

タンパク質結晶を多数利用した 高分解能構造解析

平田 邦生

理研/SPring-8 Center

メッセージ

- 今日お伝えしたいこと

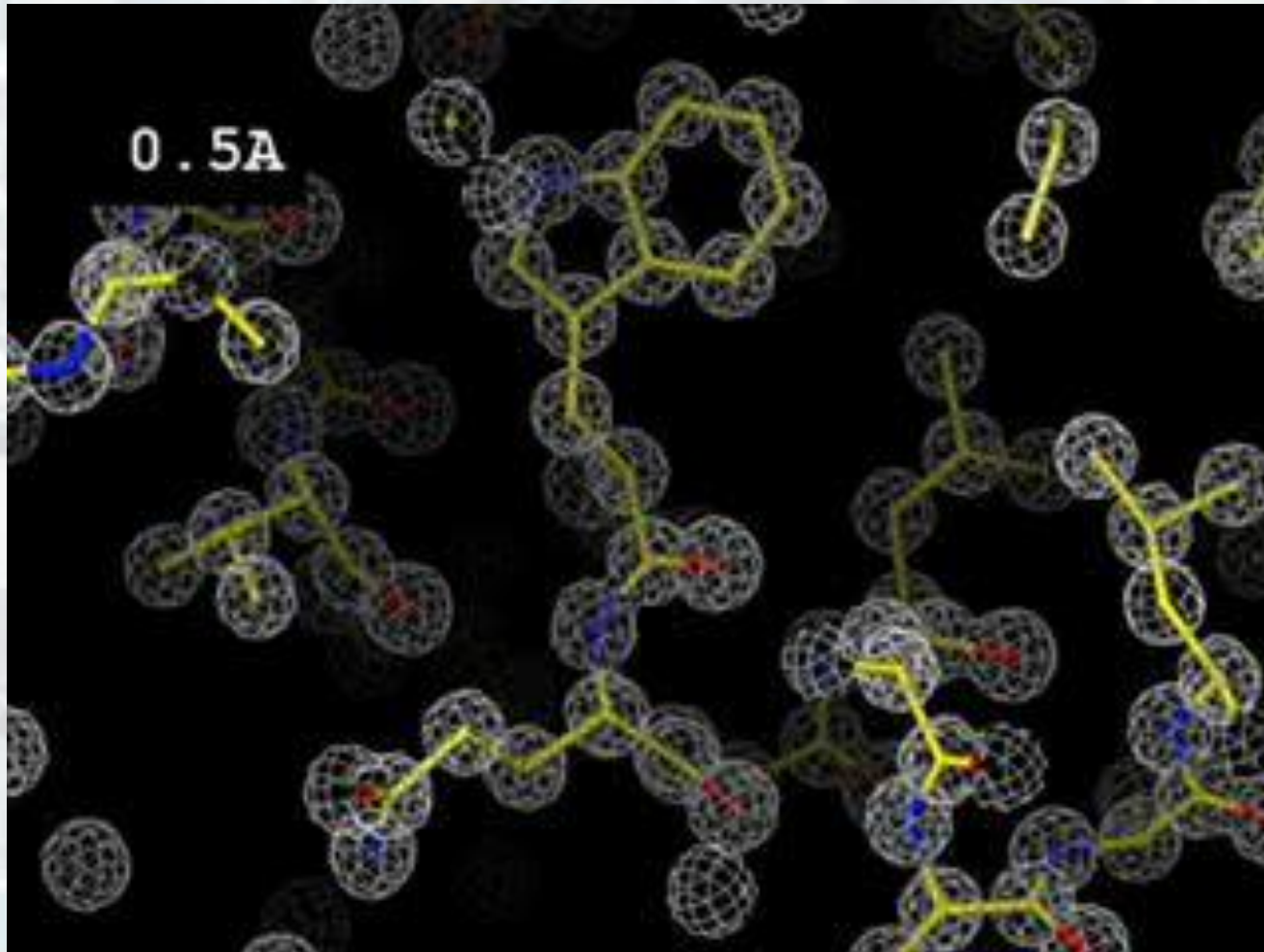
- 放射光施設→「膨大なデータ発生装置」
- タンパク質結晶構造解析
 - 人の目・人の手で処理しきれない

- 機械学習（教師なし学習）によるデータマイニング
 - “お宝構造”を見落とさないためのツール

 AI技術が我々の科学を変えつつある

高分解能という言葉

- ・ 電子密度図の空間分解能が高い という意



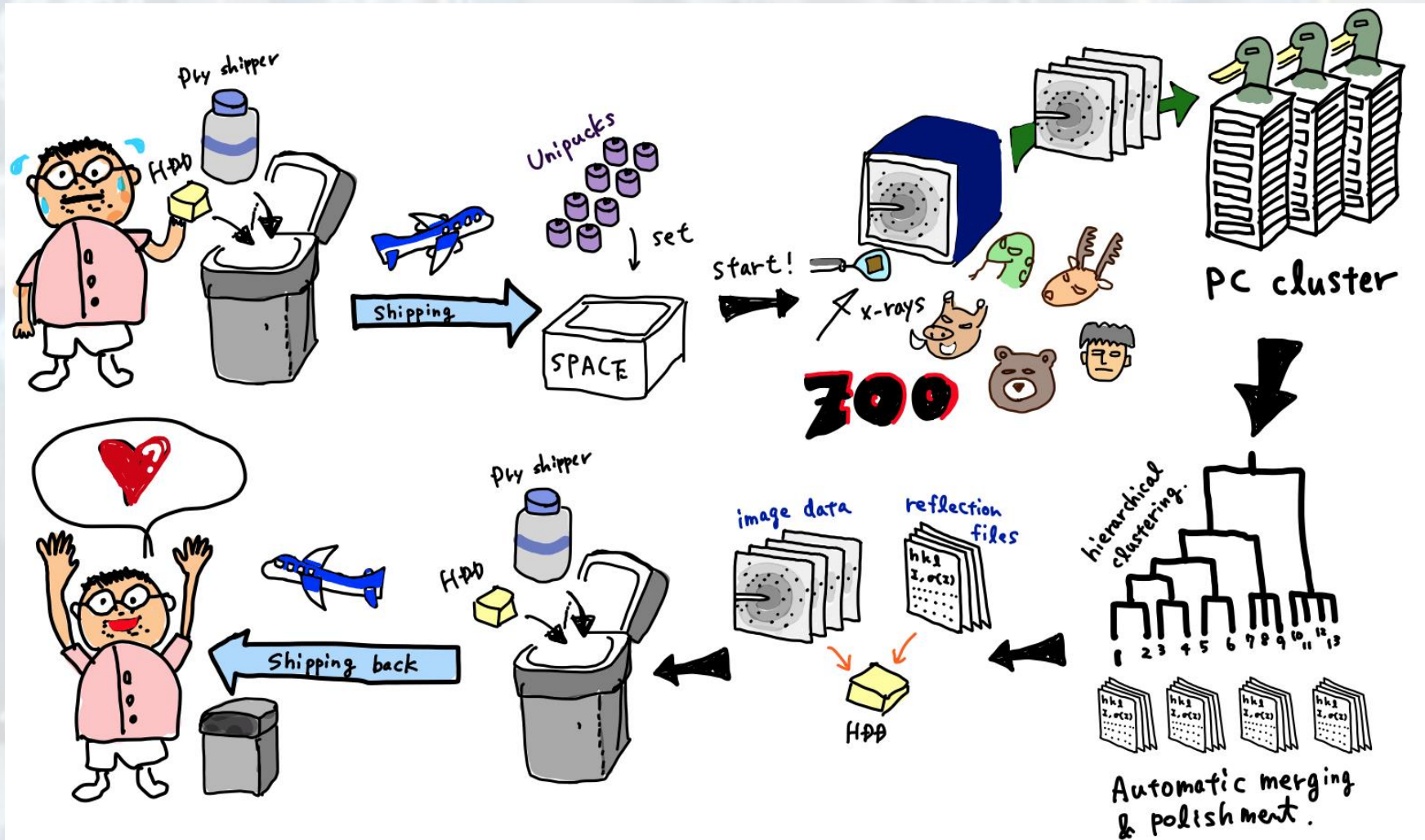
タンパク結晶回折データ収集



データ収集時間： 数秒～数十sec程度

ルーチン作業→データ収集可能

自動測定による測定の超効率化



データ収集・回折データ処理の自動化

2015~

Hirata, K., Yamashita, K. et al. (2019). *Acta Cryst D* **75**, 138–150.

Yamashita, K. et al. (2018). *Acta Cryst D* **74**, 441–449.

