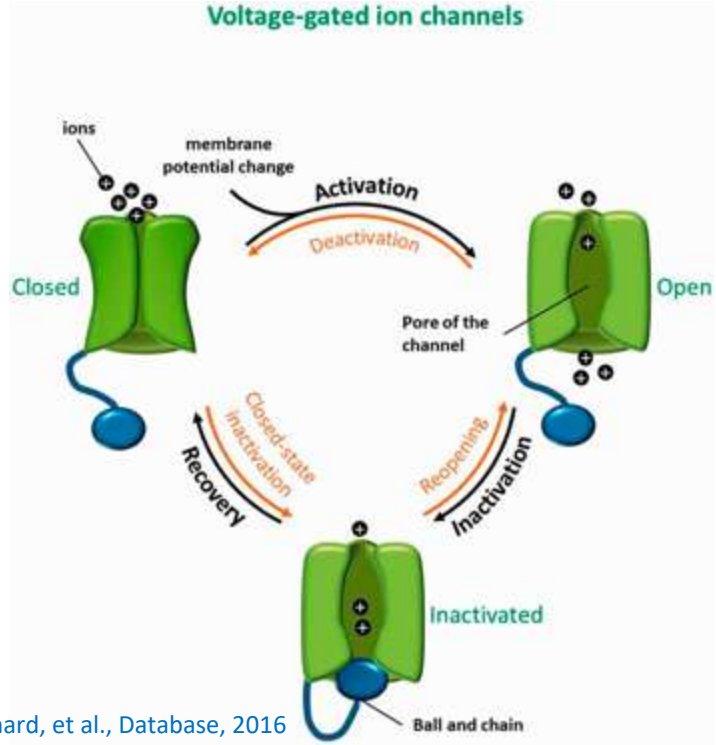
AWS ParallelClusterをハブとした 単粒子クライオ電子顕微鏡構造ベースの 化合物スクリーニング現場のIoT化

守屋　俊夫

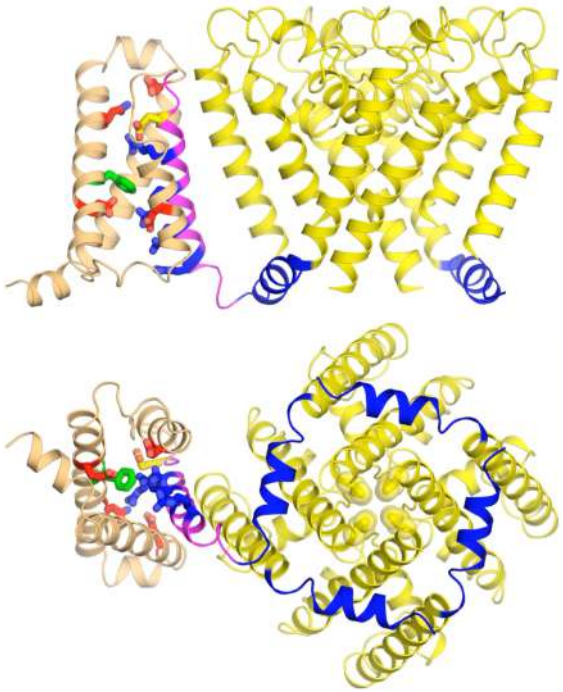
高エネルギー加速器研究機構・物質構造科学研究所・構造生物学研究センター

CBI 学会 2021年大会 企業セッションSS-22
構造生物学研究におけるクラウド活用の現在と展望
2021年10月27日 13:00 – 14:30

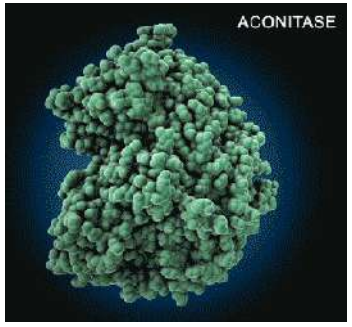
タンパク質は動くことで機能！
原子分解能レベルで動きを可視化して創薬に繋げる



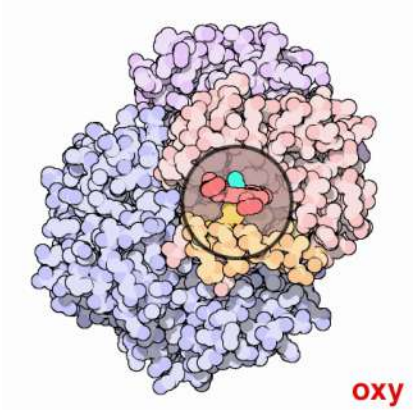
V.Hinard, et al., Database, 2016

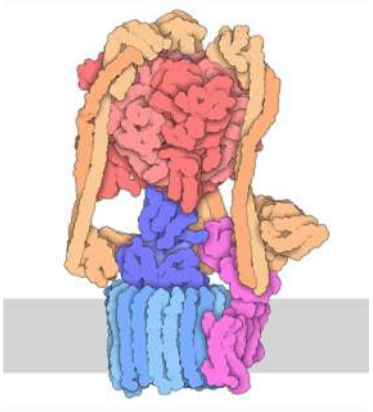


G.Wisedchaisri, et al., Cell, 2019



https://pdb101.rcsb.org/learn





大学共同利用機関法人 高エネルギー加速器研究機構 物質構造科学研究所

構造生物学研究センター

SBRC

Structural Biology Research Center

クライオ電顕　**2018年4月運転開始**

Talos Arctica (TF 200kV)

