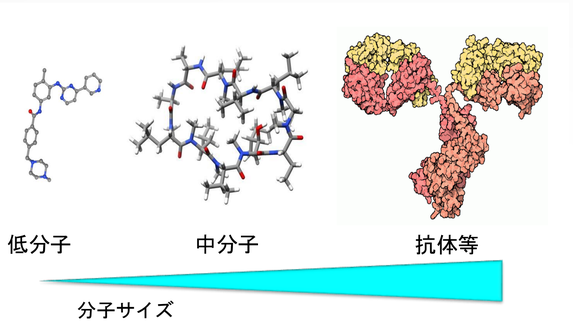
AI 駆動型データキュレーションによる持続可能な中分子相互作用統合データベースの開発

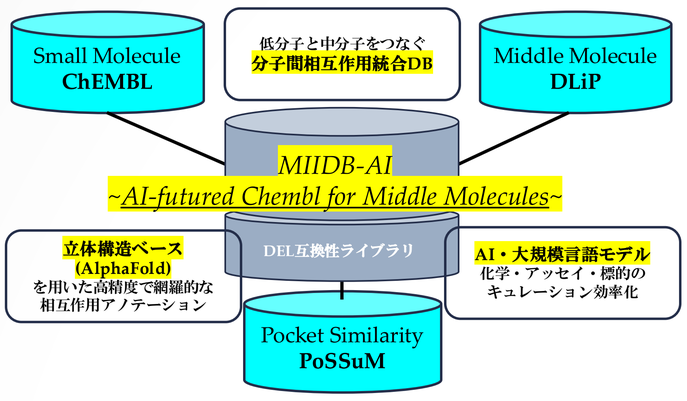
池田和由（理研）、永江翼（産総研）、米澤朋起（慶應大）、富井健太郎（産総研）

Background

- 中分子は、低分子とバイオ医薬品(抗体等)の中間に位置し、分子量が500から2,000の範囲に収まるペプチド、非ペプチド(合成化合物)、核酸が含まれる。
- 特異性と選択性が高く、副作用が少ない等の優れた特性を有し、従来の医薬品では治療が難しかった疾患領域への新たなアプローチとして、これからの創薬の幅を広げると期待されている。
- 一方、低分子に比べて中分子に関するデータは依然として不足しており、製薬企業やアカデミアの研究者にとって有用かつ容易にアクセス可能で統合された相互作用データベースの開発が求められている。

