【報告様式Ａ】

課題管理番号:24nk0101111h0005　作成／更新日:令和 6 年 10 月 20 日

日本医療研究開発機構
創薬支援推進事業・産学連携による次世代創薬 AI 開発
事後評価報告書

公開

I 基本情報

研究課題名：　（日本語）最先端の AI 技術を用いたマルチターゲット予測と構造発生を組み合わせた包括的な創薬 AI プラットフォームの開発

（英　語）Development of an integrated drug discovery AI platform combining multi-target prediction and structure generation models using state-of-the-art AI technologies

研究実施期間：令和 2 年 8 月 6 日～令和 7 年 3 月 31 日(予定)

研究代表者　氏名：(日本語) 本間　光貴

（英　語）Teruki Honma

研究代表者　所属機関・部署・役職：

（日本語）国立研究開発法人理化学研究所 生命機能科学研究センター　制御分子設計研究チーム　チームリーダー

（英　語）Team Leader, Laboratory for Structure-Based Molecular Design, Center for Biosystems Dynamics Research, RIKEN

II 研究の概要

創薬におけるヒットから開発候補品に至る過程の設計を包括的に支援できる統合創薬 AI プラットフォームについて、強化学習による構造発生 AI、グラフニューラルネットワーク(GNN)やトランスフォーマーに基づく予測 AI、連合学習（Federated Learning,FL）等の最先端技術を開発し、創薬現場で活用できるように実用化することを目指して構築した。学習データについては、公共データに加えて製薬企業 17 社から提供された 1500 万ポイント程度のオン・オフターゲット、及び薬物動態・毒性（ADMET）のデータを利用するとともに、製薬企業各社との間で FL を実施できるシステムを構築し、社内の秘匿データも学習に活用できる体制を立ち上げて、運用した。

統合創薬 AI プラットフォームには、予測 AI と構造発生 AI を組み合わせた医薬品設計機能（プロジェクト機能)、予測 AI や構造発生 AI を単体または任意の順番で組み合わせて利用する機能(ワークフロー機能)、予測 AI と構造発生 AI に加えてドッキング等のスクリプトを登録・管理する機能、FL や追加学習（fine tuning）を行う機能、遺伝子発現データ等のオミクス情報に基づくターゲット予測、構造発生を行える機能を実装した。

構築した統合創薬 AI プラットフォームは、すでに 17 社の製薬企業に移植を実施しており、ユーザーから

の要望に対応して、機能強化や利便性の向上を行うために令和6年度内に3度のアップデートを行う予定である。また、令和6年度には、創薬ブースター、理研、京大、名大の創薬ターゲット（8ターゲット以上）に対して統合創薬AIプラットフォームを利用した設計が実施された。すべての創薬ターゲットについて、テーマ提案者、合成担当者、アッセイ担当者との会議を行い、目的を明確にしたうえで、報酬スコア(reward)を構成する予測手法の選択及びfine tuning、構造発生の実施、合成化合物の選定を行った。8ターゲットで合成まで、5ターゲットで主活性の評価まで進んでおり、既知化合物と遜色のない活性を持つ化合物か、新規骨格の発見に至っている。現在、ターゲット選択性、細胞、ADMET試験を進めている。また、プロジェクト終了後にも本研究課題の成果を活用するために、民間のIT企業が事業化する検討も進んでおり、IT企業の製品に本研究課題の成果を実装するための開発も開始されている。

統合創薬AIプラットフォームは、理研グループにおける構造発生AIに基づく新規化合物提案AI、京大グループにおけるオン・オフターゲット及びADMET予測を行う化合物プロファイル予測AI、名大グループにおけるオミクス情報を含む多階層データを考慮した創薬AIに分かれて開発を進めた。以下にそれぞれのグループの研究成果の概要を述べる。

新規化合物提案AIの開発においては、モンテカルロ木探索と再帰型ニューラルネットワーク（RNN）に基づき新規性・多様性を重視した構造発生ができるChemTS、医薬品らしさを目的として置換基を網羅的に組み合わせるRLS、骨格変換を目的とした遺伝的アルゴリズムに基づくXGG法の3種類の構造発生AI手法を開発するとともに、創薬の現場での実用性を高めるために複数の項目の同時最適化を実施できる手法(DyRAMO)の開発を行った。DyRAMOは、予測AIの適用範囲（Applicability Domain, AD）を考慮することも可能であり、ADを外れない領域で予測値が向上する構造を優先的に探索できる。EGFRを題材とした検証で、上市薬を含む有望な構造発生に成功している。また、その他にも近年の上市薬について、初期ヒットからの構造発生によって上市薬そのものを含む有望な構造を発生できることを検証した。予測AIのAD指標については、従来から使われてきた学習セットに対するタニモト係数による指標では、秘匿されている学習セットには利用できないため、学習セット内の構造フィンガープリントのビット利用頻度に基づく手法（SOOD）を開発した。これにより秘匿データセットの場合でもADを把握することが可能となった。予測AIのfine tuningによる精度向上については、半教師あり学習と自己教師あり学習の手法を開発した。発生した多くの構造から合成候補を絞り込むための構造スコアリング・フィルタリングについては、医薬品らしさ、忌避構造、環骨格の希少性などの手法を開発して実装した。また、合成難易度については、GNNに基づく合成経路探索手法RetRekとともに、RetRekによる合成経路探索の成功率をスコア化したRetRek scoreを開発した。RetRek scoreは、製薬企業研究者による発生構造のアンケートにおける合成展開の妥当性の集計結果を精度良く予測することに成功している。

