氏　　名　Zhang Qi

学位(専攻分野)　博士(統計科学)

学 位 記 番 号　総研大甲第 2452 号

学位授与の日付　2023 年 9 月 28 日

学位授与の要件　複合科学研究科　統計科学専攻
学位規則第6条第1項該当

学 位 論 文 題 目　Machine learning for de novo design of functional molecules and their synthetic routes

論 文 審 査 委 員　主　査　中野 慎也
統計科学コース 教授
持橋 大地
統計科学コース 准教授
Stephen Wu
統計科学コース 准教授
吉田 亮
統計科学コース 教授
寺山 慧
横浜市立大学 生命医科学研究科 准教授

博 士 論 文 の 要 旨

氏　　名 ： Zhang Qi

論 文 題 目 ： Machine learning for de novo design of functional molecules and their synthetic routes

As molecular design plays a crucial role in drug discovery, materials science, and other related fields, finding ways to design molecules with desired properties is of great importance. In recent years, machine learning has shown promise in achieving this goal. However, the practical application of de novo molecular design is still hindered by the difficulty and cost of synthesizing computationally designed molecules. To address this issue, synthetic route design using deep neural networks has been studied as a potential solution. Despite this progress, simultaneous design of molecules and their synthetic routes is a challenge that remains unresolved. This issue has been addressed in a novel approach presented in this paper, which utilizes Bayesian inference. This approach involves designing a set of reactants in a reaction network and its topology. The design space can be extremely large, consisting of all combinations of purchasable reactants, which can be in the order of millions or more. Moreover, the designed reaction networks can adopt any topology beyond simple multistep linear reaction routes. To address the hard combinatorial problem of simultaneous molecule and route design, a powerful sequential Monte Carlo algorithm is presented, which recursively designs a synthetic reaction network by building up single-step reactions sequentially. We applied this approach to design drug-like molecules based on commercially available compounds and compared it with heuristic combinatorial search methods. The results showed superior performance in terms of computational efficiency, coverage, and novelty with respect to existing compounds. In addition to the novel approach, we provide the Python library "Seq-Stack-Reaction," which enables researchers to test and explore the approach further. An illustrative example of designing highly viscous lubricant molecules is also provided to demonstrate the effectiveness of the Seq-Stack-Reaction library. This library can help facilitate the design of complex reaction networks and identify promising synthetic routes for molecules of interest.

De novo molecular design has emerged as a promising area in drug discovery, materials science, and other fields where designing molecules with specific properties is essential. Researchers have explored various machine learning techniques to design molecules, including generative models, reinforcement learning, and evolutionary algorithms. These approaches have shown significant potential in designing molecules with desired properties. However, despite the progress made in this field, practical applications of de novo molecular design are still limited. The cost and technical difficulty of synthesizing computationally designed molecules remain a challenge, particularly in drug discovery, where the synthesis and testing of compounds can be costly and time-consuming. To address this challenge, synthetic route design using deep neural networks has been studied as a potential solution. Synthetic route design enables the prediction of synthetic routes for computationally designed molecules, thus reducing the time and cost of synthesizing molecules with desired properties. However, the simultaneous design of molecules and their synthetic routes remains a significant challenge. This is because the design space for molecules and their synthetic routes is vast, consisting of all possible combinations of purchasable reactants, which can be in the order of millions or

more. Furthermore, the designed reaction networks can adopt any topology beyond simple multistep linear reaction routes, making the design problem even more challenging. Despite these challenges, researchers continue to explore new approaches to simultaneously design molecules and their synthetic routes. Bayesian inference has emerged as a promising approach in this area. Bayesian inference allows for the simultaneous modeling of both the molecular properties and synthetic routes, enabling the design of molecules that can be synthesized using predicted synthetic routes. This approach has shown significant promise in reducing the time and cost required for designing molecules and their synthetic routes. In conclusion, de novo molecular design has the potential to revolutionize drug discovery and materials science, but practical applications are still limited. Synthetic route design using deep neural networks is a promising approach to overcome the challenges associated with synthesizing computationally designed molecules. However, designing molecules and their synthetic routes simultaneously remains a significant challenge. Bayesian inference is a promising approach to simultaneously model the molecular properties and synthetic routes, enabling the design of molecules that can be synthesized using predicted synthetic routes. Further research is needed to develop more effective approaches for de novo molecular design and synthetic route design.

