氏　　名　山田 正嗣

学位(専攻分野)　博士(情報学)

学 位 記 番 号　総研大甲第 2455 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 情報学専攻
学位規則第6条第1項該当

学 位 論 文 題 目　Subgraph-based Machine Learning for Graph Generation

論 文 審 査 委 員　主　査　杉山 麿人
情報学コース 准教授
宇野 毅明
情報学コース 教授
井上 克巳
情報学コース 教授
佐藤 健
情報学コース 教授
佐藤 寛子
チューリッヒ大学 科学職員/
情報・システム研究機構 特任准教授

# 博士論文の要旨

氏　　名　山田　正嗣

論文題目　Subgraph-based Machine Learning for Graph Generation

Designing de novo molecules for drugs and materials with desired properties is a highly challenging task due to the combinatorial problem of finding desired graphs. Molecules are essentially represented as *graphs* with node and edge attributes. In contrast, such graph structure of chemical compounds makes it challenging to generate valid molecules with the desired biochemical activity or property. Several methods have been proposed to tackle this problem of molecular graph generation. Recent advanced approaches to finding drug-candidate molecules have employed deep generative models. The basic idea of using generative models is to learn the latent representation of molecules, which enables latent vectors to be reconstructed and explore molecules that satisfy target properties in the learned latent chemical space. Exploration methods such as Bayesian optimization are used to search the latent chemical space. However, it is fundamentally difficult to reconstruct molecular graphs from the latent space and search for molecules with the desired property by extrapolation from a training dataset, as a large part of the latent space represents invalid molecules. Furthermore, there is another problem when using graph neural networks (GNNs) for embedding graphs and surrogate models for predicting target properties. Many proposed GNN models are based on message passing algorithms, which is the fundamental approach to extracting features of subgraph structures by aggregating information in neighboring nodes. This operation is related to the Weisfeiler-Lehman graph isomorphism test for discerning whether two graphs are isomorphisms. When the number of message-passing iterations increases, the problem has been reported that GNNs fail to represent node features in graphs. It is a critical problem when incorporating GNNs into graph generation algorithms.

This thesis consists of four chapters. In the first chapter of this thesis, we state the outlines of the preliminary graph theory and the peripheral areas to understand graph generation molecular graphs. Graph theory is indispensable for representing graphs appropriately. Graph generation covers a broad range of machine learning topics: deep generative models for generating data, graph kernels and graph neural networks for predicting target properties, and reinforcement learning for generating graphs to guide with target properties.

In the second chapter of this thesis, we state the molecular graph generation algorithm based on subgraphs. We propose a novel method called the MOLDR (MOLecular graph Decomposition and Reconstruction) algorithm to generate molecular

graphs by combining subgraphs mined from molecular graph datasets by using the subgraph mining algorithm and searching molecules with target properties via Monte Carlo tree search and reinforcement learning. Our method can generate molecules maximized with respect to logP and QED (Quantitative Estimation of Druglikeness) that are used as benchmarks for molecular generation. In contrast to deep generative models, MOLDR can generate molecular graphs directly so that the generating process is highly interpretable when evaluating the generative path.

In the third chapter of this thesis, we investigate how the subgraph features from message passing affect graph classification and regression. This is also highly related to generating graphs with desired properties because when generating graphs, it is necessary to prepare for some objective function to maximize or minimize target properties, which is basically a surrogate model, that is, a property prediction function trained on a graph dataset. Our main contributions are that the WL kernel outperforms the performance of classification and regression over the most fundamental graph neural networks inspired by message passing schemes, such as graph convolutional networks (GCNs) and graph isomorphism networks (GINs). We also investigate the effect of a bigger size of subgraph structures by increasing the number of message passing iterations. The performance of the WL kernel does not deteriorate even if the message passing iteration increases. In contrast, the performance of GCNs and GINs deteriorate due to a large number of parameters to train and ill-trained previous features.

In the final chapter, we summarize our contributions to this dissertation and discuss future work of molecular graph generation. Although MOLDR can generate molecules with maximized target properties, we need to investigate further the strengths and weaknesses of MOLDR on other benchmark datasets. MCTS needs more rollout to search molecules, so we need to speed up searching and apply another reinforcement learning with parallel computations to MOLDR.

（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　　**山田　正嗣**

Title
論文題目　　Subgraph-based Machine Learning for Graph Generation

本学位論文は、「Subgraph-based Machine Learning for Graph Generation」と題し、主に分子の解析を目的としたグラフ機械学習手法について、部分構造を組み合わせることで効率的な分子生成を実現する新規手法と、部分構造に基づいた分子特性予測の性能解析に関する成果を述べている。創薬や材料科学において所望の特性を持つような新たな分子構造を発見するために、分子をグラフとして捉え、グラフ機械学習を用いて分子の生成や予測をおこなう深層学習アプローチが近年盛んに研究されている。しかし、深層構造で実現される潜在空間での学習では、得られた特徴量ベクトルから分子を復元したり、分子として有効なグラフのみを生成したりすることが困難であるという問題があった。そこで本論文では、まずグラフマイニングを用いて既知の分子群を部分グラフへと分解し、それらを部品として再構成することで新しい分子を生成する、というアプローチによって、分子として有効なグラフを効率的に生成する手法を提案している。本学位論文は英語で執筆されており、全 4 章から構成されている。

