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)
研究開発実施報告書　様式

2024 年度 研究開発実施報告

概要

研究開発課題名	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
開発対象データベースの名称(URL)	MIIDB-AI: Middle molecule Interaction Integrated DataBase using AI (https://www.miidb-ai.com) (仮)
研究代表者氏名	池田 和由 (00415770)
所属・役職	理化学研究所　計算科学研究センター　ユニットリーダー (2025 年 3 月時点)

□目次

§1. 研究実施体制

グループ名	研究代表者・研究分担者氏名	所属機関・役職名	研究題目
池田グループ	池田 和由	理化学研究所・ユニットリーダー	中分子データの収集とMIIDB-AIデータベースの開発
富井グループ	富井 健太郎	産業技術総合研究所・研究チーム長	MIIDB-AIデータベースのための拡張機能の開発と実装
米澤グループ	米澤 朋起	慶應義塾大学・特任助教	MIIDB-AIデータベースに収載する新規の中分子データの取得とデータキュレーション法の開発支援

§2. 研究開発対象とするデータベース・ツール等

(1) データベース一覧

【主なデータベース】

No.	名称	別称・略称	URL
1	Middle molecule Interaction Integrated DataBase using AI（仮）	MIIDB-AI（仮）	https://www.miidb-ai.com

【その他のデータベース】

No.	名称	別称・略称	URL
1	DLiP		https://skb-insilico.com/dlip
2	PoSSuM		https://possum.cbrc.pj.aist.go.jp/PoSSuM/
3	PoSSuMAF		https://possum.cbrc.pj.aist.go.jp/PoSSuMAF
4	PreBINDS		https://prebinds.airc.aist.go.jp/

(2) ツール等一覧

No.	名称	別称・略称	URL
1	KNIME		https://www.knime.com/
2	RDKit		https://www.rdkit.org/
3	Pipeline Pilot		https://www.3ds.com/ja/products-services/biovia/products/data-science/pipeline-pilot/
4	AlphaFold		https://alphafold.ebi.ac.uk/

§3. 実施内容

(1) 本年度に計画されていた研究開発項目・タスク

【項目1】データの作成とキュレーション

【項目 1-1】中分子相互作用データの収集

・公共データからの中分子相互作用データの収集

・独自の中分子相互作用データの検討・入手

【項目 1-2】メタデータの収集

・公共データからの中分子関連メタデータの収集

・オントロジーの検討・入手

【項目 1-3】データキュレーション法の開発

・キュレーションプロトコルの作成

・データキュレーション効率化の検証

【項目 2】DB システムの設計と開発

【項目 2-1】プロトタイプ DB の設計と開発

・プロトタイプ DB の設計

【項目 3】拡張機能の開発

【項目 3-1】構造情報を利用した機能開発

・立体構造情報の利用による標的アノテーション法の開発

【項目 3-2】標的予測の精度向上

・AI 活用による中分子標的予測のベンチマーク

(2) 進捗状況

（本年度に実施した研究開発項目・タスクについて）

【項目1】データの作成とキュレーション

【項目 1-1】中分子相互作用データの収集

1) 公共データからの中分子相互作用データの収集

公共データベースから、分子量 650 以上の中分子リガンドと標的タンパク質の相互作用データを収集した。ChEMBL(version 34)と PubChem の調査から、特に活性が明確（IC50、EC50 ≤ 10 μM）な中分子化合物を優先的に収集した。中分子化合物数が 24,181 個、その相互作用データが 46,583 件、そしてタンパク質標的数は 1,354 種類（うち PPI 標的が 26 種類）であった。

2) ペプチドデータの収集

公共データベースから、天然および非天然アミノ酸を含むペプチド（3～50 残基、分子量 650～5,000）の相互作用データを収集した。まず、DLiP データベースからペプチド 117 個、ChEMBL から医薬品ペプチド 49 個、および抗菌ペプチドの活性情報 1,550 件、IEDB および PPI3D データベースから数千件のデータを収集した。これら収集したペプチド情報は、SMILES 及び HELM 配列で記述可能で