構造解析計算

GPUボックス/
コンピュータークラスタ

「部分」構造
フィッティング

単粒子解析構造

中～高分解能「全体」構造

初期3次元構造

3次元構造多形分析

X線結晶構造

高分解能「部分」構造

「部分」構造
フィッティング

X線溶液散乱解析

低分解能「全体」構造

試料準備
結晶化ロボット

Photon Factory
ビームライン
X線構造解析

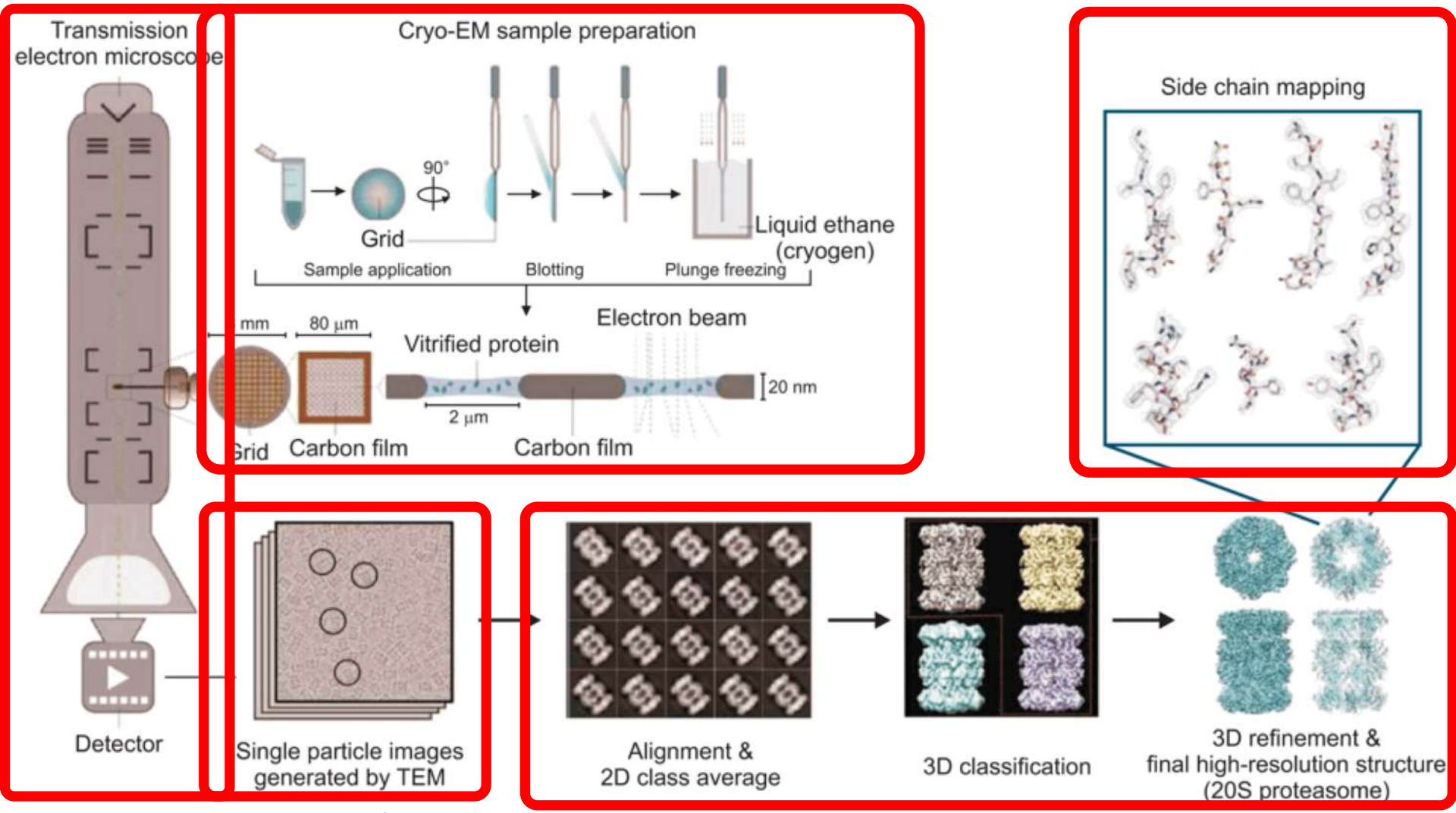
相関構造解析が可能な総合的な構造生物研究環境

クライオ電顕・単粒子解析の全体像

電顕撮影

試料準備（氷包埋法）

原子座標モデル構築



電顕像

コンピューター画像処理（3次元構造再構築）

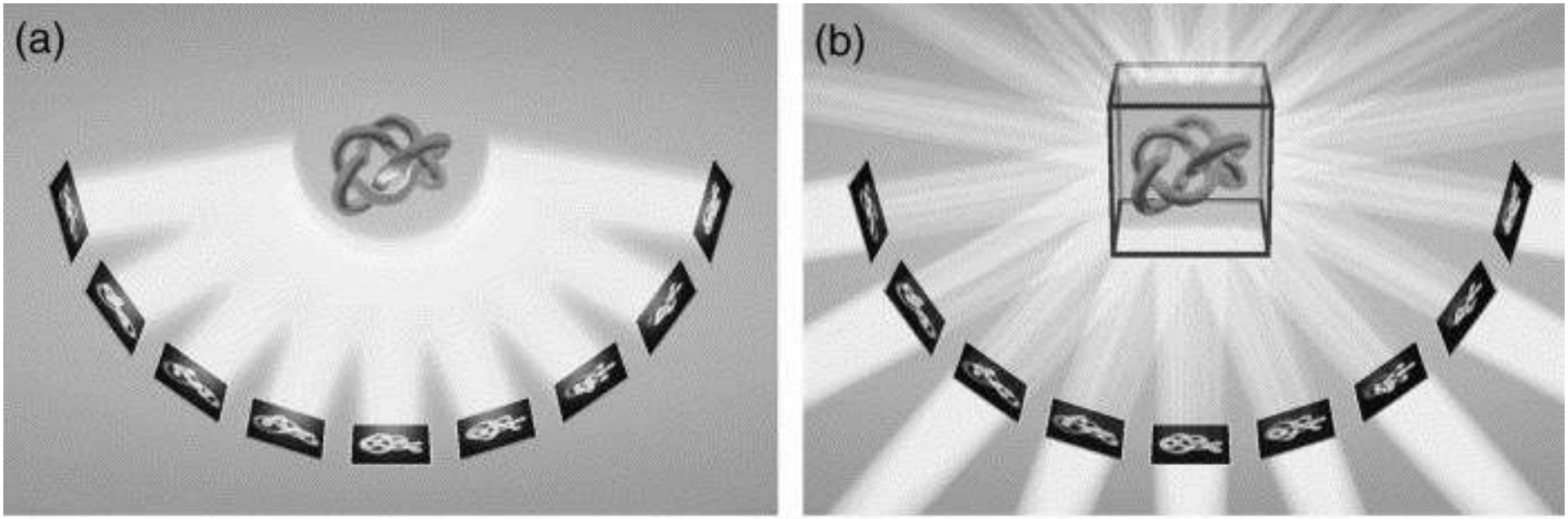
DOI: https://doi.org/10.9729/AM.2017.47.4.218

