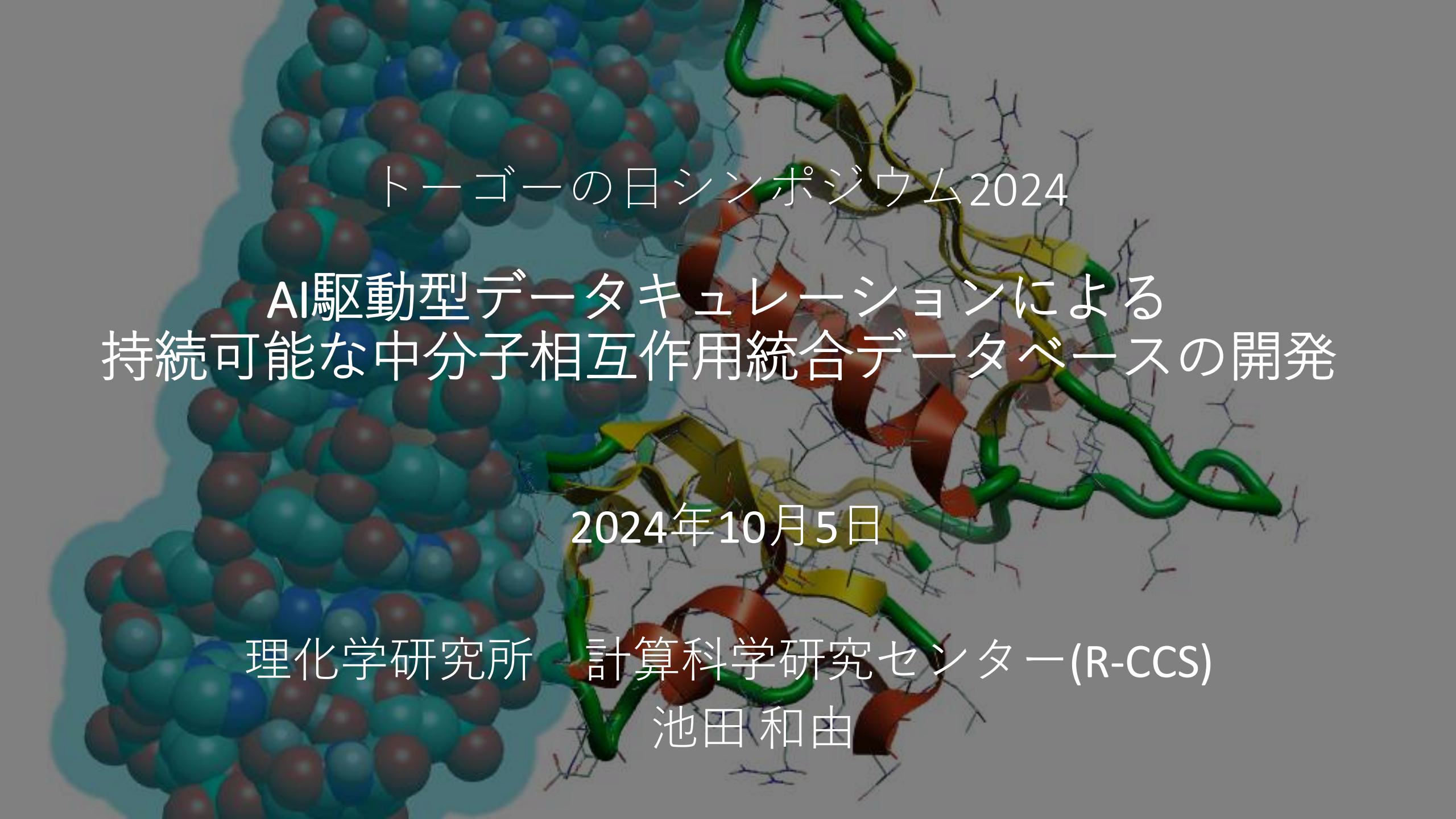


The background of the slide features a complex molecular simulation. On the left, a large, dense cluster of atoms is shown as a space-filling model, with red spheres representing oxygen, blue spheres representing nitrogen, and teal/grey spheres representing carbon and other atoms. To the right, a protein structure is depicted using a ribbon representation, with green and yellow ribbons indicating different parts of the molecule. The overall scene is set against a dark, semi-transparent grey background.