タンパク質結晶構造解析のデータ

- **1 データセットの容量 (EIGER X 9M)**

- 圧縮済(HDF) : 13 GB / dataset (4500 frames)
- 6-8 min/dataset

- 例 : 1回の実験 12 hours 1.7 TB
120 データセット

大量データ収集→すべてのデータをマジマジと見ない時代
(若者の生データ離れが加速)

複数結晶を利用したデータ統合



鴨



山下恵太郎
(現・東大)

Index/Integrate each dataset

Grouping datasets (cells)

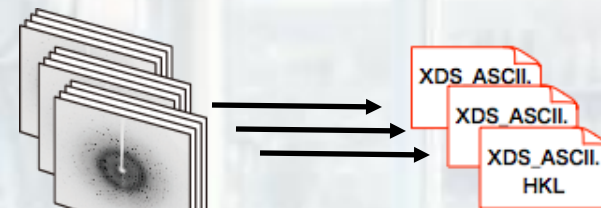
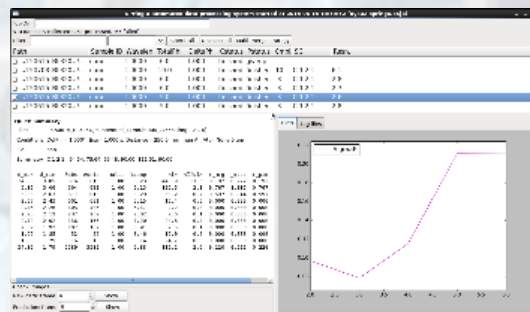
Determination of space group

Clustering (cell/intensity CC)

Selecting good cluster
(Completeness, Redundancy)

Merging datasets

Outlier detection/rejection



指数付・強度積分

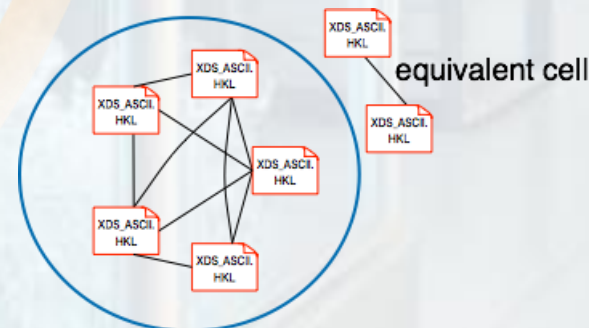
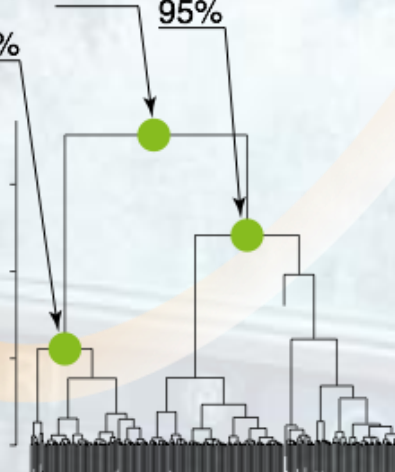
統合 & 結果表示
同型グループ

結晶格子による
等価結晶の分類

Completeness: 100%

90%

95%

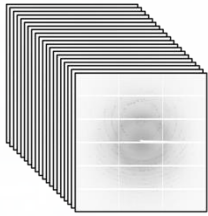


階層的クラスタリング

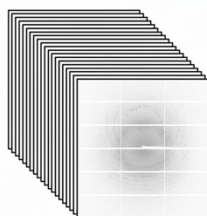
- 結晶格子定数
- データ間の強度相関

強度の相関を利用した階層的クラスタリング

データセット1

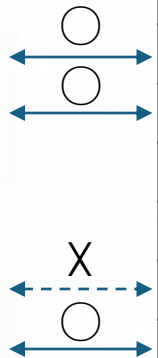


データセット2



共通？

H	K	L	I(HKL)
0	0	1	35000
1	5	5	5000
.	.	.	.
.	.	.	.
25	13	15	230
26	13	14	25
.	.	.	.



H	K	L	I(HKL)
0	0	1	34900
1	5	5	5005
.	.	.	.
.	.	.	.
25	13	12	202
26	13	14	28
.	.	.	.

大前提

- 空間群が同じ
- 格子定数もほぼ同じ

対応する反射強度がない→寄与しない

データセット間で共通に観測された反射強度で相関係数を計算 (CC)

$$CC_{ij} = \frac{\sum_h [I_i(h) - \bar{I}_i][I_j(h) - \bar{I}_j]}{\left\{ \sum_h [I_i(h) - \bar{I}_i]^2 \sum_h [I_j(h) - \bar{I}_j]^2 \right\}^{1/2}}$$

共通反射は通常
300~3000個

回折データを処理した結果の強度を利用して計算 (1組1CC)

CCを利用したHCA

測定した全データセット間のCCを総当りで計算する

	Data1	Data2	Data3	Data4	Data5	...
Data1						...
Data2	CC ₁₂					...
Data3	CC ₁₃	CC ₂₃				...
Data4	CC ₁₄	CC ₂₄	CC ₃₄			...
Data5	CC ₁₅	CC ₂₅	CC ₃₅	CC ₄₅		...
...	...	...	...	...	...	...

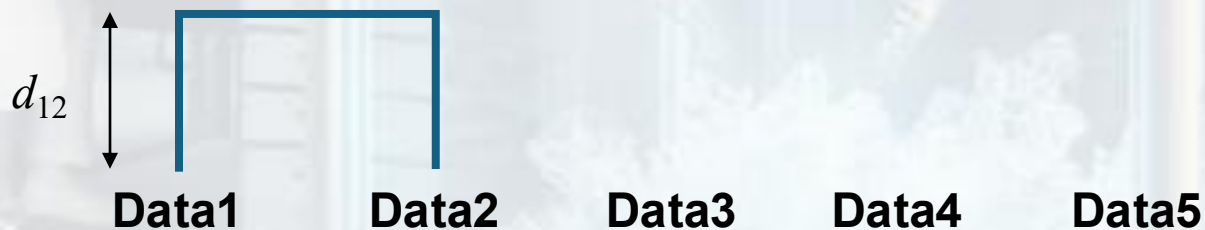
CCを利用したHCA

CCを距離に修正 $d = (1 - CC^2)^{1/2}$ (CC高い→距離d短い)

	Data1	Data2	Data3	Data4	Data5	...
Data1						...
Data2	d_{12}					...
Data3	d_{13}	d_{23}				...
Data4	d_{14}	d_{24}	d_{34}			...
Data5	d_{15}	d_{25}	d_{35}	d_{45}		...
...	...	...	...	...	...	...

CCを利用したHCA

- d_{ij} の短い=データ似ている
- 縦軸の距離が離れるほど構造は異なる



CCを利用したHCA

- d_{ij} の短い=データ似ている
- 縦軸の距離が離れるほど構造は異なる



CCを利用したHCA

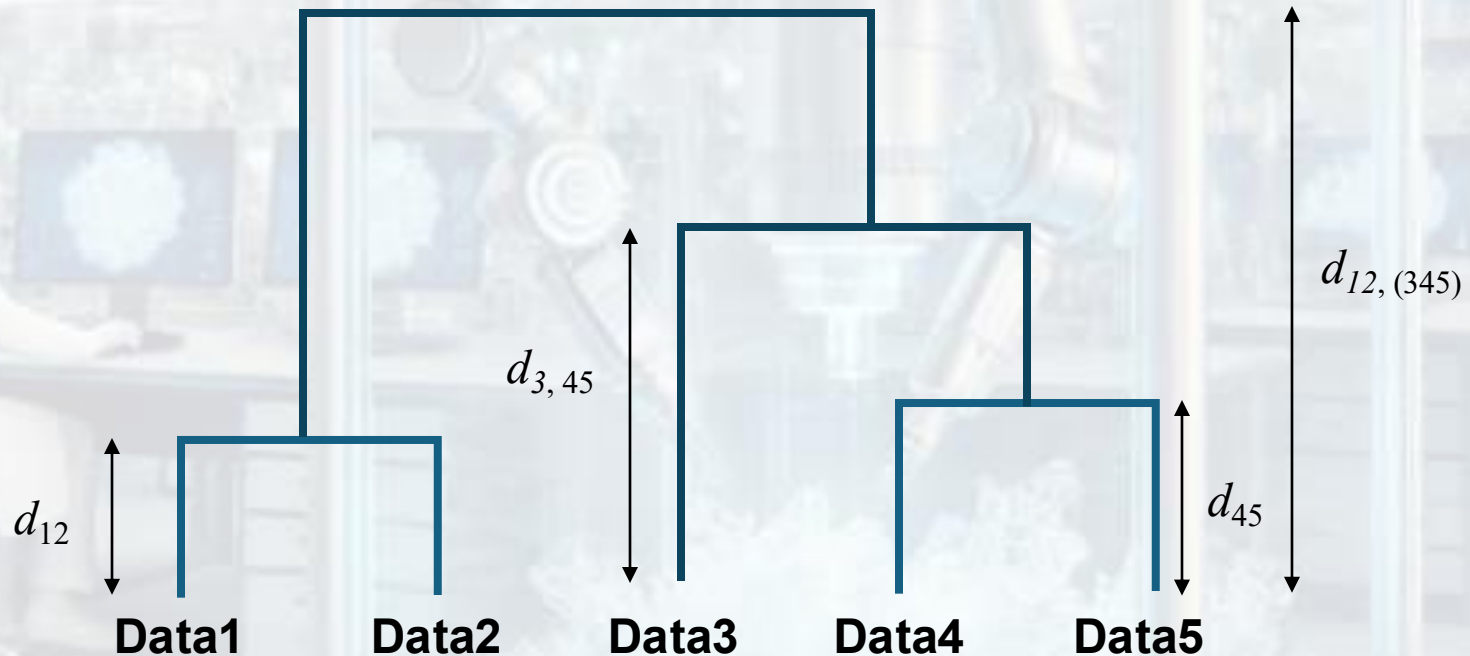
- d_{ij} の短い=データ似ている
- 縦軸の距離が離れるほど構造は異なる