3次元再構築の基礎

生体物質を透過する信号を利用
→投影像が得られる

ラドン変換
（順投影 3D→2D）

逆ラドン変換
（逆投影 2D→3D ）

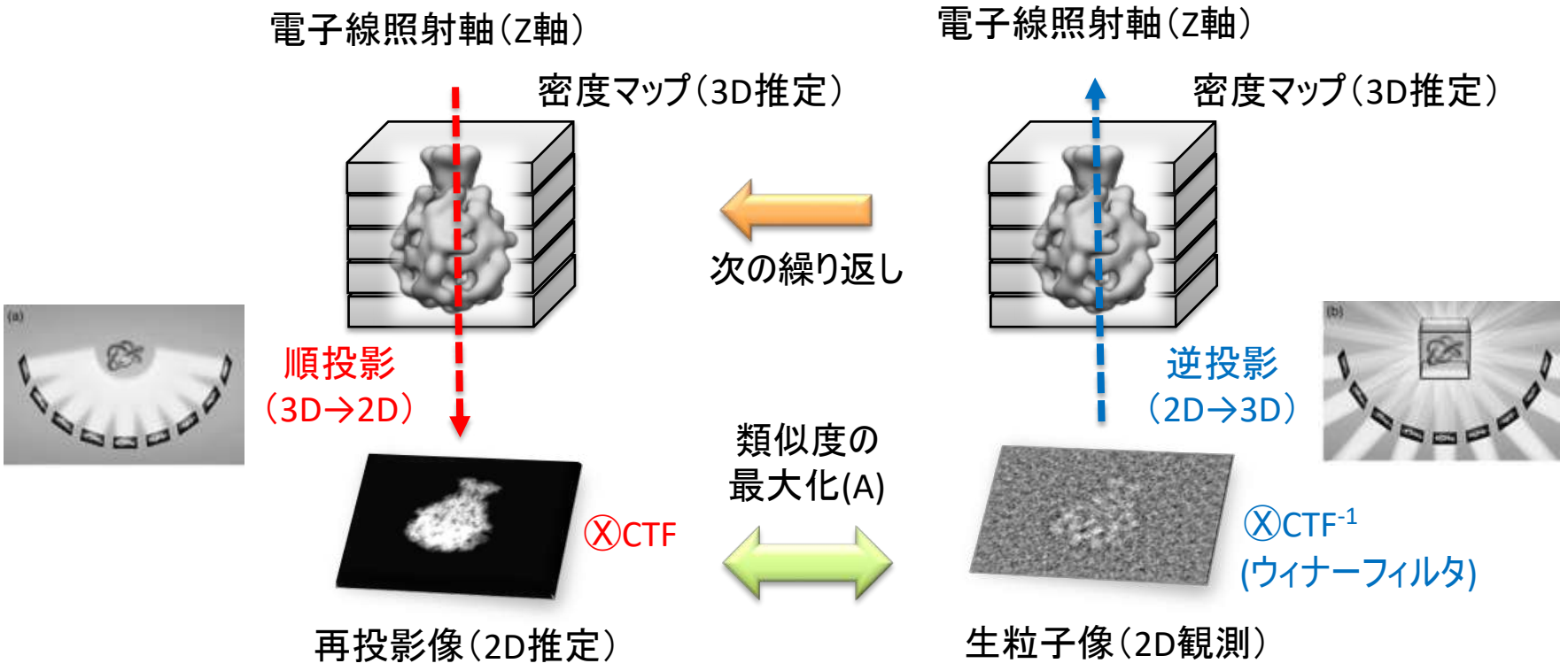


https://www.cell.com/

これが透過型電子顕微鏡がすること！
ソフトウェアでシミュレートできる

これが3次元再構築含む
単粒子解析アルゴリズムで
やりたいことの全て！

従来の3次元精密化
とその投影モデル（実空間）



参照3次元構造有りの
投影3次元角度推定

＋

3次元再構築

＝

プロジェクトマッチング法
による3次元精密化

古典的な3次元精密化
プロジェクトマッチング法（投影像適合法）

初期参照マップ（入力）

新しい参照マップ（繰返し）

注記：比較時は、入力された検索範囲と幅に従った全てのXYシフトと面内回転の組み合わせを確認

1.順投影により
再投影像を生成

2.各粒子像と全ての再投影像を比較、各粒子に最も適合する再投影像を検出し、その投影3次元角度を粒子にアサイメント => MRA

0,0,0

最適合！

粒子像（入力）

0,0,0

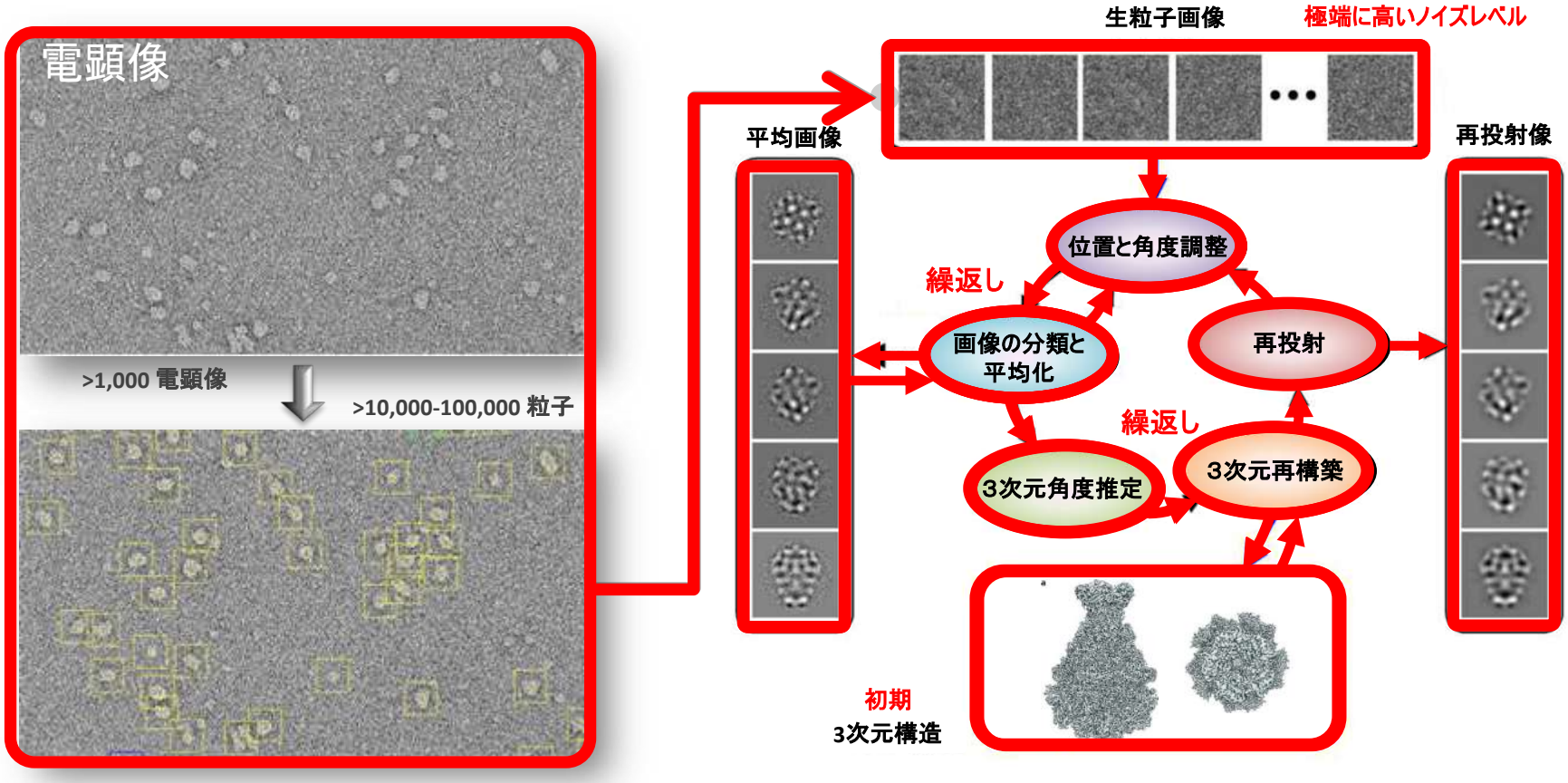
3. 逆投影による再構築

3次元密度マップ
（出力）

4. 参照マップを更新して、
全ての処理ステップを繰り返す

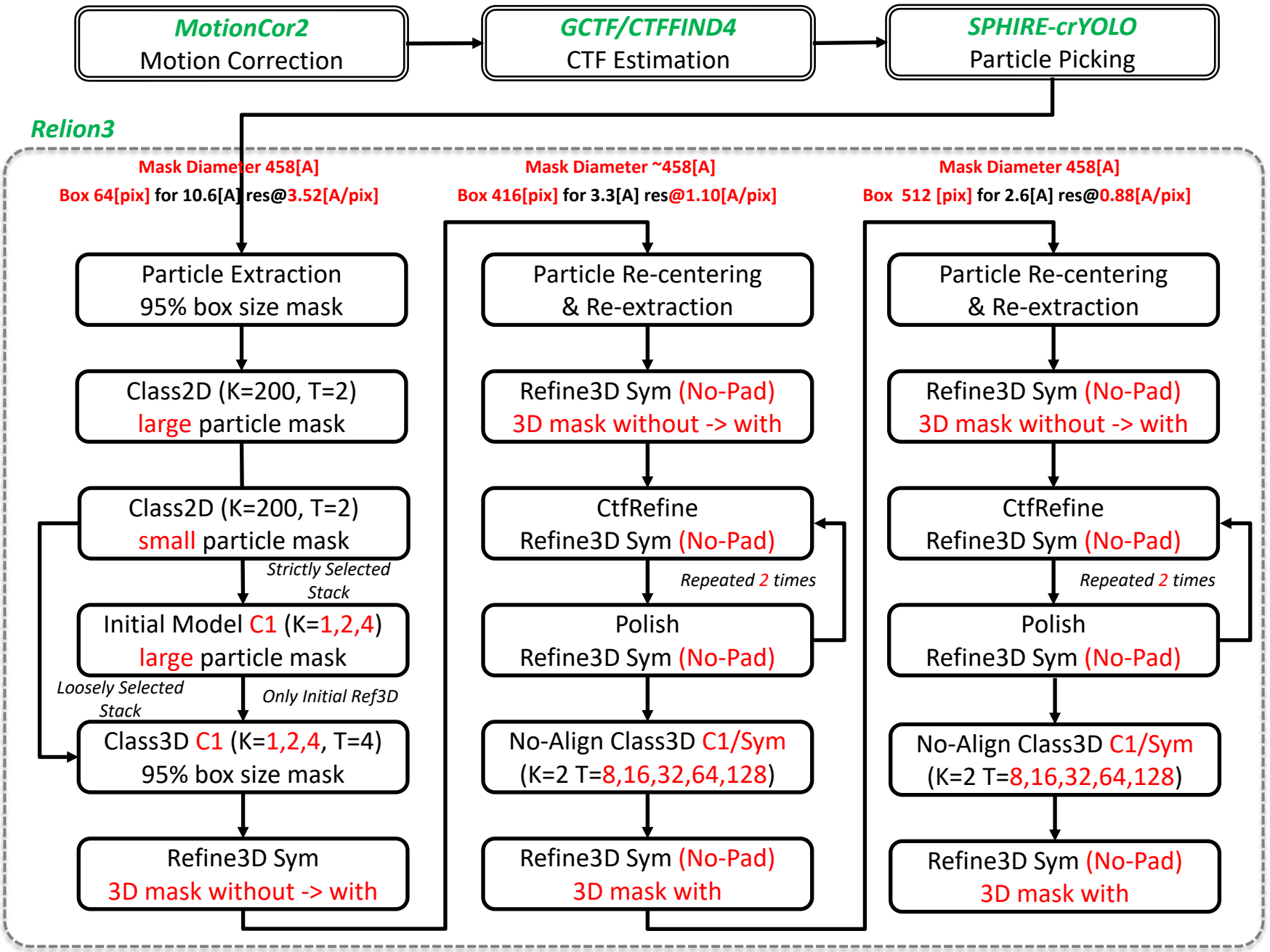
ちょっと古い単粒子解析のワークフロー（処理手順）
コンピューター画像処理

粒子拾い上げ → 初期2次元平均化 → 初期3次元構造の決定 → 3次元構造の分解能向上



単粒子解析は画像処理ヘビーな手法

特に、トモグラフィーと違い「3次元角度推定」が必要


MotionCor2
