1N-08

化学言語モデルに基づく天然物らしさの評価

坂野 晃　　古井 海里　　大上 雅史
東京科学大学 情報理工学院 情報工学系

1　序論

生体内で産生される化合物である天然物は進化の過程で獲得された複雑な構造を持ち，その多くが生物活性を示すことから，薬剤開発における出発点として利用されている．これまで，新規薬剤の探索において天然物のようなユニークで標的に特異的に結合するような構造を持つ化合物を得るために「天然物らしさ」を評価するスコアが開発されてきた．ただし，最も広く用いられている天然物らしさ評価手法は天然物データセット内のフラグメント出現頻度に基づく手法だが，部分構造間の相互作用や立体的な配置を考慮できず，既知の天然物も過小評価してしまう場合があるという課題があった [1]．また，ニューラルネットワークによる二値分類モデルも提案されており，天然物と合成化合物を 0 か 1 かで区別するため，中間的な性質を持つ化合物や合成化合物の中での天然物らしさの評価のランク付けには不向きであった [2]．

そこで本研究では，化合物の SMILES 表記を言語として扱う化学言語モデルに着目し，天然物データと合成化合物データに基づく尤度比による新たな天然物らしさ評価手法を提案する．

2　提案手法

天然物データベース COCONUT [3] と合成化合物データベース ZINC [4] を用いて GPT-2 モデル [5] を学習し，各モデルが算出する対数尤度の差分として天然物らしさスコアを定義する．提案手法の概要を図 1 に示す．モデルには GPT-2 を採用し，SMILES 文字列の次に出現する文字を予測するように学習させる．学習には，天然物データベース COCONUT から全件の 694,709 件，合成化合物データベース ZINC 22 からは天然物と同程度の数になるように 700,000 件をランダムに取得した．データは構造の類似性に基づく Bemis-Murcko スキャフォールド分割により，train/validation/test に分割した．これにより，異なる split には類似度の低い化合物が含まれることになり，モデルの汎化性能が検証できる．

図 1　モデルの概要

提案する天然物らしさスコアは，未知の化合物 x に対して天然物モデルが算出する対数尤度 $\log P_{\text{natural}}(x)$ と合成化合物モデルが算出する対数尤度 $\log P_{\text{synthetic}}(x)$ の差分として定義される．

$$s(x) = \log P_{\text{natural}}(x) - \log P_{\text{synthetic}}(x) \qquad (1)$$

ただし，尤度 $P(x) = \prod_{t=1}^{T} P(x_t|x_{<t})$ はモデルが予測するトークン列の出現確率を順に掛け合わせて得られる．さらに，シグモイド関数を用いて $(0, 1)$ の範囲に正規化した $\tilde{s}(x)$ を最終的なスコアとする．

$$\tilde{s}(x) = \frac{1}{1 + \exp(-s(x))} \qquad (2)$$

この手法により，部分構造の単純な足し合わせではなく，分子全体における原子の配列パターンや長距離の依存関係を考慮した評価が可能となる．

3　実験設定

提案手法の有効性を検証するため，以下の実験を行った．まず，テストデータセット（天然物 37,247 件，合成化合物 9,914 件）を用いて，既存手法である Ertl NP-likeness score [1] および NN score [2] と識別性能を比較した．評価指標には ROC-AUC 値を用いた．

次に，提案スコアを報酬関数として分子生成モデル REINVENT 4 [6] に組み込み，強化学習により天然物らしい化合物の生成を試みた．報酬関数は天然物らしさスコアと QED（薬剤らしさ指標）を組み合わせて設計した．生成された分子に対して，タンパク質-リガンド複合体予測モデル Boltz-2 [7] を用いて標的タンパク質 EGFR（PDB ID: 2ITY）との結合親和性を評価した．

Natural Product-Likeness Scoring Based on Chemical Language Models
Koh Sakano, Kairi Furui, Masahito Ohue. School of Computing, Institute of Science Tokyo