CCを利用したHCA

- d_{ij} の短い=データ似ている
- 縦軸の距離が離れるほど構造は異なる

※距離定義・計算方法は多様



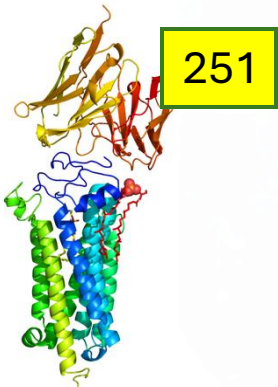
順番にクラスタを形成していきすべてリンクするまで繰り返す

似ているもの同士がグルーピングされる

※実空間ではなく逆空間（位相なし）での計算

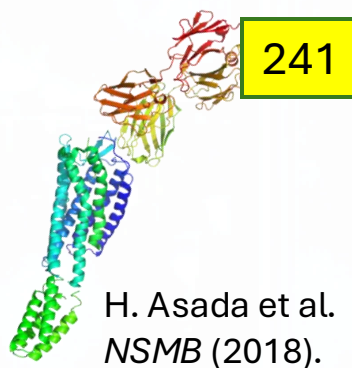
膜タンパク質の高分解能結晶構造

Orexin 2 receptor 2.0 Å



251

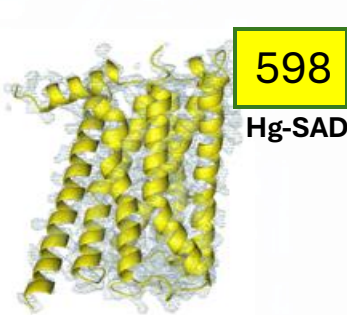
Angiotensin II receptor 3.2 Å



241

H. Asada et al.
NSMB (2018).

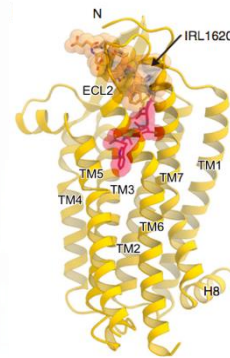
M2 muscarinic receptor 2.3 Å



598

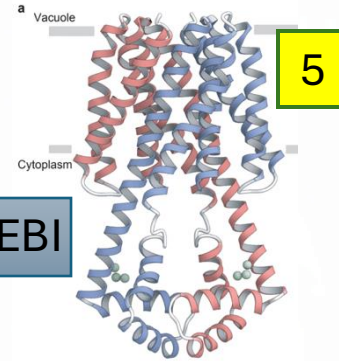
Hg-SAD

Human ETB receptor 2.0 Å



483

Ion transporter(VIT1) 2.7 Å



5

HEBI

R. Suno et al.
Structure(2018).

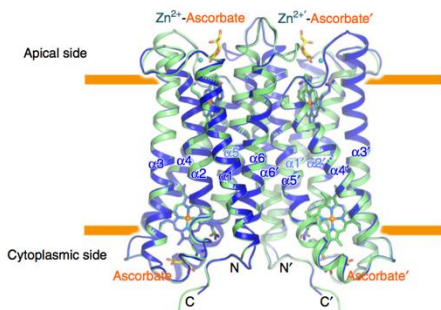
R. Suno et al.
Nat. Chem. Biol.(2018)

W. Shihoya et al.
Nat. Commun.(2018)

T. Kato et al.
Nat. Plants.(2019)

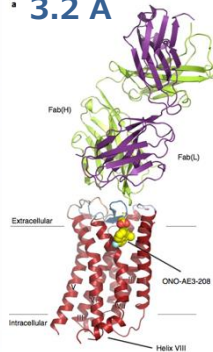
of datasets

Duodenal cytochrome b 2.6 Å



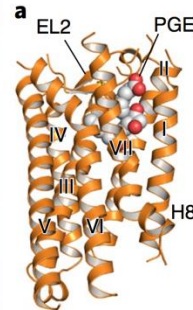
866

Prostaglandin E receptor(EP4) 3.2 Å



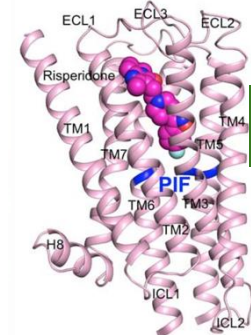
203

Prostaglandin E subtype EP3 3.2 Å



683

5-HT2A Receptor 2.7 Å



265

M. Ganasen et al.
Commun. Biol. (2018).

Y. Toyoda et al.
Nat. Chem. Biol. (2018)

K. Morimoto et al.
Nat. Chem. Biol. (2018)

K. Kimura et al.
Nat. Chem. Biol. (2019)

多数の結晶を利用することで「空間分解能」の向上→パラダイム・シフト

階層的クラスタリングの意義の変遷

評価する物理量：同型性 → **不変**

(同型性 → 空間群、結晶格子長、分子の構造)

運用初期の目的：「**異常データを棄却**」



現状：「**生化学的に意味のある多型検出**」

分子の構造変化

階層的クラスタリングによる回折データの分類

複数の結晶から収集した回折データ / 単一結晶の異なる露光点における回折データ



通常のマージ処理



異なる構造情報が平均化された
不明瞭な電子密度マップ？

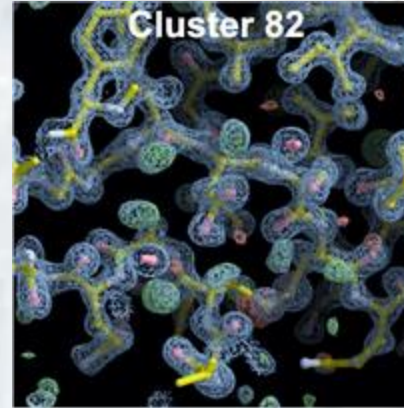
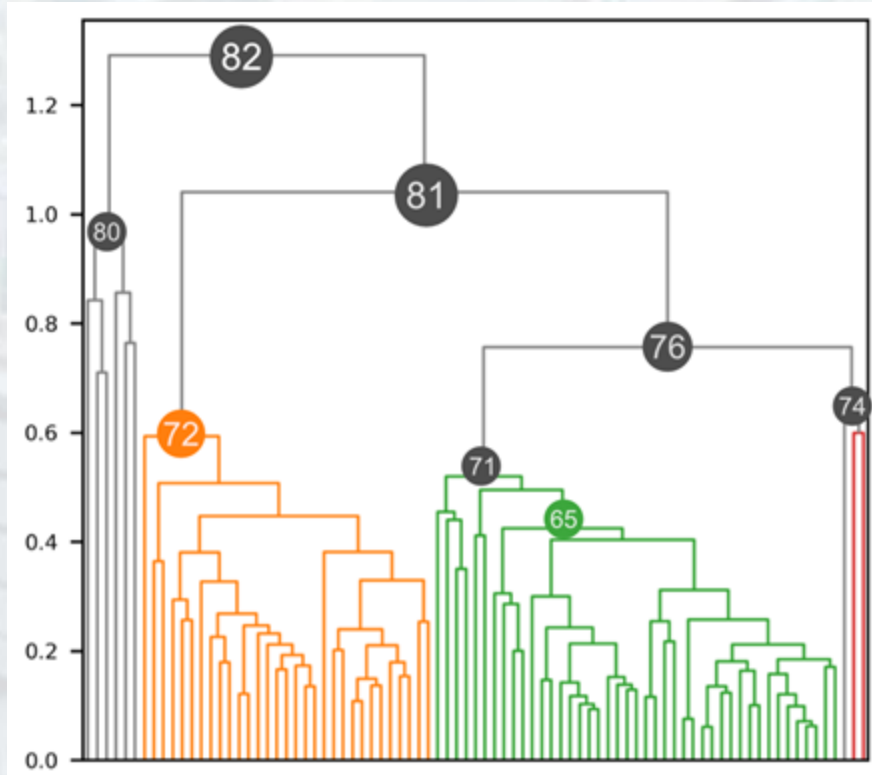
階層的クラスタリング(HCA)
→マージ処理



