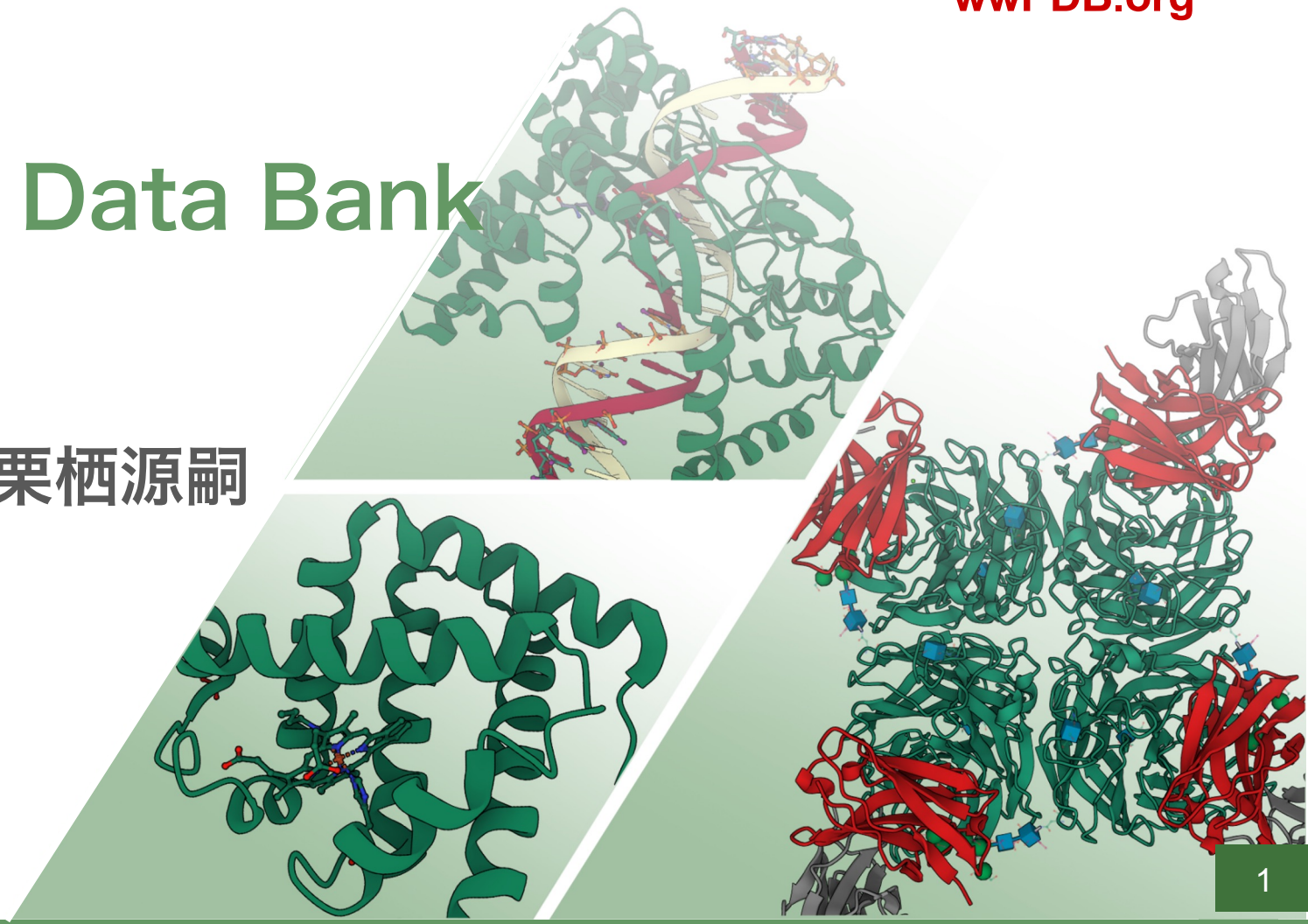
PDBj
Protein Data Bank Japan

WORLDWIDE PDB
PROTEIN DATA BANK

wwPDB.org

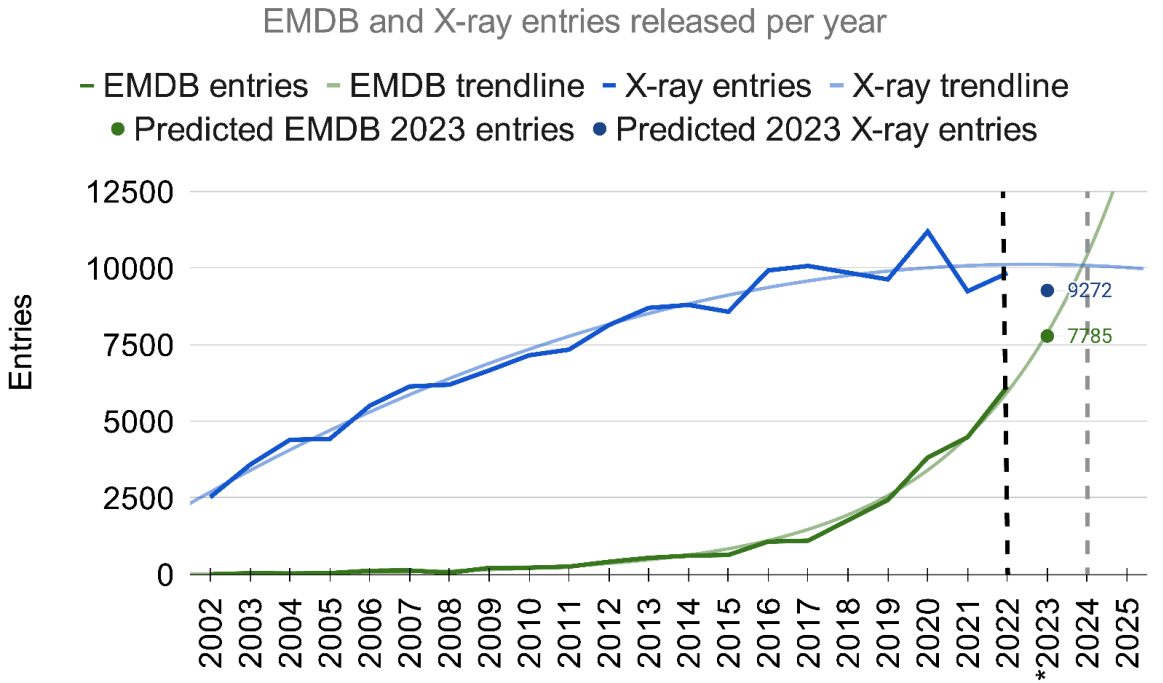
AlphaFold時代の Protein Data Bank

大阪大学蛋白質研究所　栗栖源嗣



PDB Archiveの推移（1）

- 210,000件を超える専門家が編集処理したデータをCC0 1.0 として自由に利用可能
  - ❖ 年率 ~8%でデータ量が増加
- 全世界で >400 を超える外部データベースで活用されている
- 2022年集計でクライオ電顕のエントリーが
  - ❖ 前年比60%増
  - ❖ 原子分解能 (~1Å)に到達するエントリーもで始めた



