

2025 年度年次報告書

ライフサイエンスデータベース統合推進事業 統合化推進プログラム

【研究課題名】「AI 駆動型データキュレーションによる持続可能な中分子相互作用統合データベースの
開発」

研究代表者：

池田 和由（理化学研究所 計算科学研究センター ユニットリーダー）

研究分担者：

富井 健太郎（産業技術総合研究所 人工知能研究センター 上級主任研究員）



©2025 池田 和由(理化学研究所) licensed under CC表示4.0国際

1. 研究進捗状況と成果

本年度は、AI 駆動型データキュレーションによる中分子相互作用統合データベース「MIIDB-AI」の構築に向け、データ整備、キュレーション技術、構造情報を用いた拡張機能、およびデータベース (DB) システム開発を並行して進めた。データ整備では、公共 DB 由来の相互作用情報に加え、大学・連携研究機関由来の独自中分子データの収集方針を具体化し、ペプチド、非ペプチド性中分子化合物、核酸を対象に、構造・標的・活性・配列・相互作用部位情報を整理した。相互作用データの統一的な扱いに向けて、化合物・標的・アッセイ条件・相互作用情報を関連付けるためのメタデータ項目およびオントロジー体系の検討を行った。キュレーション技術では、大規模言語モデルおよび機械学習法を活用し、論文や公開 DB から化合物情報・標的情報・アッセイ情報を抽出・分類・整理するワークフローを構築し、専門家キュレーション結果との比較により実用性を評価した。特に、構造表記の正規化・標的名、アッセイ条件の整理、実験条件の比較、信頼性の低いデータの検出など、キュレーション自動化に必要となる要素技術について有効性と課題を明確化した。一方、構造情報を用いた拡張機能開発では、PDB から中分子と周辺ポケット残基を抽出し、約 25 万件の中分子相互作用ポケットを整理した。さらに、PoSSuM の結合ポケット情報や、AlphaFold、Boltz-2 等による予測構造情報と統合し、相互作用部位推定やアノテーション付加に向けた検討を進めた。DB システム開発については、プロトタイプ的设计を基に、SMARTS/SMILES/HELM 検索、分子種別による絞り込み、検索結果一覧、3D 相互作用の可視化機能を備えた β 版を開発した。 β 版には 4 万件超の中分子データ、約 7 千種類のペプチド、1,300 種類超の標的タンパク質を統合し、外部ユーザーおよび共同研究者から得たフィードバックを基に、正式版に向けた検索・表示・外部 DB 連携の改良方針を整理した。

2. 代表的な成果事例

1. Esaki, T. and K. Ikeda. Data curation in cheminformatics: Importance and implementation. *J Cheminform*, 18(1), 43 (2026).

概要：ケモインフォマティクスにおけるデータキュレーションの重要性と実装上の課題を整理し、信頼性の高い化学・生物活性データ整備の必要性を示した。本研究で進める AI 駆動型データキュレーションの基盤となる知見である。

2. Shimizu, Y., Akabane, T., Ohta, M., Honma, T. and Ikeda, K. ChEMBL-derived benchmark dataset and computational results of Boltz-2-based binding affinity prediction. *Chem-Bio Informatics Journal*, 25, 12-17 (2025).

概要：ChEMBL 由来の結合親和性予測用ベンチマークデータセットを構築し、Boltz-2 を用いた予測性能を評価した。中分子を含む化合物-標的相互作用予測や、MIIDB-AI における解析・予測機能の高度化につながる成果である。

以上