Motion Correction
GCTF/CTFFIND4
CTF Estimation
SPHIRE-crYOLO
Particle Picking
Relion3
Mask Diameter 458[A]
Box 64[pix] for 10.6[A] res@3.52[A/pix]
Particle Extraction
95% box size mask
Class2D (K=200, T=2)
large particle mask
Class2D (K=200, T=2)
small particle mask
Strictly Selected Stack
Initial Model C1 (K=1,2,4)
large particle mask
Loosely Selected Stack
Only Initial Ref3D
Class3D C1 (K=1,2,4, T=4)
95% box size mask
Refine3D Sym
3D mask without -> with
Mask Diameter ~458[A]
Box 416[pix] for 3.3[A] res@1.10[A/pix]
Particle Re-centering
& Re-extraction
Refine3D Sym (No-Pad)
3D mask without -> with
CtfRefine
Refine3D Sym (No-Pad)
Repeated 2 times
Polish
Refine3D Sym (No-Pad)
No-Align Class3D C1/Sym
(K=2 T=8,16,32,64,128)
Refine3D Sym (No-Pad)
3D mask with
Mask Diameter 458[A]
Box 512 [pix] for 2.6[A] res@0.88[A/pix]
Particle Re-centering
& Re-extraction
Refine3D Sym (No-Pad)
3D mask without -> with
CtfRefine
Refine3D Sym (No-Pad)
Repeated 2 times
Polish
Refine3D Sym (No-Pad)
No-Align Class3D C1/Sym
(K=2 T=8,16,32,64,128)
Refine3D Sym (No-Pad)
3D mask with

国立研究開発法人 日本医療研究開発機構
Japan Agency for Medical Research and Development

AMED

BINDS

創薬等先端技術支援基盤プラットフォーム

Basis for Supporting Innovative Drug Discovery and Life Science Research

「知って、使って、進むあなたの研究」

【目標】データ解析のハイスループット化

研究遂行のボトルネックの一つが解析であることは、経験上明らか・・・

クラウド環境にハイスループットの解析環境を整えスループットを向上！

1日に10-20データの測定は、近い将来達成される

1日20構造決定を目標に！

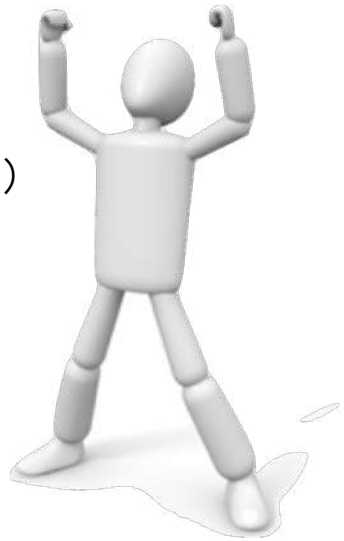
平均的な解析PC（約 150万円）では、1構造に1−2週間必要
年間20回程度のマシンタイムとすると2週間に1回のマシンタイムとなり、
次回のマシンタイムまでに結果を得ることができない。
かつ20構造決定しようとすると、20台の解析PCが必要（3000万円の投資）
＋保守の人件費＋煩雑なメインテナンス作業

ハイスループット
解析環境
TCEF Cloud (KEK)