PDB Archiveの推移（2）

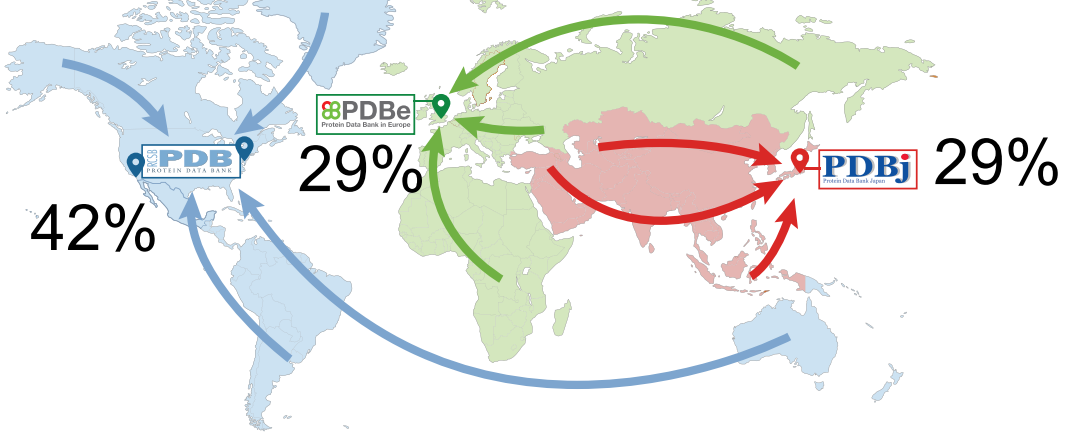
- 2022年1年間に限ると，アジア地区のデータ量増加率が他地域よりも多いので，PDBjへの登録の割合は約29%に増加
(4,758/16,344 = 0.291)

  ➡ **中国発のエントリー数の急増し，PDB Chinaがスタート**

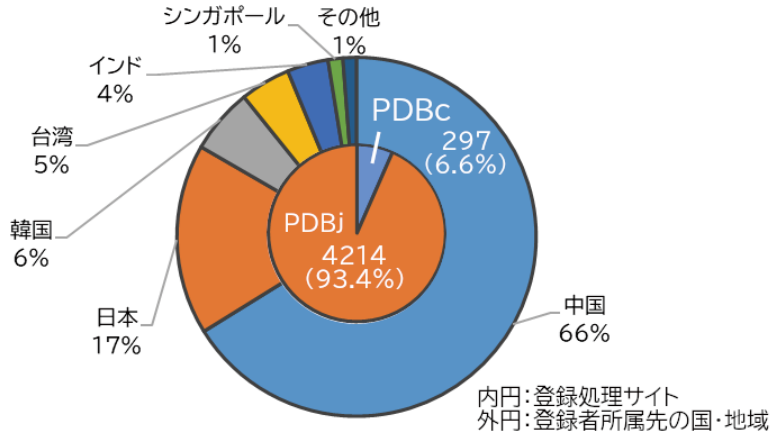
- 機械学習による予測構造を併用した構造解析の増加。

  ➡ **構造予測手法（AlphaFold等）を併用したIntegrated/Hybrid構造解析に対応した検証レポートの構築が課題**

2022年1年間の内訳

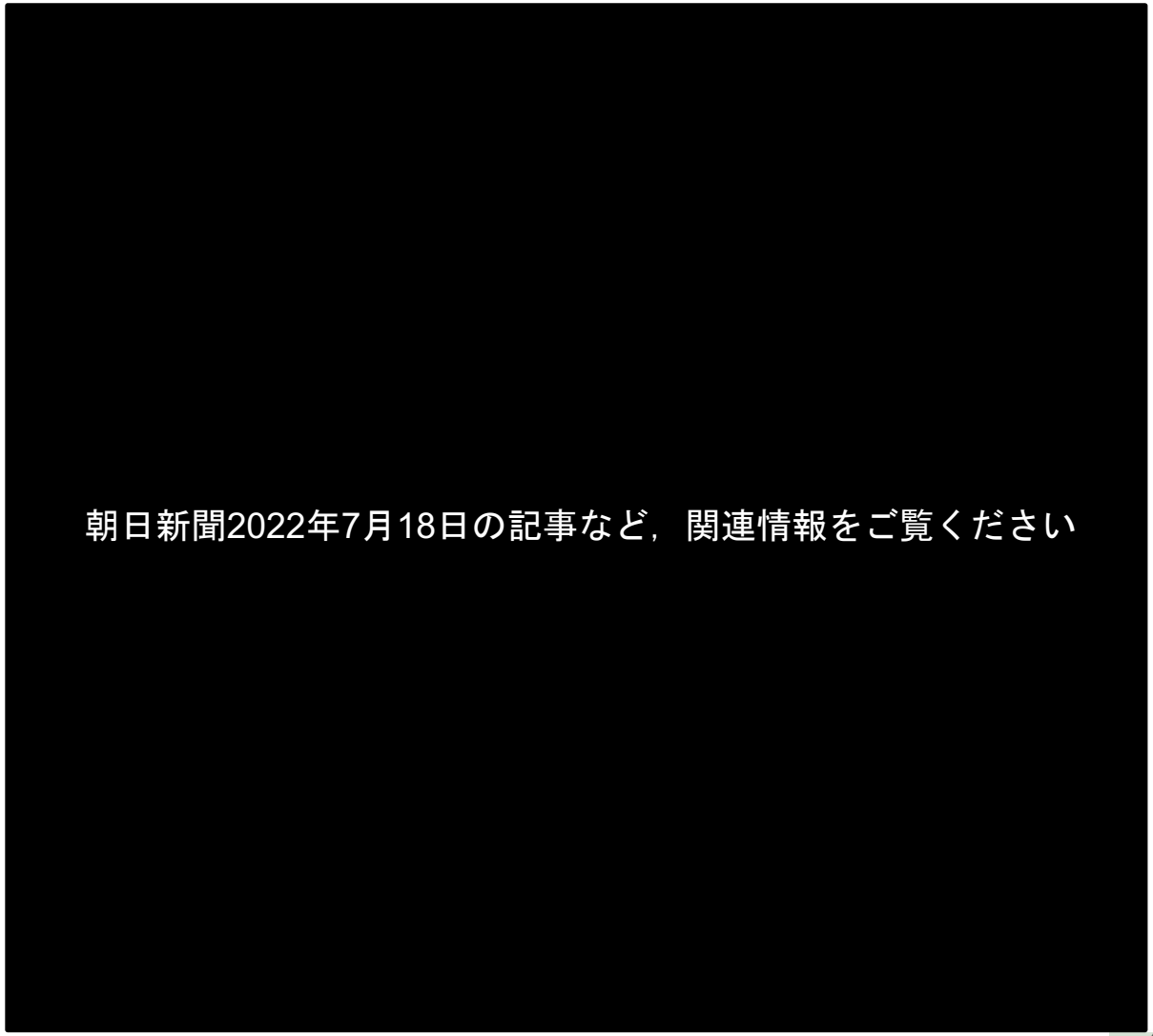


PDBj+PDBcが登録処理した
PDBエントリーの国・地域分布（2022年）



構造生物学やデータベースの現状の課題と将来展望について

- AlphaFoldなど構造予測技術の精度向上が構造生物学に様々な影響を与えている。
- 進展著しいChatGPTなどの技術をデータキュレーション等への活用は？
- 検証レポートだけではなくAIを活用してPDBの利便性の向上を図る。