あり、特に活性が明確なものを優先的に収集した。また、承認済みペプチド薬28種について、e-Drug3Dなどからの追加情報も統合した。

3) 核酸データの収集

核酸分子としては、ChEMBLから収録された238個のオリゴヌクレオチド分子について調査を行った。さらに、AptaDBからアプタマーを中心に実験的に検証された1,293件の核酸配列を収集し、このうちRNA配列の264個を抽出した。PDB内の核酸データとの整合性を検討した結果、PDBではオリゴヌクレオチドが主にChainとして登録されていることが明らかとなり、今後、検索アルゴリズムの改善を我々のグループで検討してく予定である。

4) 独自の中分子データ収集

慶應義塾大学やAMED-DISCと連携することで、独自性の高い中分子データを収集した。PPI標的指向性中分子ライブラリーとして、15,214個の中分子化合物を収集した。また、DISCライブラリーとして、市販中分子化合物カタログの70万化合物から構造フィルター、代表構造クラスタリング、膜透過性・溶解度予測を用いて、約4万個を選抜したものを入手した。

【項目 1-2】メタデータの収集

1) 公共データからの中分子関連メタデータの収集

収集したペプチド承認薬(49個)に対して、メタデータを体系的に収集・整理した。まず、Chemical Information Ontology(CHEMINF)の化学記述子オントロジー、ChEBIの生物学的・機能的分類オントロジー、WHOの解剖治療化学分類(ATC)コード、およびPoSSuMポケット情報、OMIM・Orphanetからの疾患などの付加情報を対象とした。手順として、ChEMBL IDからPubChem CIDおよびChEBI IDを特定し、PubChem REST APIを用いて分子量、LogP、TPSAなどの化学記述子を取得し、CHEMINFオントロジーにマッピングした。さらに、ChEBI SPARQLエンドポイントを用いて生物学的役割・化学分類を抽出し、ChEMBLからATCコードを調査・照合した。結果として、中分子ペプチドのPubChem CIDを約90%、ChEBI IDを約80%、ATCコードを約75%で特定できた。

2) オントロジーの検討・入手

メタデータ情報の事前調査を通じて、中分子は低分子とは異なり、ペプチドなど構造記述の複雑さなどの特徴があるが、既存のオントロジーで一定のカバーが可能である。まず、化学構造や物理化学的性質については、CHEMINFを用いて標準化された記述が可能で、PubChem CIDやChEMBL IDをキーとして統一的に情報を取得・管理できる。さらには、生物学的活性・薬理作用に関してChEMBL・ChEBIオントロジーを活用することで、機能や治療用途など関連付けに有効であることが分かった。標的・疾患に対するオントロジーとしてのOMIMとOPRHANETは、ChEMBL (Target) IDからマッピングが可能である。

以上の分析から、中分子に対して既存のCHEMINFおよびChEBIを中核としつつ、薬理・標的情報との接続にはChEMBL、さらに臨床的適応にはATCコード、実験手法にはBioAssay Ontologyを補完的に用いることで、拡張性の高いメタデータ記述が可能であるという結論に至った。今後は、中分子特有の構造・機能的特徴が既存オントロジーで十分に記述されていないことが明らかとなり、専用の中分子オントロジー整備の必要性が示唆された。

【項目 1-3】データキュレーション法の開発

1) キュレーションプロトコルの作成

大規模言語モデル（Large Language Model：LLM）を中心とした対話型 AI 技術を活用し、化学・標的・アッセイのデータキュレーションを自動化するためのツール開発に着手した。まず、OpenAI が提供する ChatGPT を用いて、化合物の SMILES や医薬品情報において重要な化学的および生物学的データを効率的に抽出・整理できるか検討した。事前調査として行った医薬品医療機器総合機構（PMDA）が提供する医薬品添付文書からの薬物動態特性（Cmax など）を自動抽出するシステムの開発と検証により、ChatGPT は表形式のデータを含め、高精度に目的の情報を抽出可能であった[1]。また、化学構造のデータキュレーションとして、SMILES の文法的に誤った表記を GPT で校正する手法を試行した。その結果、AI によって生成された SMILES に含まれる誤りを GPT が部分的に修正可能であることが明らかとなり、化学構造データのキュレーションの効率化に活用できる可能性を示した。ただし、大幅な文法修正を必要とする場合の対応には限界があり、今後の改良課題として残った。

