5ZJ-01

天然物様化合物を生成する化学言語モデルの開発

坂野 晃　　古井 海里　　大上 雅史
東京工業大学 情報理工学院 情報工学系

1 序論

微生物等が産生する天然由来化合物（天然物）は，様々な生物活性を示す有用物質として注目されている [1]．上市薬の約半分は天然物由来であり [2]，医薬品開発においても欠かせないものとなっている．天然物の生物活性は，合成化合物にはあまり見られない独特の分子構造にあるとされる [1]．実際に，2010 年頃までは天然物の分子構造を起点とした分子設計が盛んに行われていた．一方で，生化学実験のスループットが向上した現在においては，構造決定や合成コストの観点から天然物創薬を避ける傾向にある [3]．しかし，合成化合物からは得られない天然物の生物活性や構造の多様さには魅力があり，天然物に基づく創薬は重要視され続けている．

合成化合物の探索では，近年は深層学習などを利用した分子生成技術も用いられつつある [4]．計算機上で化合物を仮想的に生成し，その中から有用な候補化合物を見出す目的で使われる．しかしながら，学習に用いる分子は化合物の一般的なデータベースが用いられているため，天然物のような大きく複雑な化合物については生成が難しく，生成化合物の化学空間が限られるという問題があった [5]．

そこで本研究では，天然物に似た化合物（天然物様化合物）を生成できる分子生成モデルを提案する．天然物の化学空間を満たすような分子群を生成することで，創薬起点分子の探索や，天然物が有する分子構造の有用性の理解に繋がることを目指した．

2 手法

大規模な事前学習を行った化学言語モデルに対し，天然物データによってファインチューニングする方針とした．化学言語モデルは化合物の文字列表現から言語処理技術で事前学習したモデルであり，複数の学習済みモデルが提案されている [6, 7]．

2.1 データセット

約 40 万件の天然物を収載した COCONUT データベース [8] を用いた．前処理として SMILES の正規化と大きな化合物（原子数 > 150 または環の数 > 10）の除去を行った後，SmilesEnumerator [9] によってデータを約 9 倍にオーグメンテーションした．得られた約 360 万件のデータセットをファインチューニングに供した．

2.2 生成モデル

生成モデルとして Decoder のみの Transformer ベースのモデルを扱った．学習済みモデルとして ChemGPT [6]（GPT-Neo 利用, SELFIES 表現）と smiles-gpt [7]（GPT-2 利用, SMILES 表現）の 2 種を比較検討した．両者とも 1000 万分子程度のデータセットで事前学習されており，アーキテクチャと化合物の文字列表現方法が異なる．

3 実験および考察

3.1 生成化合物の天然物らしさ

生成化合物の天然物らしさを評価するために，既知天然物の部分構造情報から推定する Natural Product-likeness Score (NP Score) [10] を適用した．NP Score が高いほど天然物らしいことが期待される．元のモデルとファインチューニング後のモデルでそれぞれ生成した分子の NP Score の分布を図 1 に示す．ChemGPT (finetuned) は COCONUT の分布を捉えられていないが，smiles-gpt (finetuned) は COCONUT に近い分布の化合物群を生成しており，SMILES ベースの GPT モデルは天然物の分布を学習するのに適している．SMILES による学習モデルに比べて SELFIES は性能が低いという報告 [11] もあり，本結果にも影響したものと考えられる．

3.2 リガンドドッキングによる評価

薬剤候補として有用な化合物が生成できているかをリガンドドッキング計算によって検証した．標的タンパク質として上皮成長因子受容体 (EGFR)

Development of a chemical language model for generating natural products
Koh Sakano, Kairi Furui & Masahito Ohue, Department of Computer Science, School of Computing, Tokyo Institute of Technology

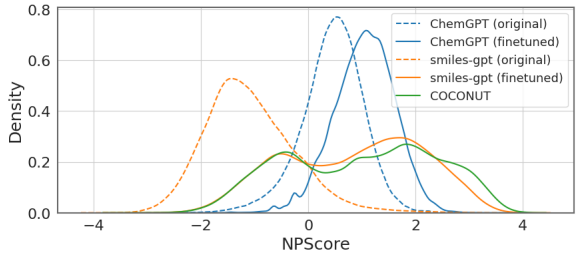

図 1　生成分子および COCONUT 化合物の NP Score の分布

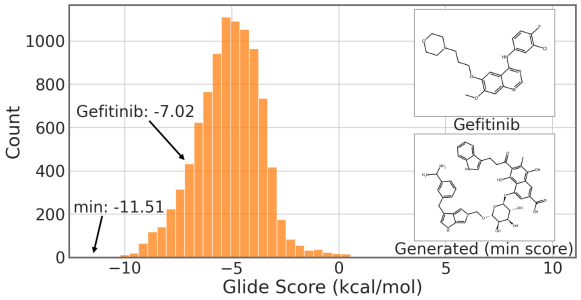

図 2　生成分子の EGFR に対するドッキングスコア（光学異性体等を含む 12,407 件）の分布

を取り上げた．標的構造は PDB ID: 2ITY を用い，smiles-gpt (finetuned) で生成した分子からランダムに 1,000 分子を抽出して Schrödinger Glide [12] SP mode でドッキング計算を実施した．