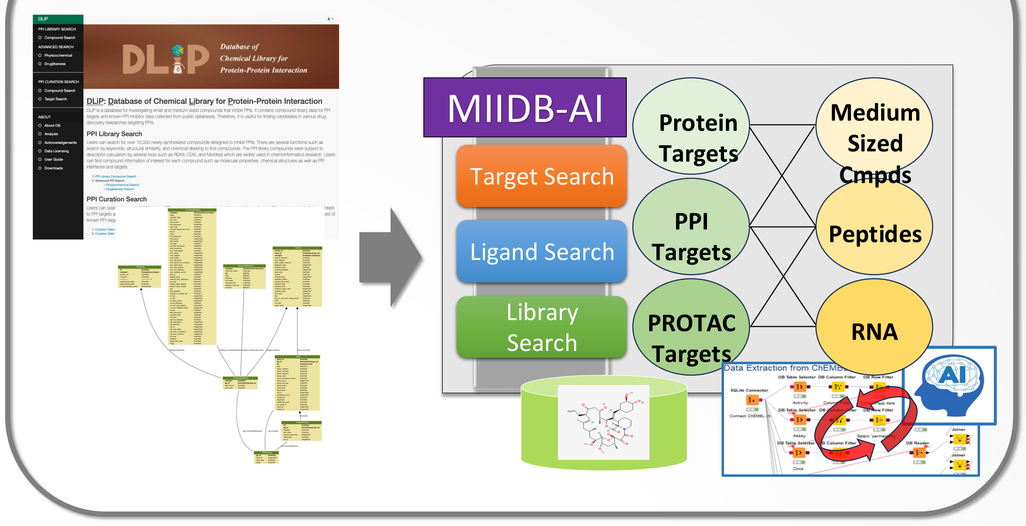


中分子相互作用データベースの構想



- AIを活用することで、標的とリガンド間の相互作用を高精度に予測する。
- 新規中分子薬候補の発見の効率化で、次世代医薬品開発への貢献を目指す。

DBスキーマ・インターフェース開発



- データベース開発
- AIを活用したキュレーション
- データ取集・統合

公共DBからのデータ取集

- 公共データベース(PDB, ChEMBL等）から中分子を抽出かつ中分子に関する独自の実験データを収集・統合する。

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

Ligands	# of Molecules
Mid-Compounds	17,000
Cyclic Peptide-like	1,200
RNA Aptamers	1,500

Targetキュレーション自動化技術の開発

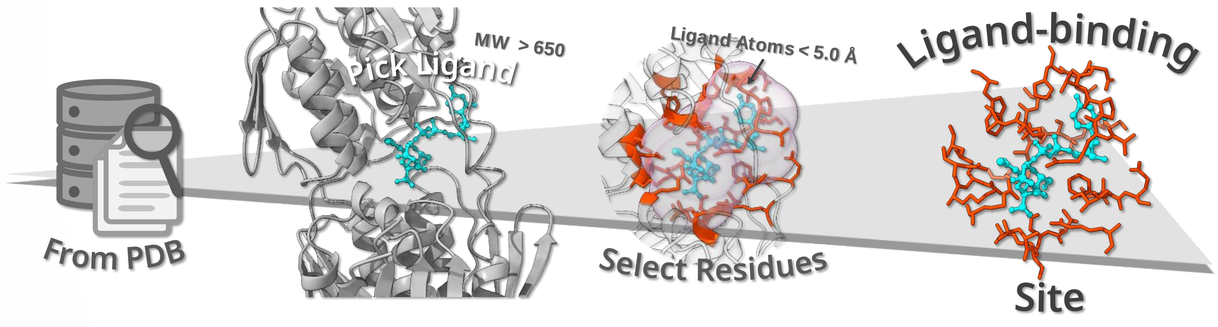
（手順）
1. OpenAI APIからChatGPT起動

2. 最適化(ファインチューニング)のプロセス

3. 予測結果の精度検証

- 対話型AI(GPT-3ベース)を利用した標的予測について理解し、これを効率化することを目的とする。ユーザーがAIとの対話を通して、化合物構造(SMILES)を入力し、標的の予測できるかを検証する。
- 方法：ChatGPTのファインチューニングを行う。5,000件の標的-リガンドデータ(訓練用:検証用=1:1)で訓練。
- ベースモデル：gpt-3.5-turbo-1106
- タンパク質言語モデルによる改良を実施。

Materials and Methods



Protein Data Bank[1] からの抽出

- Number of Distinct Protein Entities > 0
- Number of Distinct Non-polymer Entities > 0
- Refinement Resolution < 4.0 Å
- Chemical Component Molecular Weight > 650

合計 17,151 PDB エントリー(2024/08/25)