2) データキュレーション効率化の検証

中分子を対象とした標的予測の効率化を目指した。具体的には、ChatGPT を用いた PPI 標的予測システムを開発した。このシステムでは、ベースモデルとして GPT 3.5 Turbo を採用し、標的タンパク質とリガンド分子の相互作用データ計 5,000 件を用いてファインチューニング（最適化学習）を実施した。モデルの訓練および検証は、データを 1:1 の割合で分割して行った。開発したモデルでは、リガンド分子の SMILES を入力すると、それに対応する PPI 標的タンパク質を予測することが可能である。具体的な手順としては、OpenAI の API を介して ChatGPT を起動し、ユーザーからリガンド化合物の SMILES を含む質問を送信する。ChatGPT は受け取った情報を基に PPI 標的予測を行い、予測結果を返却する。現在、この予測システムの精度評価と改善を継続的に実施している。今後、精度および信頼性のさらなる向上を目指す。

【項目 2】DB システムの設計と開発

【項目 2-1】プロトタイプ DB の設計と開発

1) プロトタイプ DB の設計

中分子ごとに整理された検索機能を搭載したプロトタイプ DB を設計し、ユーザーの操作性を高めるためのインターフェース設計を行った。まず、中分子の分子情報、標的情報、さらに相互作用情報を体系的に登録・整理できるよう、データベーススキーマを設計した（図1）。このスキーマは、分子情報を管理する「molecule dictionary」テーブル、分子構造の情報を扱う「compound structure」テーブル、標的に関する情報を保持する「target」テーブル、分子と標的間の相互作用情報を記録する「activity」テーブル、および相互作用の設計仮定情報を意味する「(ppi) target design assumption」テーブルを含む構成とした。これらのテーブルを相互に関連付け、データの一元的な管理を実現した。

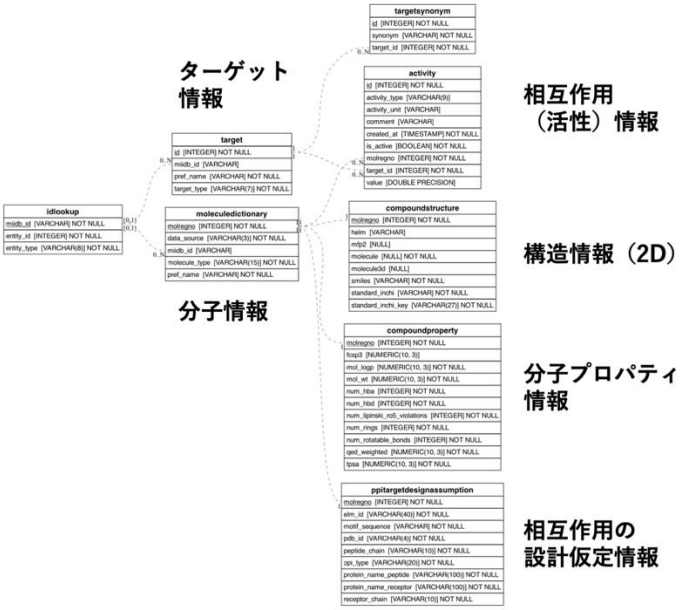

図1：プロトタイプ DB におけるデータベーススキーマ。

次に、プロトタイプ DB としての開発を進めるにあたり、ユーザーが必要な情報を直感的かつ迅速に検索できるシンプルなインターフェースを設計・実装した。特に、分子種別（非ペプチド[MEDIUM COMPOUND]、ペプチド[PEPTIDE]、核酸[NUCLEOTIDE]）を明示的に区別して検索できるように設計したほか、化合物の構造検索（SMILES）および配列検索（HELM）の機能を搭載した。これにより、収集したデータへのアクセス性と操作性を向上させた。