ドッキングの評価値である Glide Score（低いほど良い）は図 2 のようになった．EGFR 阻害薬である Gefitinib の Glide Score が −7.02 kcal/mol [13] であるのに対し，より良いスコアの化合物が複数確認できた（最良は −11.51 kcal/mol）．

さらに，生成分子それぞれについて COCONUT の全化合物との類似度（半径 2，2048 bit の Morgan Fingerprint による Tanimoto 係数）の平均値を求め，Glide Score との関係を図 3 に示す．類似度と Glide Score には緩やかな相関が見られ，天然物との類似度が高いほど Glide Score が良い傾向が確認された．

4　結論

本研究では，化学言語モデルを天然物によってファインチューニングすることで，天然物様化合物を生成する分子生成モデルを構築した．生成した分子が天然物の分布に近いことを確認し，リガンドドッキングによって生成分子の中から実際に有用物質が探索できる可能性の一例を示した．今後，生成した分子の合成可能性や立体配座の性質の検討など，本研究で考慮できなかった医薬品開発における

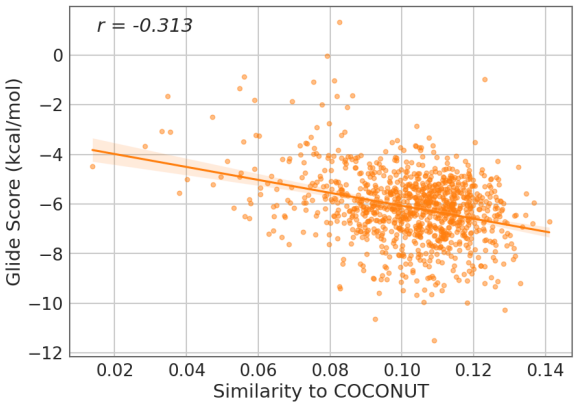

図 3　Glide Score と天然物に対する類似度の関係

要素も加味した手法開発を目指す．

謝辞　本研究は，JST 創発的研究支援事業 (JPMJFR216J)，科研費 学術変革領域研究 (A) (23H04887) の支援を受けて行われた．

参考文献

[1] Dias AD, *et al.* A historical overview of natural products in drug discovery. *Metabolites*, 2(2), 303–336, 2012.
[2] Demain AL. Importance of microbial natural products and the need to revitalize their discovery. *J Ind Microbiol Biotechnol*, 41(2), 185–201, 2014.
[3] 佐藤文治. 天然物創薬の再興のために—次世代天然物化学技術研究組合における天然物ライブラリーの相互利用. ファルマシア, 50(2), 127–131, 2014.
[4] Bilodeau C, *et al.* Generative models for molecular discovery: Recent advances and challenges. *WIRES Comput Mol Sci*, 12(5), e1608, 2022.
[5] Jin W, *et al.* Hierarchical generation of molecular graphs using structural motifs. In *ICML 2020*, 4839–4848, 2020.
[6] Frey CN, *et al.* Neural scaling of deep chemical models. *Nat Mach Intell*, 5, 1297–1305, 2023.
[7] Adilov S. Generative pre-training from molecules. *ChemRxiv*, 10.26434/chemrxiv-2021-5fwjd, 2021.
[8] Sorokina M, *et al.* COCONUT online: Collection of Open Natural Products database. *J Cheminform*, 13(1), 2, 2021.
[9] Bjerrum EJ. SMILES enumeration as data augmentation for neural network modeling of molecules. *arXiv*, 1703.07076, 2017.
[10] Ertl P, *et al.* Natural product-likeness score and its application for prioritization of compound libraries. *J Chem Inf Model*, 48(1), 68–74, 2008.
[11] Ghugare R, *et al.* Searching for high-value molecules using reinforcement learning and transformers. In *AI4Mat—NeurIPS 2023 Workshop*, 2023.
[12] Friesner RA, *et al.* Glide: a new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. *J Med Chem*, 47(7), 1739–1749, 2004.
[13] Ochiai T, *et al.* Variational autoencoder-based chemical latent space for large molecular structures with 3D complexity. *Commun Chem*, 6(1), 249, 2023.