適切にデータを分類できれば
構造多型を抽出できるのではないか？

階層的クラスタリング (HCA) による構造多型の分類

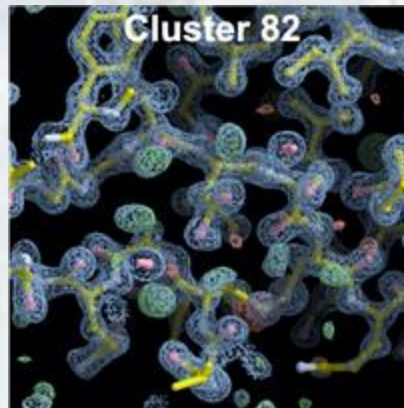
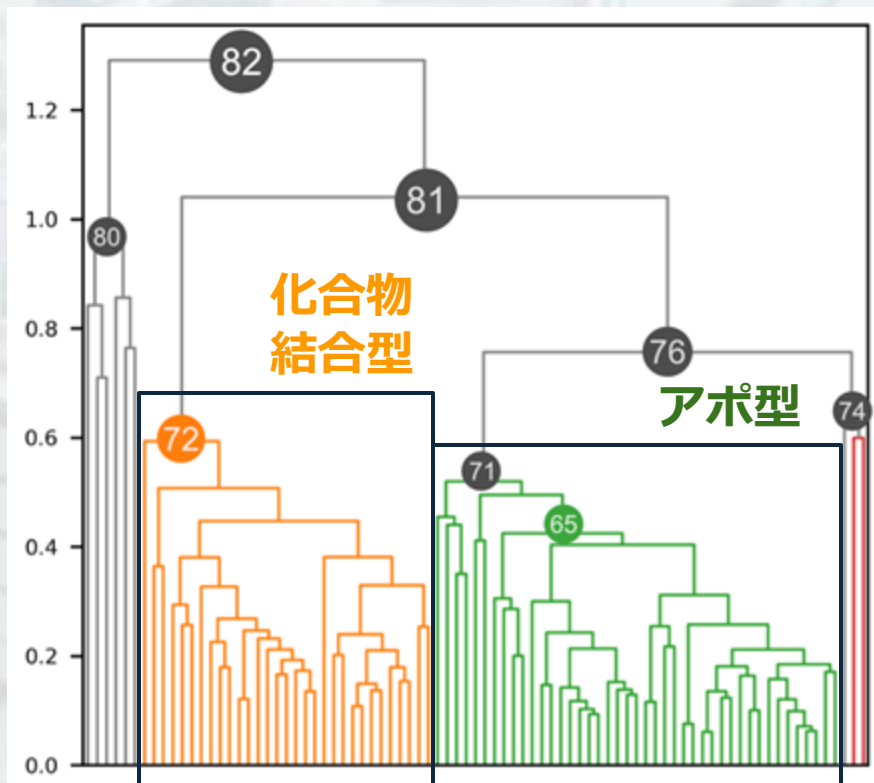
アポ型と化合物結合型のトリプシンの回折データを混ぜた→階層的クラスタリング



すべてのデータを
マージした結果では
化合物の電子密度が消失

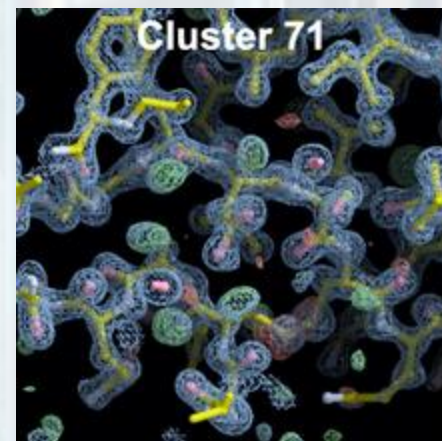
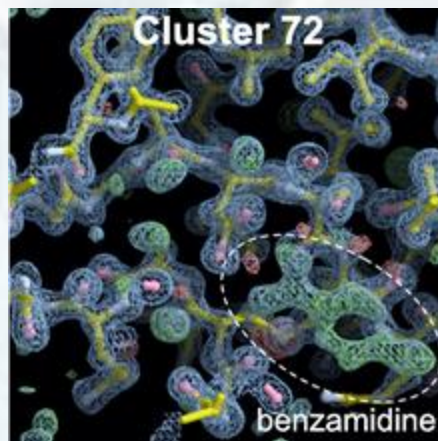
階層的クラスタリング (HCA) による構造多型の分類