To address the challenge of simultaneously designing molecules and their synthetic routes, we present a novel approach that uses Bayesian inference to optimize the design of both molecules and their synthetic routes. Our approach involves designing a set of reactants in a reaction network and its topology. The design space is vast, consisting of all possible combinations of purchasable reactants, which can be in the order of millions or more. Moreover, the designed reaction networks can adopt any topology beyond simple multistep linear reaction routes. To solve this hard combinatorial problem, we present a powerful sequential Monte Carlo algorithm that recursively designs a synthetic reaction network by sequentially building up single-step reactions. Our method also incorporates a sequence-based approach that enables us to consider the entire set of reactions needed to synthesize a given molecule. This approach helps us to design molecules with desirable properties and identify promising synthetic routes for these molecules. To evaluate our method, we applied it to design drug-like molecules based on commercially available compounds and compared it with heuristic combinatorial search methods. Our proposed method demonstrated superior performance in terms of computational efficiency, coverage, and novelty with respect to existing compounds. Moreover, the Seq-Stack-Reaction library that we provide enables researchers to test and explore our approach further. An illustrative example of designing highly viscous lubricant molecules is also provided to demonstrate the effectiveness of the Seq-Stack-Reaction library. The library can be used to facilitate the design of complex reaction networks and to identify promising synthetic routes for molecules of interest.

Although our approach provides a promising solution to the challenge of simultaneously designing molecules and their synthetic routes, there are still limitations that need to be addressed. For example, the scalability of the algorithm to larger design spaces needs to be improved. Furthermore, the accuracy of the generated synthetic routes needs to be further validated experimentally. To address these limitations, further research is required.

（別紙１ Separate Sheet 1）

Results of the doctoral thesis defense

博士論文審査結果

Name in Full
氏 名　Zhang Qi

Title
論文題目　Machine learning for de novo design of functional molecules and their synthetic routes

2023 年 8 月 16 日午前 10 時から約 2 時間にわたり Zhang Qi 氏の博士論文審査委員会を開催した．1 時間の公開発表と質疑応答，さらに約 1 時間の審査委員のみによる審査を行った結果，審査委員会は本論文が学位の授与に値すると判断した．

[論文の概要]

出願論文は英語で執筆されており，6 章 58 頁からなる．所望の特性を有する分子とその合成経路を予測する問題をベイズ推論の枠組みで定式化し，合成反応ネットワークと反応物集合の上に定義された離散事後確率分布から効率的にランダムサンプリングを行うために，単一ステップの反応を逐次的に積み上げて合成反応ネットワークを再帰的に設計する逐次モンテカルロ法を提案している．

各章の概要は，以下の通りである．

第 1 章では，機械学習を用いた分子設計や合成経路設計に関する近年の動向を概説し，問題意識と研究の学術的意義及び貢献を説明している．

第 2 章では，分子構造の標準データ形式（文字列）である SMILES 記法について解説し，深層言語生成モデルを用いた分子設計や合成経路予測の関連研究をまとめている．

第 3 章では，所望の特性を持つ分子構造と合成反応ネットワークを同時に設計するタスクをベイズ推論の枠組みに従って定式化している．探索空間はネットワークのグラフ構造と与えられた反応物集合の全ての組合せから構成されるため，標準的な手法では事後分布の近似が困難であることが論じられている．

第 4 章は，提案手法について述べた章である．具体的には，合成反応ネットワークのサンプリング効率を向上するための再帰アルゴリズム，Transformer を用いた合成反応予測の計算コスト削減を実現するための代理モデルの導入，Generative Topographic Mapping を用いた類似構造検索の高速化，非同期並列計算アルゴリズム等の数理的アイデアや計算手法が説明されている．

第 5 章では，薬剤分子設計タスクにおける物性予測，Transformer に基づく合成反応予測，分子設定・合成経路予測の性能評価実験の結果が報告されている．また，予測された合成反応ネットワークに対する有機合成化学の専門家による妥当性評価の結果が報告されている．

第 6 章は，まとめの章である．

[論文の評価]

本研究の学術的貢献は以下の通りである．

(1) 所望の特性を持つ分子とその合成反応ネットワークを予測する問題に対し，ベイズ推論の枠組みを用いて同時に解決する手法を提案した．

(2) 設計変数は，反応ネットワークを構成する反応物集合の組合せとネットワーク構造からなる．購入可能な反応物質の数は数百万以上のオーダーとなることもある．この組合せ問題に対し，独自の逐次モンテカルロ法のアルゴリズムを開発し，既存手法では発見できなかった多様な合成経路を安定的に検出できることを実証した．

本研究は，技術面でいくつかの改善すべき点が残されているが，有機合成化学において重要な学術的貢献をもたらす可能性を持つ．また，統計科学の面でも学術的新規性が十分に認められる．以上の理由により，統計科学分野の博士論文の研究として十分に高い水準に達していると判断する．

[その他]

第 3 章，第 4 章，第 5 章の内容をまとめた論文が査読付きジャーナル Science and Technology of Advanced Materials: Methods 誌（第一著者）に掲載されている．この論文は同誌の Editor's Choice に選ばれている．