対象となるリガンド

- 単独の残基でentity typeが non-polymerとなっているもの

結合部位の定義

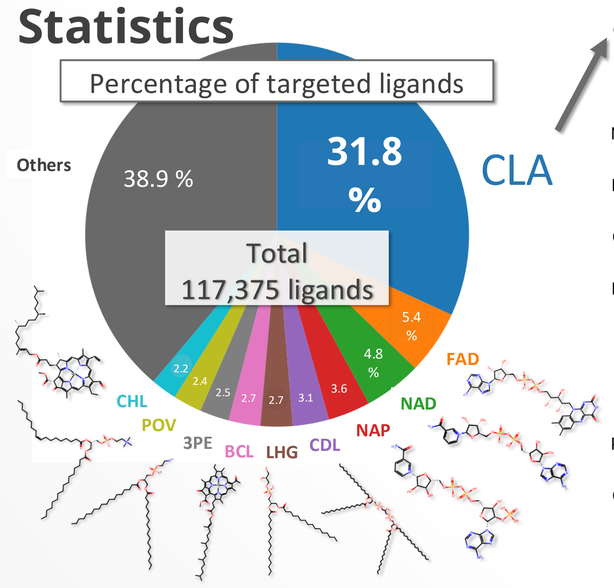
- 各リガンド重原子から5.0 Å 以内に重原子が含まれる残基
- polymer typeがPeptideLの残基を必ず一つ以上含む（リガンド単独のエントリーなどは×）
- モデルNo.が先頭のものが対象

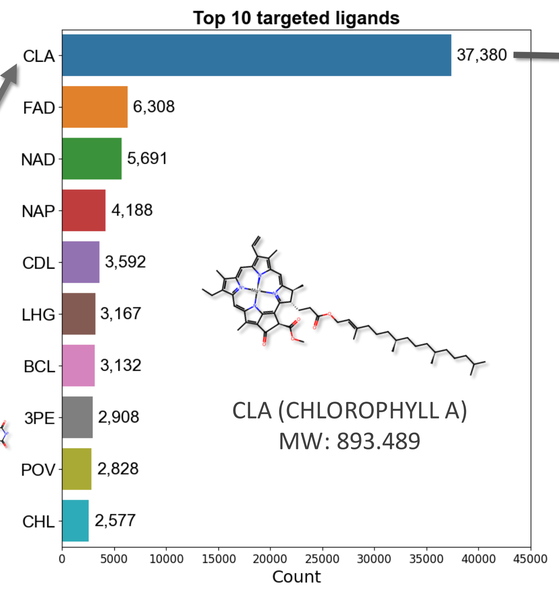
ファイルフォーマット

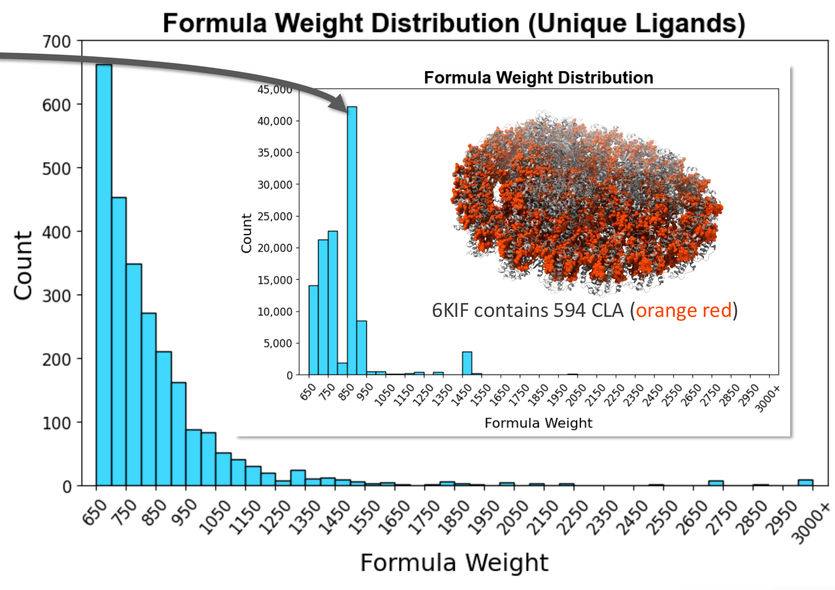
- mmCIF ファイルフォーマット
- ヘッダーは基本的に元のものを保持

Results

Statistics







分析対象のリガンドは2,564種類に及び、結合部位数は117,375に達した。

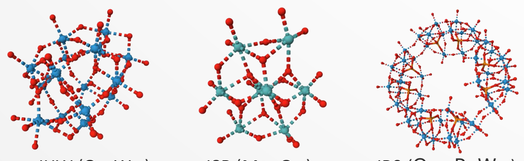
このうち、CLAが全体の30%超を占める結果となった。CLAのような一般的な生体分子は、単一エントリーに多数含まれることがあり、この高い割合につながっている。