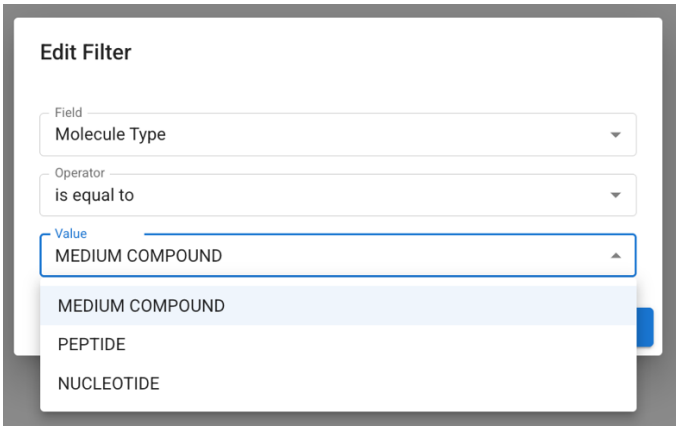

図2：プロトタイプ DB のインターフェースの検索・フィルター画面（現時点では非公開）。

ペプチドの HELM 配列を入力して、検索を行った結果、表示される画面を以下に示す。この画面では、検索条件に合致した中分子の構造が視覚的に表示されるほか、各分子に関するプロパティ情報が一覧として提供される。具体的には、分子の識別番号（MIIDB-ID 仮）、分子構造、分子量（MW）、LogP 値、水素結合アクセプター数（HBA）、水素結合ドナー数（HBD）や、医薬品らしさの評価指標 QED（Quantitative Estimate of Drug-likeness）スコアなどが示される。また、相互作用情報として生物・薬理学的情報（PPI などの相互作用情報の詳細）が表示される。これにより、ユーザーは対象中分子とそれに類似している中分子の特性および相互作用の内容を迅速に把握することが可能である。

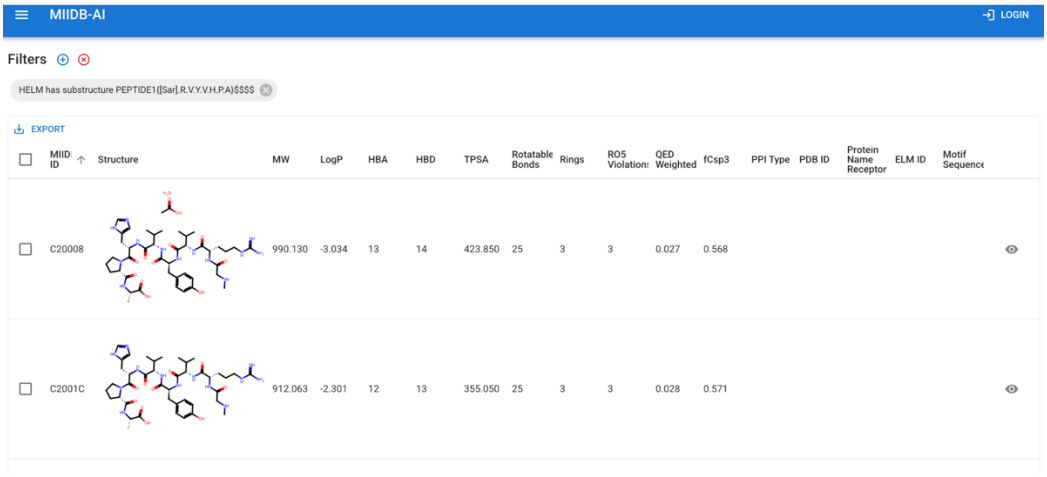

図3:プロトタイプ DB の検索結果画面。中分子化合物の検索結果の例。

次に、三次元相互作用 Viewer では、中分子(中央部にオレンジ色で示される)と、それと物理的に接触しているタンパク質の相互作用表面(周囲にグレー色で示される)を可視化する。この Viewer を利用することで、ユーザーは活性情報に加えて、分子とタンパク質間の実際の接触状態や結合ポーズを三次元的に観察でき、中分子がどのような立体的配置で標的と相互作用しているかを把握し、結合特性や活性機構の理解を深めることが可能となる。