アポ型と化合物結合型のトリプシンの回折データを混ぜた→階層的クラスタリング



すべてのデータを
マージした結果では
化合物の電子密度が消失

データを分類すると…



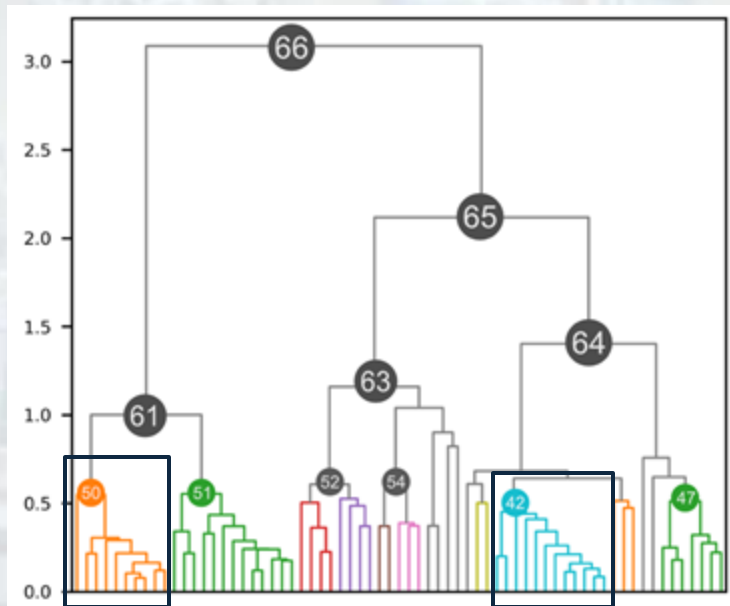
混ぜると消える→分類して足すと見える

薬剤候補物質探索にも有利

実サンプルへの適用例その1：HypD

HypD: [NiFe]ヒドロゲナーゼ成熟化因子

分子研 村木先生(現・慶応大)との共同研究

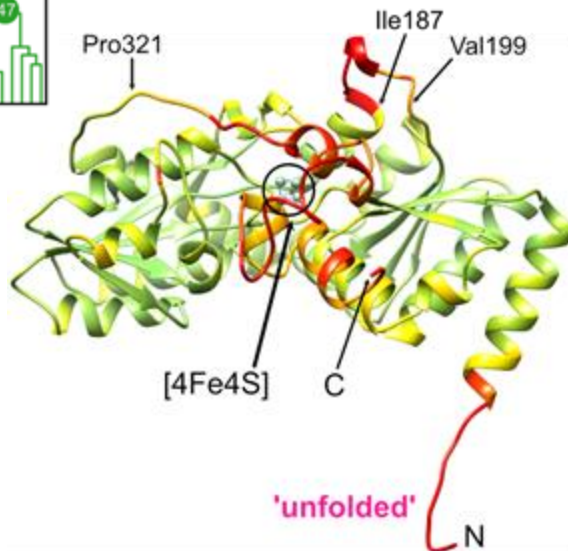


同じ結晶化条件で得られた複数結晶

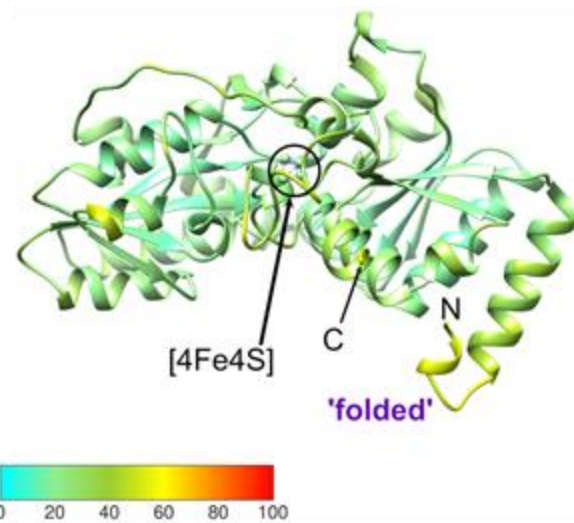
→データ収集→クラスタリング解析

→N末端の構造と[4Fe4S]クラスター周辺環境の大きく異なる2状態を同定

Cluster 50



Cluster 42

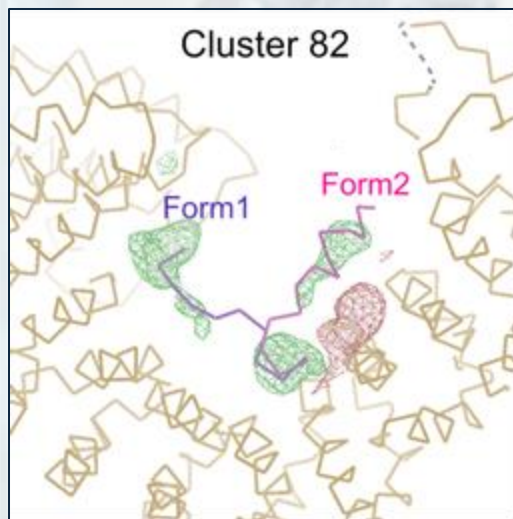
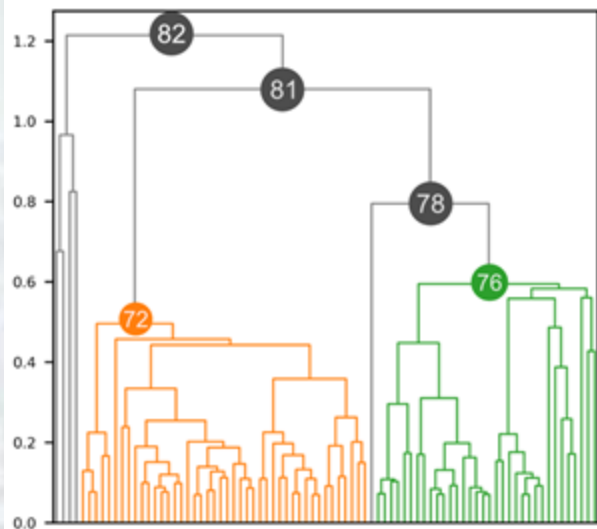


HCAにより大量データの中に多型を検出！

適用例その2 : Transportin-1 peptide complex

核輸送タンパク質Transportin-1とペプチドの複合体

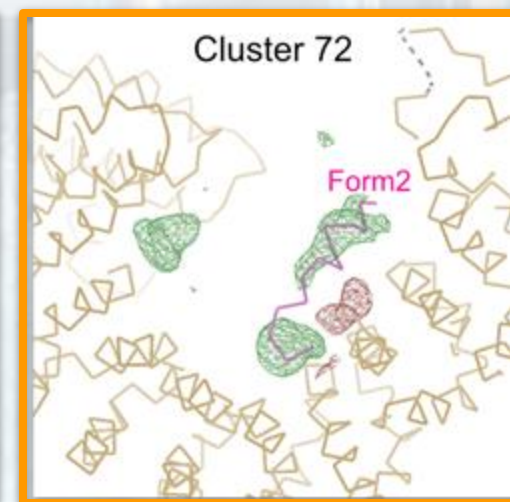
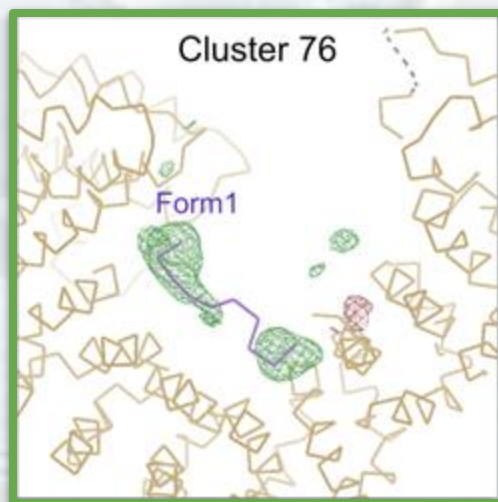
NAIST 藤間先生との共同研究