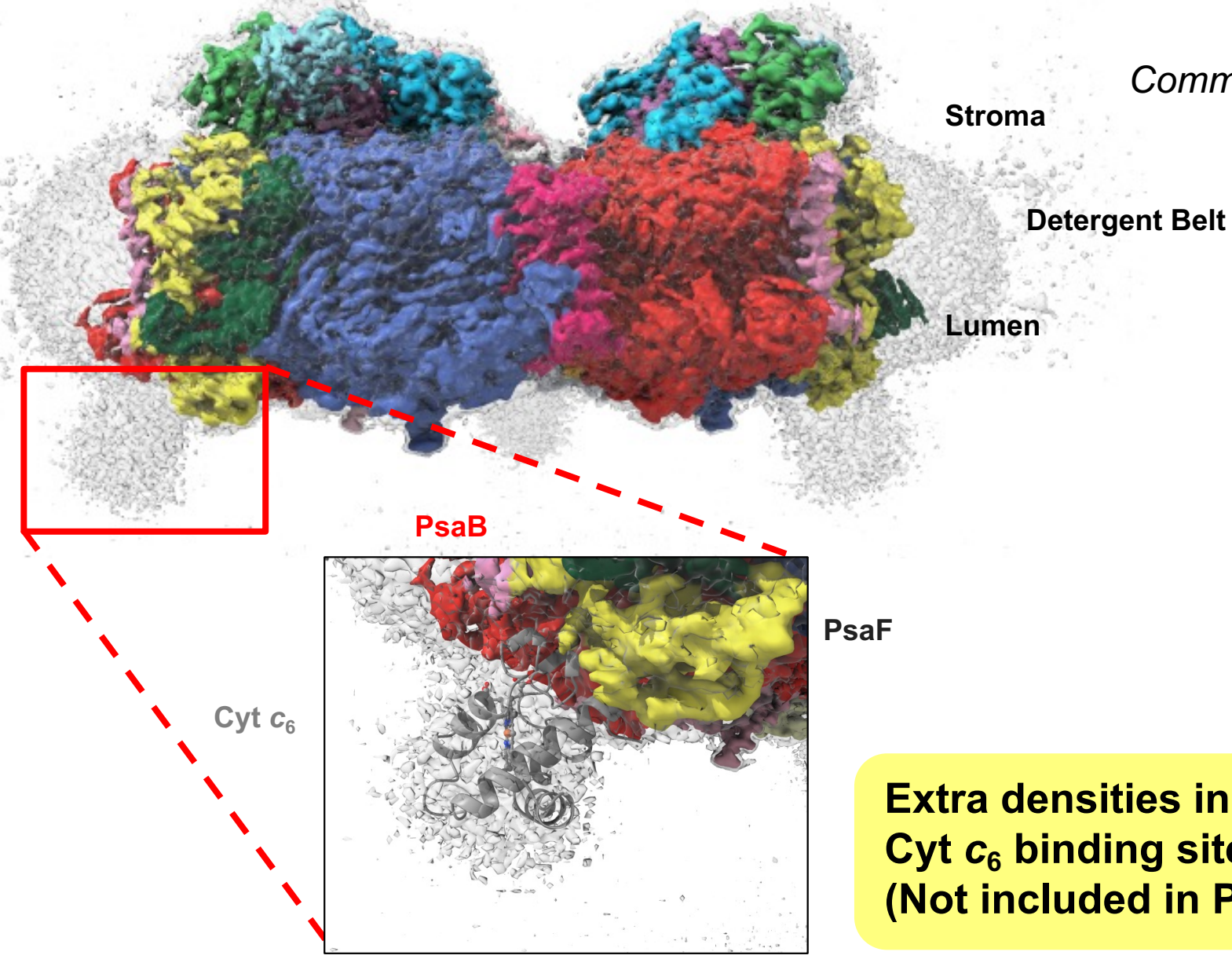


構造生物学やデータベースの現状の課題と将来展望について

- AlphaFoldなど構造予測技術の精度向上が構造生物学に様々な影響を与えている。
- 進展著しいChatGPTなどの技術をデータキュレーション等への活用は？
- 検証レポートだけではなくAIを活用してPDBの利便性の向上を図る。

Cryo-EM structure of PSI-Fd-Cyt c6 complex at 1.97 Å

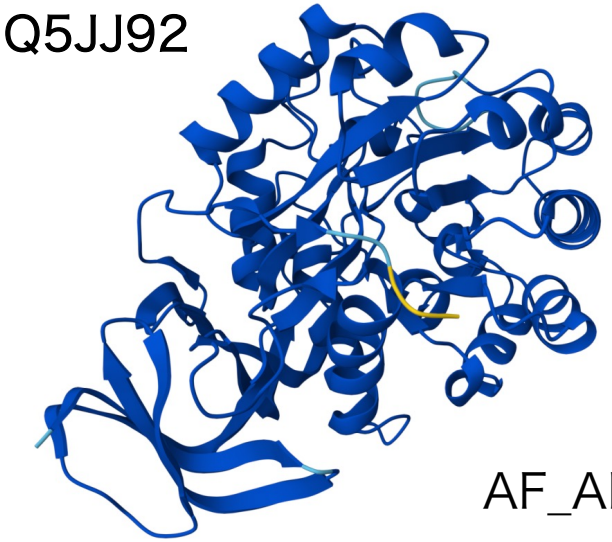
Commun. Biol. (2022)



Extra densities indicate Cyt c_6 binding site. (Not included in PDB model)