トーゴの日シンポジウム2024

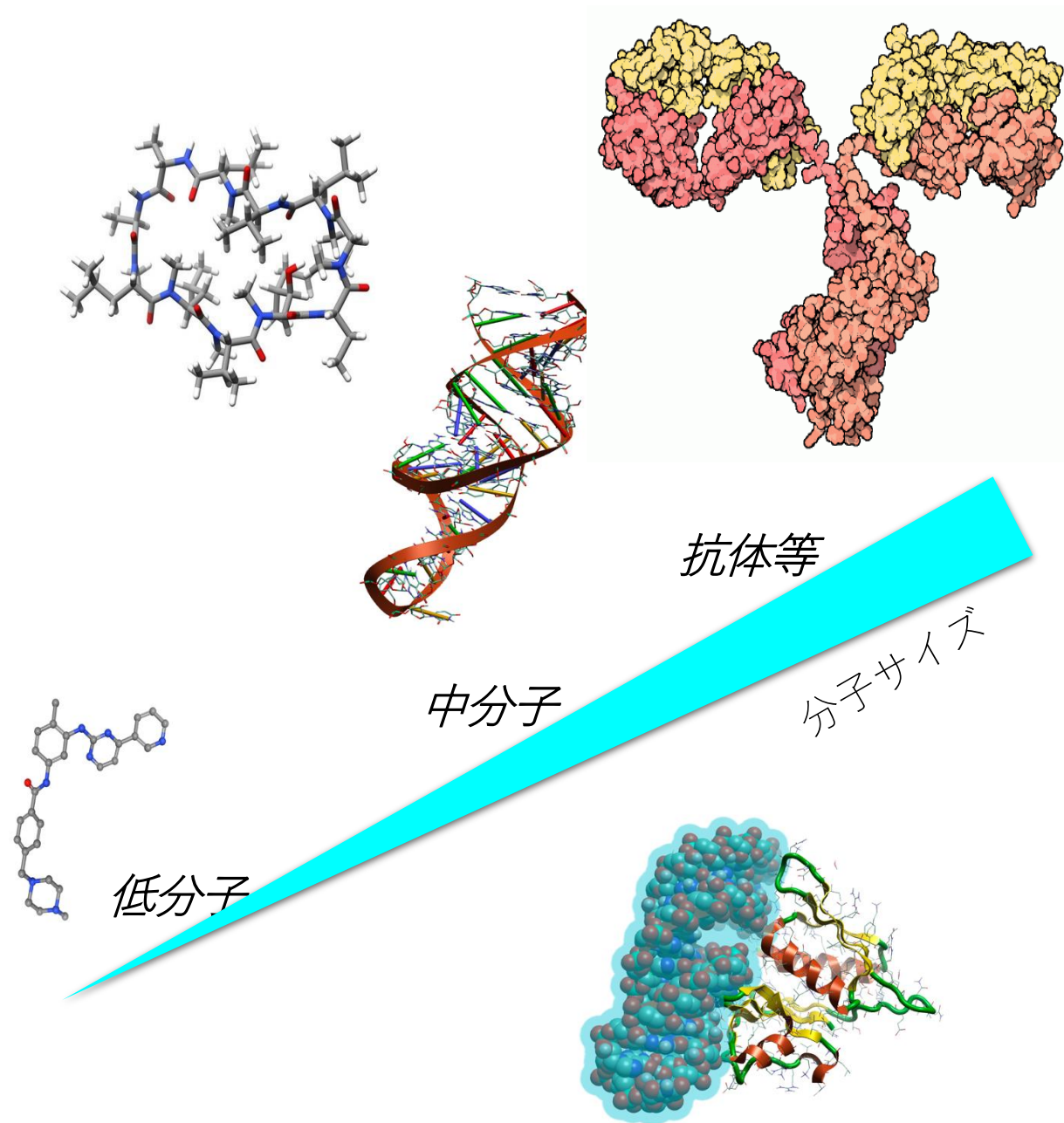
AI駆動型データキュレーションによる
持続可能な中分子相互作用統合データベースの開発