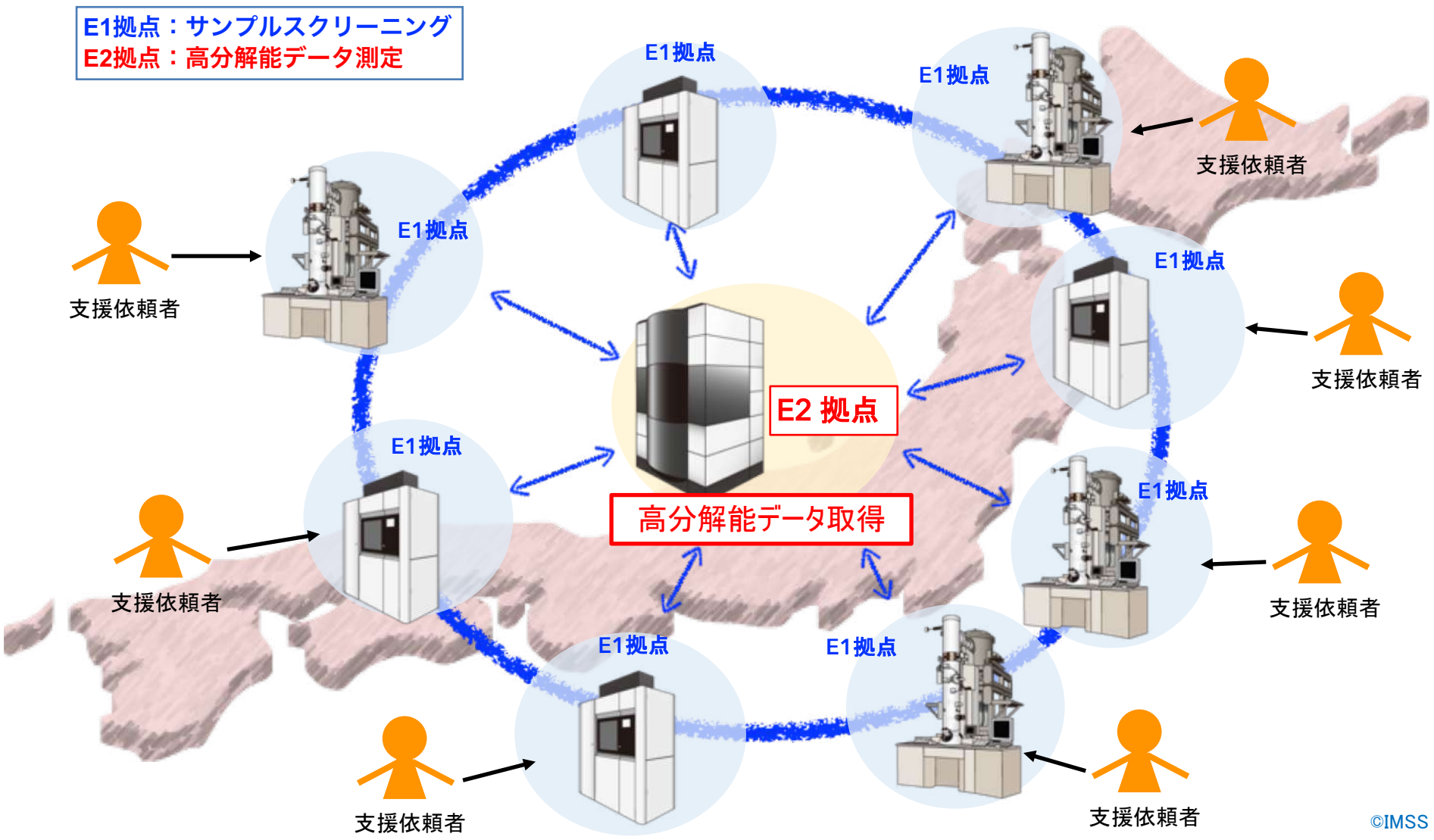
現状では１構造17万円の試算（400 x 17万円 = 6,800万円）
５万円/構造までは可能（400 x 5万円 = 2,000万円）