2種類のペプチド結合様式

クラスタリングしてみると…

ペプチドとの共結晶化を
複数の異なるpHで行った結果、
優勢な結合様式がpH依存的に変わる
ことがわかった

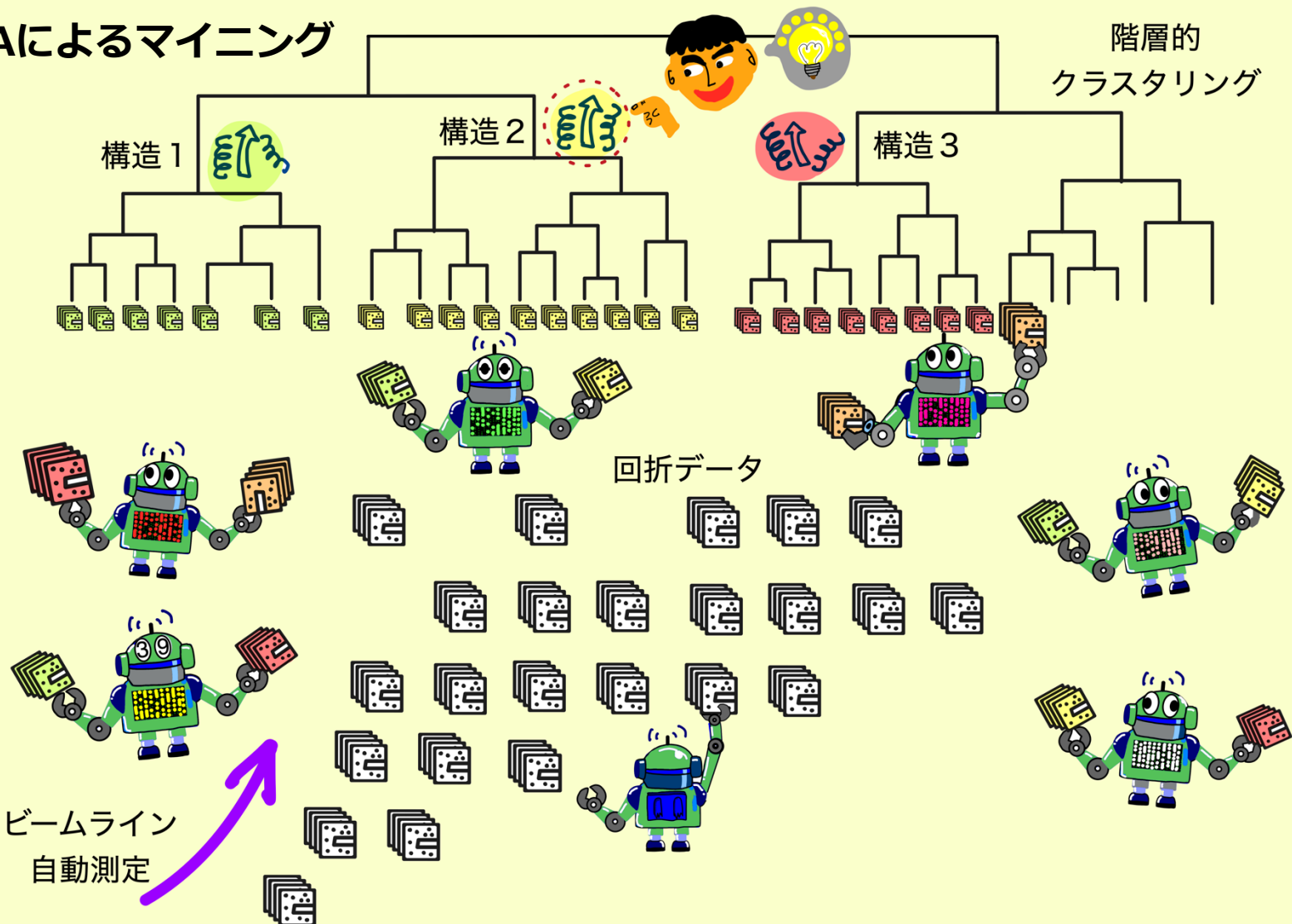


HCA→多型構造検出→生物学的に重要な知見へ

膨大なデータから多型構造検出

HCAによるマイニング

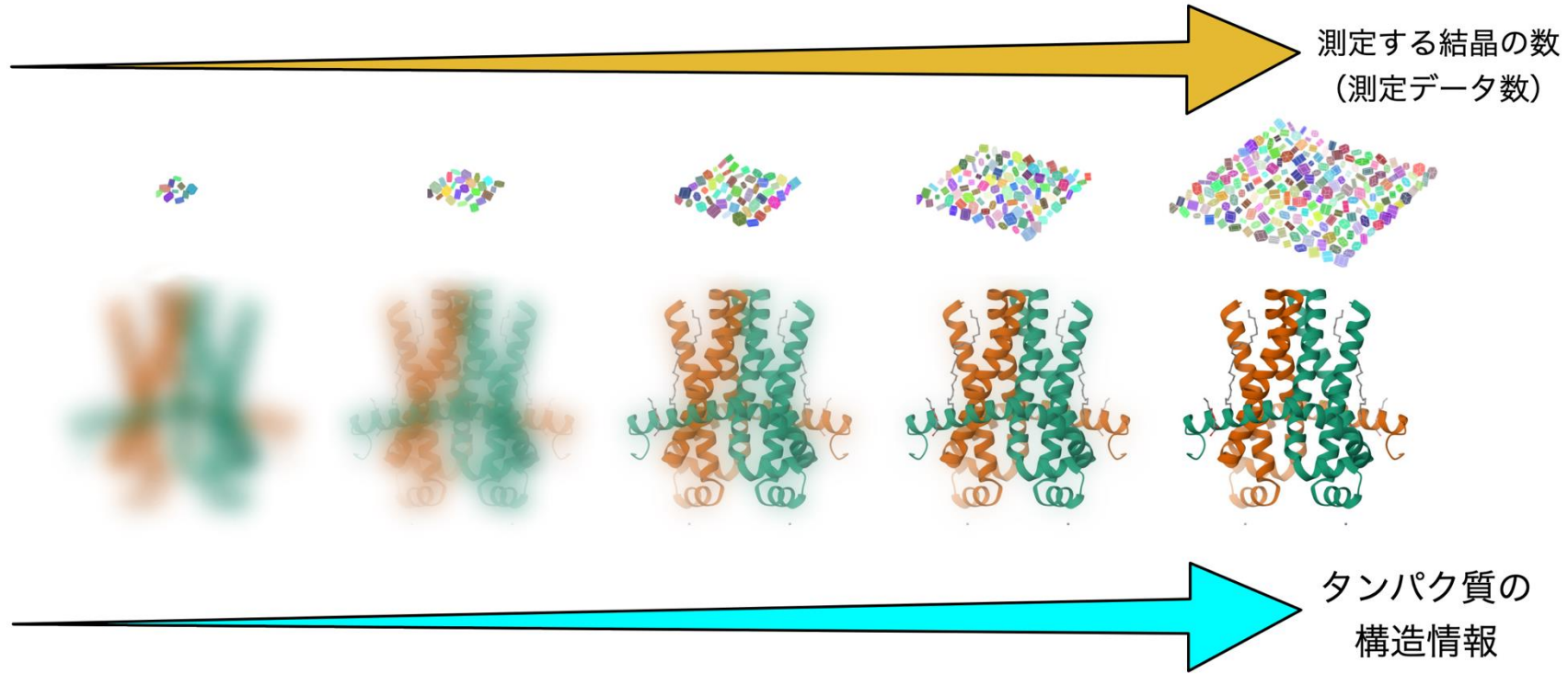
階層的
クラスタリング



膨大なデータ発生

Matsuura, H. *et al.* (2023). *Acta Cryst D* **79** 909-924

多数結晶からのデータを統合



放射線損傷による露光量/結晶の制限

→複数結晶統合による分解能向上（実績）

疑問：限界があるのか？

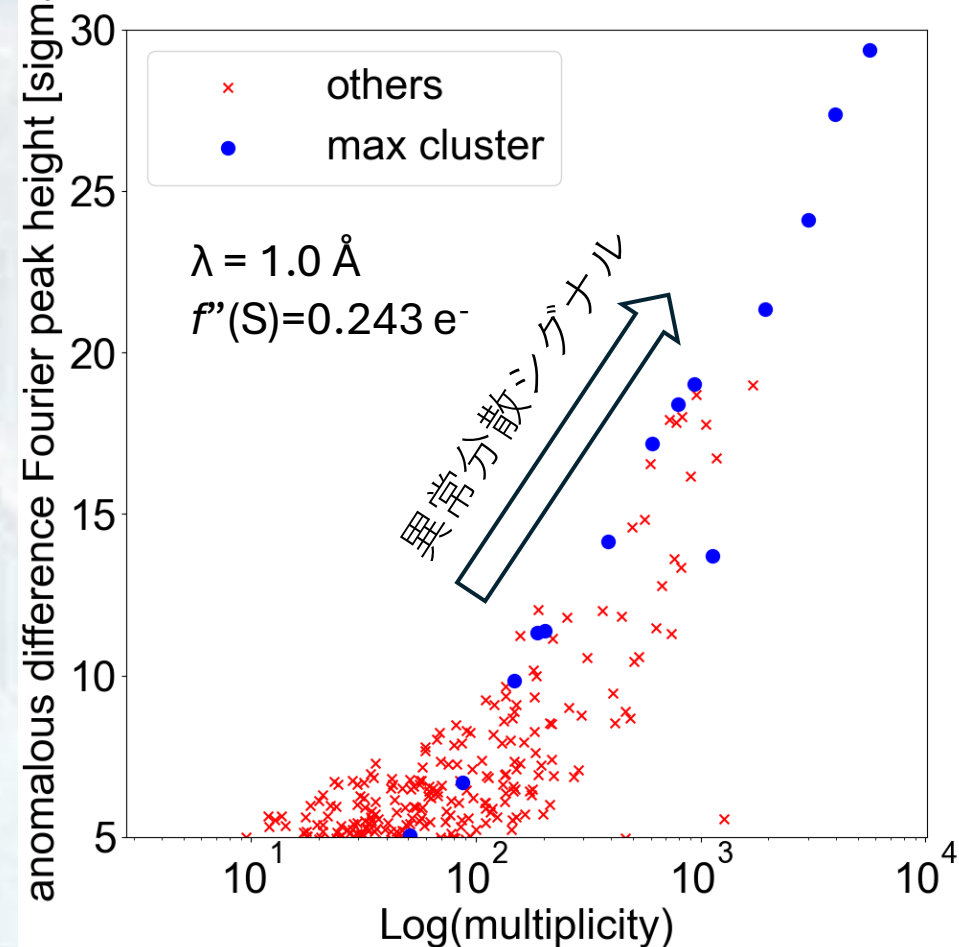
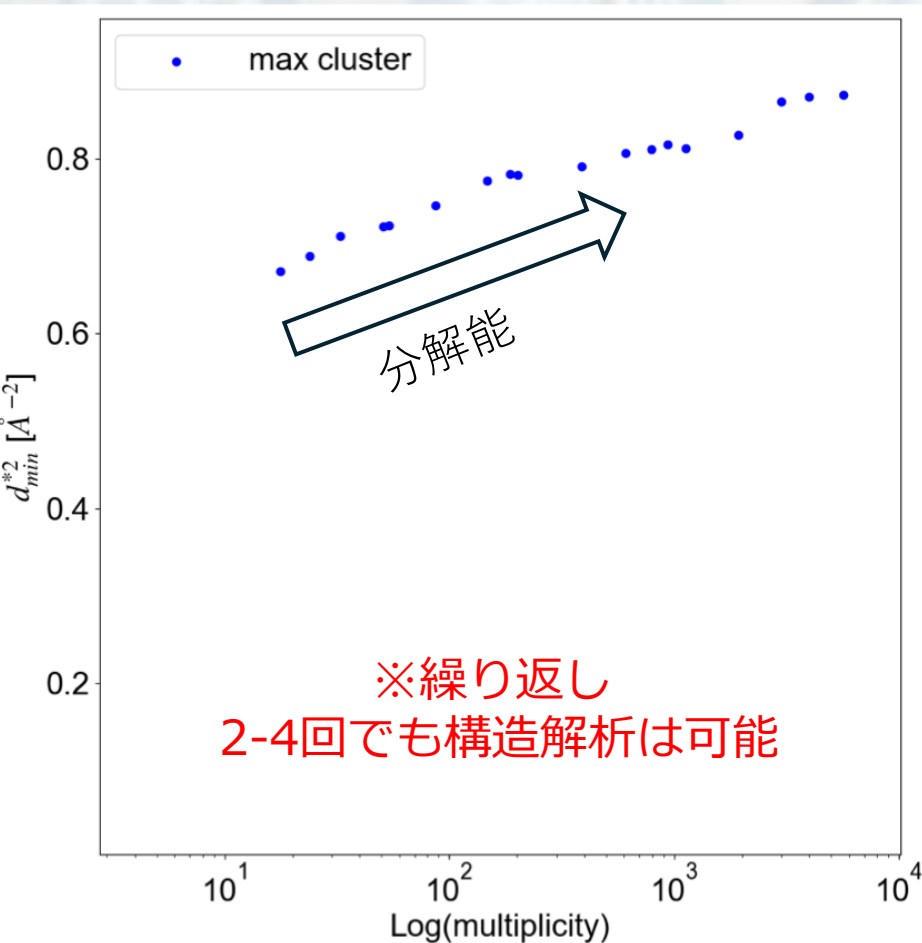
多数データ→微弱シグナル積算→高分解能化