第 1 章「Introduction」では、まず研究の背景として、グラフを対象とした機械学習を導入し、代表的な手法であるグラフカーネルやグラフニューラルネットワーク（GNN）といった技術を紹介している。その後、分子をグラフとしてモデル化した分子グラフを対象とした生成問題を導入するとともに、深層学習や強化学習を用いた分子グラフ生成アプローチについて関連研究を概観し、本論文の主要な貢献についてまとめている。

第 2 章「Molecular Graph Generation」では、所望の特性を持つ分子グラフを生成するための手法 MOLDR（MOLecular graph Decomposition and Reassembling）を提案している。まずグラフマイニングを分子グラフデータベースに適用することで、重要な分子の部分構造のみを部分グラフ群として効率的に取り出す。ただし、分子グラフにおいて例えば環構造を切断してしまうと、得られる部分グラフは分子構造として意味をなさない。そこで、あらかじめ分子グラフに木分解を適用し、分子構造として意味を持つ部分構造をまとめておくことで、グラフマイニングにおける不要な部分グラフの列挙を回避している。得られた部分グラフを部品として組み合わせることで様々な分子を生成することができるが、部品の組み合わせには膨大な候補がある。そこで提案手法では、モンテカルロ木探索と強化学習を用いるというアイデアによって、効率の良い所望の特性を持つ分子の生成を実現している。このように、潜在空間を経由せず、一貫して直接分子グラフを操作するため、解釈性の高さを保ったまま明示的に新規の分子グラフを生成できるという利点がある。さらに実データを用いた実験によって、親水性の指標である Penalized logP や分子らしさの指標である QED において、既存手法より優れた分子が生成できることを示している。

第 3 章「Substructure-based Machine Learning」では、部分グラフが持つ情報の重要度やそ

の抽出方法について、より一般にグラフの分類や回帰問題のもとで検証している。グラフを扱う機械学習手法として GNN が盛んに利用されているが、部分グラフが持つ情報をニューラルネットワークへ取り込むために、メッセージパッシングと呼ばれる機構が共通して用いられている。メッセージパッシングでは、グラフの各ノードにおいて近隣ノードの情報を集約する、という操作を繰り返し行うことで、任意のスケールにおける部分グラフ情報を取り込むことを可能としている。しかし、サイズの大きい部分グラフを考慮するためにメッセージパッシングを何度も繰り返すと、各ノードが持つ情報が過剰に平滑化され、分類や回帰の精度が低下してしまう可能性がある。この現象を詳細に解析するために、代表的な GNN である GCN（グラフ畳み込みニューラルネットワーク）及び GIN（グラフ同型ネットワーク)、そして代表的なグラフカーネル手法でありメッセージパッシングと同様の機構を持つ Weisfeiler–Lehman（WL）カーネル及び共通部分グラフを明示的に数え上げるグラフレットカーネルを様々な種類のグラフデータへ適用し、メッセージパッシングがもつ効果を実験的に検証している。その結果、多くの場合、メッセージパッシングは部分グラフが持つ大域的な情報をうまく取り込むことができず、GNN の性能が不安定になってしまうことを実験的に示している。さらに、ニューラルネットワークを用いない WL カーネルが、グラフの分類や回帰においてより安定かつ高精度であることも示している。

第 4 章「Summary and Future works」では、本論文の貢献をまとめ、提案手法の限界や欠点、そして今後の課題や展望について述べている。言語や画像を対象とした生成モデルが機械学習分野で急速に発展しているが、グラフのような複雑な離散構造を備えた対象の生成はより困難な課題と考えられる。本論文は、グラフ機械学習の基礎をなす部分グラフ情報の取り込みや利用についての成果であり、今後のグラフ機械学習の発展およびその応用における重要な知見である。

公開発表会では、博士論文の章立てに従って発表がおこなわれ、その後におこなわれた論文審査会及び口述試験では、審査員からの質疑に対して適切に回答がなされた。質疑応答後に審査委員会を開催し、審査委員で議論をおこなった。審査委員会では、出願者が情報学分野の十分な知識と研究能力を持つと認められるとともに、博士研究が分子グラフ生成問題において十分な新規性を有しており、かつ学術的にも優れた貢献であることが評価された。

以上を要するに本学位論文は、部分グラフに基づくグラフ機械学習問題を精緻に議論し、かつ実効的な手法を提案したものであり、今後の同分野の発展に影響を与えうる独創的かつ完成度の高い研究成果である。また、本学位論文の成果は、学術雑誌論文 2 件として発表され、学術的な貢献も認められる。以上の理由により、審査委員会は、本論文が学位の授与に値すると判断した。