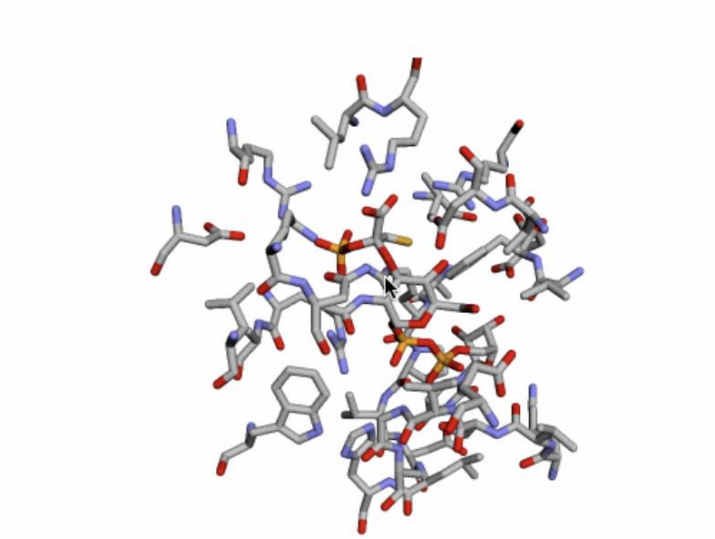

図4:三次元相互作用 Viewer。

以上の実施により、計画時に設定したプロトタイプ DB 設計を予定通り達成した。今後の予定として、2025 年度内の公開版リリースに先立ち、システムの機能面での改善点を早期に特定しより完成度の高い公開版 DB を提供する目的で、利用者を限定したβ版による試行運用を実施する。具体的には、共同研究を行っているアカデミア研究者や製薬企業関係者を中心にβ版の提供とテストを実施し、実際の研究現場からのフィードバックを通じてインターフェースや機能面の問題点を洗い出し、改善を図る予定。核酸などの中分子の取り扱いや予測(Prediction)データの表示機能の強化を図るとともに、公開版に向けた DB システム・インターフェース開発課題の洗い出しを行う。

【項目 3】拡張機能の開発

【項目 3-1】構造情報を利用した機能開発

立体構造情報の利用によるアノテーション法などの開発に向け、まず PDB に登録されている実験的に決定された一定の解像度(4.0 Å)より良いタンパク質と中分子(分子量>650)の複合体構造を対象とし、情報抽出と解析を行なった。中分子の結合部位の定義は PoSSuM 構築時のものを参考にし、117,375 結合部位を同定した。ここには 2,564 種の中分子が含まれており、分子量が 1,500 を超えるものの頻度は所謂「ロングテール」のような分布であることが明らかとなった。これら分子は、有機化合物がほとんどであるが、無機化合物も含まれており、それらを分けた上で、結合部位と中分子の座標データを代表機関に提供した。このデータは、今後データベースの一部となる予定の基礎的データであり、研究代表者と協議の上 PDBx/mmCIF 形式で統一した。また、PoSSuMAF の推定データセットも代表機関に提供した。

上記データの分析から、クロロフィル a が全体の 3 割程度を占めており、また、エントリーあたりの分子数が"bulk"分別の大まかな指標となり得ることが明らかになった。また更なるデータの取捨選択や drug-like や cofactor などのような分類やアノテーション付加に向け、DrugBank や Chemical Component Dictionary (CCD), Biologically Interesting molecule Reference Dictionary (BIRD) あるいは、PDBe のアノテーションツールなどの利用を検討した。ただし、いずれも被覆率の観点からみると高い割合ではなく、それらの併用や活用に向けた協議を研究代表者と続けている。

ペプチドや核酸などについては取り扱いがやや難航している。ペプチドについては、適切な対象の選択が、化合物の場合と比して困難なものであった。協議の結果、少なくとも当面、承認薬あるいは抗菌ペプチドなどを主な対象とすることとなったものの、阻害剤(候補)などにも拡張すべく、データ収集と分析を行なっている。核酸については、主にアプタマーなどを念頭に置いていたが、その件数は構造未知のものを含めても限定的であるため、推定データの構築は容易ではなさそうな見込みである。今後、この課題について代表機関と相談の上、解決案を検討する。

立体構造情報を用いたアノテーション開発においては、これまで主に PDBe および wwPDB の関連ツールを利用してきた。今後の開発強化を目的として、PDBj の各種ツールについても調査を進めており、PDBj の関係者と連絡をとりながら連携を深めていく予定。

【項目 3-2】標的予測の精度向上

より信頼度の高い推定やアノテーションの充実に向け、実験的に決定された既知データや推定データの結合部位やポーズについて、AlphaFold3 による予測結果の比較を行った。その結果、結合部位やポーズが比較的正確に予測された場合でも、構造に歪みが生じる化合物がみられた。また、信頼度の高い推定データの提供に適した指標のさらなる調査が必要となった。