2024年10月5日

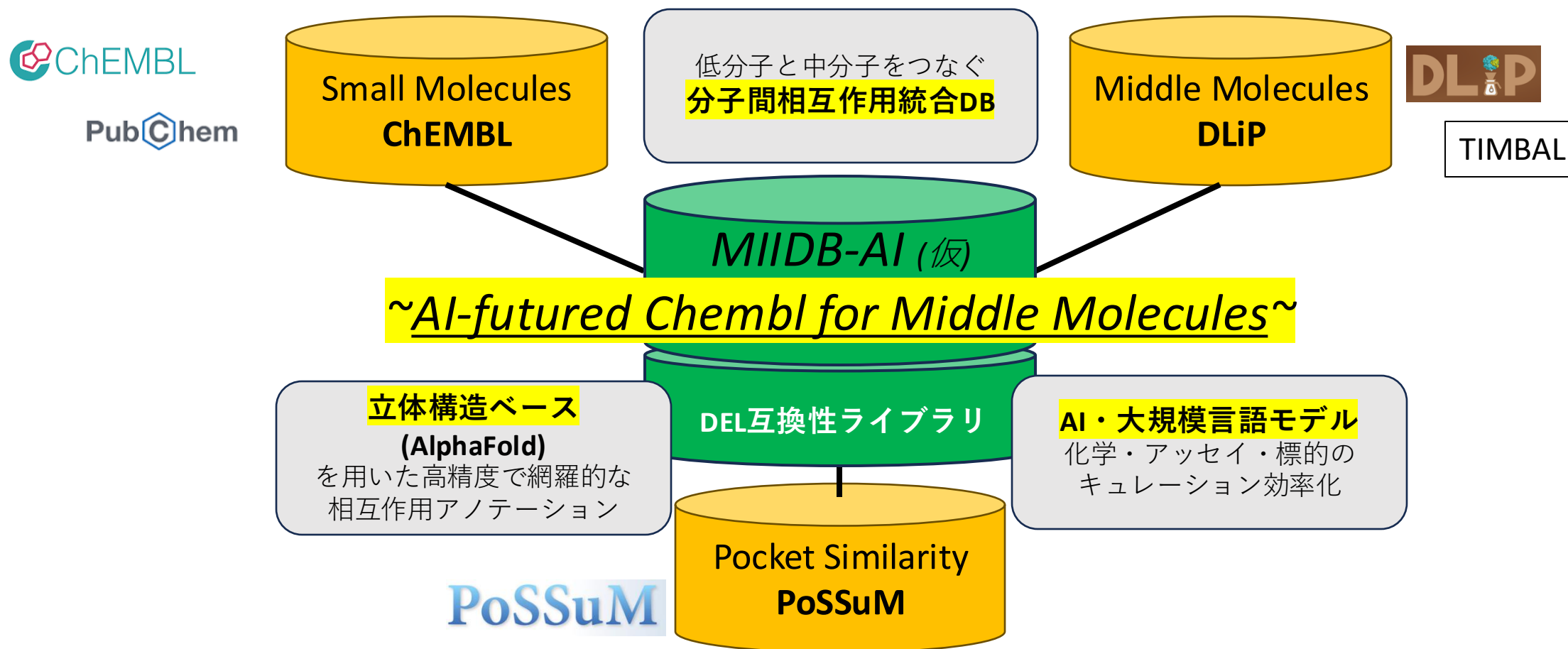
理化学研究所 計算科学研究センター(R-CCS)
池田 和由

背景

- 中分子は、低分子とバイオ医薬品の間位置する。ペプチドや核酸、中分子合成化合物などが該当する。
- 特異性が高く、副作用が少ないことから、従来の医薬品では治療が難しかった疾患領域に対する新しいアプローチとして期待されている。
- 中分子に関するデータは不足している。製薬企業やアカデミアにとって有用な相互作用データリソースが未だない。
- 中分子領域で医薬品開発に直結する相互作用データを収集・統合した一元的なデータベースの開発に着手した。

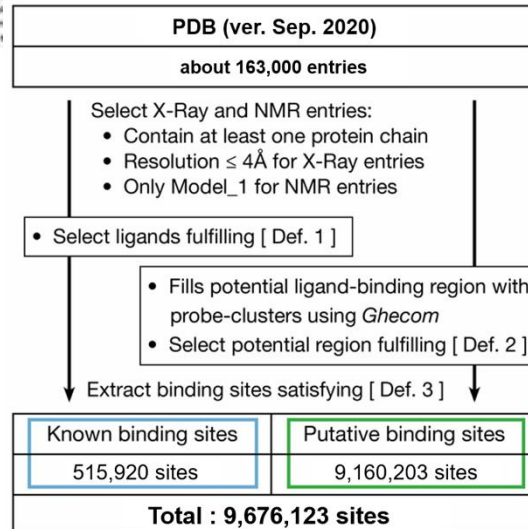


データベースのコンセプト

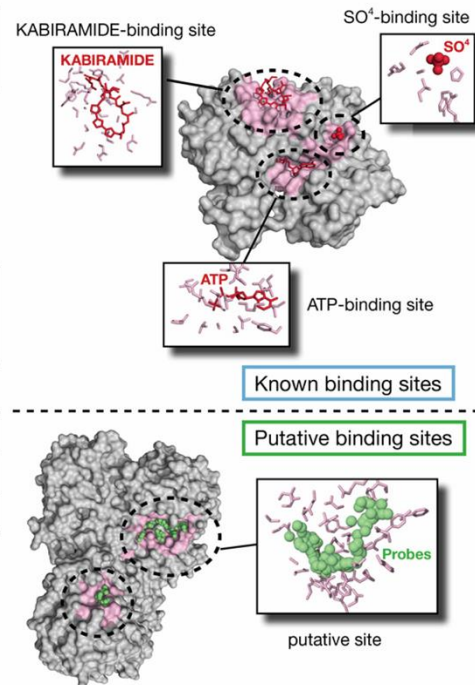


- 相互作用の「質」を担保しつつ、AIを活用することで効率化させる。
- 新規中分子薬候補の発見、次世代医薬品開発へ貢献する。

網羅的リガンド結合部位情報



Ito et al. 2012, 2015.



