TSUBAME 共同利用　令和7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 a large-scale latent space for compounds using deep learning

利用課題責任者
榊原康文

所属
北里大学未来工学部
URL　https://www.kitasato-u.ac.jp/fr-eng/laboratory/ai/

邦文抄録（300 字程度）

本研究では、TSUBAME の GPU 計算資源を用いて、フラグメント木構造 VAE である FRATTVAE の開発と、そのタンパク質条件付き拡張によるタンキラーゼ阻害剤探索を進めた。FRATTVAE ではフラグメント分割と tree-transformer を組み合わせ、大規模かつ複雑な化学構造を扱える生成基盤を構築し、分布学習、条件付き生成、分子最適化、スケーラビリティを検証した。さらに ESM-2 と統合したモデルを BindingDB 約 20 万件で学習し、タンキラーゼ阻害剤候補 155 件で fine-tuning した結果、既存法 AlphaDrug を上回る性能と有望な阻害剤候補を得た。

英文抄録（100 words 程度）

We developed FRATTVAE, a fragment-tree variational autoencoder for large and complex molecular generation, and further extended it to protein-conditioned molecule design for tankyrase inhibitor discovery. FRATTVAE combined fragment tokenization with a tree-transformer architecture and showed strong performance in distribution learning, conditional generation, molecular optimization, and large-scale training. We then integrated FRATTVAE with the protein language model ESM-2, trained the model on about 200,000 protein-ligand pairs from BindingDB, and fine-tuned it on 155 tankyrase-focused compounds. The resulting model outperformed AlphaDrug and generated promising tankyrase inhibitor candidates.

Keywords: FRATTVAE, tree-transformer, ESM-2, conditional molecular generation, tankyrase inhibitor

背景と目的

創薬における分子生成 AI では、標的タンパク質に応じて有望な化合物を効率よく提案することが重要である。一方、従来の SMILES ベース生成法は複雑な化学構造の表現や大規模分子空間の学習に制約があり、標的依存性の取り込みも十分ではない。

そこでまず、分子をフラグメント列ではなく木構造として扱う FRATTVAE を開発した。FRATTVAE はフラグメント分割、tree positional encoding、transformer 型 VAE を組み合わせることで、天然物や高分子を含む複雑分子の生成を可能にすることを目指した。

さらに本研究では、この FRATTVAE を化合物生成基盤としてタンパク質言語モデル ESM-2 と統合し、配列条件付きに化合物を生成する手法へ拡張した。制がん剤標的として注目されるタンキラーゼを対象に、基盤モデル開発と標的特異的応用の両面から有効性を検証した。

概要

FRATTVAE の基盤開発では、分子をフラグメント木へ変換し、その木構造を transformer ベースのエンコーダ・デコーダで扱う設計を採用した。基盤論文では MOSES、GuacaMol、Polymer、SuperNatural3、ZINC250K の 5 種類のデータセットを用い、分布学習、条件付き生成、構造最適化、医薬化学的妥当性、大規模学習性を検証した。

その上で、FRATTVAE のデコーダに対して ESM-2 から得たタンパク質配列埋め込みを cross-attention で入力し、

タンパク質条件付き分子生成を実現した。学習は、化合物事前学習モデルの利用、BindingDB 約 20 万件のタンパク質-化合物対による本学習、Tankyrase-1 と阻害剤候補 155 件による fine-tuning の 3 段階で行った。
標的特異的生成の評価では AutoDock Vina による Docking Score、High Affinity、タニモト類似度、Uniqueness を用い、各タンパク質あたり 60 秒以内で生成から選択までを完了する条件で比較した。基盤モデルとしての FRATTVAE 開発と、下流応用としてのタンキラーゼ阻害剤探索を一体的に進めた。

結果および考察

まず FRATTVAE の基盤開発では、分布学習・条件付き生成・分子最適化の各面で有効性を確認した。特に大分子・複雑分子を含むデータセットでも再構成性能と FCD が良好であり、フラグメント木構造表現の有効性が示された。表 1 に、基盤論文で得られた主要成果を本報告用に整理して示す。

