TSUBAME 共同利用　令和6年度　学術利用　成果報告書

利用課題名　深層学習を用いた大規模化合物潜在空間の構築
英文：Construction of Large-scale Chemical Latent Space using Transformer-based Models

利用課題責任者
Yasubumi Sakakibara

所属
Keio University
URL https://www.st.keio.ac.jp/tprofile/bio/sakakibara.html

邦文抄録（300 字程度）
化学空間の膨大さにより、望ましい分子特性を持つ化合物の探索には高効率な計算手法が求められている。本研究では、深層学習を用いた大規模化合物潜在空間の構築を目的とし、分子をフラグメント単位の木構造として扱う Tree-Transformer ベースの変分オートエンコーダ（FRATTVAE）を開発した。FRATTVAE は、フラグメントをトークン化し、自己注意機構により分子構造の特徴を学習することで、従来手法を超える再構成精度と分子生成性能を実現した。さらに、分子特性を条件として付与することで、目的に応じた分子設計が可能な Conditional FRATTVAE（C-FRATTVAE）を開発した。提案手法は、従来困難であった大規模分子や天然物の潜在空間を効率的に構築し、創薬・材料科学における分子設計の加速に寄与する。
英文抄録（100 words 程度）
The vastness of chemical space necessitates efficient computational methods for compound exploration. This study aims to construct a large-scale latent space for chemical compounds using deep learning. We propose a Fragment Tree-Transformer based Variational Autoencoder (FRATTVAE), which represents molecules as tree structures with fragment-based tokenization. Utilizing self-attention mechanisms, FRATTVAE surpasses existing methods in reconstruction accuracy and molecular generation performance. Additionally, we introduce Conditional FRATTVAE (C-FRATTVAE) for property-guided molecular design. Our approach enables the efficient construction of latent spaces for large and complex molecules, accelerating molecular discovery in drug development and materials science.
Keywords: 5つ程度
Chemical space, Tree Transformer, Variational autoencoder (VAE), Molecular generation, Fragment tokenization

背景と目的

化学空間は極めて広大であり、特に創薬や材料科学において有用な化合物の探索には効率的なアプローチが不可欠である。従来の分子設計手法では、仮想スクリーニングや QSAR モデルが活用されてきたが、それらは計算コストが高く、探索範囲が限定されるという課題があった。この問題を克服するため、深層学習を用いた化合物の潜在空間（latent space）の構築と活用が注目されている。 既存の分子生成モデルは、SMILES 表記を用いた言語モデルや、分子をグラフとして扱う手法が主流である。しかし、SMILES ベースのモデルは表記の不規則性により化学的妥当性の低い構造を生成しやすく、グラフベースの手法は大規模分子の扱いが困難であるという課題がある。 本研究では、大規模化合物潜在空間を構築し、より汎用的な分子生成を可能にすること を目的とし、分子をフラグメント単位で木構造として扱う Tree-Transformer ベースの変分オートエンコーダ（FRATTVAE）を提案する。本手法により、従来手法では扱いにくかった大規模分子や天然物を含む多様な化学空間を効率的に学習・探索することを目指す。

概要

本研究では、大規模な化合物潜在空間の構築を可能とするために、FRATTVAE を開発した。FRATTVAE は、分子をフラグメント単位の木構造として表現し、Transformer の自己注意機構を活用することで、分子全体の特徴を効率的に学習する。フラグメントベースのトークン化を導入することで、従来の SMILES 表記では困難であった大規模分子や複雑な分子を扱うことが可能となった。 また、分子の特性を条件として制御可能な Conditional FRATTVAE（C-FRATTVAE）を開発し、目的に応じた分子生成を実現した。ZINC250K、MOSES、

GuacaMol、Polymer、天然物データセットを用いた評価実験の結果、FRATTVAE は従来手法を超える再構成精度および分子生成性能を示した。

結果および考察

FRATTVAE の性能を評価するため、分布学習と分子最適化を実施した。

分布学習: FRATTVAE は、ZINC250K や MOSES などのデータセットにおいて、既存手法（JTVAE、MoLeR など）と比較して高い再構成精度（約 0.79）および Fréchet ChemNet Distance（FCD）の向上を達成した。特に、従来手法では扱いにくかった天然物データセットにおいても、高い妥当性（validity）を維持しつつ、多様な分子を生成可能であることを確認した。

分子最適化: 分子特性の最適化タスクでは、FRATTVAE が従来手法よりも効率的に目標特性を向上させることができた。特に、フラグメントレベルでの構造変更が可能なため、一度の最適化で大幅な構造変化を実現し、収束速度の向上を確認した。

条件付き生成: C-FRATTVAE を用いた分子特性制御では、複数の特性を同時に考慮した分子生成が可能であり、創薬や材料科学における応用の可能性が示された。

これらの結果から、FRATTVAE は大規模化合物潜在空間を構築し、従来手法では困難であった大規模分子や天然物の効率的な探索を可能にすることが明らかとなった。

まとめ、今後の課題

本研究では、深層学習を用いた大規模化合物潜在空間の構築 を目的とし、分子をフラグメントベースの木構造として処理する Transformer ベースの変分オートエンコーダ FRATTVAE を開発した。分布学習および分子最適化のタスクにおいて、既存手法を超える再構成精度と分子生成性能を達成し、さらに C-FRATTVAE を導入することで分子特性の制御が可能であることを示した。

今後の課題として、以下の点が挙げられる。

スケーラビリティの向上: より大規模なデータセットへの適用を可能にし、潜在空間の一般化性能を向上させる。

フラグメント分割の最適化: さらに化学的に有意義なトークン化手法を開発し、生成精度を向上させる。

実験的検証: 生成した分子の合成可能性や生物活性を実験的に評価し、実際の創薬・材料設計への応用を進める。

本研究の成果は、創薬や材料科学における分子設計を加速し、より効率的な化学空間の探索を可能にする新たなアプローチを提供するものである。