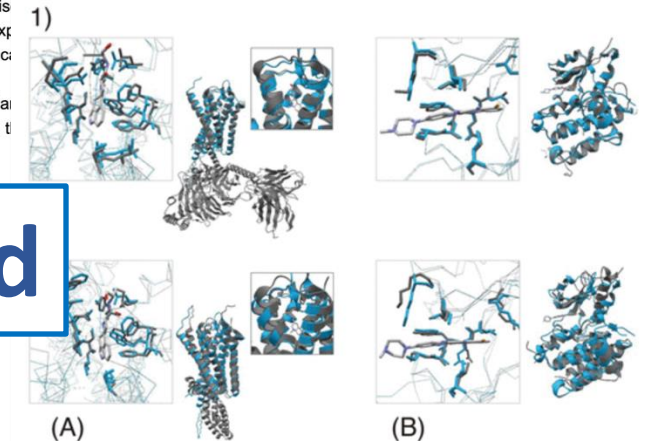
About PoSSuMAF

PoSSuMAF, a database of putative ligand-binding pockets on AlphaFold models of human proteins that are similar to known ligand-binding sites, stored in the **database of AlphaFold**, as of December 17th, 2021, was developed by applying **PoSSuM scheme**. The result was compiled as a relational database Pocket Similarity Search using Multiple-Sketches (PoSSuM) AF, which includes all the dis and Gene ontology). It enables rapid ex binding sites on PDB structures. Users c

i) "Search K" is useful for finding similar known ligand-binding site registered in t query.

AlphaFold

Tsuchiya et al. 2023.

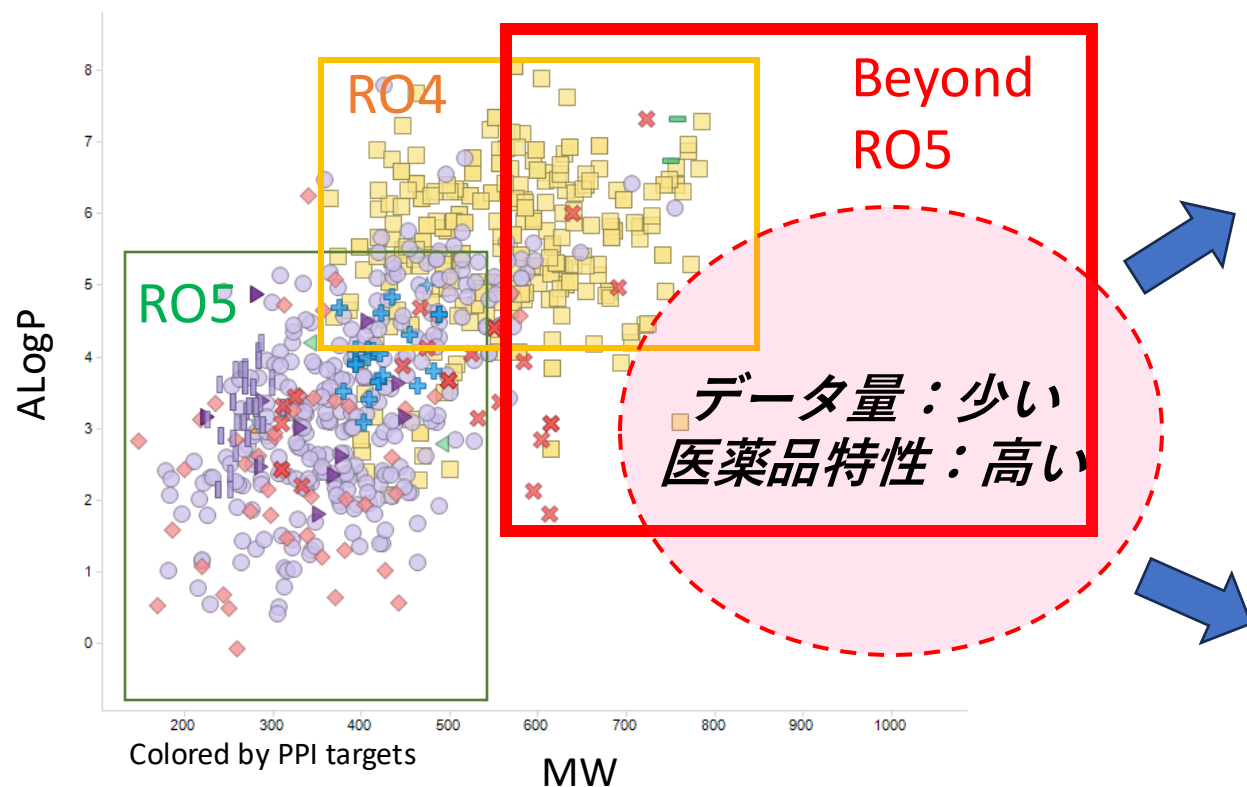


PDB中のポケットの網羅的な類似性データベース

AlphaFoldモデルのポケット間類似性
<https://possum.cbrc.pj.aist.go.jp/PoSSuMAF/>

新規の中分子相互作用データの取得

既知の活性中分子ケミカルスペース

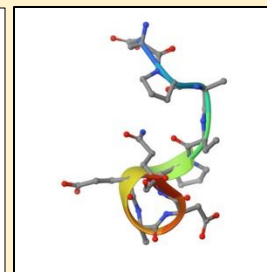
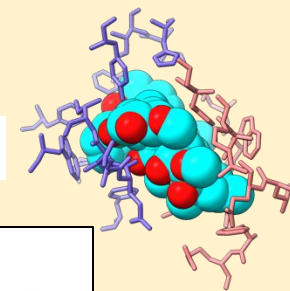