Alpha Foldで予測された様々な構造

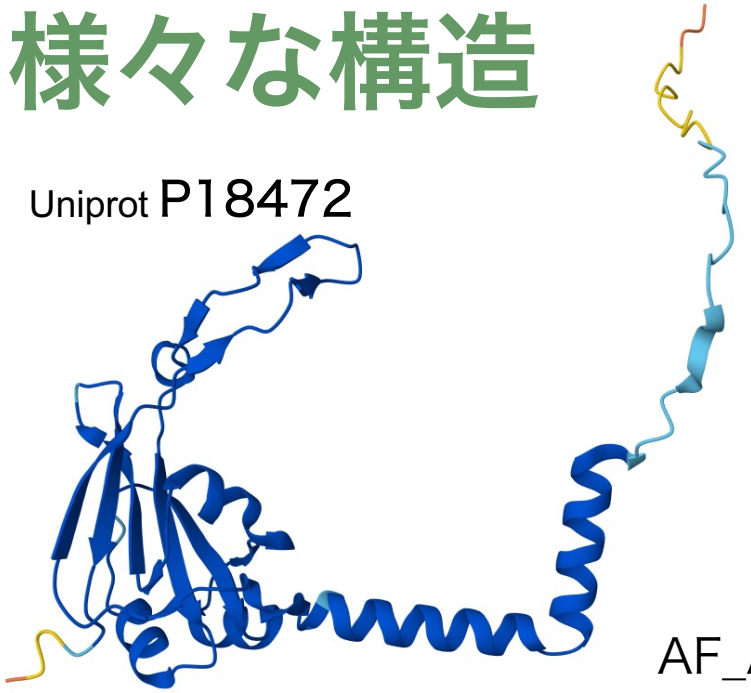
Uniprot Q5JJ92



AF_AFQ5JJ91F1

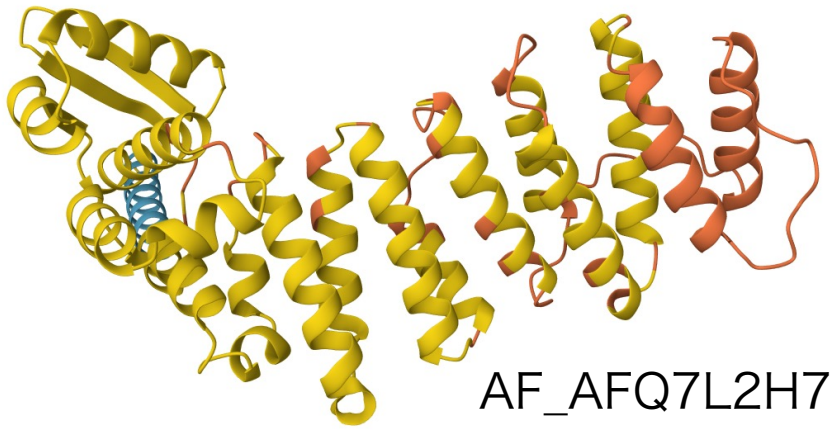
High pLDDT

Uniprot P18472



AF_AFP18472F1

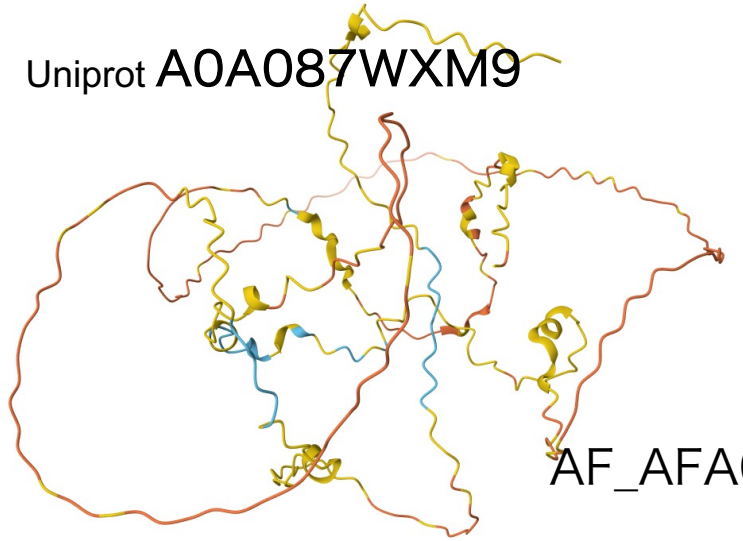
Uniprot Q7L2H7



AF_AFQ7L2H7F1

Low pLDDT

Uniprot A0A087WXM9



AF_AFA0A087WXM9F1

構造生物学やデータベースの現状の課題と将来展望について

- AlphaFoldなど構造予測技術の精度向上が構造生物学に様々な影響を与えている。

- 進展著しいChatGPTなどの技術をデータキュレーション等への活用は？

- 検証レポートだけではなくAIを活用してPDBの利便性の向上を図る。

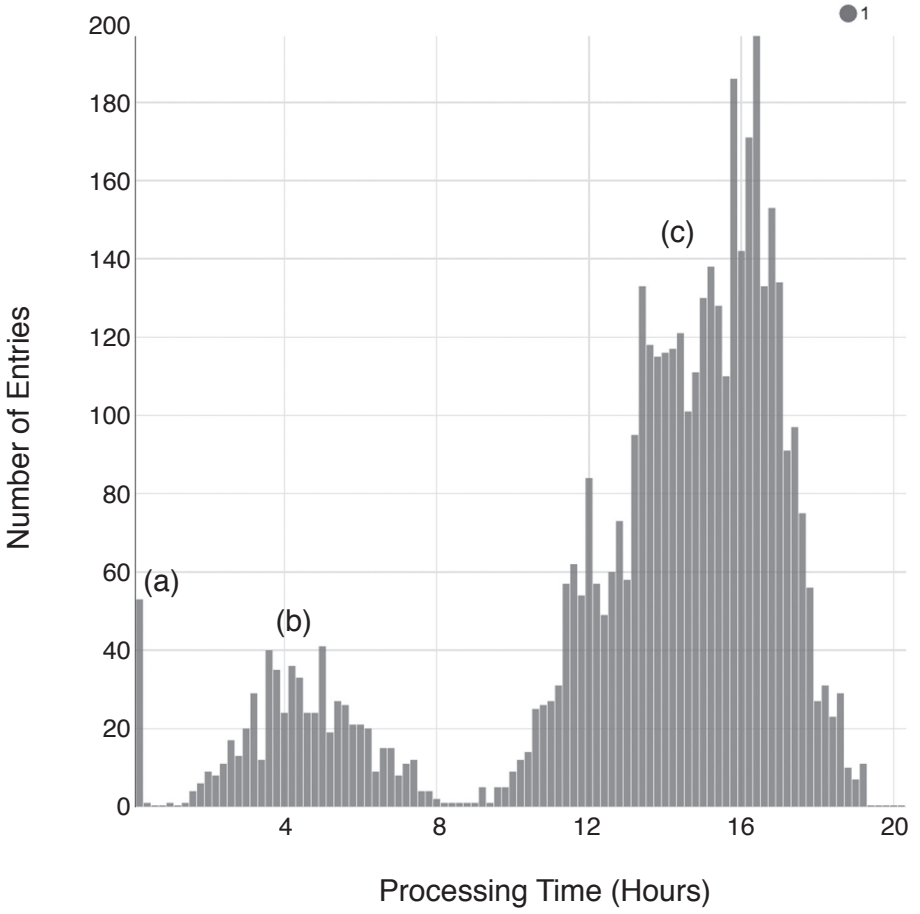
wwPDBで共同開発しているOneDepシステムにAIを導入予定

<u>1エントリーのアノテーションに要する時間</u>

(a) ~1hr:　修正依頼の必要ない単純な構造

(b) ~4 hrs: 修正依頼を必要としない複合体などの複雑な構造

(c) ~15 hrs: 登録者からの修正を必要とする構造



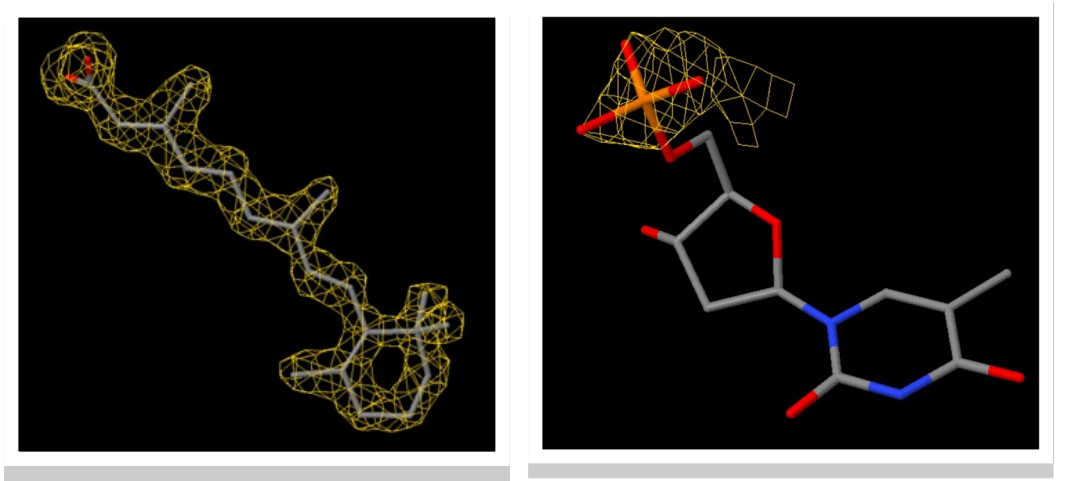
20年におよぶアノテーションの記録を学習データとしてアノテーションの更なる効率化を計る.