KEKよる高度化

2万円/構造が目標（400 x 2万円 = 800万円）

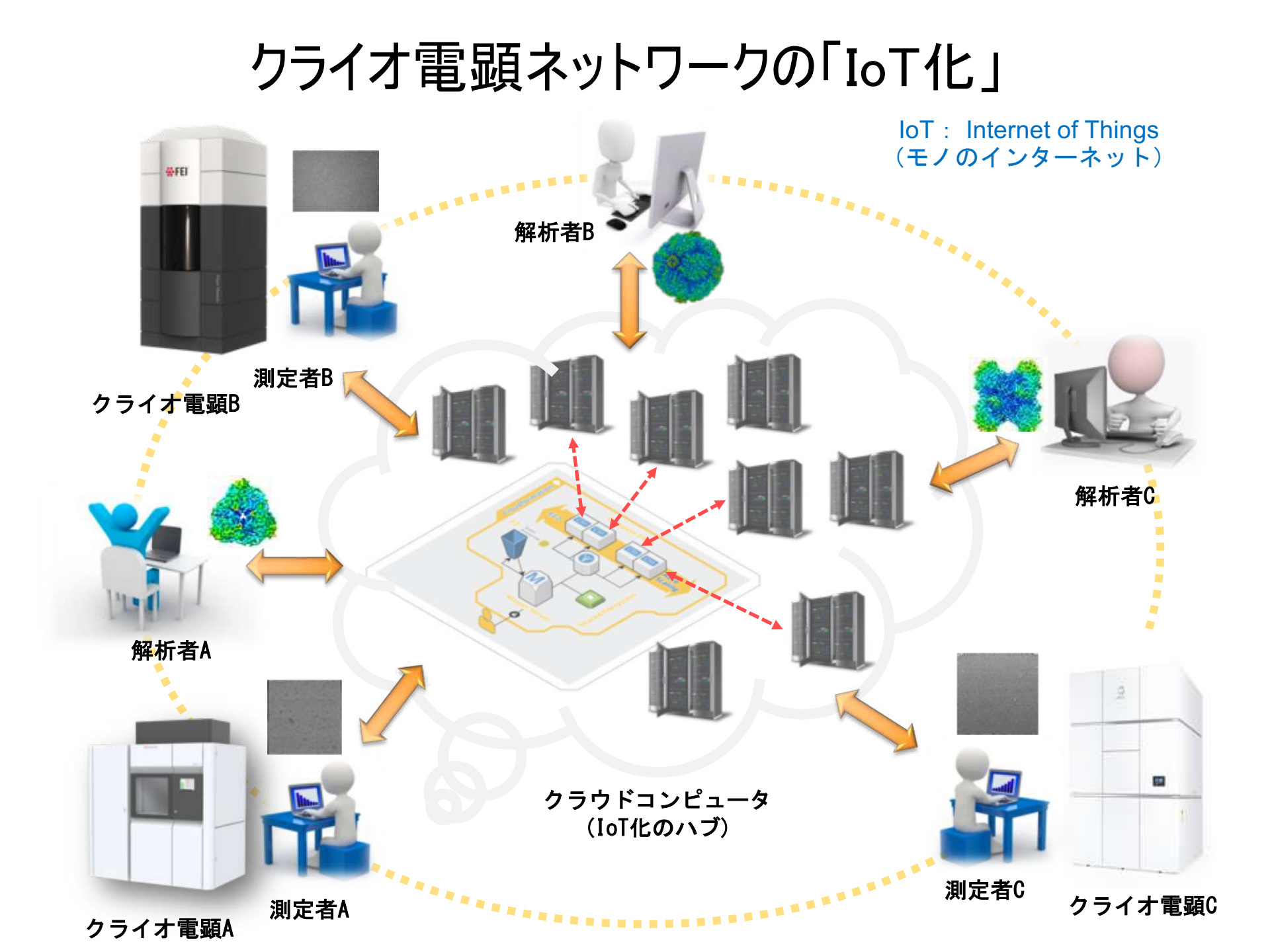


クライオ電子顕微鏡共用ネットワーク



クライオ電顕ネットワークの「IoT化」

IoT： Internet of Things
（モノのインターネット）



IoT化のハブとしてクラウドを導入する理由

現状では不可能な大規模な計算を可能に！

- 必要な時だけ台数を増して、処理期間を一気に短縮したい！
  - 化合物スクリーニング
  - プロテインダイナミクス
  - パラメータ最適化
- 講習会や勉強会などの実習で多数のユーザーが同時に使用できるようにしたい！

ユーザーサポートの促進！

- 研究室に計算機のないユーザーが、KEKが用意したソフトウェアインストール済みの解析計算環境をそのまま即座に使用できるようにしたい！
- KEK主催のRELION初心者講習会で学んだ最適な並列計算設定が、ユーザーが自分のデータセットの解析する時でも、そのまま使えるようにしたい！