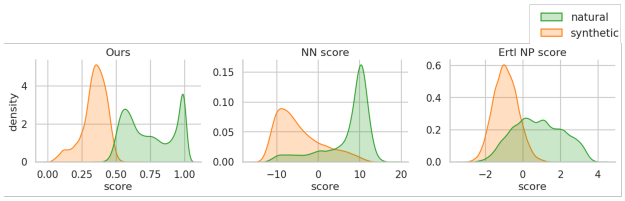

図 2　各スコアの分布

表 1　各モデルの ROC-AUC 値の比較

モデル	ROC-AUC
Ertl NP score	0.8774
NN score	0.9040
提案手法	0.9986

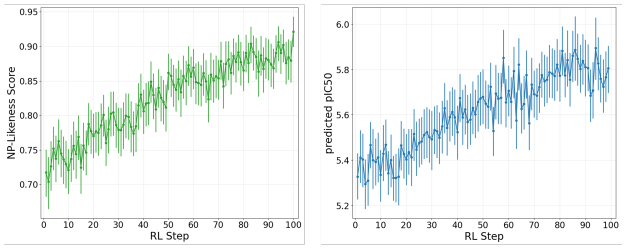

図 3　強化学習において生成された化合物の天然物らしさスコアおよび EGFR に対する pIC50 の予測値のステップごとの推移

4　結果および考察

図 2 にテストデータに対して各手法で予測したときのスコアの分布を示す．Ertl NP score が既存の天然物に対しても低いスコアをつけてしまったり，二値分類モデルである NN score が 10 と-10 あたりに二極化しているのに対し，提案手法は天然物と合成化合物を正確に分離できていることがわかる．また，スコアの分布は 0 と 1 に集中せず連続な山を形成しており，天然物らしさを連続的に評価できている．

表 1 に各手法の ROC-AUC 値を示す．提案手法の AUC 値は 0.9986 となり，NN score，Ertl らの NP score と比較して高い識別性能を示した．これは，提案手法が分子全体の原子の配列パターンを文脈として正確に捉えられていることを示す．

図 3 に強化学習において生成された化合物のステップごとの天然物らしさスコアと EGFR との結合親和性の予測値の推移を示す．スコアは初期の約 0.70 から学習が進むにつれて増加していき，最終的に平均 0.90 を超えた．これは，生成モデルが提案スコアによる報酬を受け取り，天然物らしさスコアが高くなる化学空間へと探索範囲を適応させたことを示す．生成された分子の構造を比較すると，学習初期には直鎖状の構造やハロゲンなど合成化合物に頻出する部分構造が多く見られたが，終盤では多環式骨格や縮合環構造が増加し，天然物に特徴的な複雑な構造が生成されていることが確認された．

EGFR に対する予測 pIC50 の値も同様にステップごとに向上している．強化学習の報酬関数には結合親和性に関する項を含めていないにもかかわらず，pIC50 の値が上昇していることから生成モデルを天然物らしさスコアによって最適化することで，結果として生物活性の向上に寄与する可能性を示している．ただし，ステップごとに生成される化合物の分子量は大きくなり，脂溶性も高くなっていることが確認でき，一般的な低分子医薬品とは異なる化学空間の探索を行っている．

5　結論

本研究では，化学言語モデルに基づく天然物らしさ評価手法を開発した．提案手法はテストデータセットにおいて既存手法よりも高い精度で天然物と合成化合物を分離でき，本スコアを報酬関数として強化学習型分子生成モデルに組み込むことで，高い天然物らしさを持つ新規化合物の生成に成功した．生成された化合物は複雑な環構造などの天然物様の特徴を有しており，EGFR との結合親和性予測においても，天然物らしさスコアが高い化合物ほど高い親和性を示す傾向が確認された．以上の結果より，本研究で開発した手法は天然物化学空間の探索において有用な指標となり，新規医薬品候補化合物の設計，発見に貢献し得ることが示唆された．

謝辞　本研究は JST 創発的研究支援事業 (JPMJFR216J)，JSPS 科研費（JP23H04880, JP23H04887, JP24KJ1091），JST ACT-X (JPMJAX25LB) の支援を受けて行われた．

参考文献

[1] 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.
[2] Menke J, *et al.* Natural product scores and fingerprints extracted from artificial neural networks. *Comput Struct Biotechnol J*, 19, 4593–4602, 2021.
[3] Chandrasekhar V, *et al.* COCONUT 2.0: a comprehensive overhaul and curation of the collection of open natural products database. *Nucleic Acids Res*, 53(D1), D634–D643, 2025.
[4] Tingle BI, *et al.* ZINC-22: A free multi-billion-scale database of tangible compounds for ligand discovery. *J Chem Inf Model*, 63(4), 1166–1176, 2023.
[5] Radford A, *et al.* Language models are unsupervised multitask learners. *OpenAI Blog*, 2019.
[6] Loeffler HH, *et al.* Reinvent 4: Modern AI-driven generative molecule design. *J Cheminform*, 16(1), 20, 2024.
[7] Passaro S, *et al.* Boltz-2: Towards accurate and efficient binding affinity prediction. *bioRxiv*, 2025, doi:10.1101/2025.06.14.659707.