構造生物学やデータベースの現状の課題と将来展望について

- AlphaFoldなど構造予測技術の精度向上が構造生物学に様々な影響を与えている。
- 進展著しいChatGPTなどの技術をデータキュレーション等への活用は？
- 検証レポートだけではなくAIを活用してPDBの利便性の向上を図る。

分解能だけで選別していませんか？

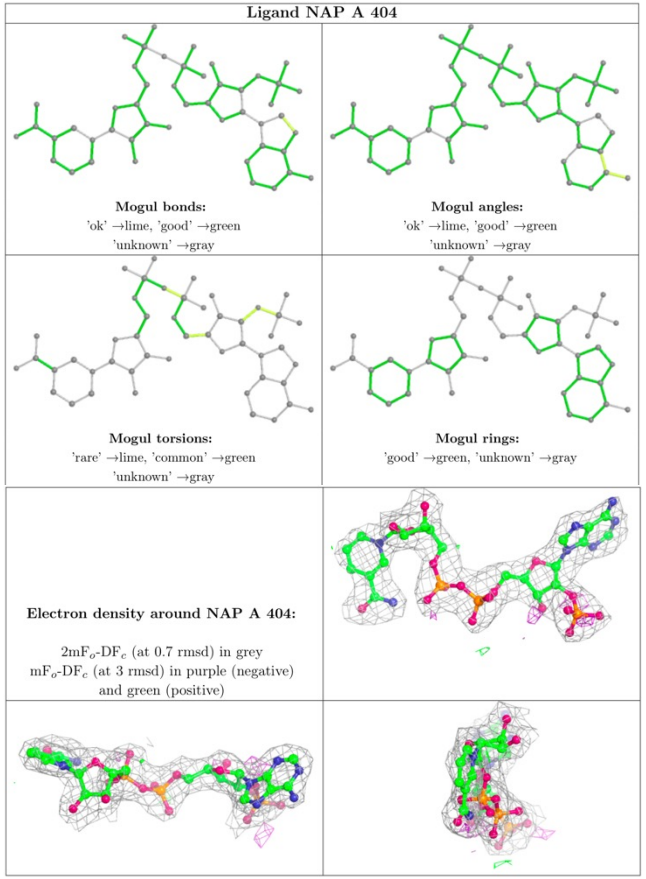
- 原子座標だけだと厳密な化合物を区別できないので，編集作業中に厳密に化合物を特定します
- 特に結合次数やキラリティー，原子の価数を厳密にチェック
- しかし，下のような実験データで判断した構造情報でよいのでしょうか？



同じ2Å分解能でも異なる実験データとの一致度の例
左：RSR=0.10, CC=0.95，右：RSR=0.41，CC=0.70

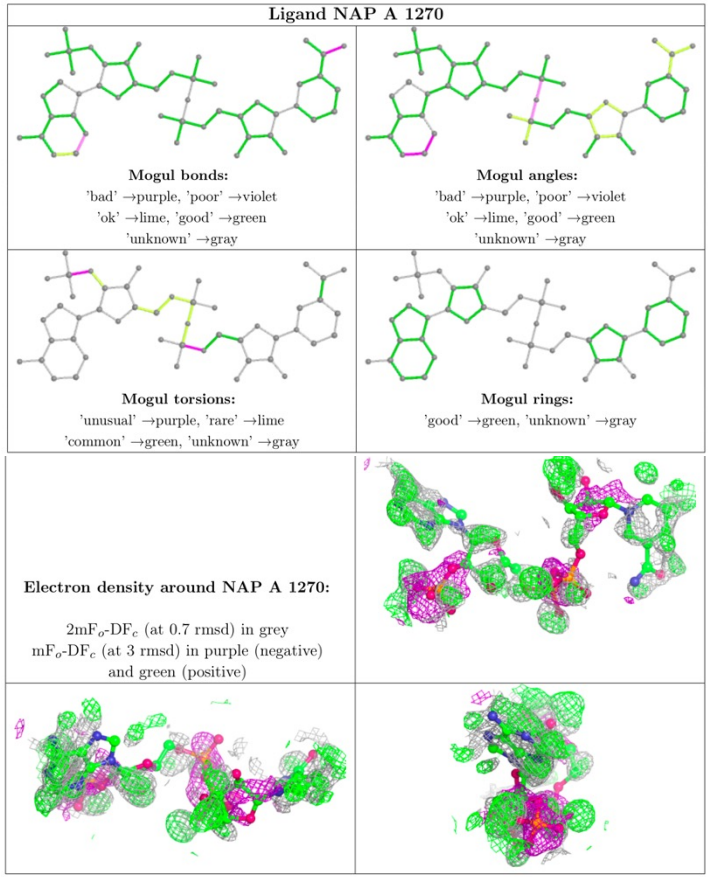
化合物情報を構造化学的・実験科学的に評価

Mol	Type	Chain	Res	Atoms	RSCC	RSR	B-factors(Å^2)	Q<0.9
3	NAP	A	404	48/48	0.96	0.14	31,43,66,70	0



PDB entry 5zix (Better data quality)

Mol	Type	Chain	Res	Atoms	RSCC	RSR	B-factors(Å^2)	Q<0.9
3	NAP	A	1270	48/48	-0.06	0.67	87,96,100,100	0

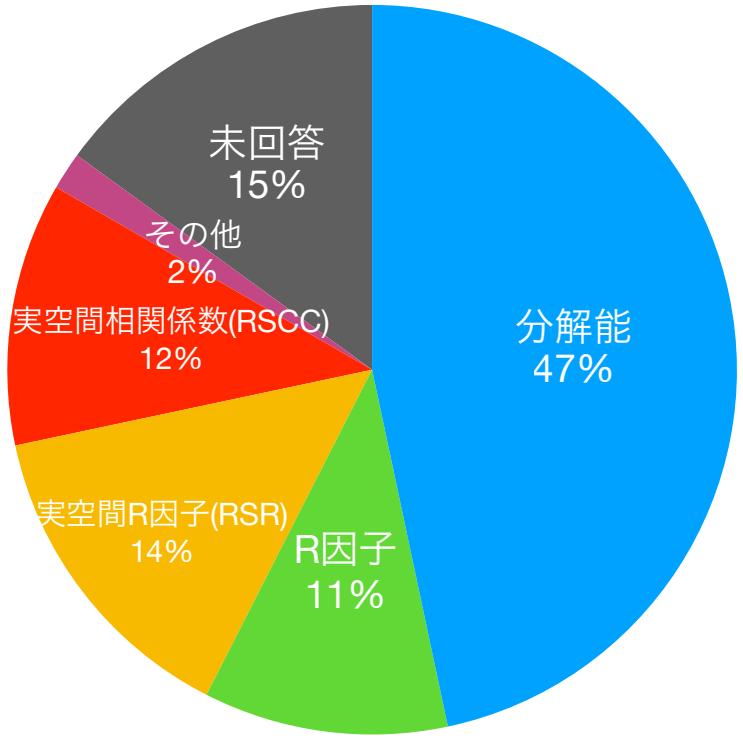


PDB entry 1zk4 (Worse data quality)