同様にFADなどの補因子、さらに脂質分子も頻出した。加えて無機クラスターなども見受けられた。これらは医薬品としての適性に乏しいため、適切な基準を設けて除外する必要性が示唆される。

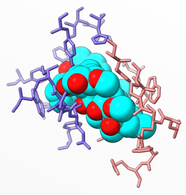
一方で、そのような選別基準の設定には慎重な検討が求められることも明らかとなった。

Removal Candidates... ?Drug-like from DrugBank[3]

Inorganic clusters



Pharmacologically active (54 sites)

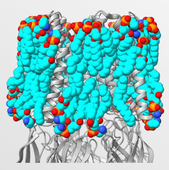


RAPAMYCIN (Sirolimus)
6M4U RAP-P42345

Target: mTOR (P42345)

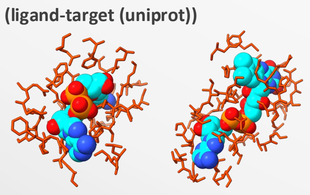
Active:
mTOR inhibitor immunosuppressant

Lipids
(in membrane proteins)



Proton-gated ion channel
(8WCR-PEE)

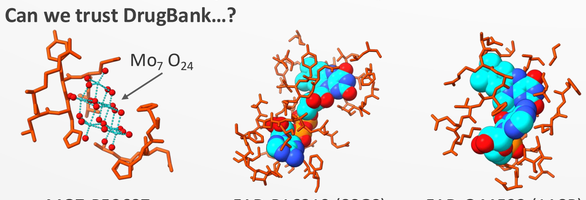
Cofactor-like[2]
(ligand-target (uniprot))



NAD-P20000
(1A4Z)

FAD-M1VMF7
(8XCM)

Unknown activity (3,747 sites)
Can we trust DrugBank...?



MO7-P52687
(1P0Z)

FAD-P16219 (8SGS)
drug-like (not cofactor-like)

FAD-Q44532 (1A8P)
drug-like & cofactor-like

Future Work

Curation

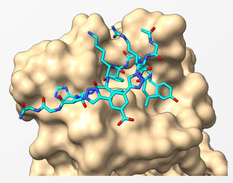
PDBを網羅的に調査した結果、10万を超えるリガンドの結合部位を特定することに成功した。同定されたリガンド化合物は多岐にわたるものの、その全てが中分子医薬品の候補として適切とは限らない。

今後のデータの選別においては、化合物の特性のみならず、結合対象となるタンパク質の性質も考慮に入れつつ、詳細なアノテーションを行い、適切なフィルタリング基準を適用する必要性が示唆された。

Additional data (peptides and nucleic acids)

本研究で収集したリガンドの大半は低分子化合物に分類される。一方、中分子医薬品の領域ではペプチドや核酸などのモダリティも重要であり、これらのデータも収集する必要性が認識された。

これらの追加データ収集に際しては、PDBが提供する"Biologically Interesting Molecule Reference Dictionary"や、RNAアプタマーのデータベースである"AptaDB[4]"などの活用を検討している。



PRD_002559 (6q38)
Stapled peptide

References

1. Nucleic Acids Res 2023 Jan 6;51(D1):D488-D508. doi: 10.1093/nar/gkac1077.
2. Bioinformatics. 2019 Sep 15;35(18):3510-3511. doi: 10.1093/bioinformatics/btz115
3. Nucleic Acids Res. 2024 Jan 5;52(D1):D1265-D1275. doi: 10.1093/nar/gkad976.
4. RNA. 2024 Feb 16;30(3):189-199. doi: 10.1261/rna.079784.123.

CC BY NC ND