ChEMBL中のPPI阻害剤(抜粋)

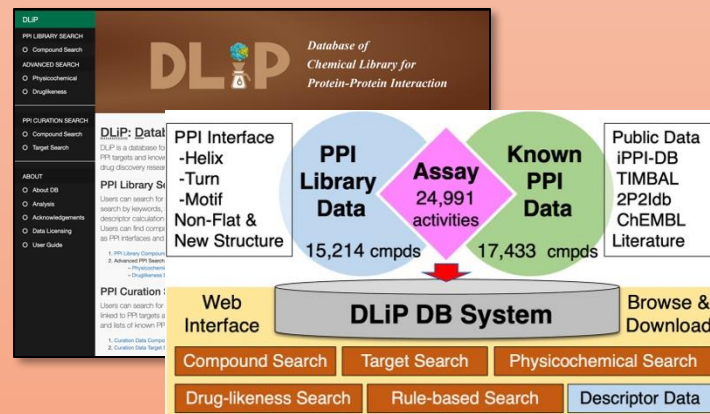
公共DBからのデータ抽出



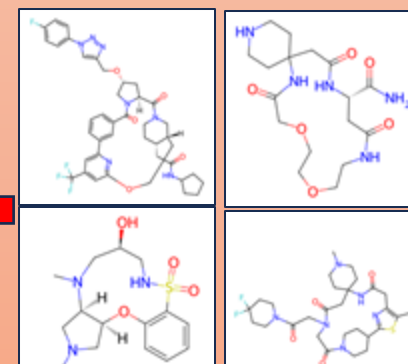
DRUGBANK



中分子ライブラリーの利用



ikeda et al. 2023.



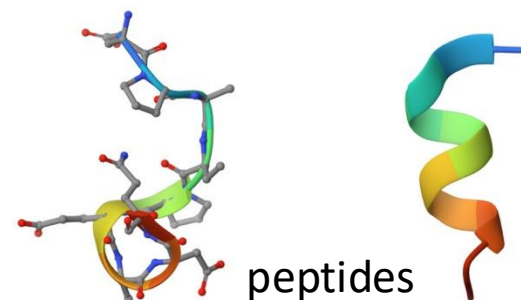
2種類の標的に対する
スクリーニング結果

公共DBからの中分子データの取得

- 公共DB (PDB, ChEMBL, DrugBank等)から中分子リガンドと標的(タンパク質)の情報を収集した。

Targets	# of Targets
Protein Targets	~10,000
PPI Targets	101
PROTAC Targets	280

Ligands	# of Molecules
Mid-Compounds	>17,000
Cyclic Peptides (Ike)	>1,200
RNA Aptamers	~1,500



4ch9_C EPEEPADQHQ	Serine/threonine-protein kinase WNK4	3lnz_D TSFAEYWALLSP	synthetic peptide
----------------------	--------------------------------------	------------------------	-------------------

- 立体構造に基づく相互作用データの取得方法の検討

PDBからデータ収集:

- 解像度4Å以下の構造データ
- mmCIFフォーマットで取得
- 分子量650以上の分子

↓
リガンドの重原子から5.0Å以内に存在するアミノ酸残基を結合部位として定義

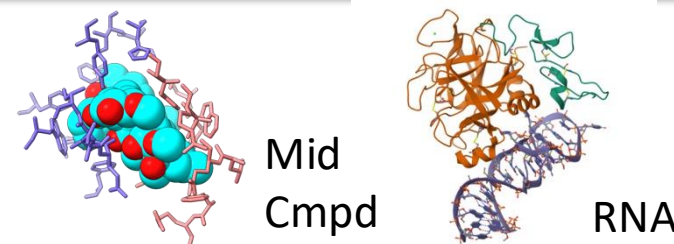
結果:

収集リガンド数: 2,564種類
結合部位数: 117,375個

↓
リガンド-タンパク質の相互作用の大規模なデータセットが作成された。

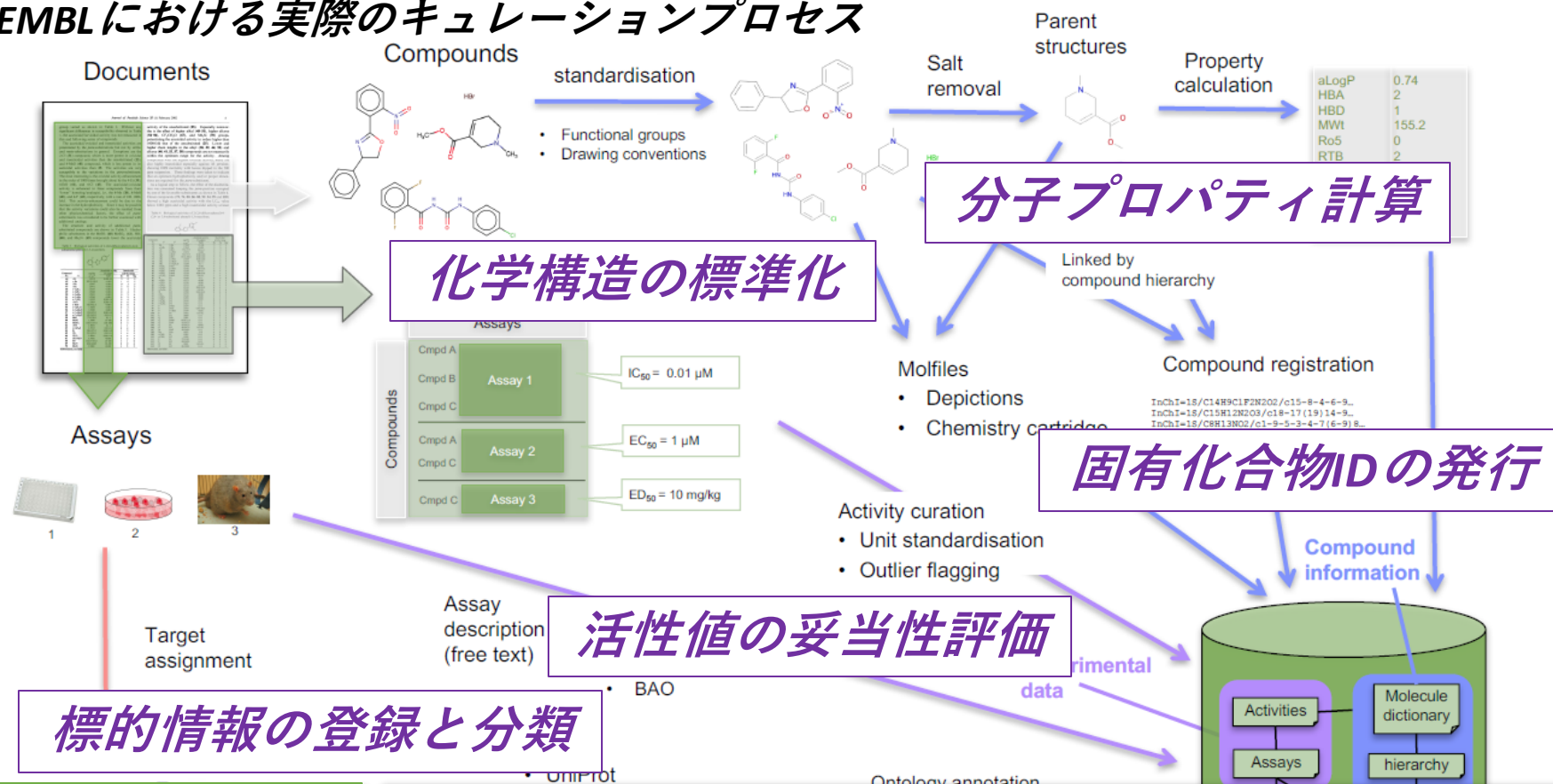
今後の方向性:

- まだ医薬品適性が低いものが含まれている。適切な選別基準を導入する必要あり。



従来のマニュアルキュレーションの流れ

ChEMBLにおける実際のキュレーションプロセス



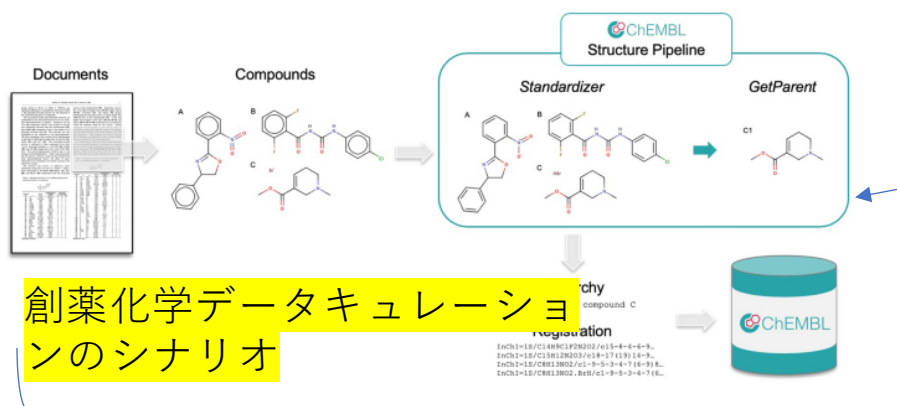
① 化合物(リガンド) キュレーション

② 標的(タンパク質) キュレーション

③ アッセイ(実験条件) キュレーション

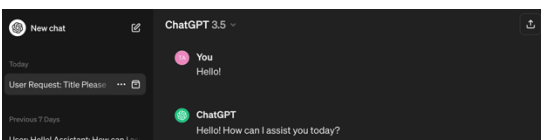
データベースの管理・更新コストを下げる
持続可能なDB開発を目指す

LLMデータキュレーション技術の開発と検証



対話型AI

(手順)
1. OpenAI APIからChatGPT起動



```
import openai
completion = client.chat.completions.create(
    model='<model_name>'
    messages=[
        {'role': 'system', 'content': 'Chatbot for PPI target prediction.'},
        {'role': 'user', 'content': 'What are the PPI targets for the following compounds? \n <SMILES>'}
    ]
)
target_name = completion.choices[0].message.content
```