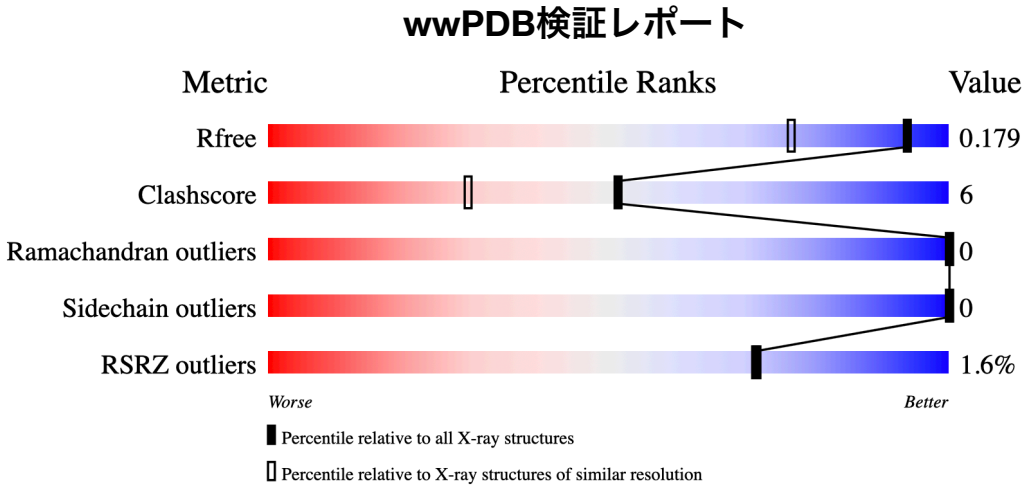
PDB IDが複数ヒットしたらどれを選ぶ？

昨年度のセミナーのアンケート結果より

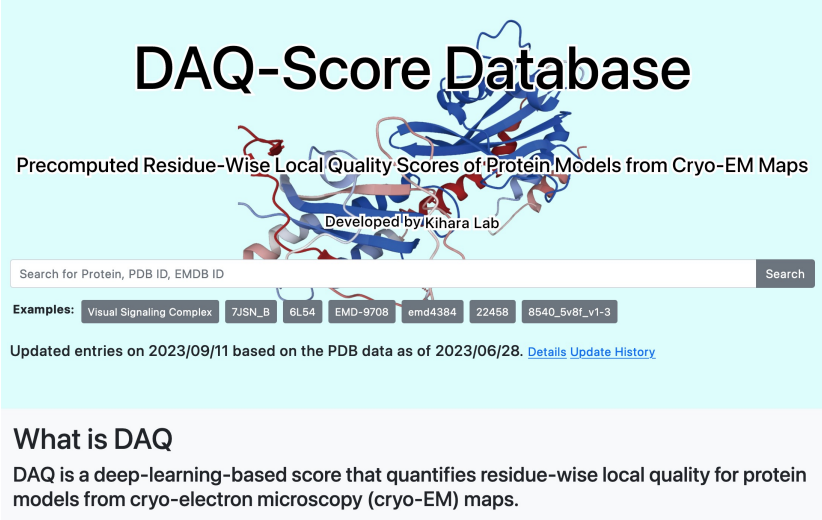
＞　データセットの選抜の指標を教えてください。



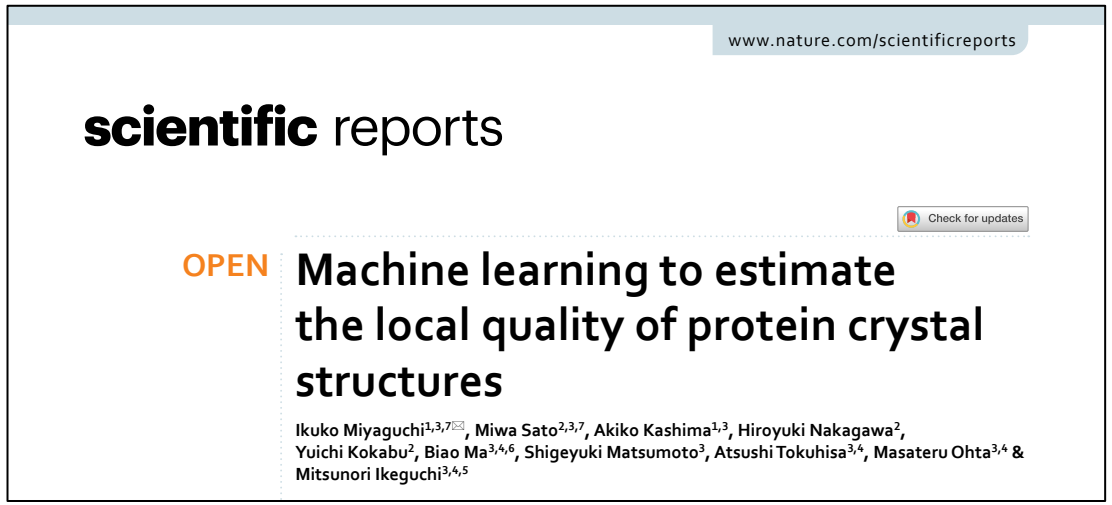
- 興味のある配列領域、リガンドが存在する？
- 実験手法を選択する？
- 立体障害、異常な二面角はない？
- 実験データの解像度
  - 分解能、FSC
- 実験データに忠実なモデル？
  - R因子、Atom Inc., Q-score
- 残基ごとの局所的な指標を活用する？
  - RSR, RSCC, Q-score



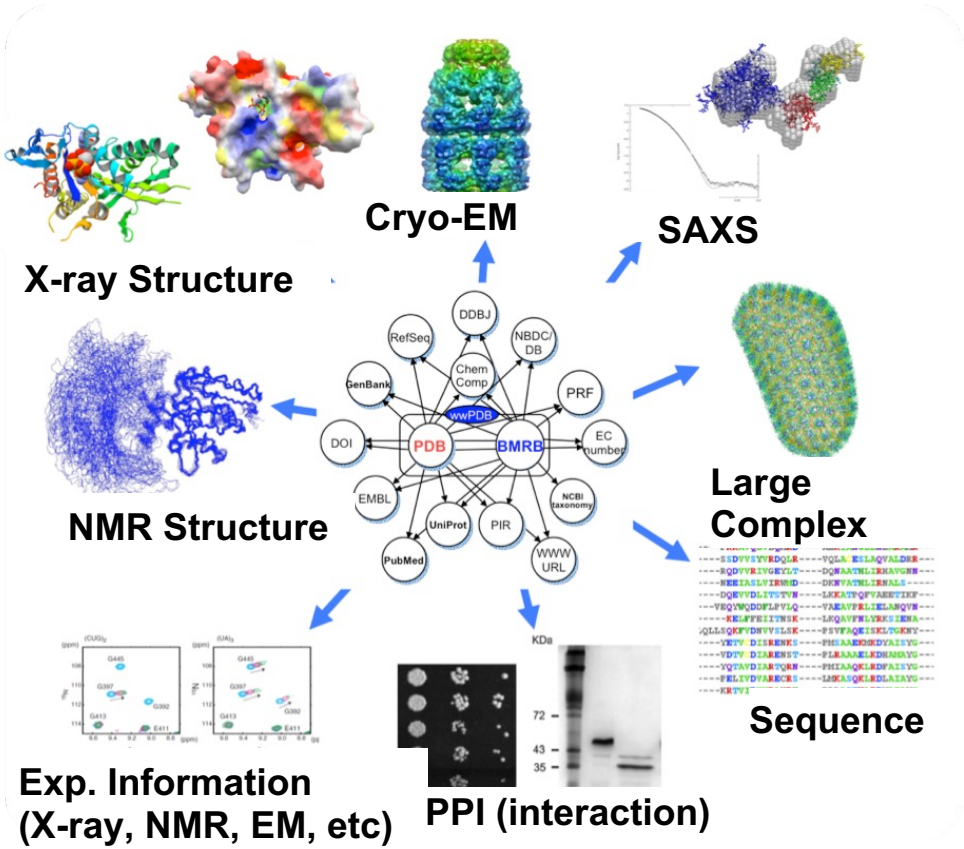
DAQ-scoreおよびDAE-mapとの連携



https://daqdb.kiharalab.org/



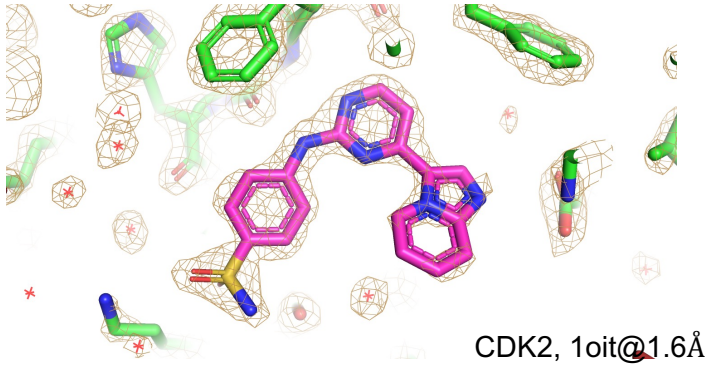
Sci Rep **11**, 23599 (2021)



