業の発展の原動力となってきた．本研究は，技術面においていくつかの改善すべき点が残されているが，有機合成化学に対して重要な学術的貢献をもたらす可能性を持つ．また，統計科学の観点においても学術的新規性が十分に認められる．

[その他]

第 3 章，第 4 章の内容をまとめた論文が査読付きジャーナル Journal of Chemical Information and Modeling 誌（Impact Factor 2019: 4.549，第一著者）に掲載されている．