2. 最適化(ファインチューニング)のプロセス

USER	What are the PPI targets for the following compounds? <chem>COc1ccc(S(=O)(=O)N(CC(=O)O)c2ccc(N[C@@H](C(=O)O)[C@@H](C)OC(C)(C)C)c3ccccc23)cc1</chem>
ASSISTANT	Integrins

3. 予測結果の精度検証と改良

① Chemical Curation

- 化学構造標準化プログラムの自動生成・修正 LLM(GPT)
- DEL互換性ライブラリーへの変換 Rule-Based・LLM

難易度

低

② Target Curation

- AFによる標的ドメインアノテーション
- 標的指向性予測や標的クラス分類

Rule-Based
ML/DL

中

③ Assay Curation

- Assay Descriptionの要約・実験条件抽出
- Assay Typeの分類モデル

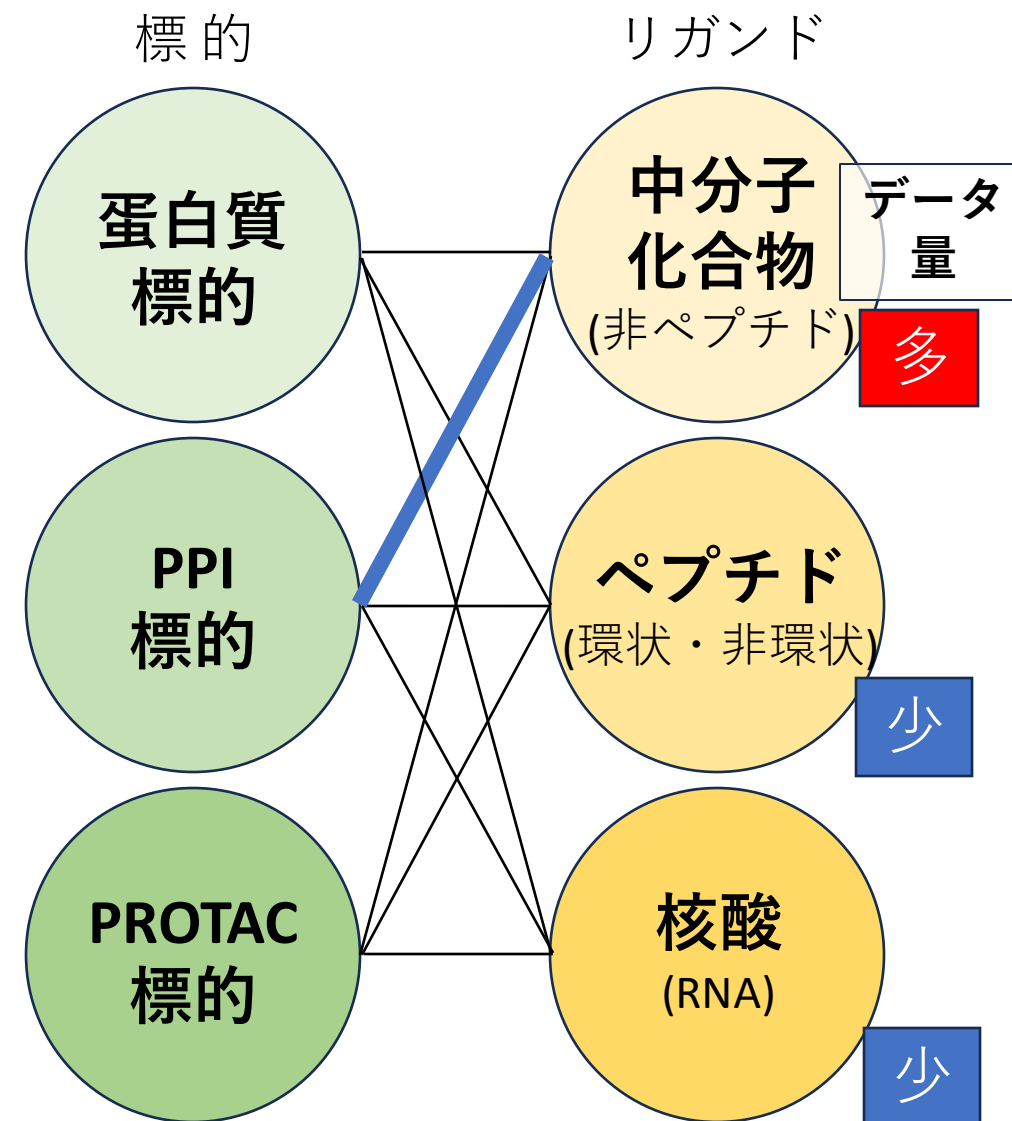
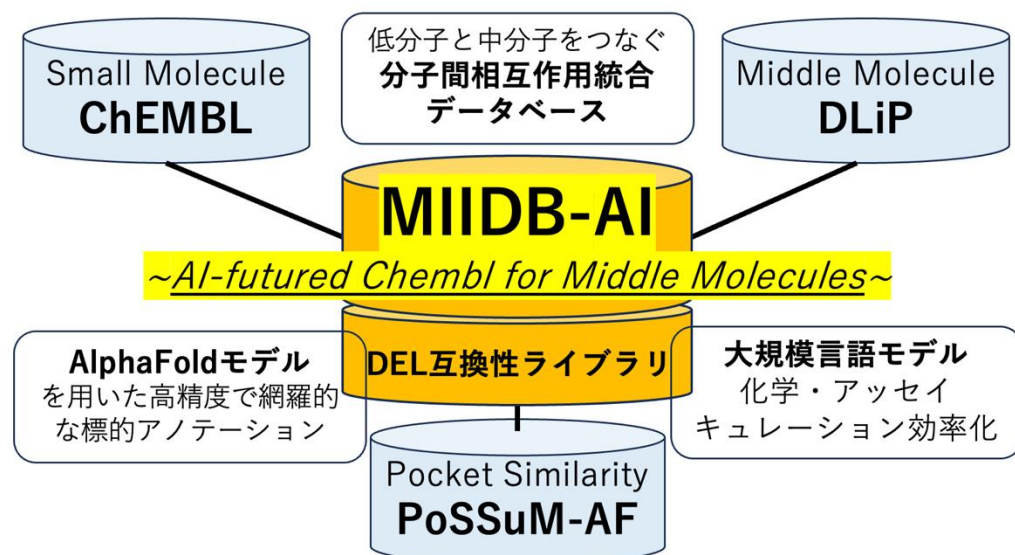
LLM(GPT)
LLM(BERT)