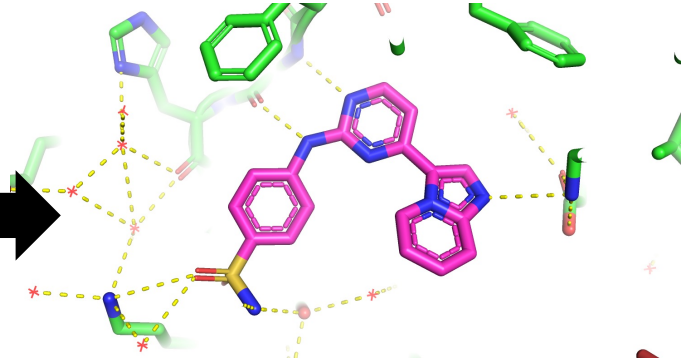
Kinjo et al. (2012) Nucl. Acids Res. 40, D453-D460.
Yokochi et al. (2016) J. Biomed. Semantics, 7:16.

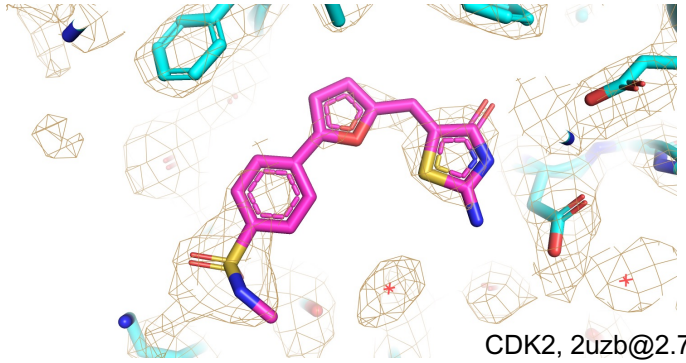
DAE mapによる検証のBackground

Structural biologists



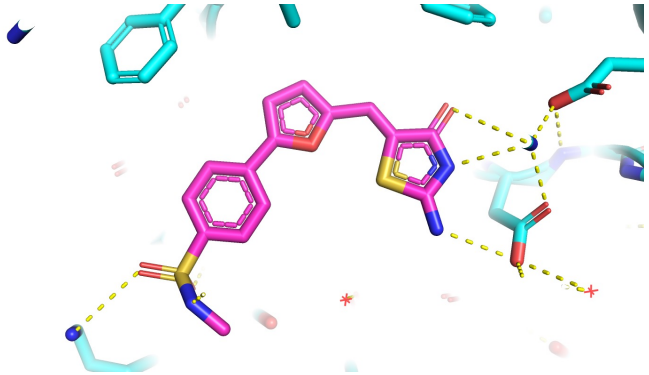
CDK2, 1oit@1.6Å





CDK2, 2uzb@2.7Å

Coordinates are determined based on electron density.



Information on electron density tends to be ignored.

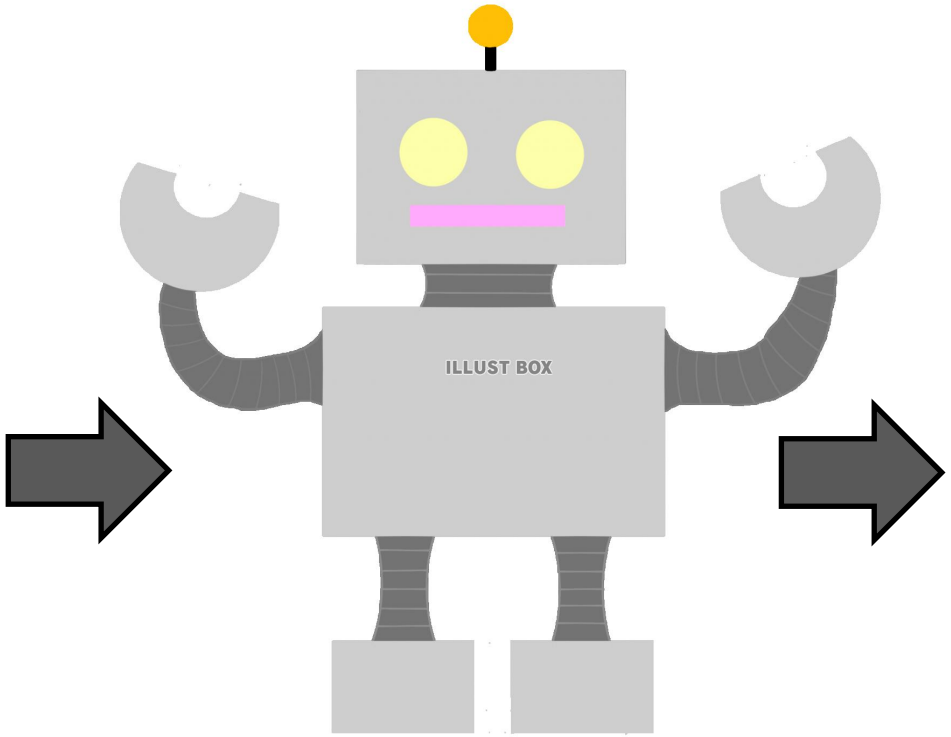
Medicinal chemists
Computer chemists

We want an objective metric for determining coordinates.

「田辺三菱製薬（株）の宮口郁子さんから使用許可をいただいて紹介しています」

An evaluation method we want

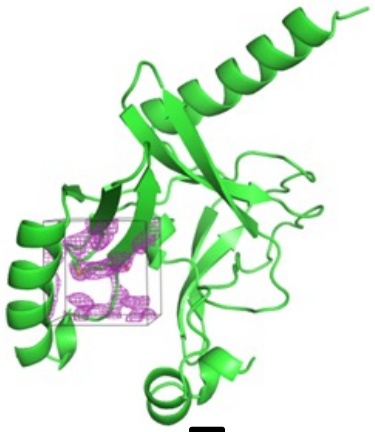
Input data
- Coordinates
- Electron density map



Output
Coordinate evaluation metrics
- They should evaluate how well the coordinates fit the correct structure in a resolution independent manner.
- "Correct structure" is the structure determined at high resolution.
- "High resolution" is higher than 1.5Å resolution.

The issue could be solved by machine learning.

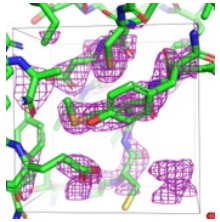
How machine learning solves the problem.



Training data
- Coordinates
- Electron density maps

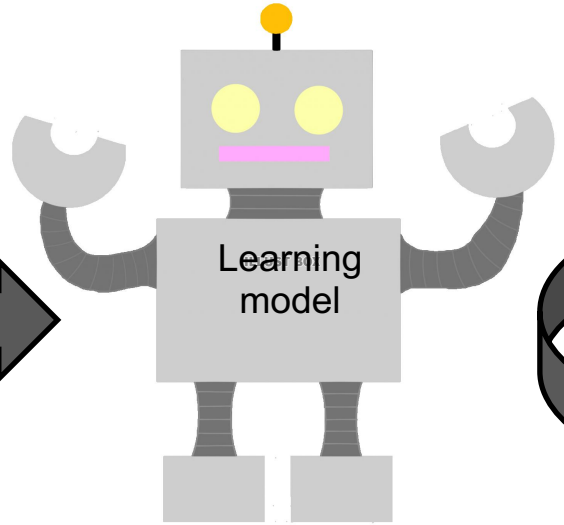
at high & low resolution

- The learning unit is an amino acid.
- A cubic box containing an amino acid of interest is cut out.