表 1　FRATTVAE 基盤開発で得られた主要成果

観点	主な結果	意義
分布学習	5 種のデータセットで高い再構成性能と FCD を達成	複雑分子や大分子にも適用可能
条件付き生成	特性条件を与えた分子生成に対応	実用的な分子設計へ拡張可能
構造最適化	GuacaMol 20 課題中 12 課題で MoLeR を上回る	探索性能を確認
医薬化学的妥当性	上位生成分子の構造アラートは約 22%	MoLeR(約 35%)より低アラート
大規模学習	1,225 万分子・1.03B params・約 120 万フラグメント	TSUBAME を活かした拡張性を実証

FRATTVAE は GuacaMol の goal-directed optimization 20 課題中 12 課題で MoLeR を上回り、上位生成分子の構造アラートも約 22%に抑えられた。さらに約 120 万フラグメント、1,225 万分子、1.03 billion parameters 規模まで拡張可能であり、TSUBAME を活用した大規模化学モデルとしての有効性が示された。
次に、FRATTVAE を ESM-2 と統合した標的特異的生成モデルの結果を表 2 に示す。提案法は AlphaDrug に対して Docking Score と High Affinity で優位であり、特に事前学習なし 1 層モデルでは Docking Score -10.2、High Affinity 51%を達成した。化合物側事前学習の導入直後は性能低下が見られたが、conversion layer を 3 層に増やすことで改善し、基盤表現を下流タスクへ接続する際の容量設計の重要性が示された。

表 2　タンパク質条件付き生成モデルと AlphaDrug の比較

モデル	条件	Docking Score↓	High Affinity↑	Tanimoto↑	Uniqueness↑
提案法	なし・1 層	-10.2	51%	0.14	94%
提案法	あり・1 層	-9.7	46%	0.12	82%
提案法	あり・3 層	-10.0	50%	0.14	80%
AlphaDrug	-	-9.3	29%	0.13	99%

タンキラーゼ阻害剤探索では、fine-tuning なし/ありの生成候補がそれぞれ Docking Score -7.4/-7.5 を示し、既知阻害剤 5zqr の-6.8 を上回った。特に fine-tuning 後の候補は、既知阻害剤と類似した結合様式を保ちながら、アデノシンポケットに加えてニコチンアミドポケットとも相互作用した。

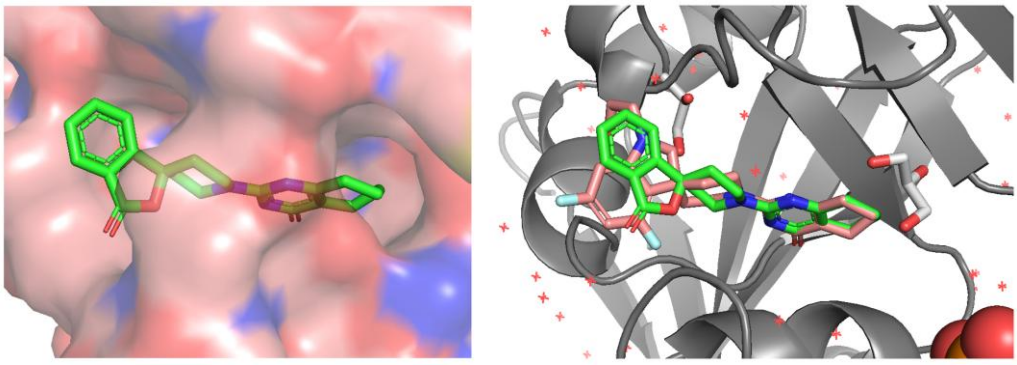

図 1　fine-tuning 後に得られた候補化合物の結合様式（上）と相互作用解析（下）

図 1 に示す AlphaFold3 解析でも、生成候補はポケット内に収まり、重要残基との相互作用が確認された。さらに、condition として与えた QED と生成分子の QED には相関係数 0.76 が得られ、基盤モデルとして開発した FRATTVAE が物性制御と標的特異性の両立に有効であることが分かった。

まとめ、今後の課題

本研究では、FRATTVAE という大規模・複雑分子向け生成基盤を開発し、その上に ESM-2 を組み合わせた標的特異的分子生成モデルを構築した。基盤開発では分布学習・最適化・スケーラビリティの有効性を示し、応用段階ではタンキラーゼ阻害剤探索で既存法を上回る性能と有望候補を得た。今後は、タンパク質立体構造情報の直接利用、活性指標とドッキング評価の整合性向上、および実験系による候補化合物の検証を進める。