審 査 の 結 果 の 要 旨

氏　　名　　井上　貴央

本論文は「Molecular Structure Generation Using Small Chemical Dataset (小規模の化学データセットを利用した分子構造生成)」と題し、興味のある物性を持つ新規化合物の構造設計に利用される統計モデルベースの構造生成器に対して、統計モデルの訓練に用いるサンプル数が少ない場合においても、構造生成を良好に行うための手法に関する研究結果をまとめたものである。本文は英語で書かれており、五つの章から構成されている。

第 1 章では、研究背景と研究目的を述べている。研究背景では、新規化合物設計の効率化のために利用される定量的構造物性相関 (QSPR) モデルについて基本的な概念を説明した後、本論文の主題である統計モデルベースの構造生成器について、ルールベース構造生成器・深層構造生成器それぞれの既往研究とその問題点を概説している。また、構造生成器が持つべき特徴として、(1) 目的物性を持つ分子を多く生成できること、(2) 多様な分子構造を生成できること、(3) 安定で合成しやすい分子構造を多く生成できること、(4) 分子構造の生成効率が良いことの四つを挙げている。構造生成器を実応用する際の課題として、統計モデルの訓練に利用できるサンプルが少ない場合に、モデルの過剰適合のために上述の (1), (2) のような特徴を十分に達成できない可能性があることを指摘している。これを受けて、小規模化学データセットを「1,000 件程度の興味のある性質を持つ化合物からなるデータセット」と定義した上で、本論文の目的を小規模化学データセットにおいても、上述の (1)–(4) の特徴を満たせるような性能の良い構造生成を可能にすることと設定している。

第 2 章では、ルールベース構造生成器 DAECS を基礎にして、目的物性を持つ多様な分子構造を生成するための手法について述べている。深層学習に依らない統計モデルはモデルパラメータが少ないために過剰適合しにくいという点に着目し、非深層 QSPR モデルを利用してルールベースの構造生成を行う構造生成器として、特徴 (1) を満たすことが報告されている DAECS を選定している。特徴 (2) を達成するため、多段階の構造変化を 1 ステップで行える汎用的な構造修正ルールを二つ追加し、分子構造の分布を表す 2 次元マップ上でのクラスタリングを利用した修正構造の選択アルゴリズムを提案している。さらにその有効性を、500 件程度のヒスタミン H_1 受容体に対するリガンドデータセットを利用したケーススタディにより検証し、ターゲット近傍の生成構造の多様化が達成されることを示している。

第 3 では、小規模化学データセットでグラフベース深層構造生成器を訓練するための手法について述べている。すなわち、グラフデータを入力に取るグラフニューラルネットワーク（GNN）の過剰適合を防ぐための手法として画像データなどでも広く利用されているデータ拡張に着目し、分子グラフデータに対するデータ拡張手法を新たに提案している。この手法は、GNN のメッセージパッシング操作中に頂点に対して、標準正規乱数からなる摂動ベクトルを一定確率で加えることでデータ拡張を実施ものである。提案手法の有効性は、QM9 データセットを用いた回帰タスクと PCBA データセット・Tox21 データセットを用いた分類タスクの三つの物性予測タスクで検証しており、GNN のリードアウト操作直前での摂動により小規模データセットでの予測性能が改善されることを示している。また、訓練サンプル数や摂動確率を変化させた場合の予測性能への影響についても検討している。

第 4 章では、グラフデータを利用する深層構造生成器 Junction Tree Variational Autoencoder（JT-VAE）を基礎にして、性能良く構造生成する手法について述べている。大規模な分子構造データセットによるネットワークの事前訓練の結果を利用する転移学習に加え、第 3 章で設計されたデータ拡張手法を JT-VAE の GNN による特徴抽出部分に適用することで、転移学習の過剰適合を抑える効果を高めている。事前学習用の ZINC データセットと 1,000 件の化合物からなる PCBA データセットを用いたケーススタディにおいて、ジャンクション木の特徴抽出を行う GNN でのデータ拡張により、生成構造群のデータセットとの類似度や部分構造の新規性といった生成指標が改善しており、構造生成性能が大きく改善することを示している。また、PCBA データセットの訓練サンプル数を減らした場合の生成性能についても検討しており、少なくとも 500 件の訓練サンプルがあるのが望ましいとの結論に至っている。

第 5 章では、各章に記載された結果が総括している。研究成果とその意義を述べた上で、今後の課題として上述の (3), (4) の特徴への対応、さらに小規模なデータセットを用いる場合についての対応を挙げている。

以上、本論文は、小規模化学データセットを用いた構造生成に関する成果をまとめたものであり、構造生成器の実応用上の課題を克服するための新たな手法の提案、その手法の適用範囲に関する検討について詳述している。これら一連の研究成果により、データを有効活用した新規化合物の設計プロセスの効率化が達成されると期待され、ケモインフォマティクスおよび化学システム工学の進展に大きく貢献するものであると判断される。

よって本論文は博士（工学）の学位請求論文として合格と認められる。