Objective variables
=Evaluation metrics
- need to be newly defined.
- should indicate how well the coordinates fit the correct structure in the box

Training



input

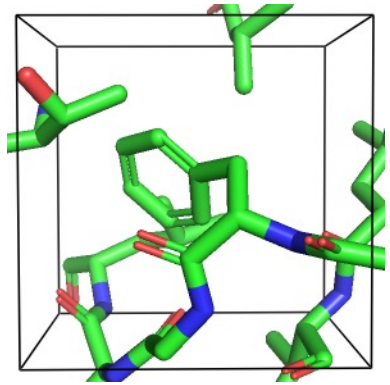
Structure to be evaluated
- Coordinates
- Electron density map

Evaluation metric

prediction

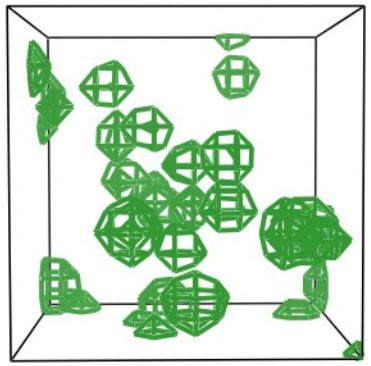
＊If the learning was successful, the evaluation metrics can be predicted without high-resolution data.

Definition of the evaluation metric, bCC

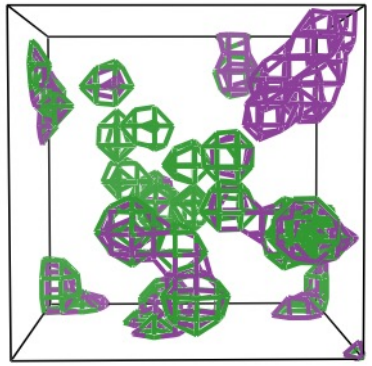


Structure to be evaluated

B=2.0
converted to
electron density

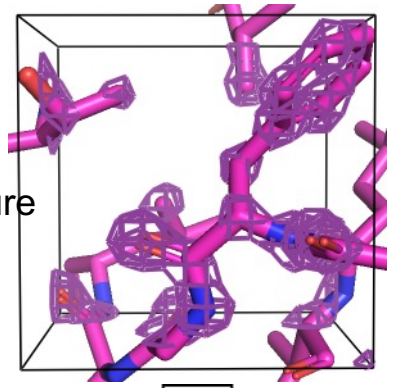


Atom map

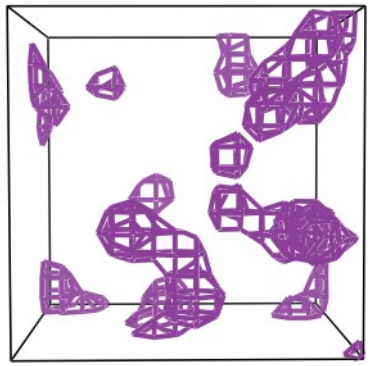


Box Correlation Coefficient, bCC
(Evaluation metric)
(Objective variable)

Correct high-resolution structure
for reference



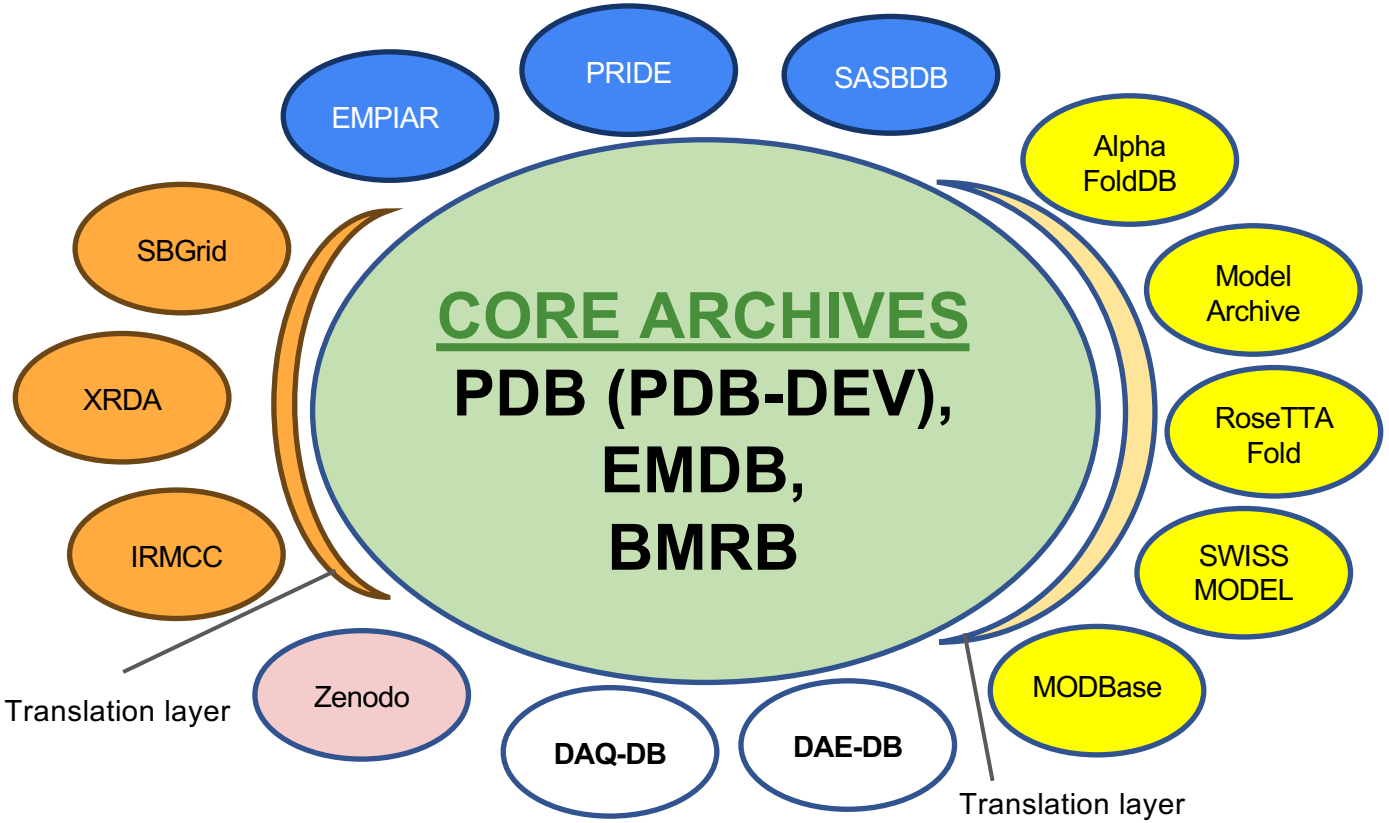
electron density map



2mFo-DFc map

構造データのEcoSystem

- 構造予測モデルアーカイブ
- 相互作用やイメージデータアーカイブ
- 実験データアーカイブ
- 一般的なアーカイブ



○構造評価指標データベース

WORLDWIDE PDB PROTEIN DATA BANK

The worldwide Protein Data Bank

www.wwPDB.org • info@wwPDB.org

PDBj Protein Data Bank Japan

as a member of WORLDWIDE PDB PROTEIN DATA BANK

RCSB PDB PROTEIN DATA BANK | BMRB Biological Magnetic Resonance Data Bank | PDBe Protein Data Bank in Europe | EMDB Electron Microscopy Data Bank | PDBj Protein Data Bank Japan