化合物プロファイル予測AIにおいては、AI学習用のデータベースとしてChEMBL, PubChem, DruMAP等の公共データを統合し、企業データについてセキュリティに配慮しながら学習できるシステムを開発した。製薬企業のデータとしては、500種類程度のオン・オフターゲット、30種類程度のADMETについて延べ1500万データポイント程度の提供を受けた。学習手法としては、GNNに基づく独自の学習手法であるkMolを開発し、マルチタスク学習、マルチモーダル学習を行った。マルチモーダル学習による予測AIについては、予測対象の化合物の構造式とターゲットのアミノ酸配列を入力とすることで任意のターゲットに対する予測結果を得ることができるため、学習に投入した500種類以外のターゲットに対しても予測可能である。また、AlphaFold2型の記述子やTransformer型の学習ができる手法も開発しており、単純なGNNよりも良好な予測精度を示した。公共データのみの場合と企業データを加えた場合の学習効果の検証を行った結果、判別モデル（ROCAUC値で比較）及び回帰モデル（決定係数で比較）において、多くのターゲットにおいて企業データの追加によって有意な改善が認められ、一部では顕著な改善を示した。企業内に秘匿されているデータを学習できるFLについて、PyTorchによるシステムを開発し、製薬企業に導入して、インターネット経由で正常に動作することを確認した。

オミクス情報を含む多階層データを考慮した創薬AIについては、オミクス情報(特に遺伝子発現データ)を用いたターゲット予測AIを開発し、欠損値補完にはテンソル分解法を導入した。特に、LINCSデータベースの化合物を用いた階層的クラスタリングによる適用範囲を考慮した予測モデルを構築した結果、オミクスデータを用いることによって化合物構造情報に依存せず、ターゲット予測に有効であることが示された。学習データとしては、LINCSデータベースに加えて、遺伝子発現プロファイルの新規測定を行い、令和6年度までに約3000種類の化合物のデータを取得した。遺伝子発現プロファイルに基づいて構造発生する手法については、トランスフォーマーを導入した敵対的生成ネットワーク(GAN)を用いて学習する手法(TransGAN)及び変分オートエンコーダ(VAE)とRNNを組み合わせた手法(Gx2Mol)を開発し、既知リガンドを高精度で発生させることができることを確認した。また、合成経路を考慮できる構造発生AIであるcasVAEも開発した。ビシクロペンタンなどの未踏ケミカルスペースの合成法を開発し、物性や毒性のプロファイルが良好であることを検証した。

The Integrated Drug Discovery AI Platform was developed to comprehensively support the drug discovery process from hit identification to drug candidates. This platform incorporates cutting-edge technologies such as Monte Carlo Tree Search, Transformer-based structure generation AI, Graph Neural Networks (GNN) for predictive AI, and Federated Learning (FL) to enable practical use in drug discovery. The system utilizes data from both public sources and approximately 15 million on-/off-target, ADME, and toxicity data points provided by 17 pharmaceutical companies. Additionally, a secure system was established to perform FL, allowing internal confidential data to be used for training.

Key functions of the platform include project-based drug design combining predictive and structure generation AI, customizable workflows, management of scripts (such as docking), FL and fine-tuning capabilities, and target prediction/structure generation based on omics data like gene expression profiles. The platform has already been deployed to 17 pharmaceutical companies, with three updates scheduled for fiscal year 2024 to enhance functionality based on user feedback. It has been applied to design processes for eight or more drug discovery targets from institutes like AMED, RIKEN, Kyoto University, and Nagoya University, resulting in the synthesis and activity evaluation of new compounds.

The platform development involved collaboration with several groups:

New Compound Proposal AI (RIKEN): RIKEN group developed ChemTS for novelty, RLS for drug-likeness, and XGG for scaffold transformation, as well as DyRAMO for multi-objective optimization considering the Applicability Domain (AD) of predictive AI. A method called SOOD was developed to handle hidden data sets, improving AD identification for confidential datasets. Fine-tuning techniques such as semi-supervised and self-supervised learning were also implemented. Additionally, scoring methods for drug-likeness, rare scaffolds, and synthetic feasibility (RetRek score) were created.

Compound Profile Prediction AI (Kyoto University): Kyoto University group integrated public databases like ChEMBL, PubChem, and DruMAP with pharmaceutical companies' data for secure training. The GCN-based learning system kMol was developed for multitask and multimodal learning, allowing predictions for various targets. Performance improvements were observed across multiple targets, with the successful implementation of FL using PyTorch.

Omics-Based AI (Nagoya University): Nagoya University group developed a target prediction AI using gene expression data with tensor decomposition for data completion. Approximately

3,000 compounds' gene expression profiles were collected, and AI models such as TransGAN and Gx2Mol were developed to generate high-precision ligand structures. The casVAE model was also developed to consider synthetic pathways in the structure generation process.

The platform's success in drug discovery has led to ongoing commercialization efforts by private IT companies.