東京科学大 安部先生のご厚意

Polyhedral protein crystal: 3-5 μm size

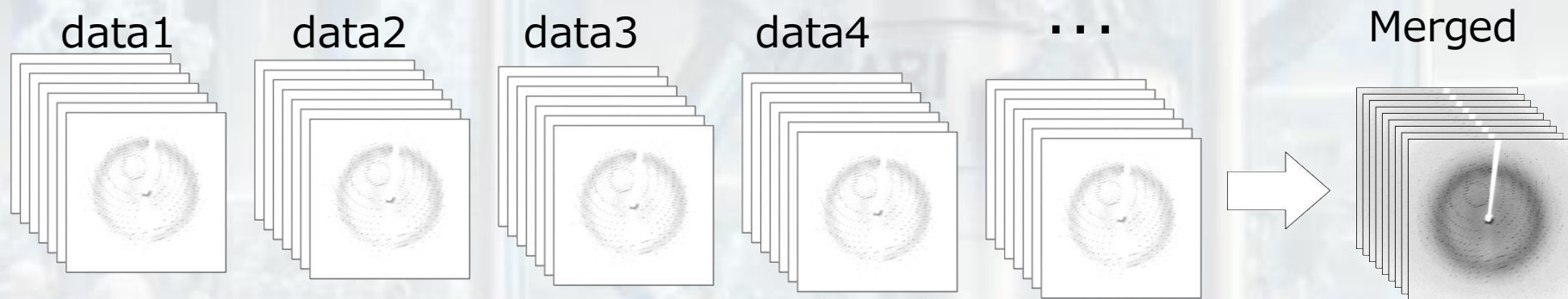
Max cluster=the largest isomorphous group

