TSUBAME 共同利用 令和 6 年度 学術利用 成果報告書

利用課題名 医薬プロセス最適化と新薬開発を効率化・加速する製剤処方設計 AI の開発
英文：Development of AI for drug formulation design to streamline and accelerate pharmaceutical process optimization and new drug development

利用課題責任者
Yasushi Okuno

所属
Department of Biomedical Data Intelligence,
Graduate School of Medicine, Kyoto University
http://clinfo.med.kyoto-u.ac.jp/

邦文抄録（300 字程度）

本研究では、マルチモーダルな生物医学データを統合し、AI および機械学習を活用した創薬支援モデルの構築を試みた。TSUBAME を活用し、ESM-2 によるタンパク質配列解析、Geneformer を用いた遺伝子摂動シミュレーション、グラフ対照学習による遺伝子制御ネットワーク解析、大規模トキシコゲノミクスデータを用いたマルチモーダルモデルによるラット毒性病理所見予測、さらに CT/MRI 画像を用いた医療画像解析を実施した。これらの取り組みにより、複数のモダリティを統合した解析手法の有用性が示唆され、今後の創薬予測モデルの基盤となり得る知見が得られた。

英文抄録（100 words 程度）

This study explored the construction of drug discovery support models by integrating multimodal biomedical data using AI and machine learning. Leveraging TSUBAME, we conducted protein sequence analysis with ESM-2, gene perturbation simulation with Geneformer, graph contrastive learning for regulatory networks, multimodal model construction to predict rat toxicopathological findings from toxicogenomics data, and medical image analysis using CT/MRI. The results suggest the potential utility of multimodal data integration in predictive modeling and provide a basis for future development of drug discovery frameworks.

Keywords: 5つ程度
Drug Discovery,
Graph Neural Network,
Medical Imaging,
Multimodal Data,
Protein Language Model

背景と目的

創薬研究においては、タンパク質の構造情報や化学的性質、生物学的作用機序、さらには臨床データといった多様なモダリティのデータを統合的に解析する必要がある。しかし、従来のシングルモダリティベースの解析手法では、新規治療薬の探索に限界があり、予測精度や解釈性の面でも課題が残る。

本プロジェクトでは、マルチモーダルな生物医学データを統合し、AI・機械学習を用いた高精度な創薬予測モデルの構築を目的とした。TSUBAME を用いた大規模計算により、従来手法では実現困難であった大規模かつ高次元のデータ統合と解析が可能となり、精度の高い創薬予測モデルの基盤となり得る成果を得た。

概要

本課題では、以下の 4 つの研究を実施した。

①ESM-2 によるシーケンス特徴量解析

タンパク質・ペプチド複合体の構造適合性を対象に、言語モデル ESM-2 を用いたシーケンス特徴量の抽出と評価を実施。従来手法と比較して分解能およびモデリング精度の向上が見られた。

②Geneformer を用いた遺伝子摂動シミュレーション

トランスクリプトームデータを対象に、ファインチューニング済み Geneformer による遺伝子発現摂動への応答予測を行い、重要遺伝子および疾患関連細胞種の同定を試みた。

③グラフ対照学習による遺伝子制御ネットワーク解析

TCGA および LINCS データをもとに、乳がん患者および細胞株のグラフ構造と臨床データを統合し、Graph Neural Network の Pretraining・Finetuning を実施。6 手法の比較による有効性評価も行った。

④マルチモーダルによる薬剤障害病理所見予測

大規模トキシコゲノミクスデータを活用し、遺伝子発現データと病理組織画像からラット毒性病理所見を予測するためのマルチモーダルモデルを構築した。

⑤医療画像解析

脳神経外科・整形外科・消化器外科領域における CT/MRI 画像を対象に、病変セグメンテーション、疾患予測モデルの構築、GAN による画像生成などを行った。

結果および考察

ESM-2 を用いた解析では、言語モデルによるシーケンス表現の分解能が高く、タンパク質複合体の構造適合性のモデリング精度が向上した。

Geneformer による遺伝子摂動解析では、潜在特徴量の変化から疾患関連の重要遺伝子や細胞種の推定が可能であった。

グラフ対照学習においては、提案手法が従来手法に比べて、臨床データを用いた下流タスクにおいて高い予測性能を示した。特にサブタイプ分類や生存時間予測において、事前学習が効果的であることが確認された。

大規模トキシコゲノミクスデータを活用したマルチモーダルモデルによるラット毒性病理所見予測については、マルチモーダルモデルが既存の単一モーダルモデルと比較して高い予測性能を示したことから、新たな薬剤性臓器障害予測のアプローチの可能性を示した。

医療画像分野では、CT/MRI を用いた病変検出の精度向上に加え、GAN による画像生成が可能であった。

まとめ、今後の課題

本研究により、マルチモーダルデータを統合した創薬支援モデルの構築可能性が示された。とりわけ、言語モデルやグラフニューラルネットワークの応用により、従来手法を凌駕する予測精度を得ることができた。

今後の課題としては、より多様なデータセットへの対応、臨床現場への応用を見据えたモデルの汎化性能の向上などが挙げられる。TSUBAME の計算資源を活用することで、これらの課題解決に向けたさらなる高精度化と高速化を目指す。