高

- ✓ 対話型AIによる標的予測の効率化技術を開発。
- ✓ 方法: ChatGPTのファインチューニングを実施。5,000件の標的-リガンドペアデータを使用し、訓練と検証データを1:1の割合で分割してモデルを学習させた。
- ✓ ベースモデル: gpt-3.5-turbo-1106を使用。

MIIDB-AIのコンテンツ

- 10億個超の世界最大の中分子データを搭載。
- 既知の活性中分子として10万個超（ペプチド:1万個、非ペプチド:10万個、核酸:500個）
- 合成可能な化合物（1,000万個）、AI生成のDEL互換性中分子ライブラリー（10億個超）
- 標的の生物学的特性(タンパク質標的・PPI標的・PROTAC標的)や相互作用情報（Known・Putative・Predicted）を付加。

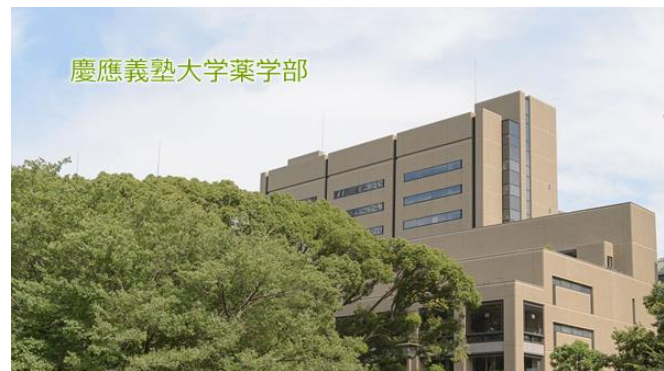


MIIDB-AIのコンテンツ

(※青線：現DLiPの対象は一部でしかない)

研究体制

No.	グループ名	氏名	実施機関	研究題目
1	池田グループ	池田 和由	理化学研究所	中分子データの収集とMIIDB-AIデータベースの開発
2	富井グループ	富井 健太郎	産業技術総合研究所	MIIDB-AIのための拡張機能の開発と実装
3	米澤グループ	米澤 朋起	慶應義塾大学	MIIDB-AIのためのデータキュレーション法の開発支援



まとめ・今後の課題

- 本データベースを構築することにより、中分子標的とそのリガンド間の相互作用データを補完し、その相互作用を予測する技術開発を支援する。
- PPIなどの従来創薬が困難な標的分子の相互作用を収載し、次世代創薬研究への進展に貢献する。
- 中分子と相互作用する大規模データセットを作成した。今後はキュレーションプロセスの高精度化と効率化を進める。

謝辞

- 永江翼（産総研）
- 富井健太郎（産総研）
- 米澤朋起（慶應大）
- 大澤匡範（慶應大）