5° x ~12,000 datasets

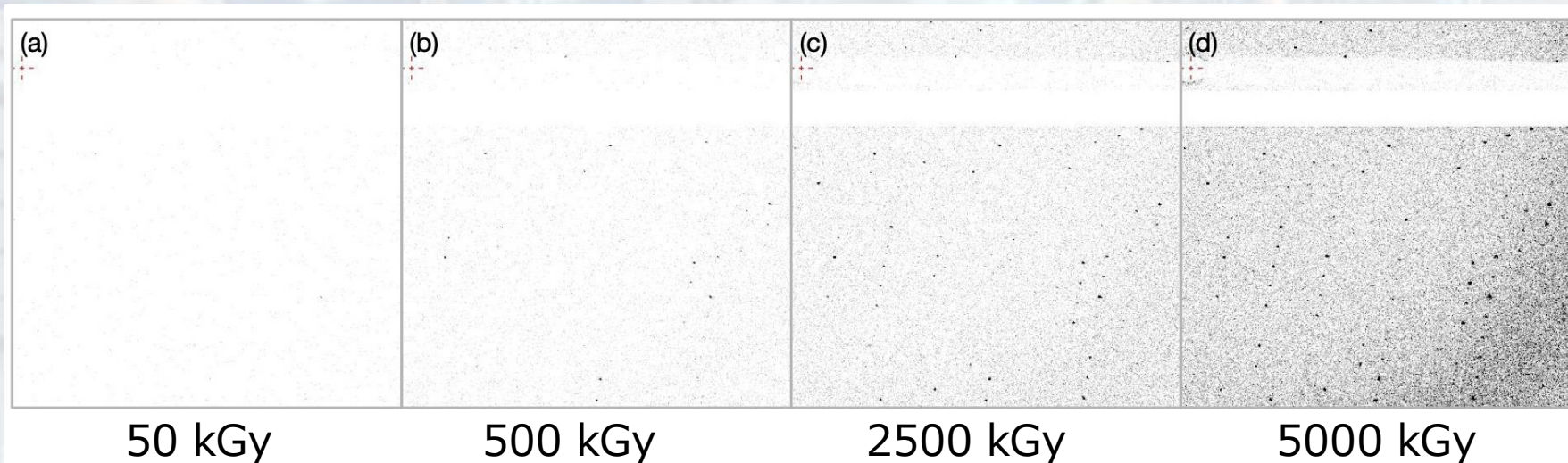


冗長度 ~5,600 : 構造情報増加が継続

微弱なシグナルの積算



低Doseデータの積算→シグナルの検出



50 kGyのデータを100枚収集して少しずつ積算

測定に利用した光子数増→分解能向上

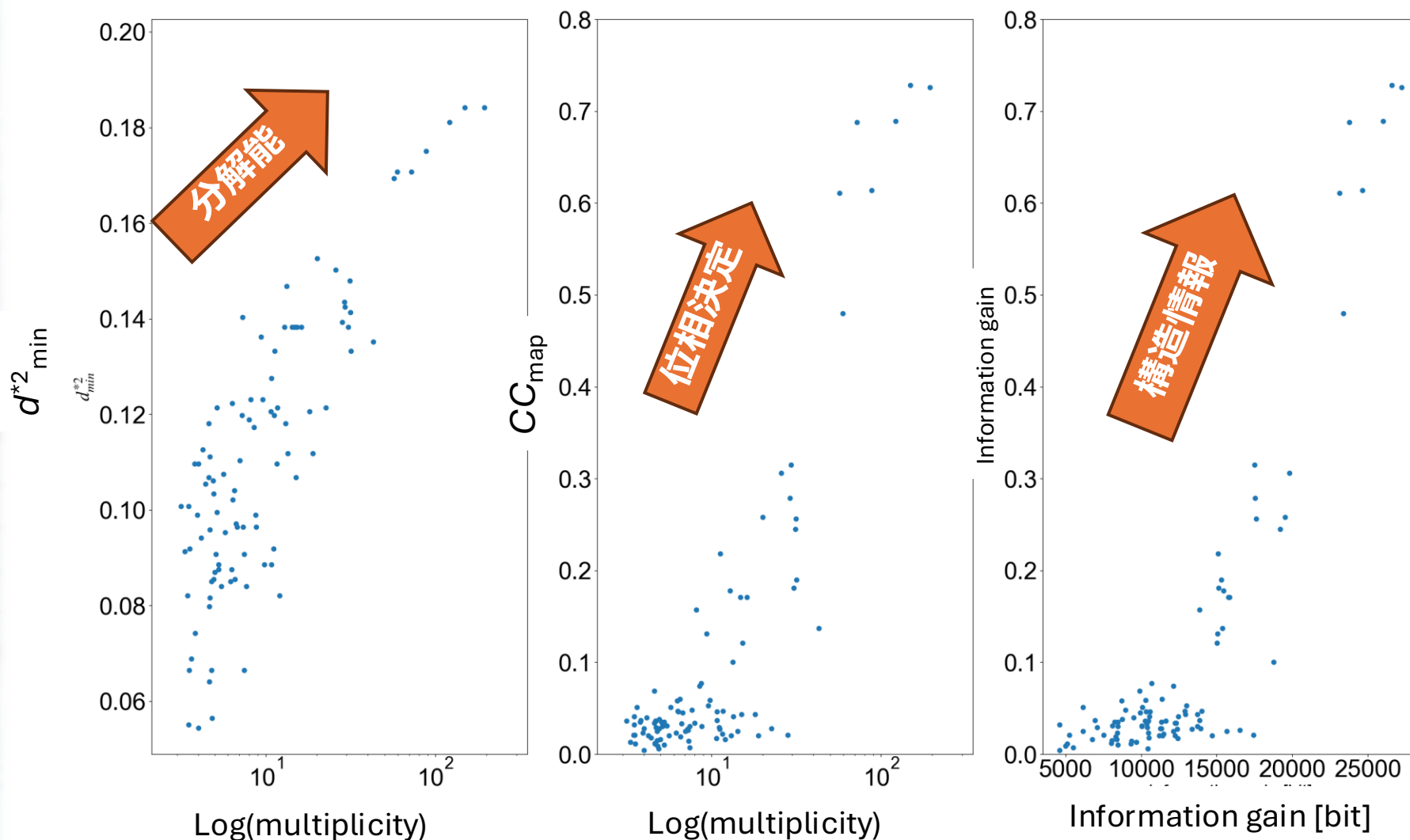
膜タンパク質位相決定精度向上



Seを利用した位相決定

復旦大学 服部教授のご厚意による

bacterial CNNM/CorC
Mg²⁺ transporter

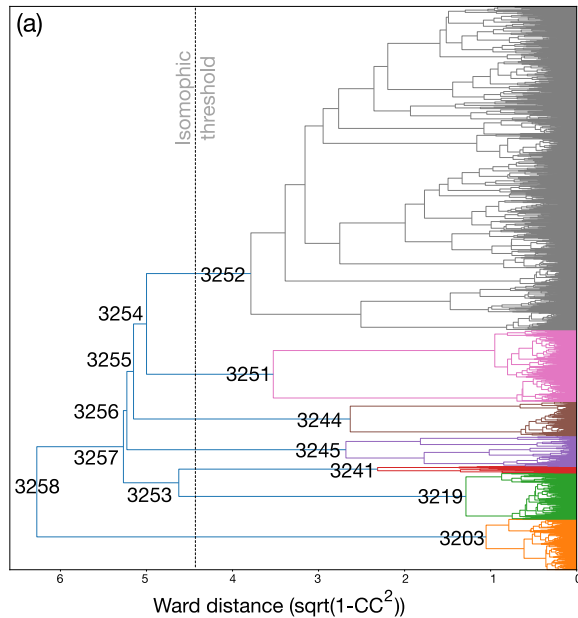


高難度試料においても真

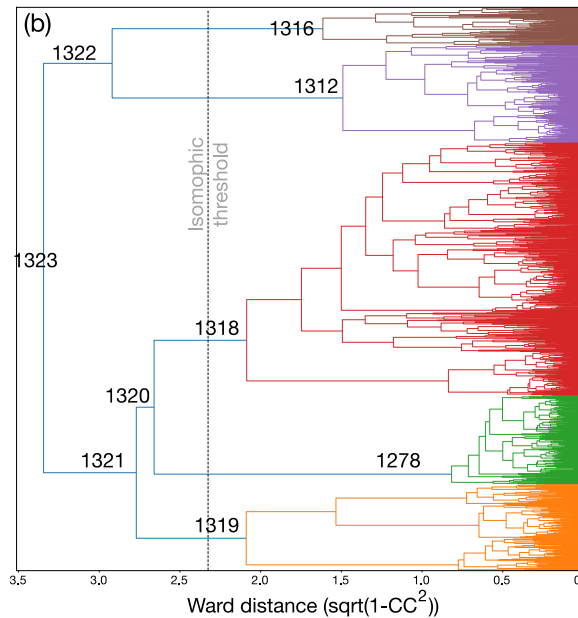
データ分類の重要性@データ統合

デンドログラムを見て異常データの棄却・多型抽出

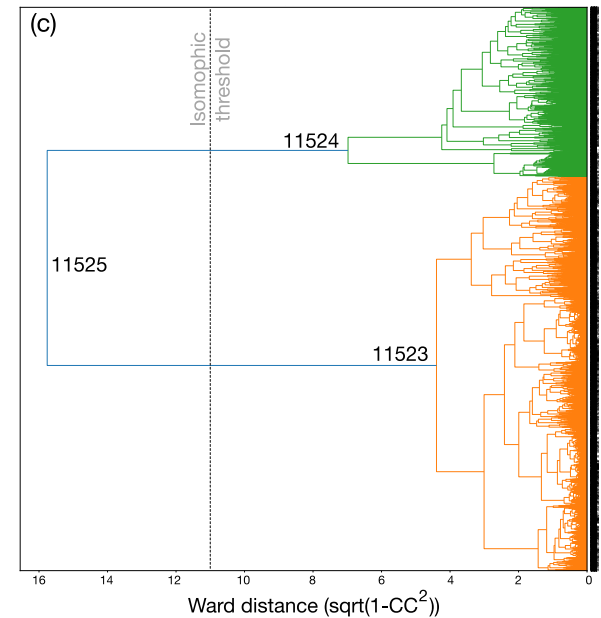
2 型アンジオテンシンII受容体



Mg²⁺トランスポーター
CNNM/CorC

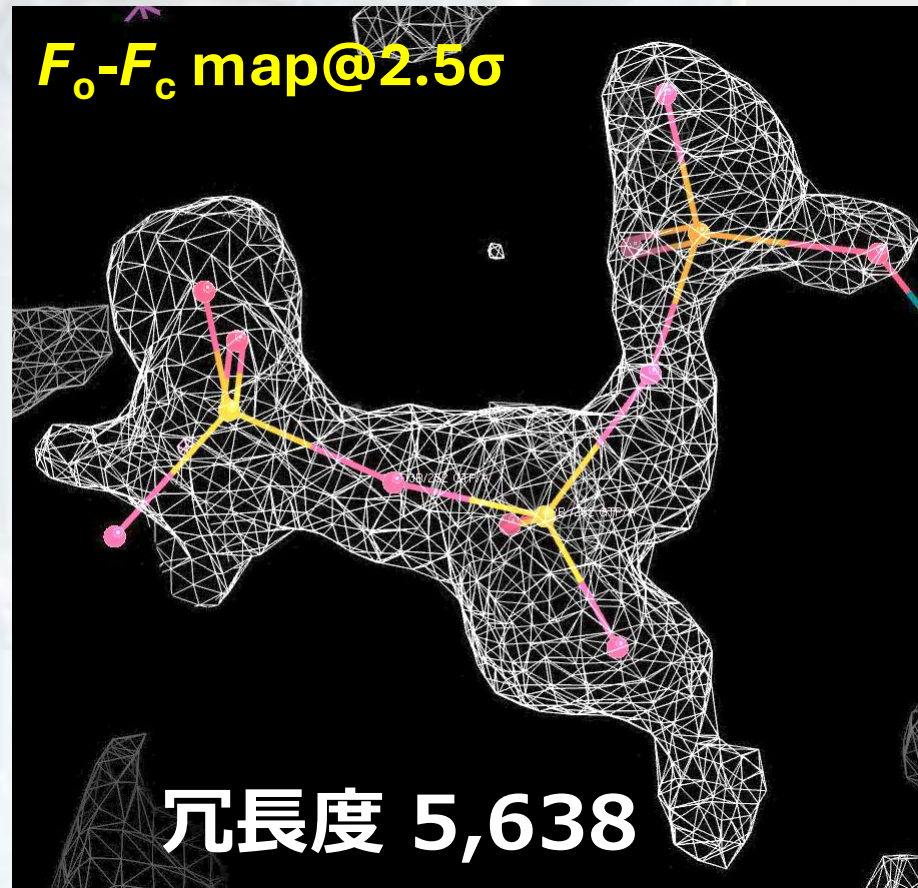
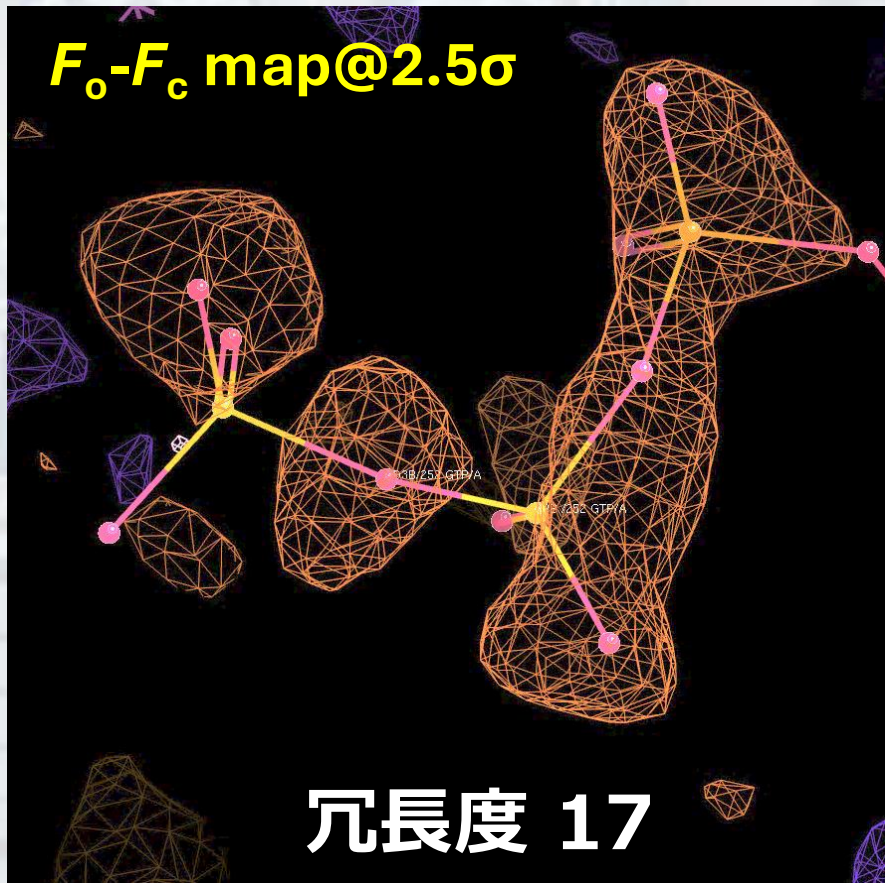


Polyhedra protein



AI技術で「見るべき集合に圧縮」→宝探し

多角体タンパク質: GTP分子の電子密度図の改善



位相が正しい→解釈しやすい電子密度図
→正しい構造へ近づく

さらに質の高い構造生命科学へ

• 構造と機能の関係を解き明かす

- AlphaFoldなどの構造予測ではまだ解き明かせないものがあるのでは？
- PDB登録情報 X線全体 197,459構造
 - 水分子が見える分解能 (1.8Å) 以上の構造 33% (66,042構造)
 - 水素原子が見える分解能 (1.1Å) 以上の構造 2% (4,150構造)

• 多型構造の自動検出（実装済み）

- 自動データ収集→HCA→多型構造が検出されたら確認

• 高分解能構造解析

- 結晶の回折能に規定されてきた→データ数（結晶数）でも向上
- 結晶化・データ収集の迅速化・効率化により
 - タンパク質をとりまく水分子の位置や動き
 - 水素原子の可視化→水素原子位置・プロトン化状態の可視化

• 時分割構造解析（SPring-8 / SACLA）

まとめ

- タンパク質結晶構造解析

- 自動測定・自動処理→機械学習データマイニング

- “データ量爆発→AI技術で宝探し！”

- 放射光 × AI が拓く未来

- データ発生装置 × AI = 新たな発見を量産