§4. 成果発表等

(1) 原著論文発表

① 論文数概要

種別	国内外	件数
発行済論文	国内(和文)	1 件
	国際(欧文)	0 件
未発行論文	国内(和文)	0 件
	国際(欧文)	0 件

② 論文詳細情報

1. 江崎剛史、小川慶子、池田和由、添付文書から薬物動態特性を抽出する方法の検討:対話側 AI システムの活用、医薬品情報学(26 巻 2 号 P80～P91)、2024 年

(2) その他の著作物(総説、書籍など)

1. 江崎剛史、渡邉怜子、熊澤啓子、土井雄貴、小川慶子、池田和由、GPT(Generative Pre-trained Transfer)を化学分野で活用する方法の検討、CBI 学会誌、第 13 巻第 1 号、P28-P31、2025 年)

(3) 国際学会および国内学会発表

① 概要

種別	国内外	件数
招待講演	国内	4 件
	国際	0 件
口頭発表	国内	1 件
	国際	0 件
ポスター発表	国内	2 件
	国際	0 件

② 招待講演

〈国内〉

1. Kazuyoshi IKEDA、Accelerating Middle Molecule Drug Discovery with Evolving Next-Generation AI Platform and Databases、The R-CCS Café, オンライン、2024 年 7 月 19 日

2. 池田和由、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の開発、トーゴーの日シンポジウム 2024、東京、2024 年 10 月 5 日

3. 池田和由、Transforming Medicinal Chemistry: Accelerating Lead Discovery and Development with AI、慶應医学賞サテライトシンポジウム、東京、2024 年 11 月 21 日

4. 池田和由、データベースと AI の相互進化による医薬品設計の革新、MOLSIS 創薬 DX セミナー、オンライン、2025 年 2 月 27 日

〈国際〉
なし

③ 口頭講演

〈国内〉
1. Kazuyoshi IKEDA, Tomoki YONEZAWA, Masanori OSAWA, Tsubasa NAGAE, Kentaro TOMII、Development of a sustainable database for middle molecules using AI-driven data curation、CBI 学会 2024 年大会、東京、2024 年 10 月 28 日-31 日

〈国際〉
なし

④ ポスター発表

〈国内〉
1. 永江翼、池田和由、富井健太郎、中分子相互作用データベース開発に向けたデータ収集・統合・解析、統合化推進プログラム研究交流会、東京、2024 年 10 月 4 日

2. 永江翼、池田和由、富井健太郎、中分子創薬研究のためのタンパク質-リガンド相互作用データベースの構築:Protein Data Bank の包括的解析、日本薬学会第 145 年会、福岡、2025 年 3 月 27 日(木)

〈国際〉
なし

(4) 知的財産権の出願 （国内の出願件数のみ公開）

① 出願件数

種別		件数
特許出願	国内	0 件

② 一覧

国内出願

なし

(5) 受賞・報道等

① 受賞

なし

② メディア報道

なし

③ その他の成果発表

なし

§5. 主要なデータベースの利活用状況

1. アクセス数

① 実績

表 1 研究開発対象の主要なデータベースの利用状況

名称	種別	2024 年度（月間平均値）
MIIDB-AI（仮）	訪問者数	未公開
	訪問数	未公開
	ページ数	未公開

② 分析

開発中のため未公開。

2. データベースの利用状況を示すアクセス数以外の指標

なし。

3. データベースの利活用により得られた研究成果（生命科学研究への波及効果）

なし。

4. データベースの利活用によりもたらされた産業への波及効果や科学技術のイノベーション（産業や科学技術への波及効果）

なし。

§6. 研究開発期間中に主催した活動(ワークショップ等)

(1) 進捗ミーティング

年月日	名称	場所	参加人数	目的・概要
2025 年 9 月 25 日	チーム内ミーティング(非公開)	産総研臨海副都心センター別館	3 人	研究進捗報告のためのミーティング

(2) 主催したワークショップ、シンポジウム、アウトリーチ活動等

年月日	名称	場所	参加人数	目的・概要

以上

別紙１　既公開のデータベース・ウェブツール等

No.	正式名称	別称・略称	概要	URL	公開日	状態	分類	関連論文
1								