産学連携の促進！

- クライオ電顕・単粒子解析のシステム開発とサービス化は、産業界へ引き継ぐ時期に入っているので、産学連携を促進してスムースな引継ぎを実現するための共通プラットフォームが欲しい！

GoToCloudシステムの開発とその構成

山田悠介
(KEK)

Nextcloud Download
（ファイルサーバー）

ユーザーによる
Movieのダウンロード

注記：オン・ザ・フライ処理はAWSに入れることも簡単にできる。しかし、KEKではこちらの構成の方が現実的で柔軟性が高い！

守屋俊夫
(KEK)

山本美里
(KEK)

MXデータを自動転送
（担当：山田）

Movieを自動転送
（担当：山田）

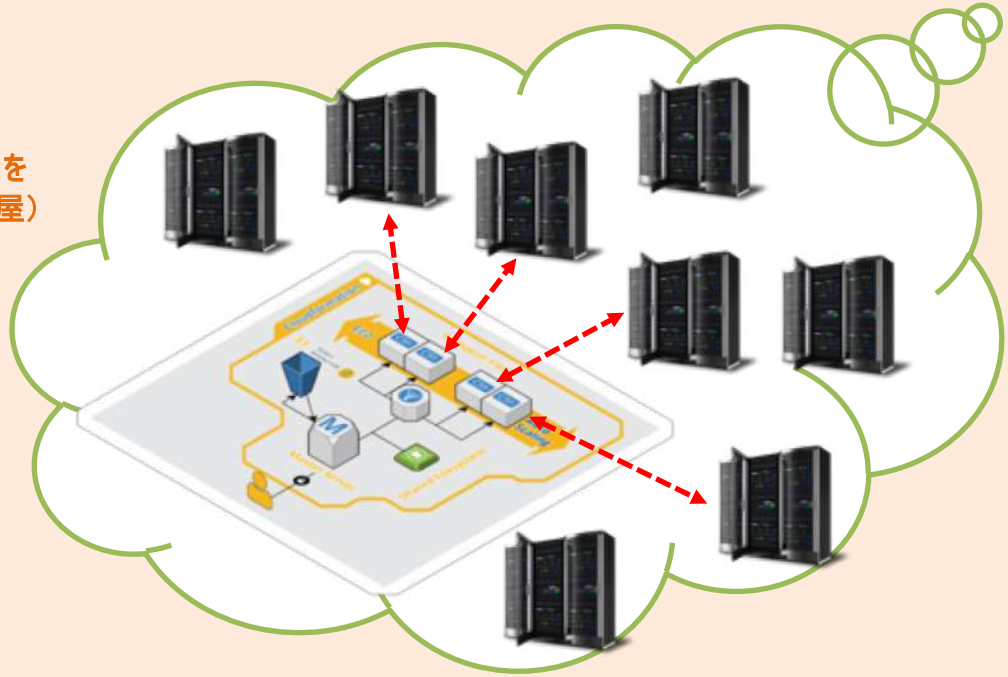
ユーザーによる
解析出力のダウンロード

GPFSシステム@KEK

[1] オン・ザ・フライ処理@KEK

Movie及び解析出力を
自動転送（担当：守屋）

[2] GoToCloud@AWSクラウド



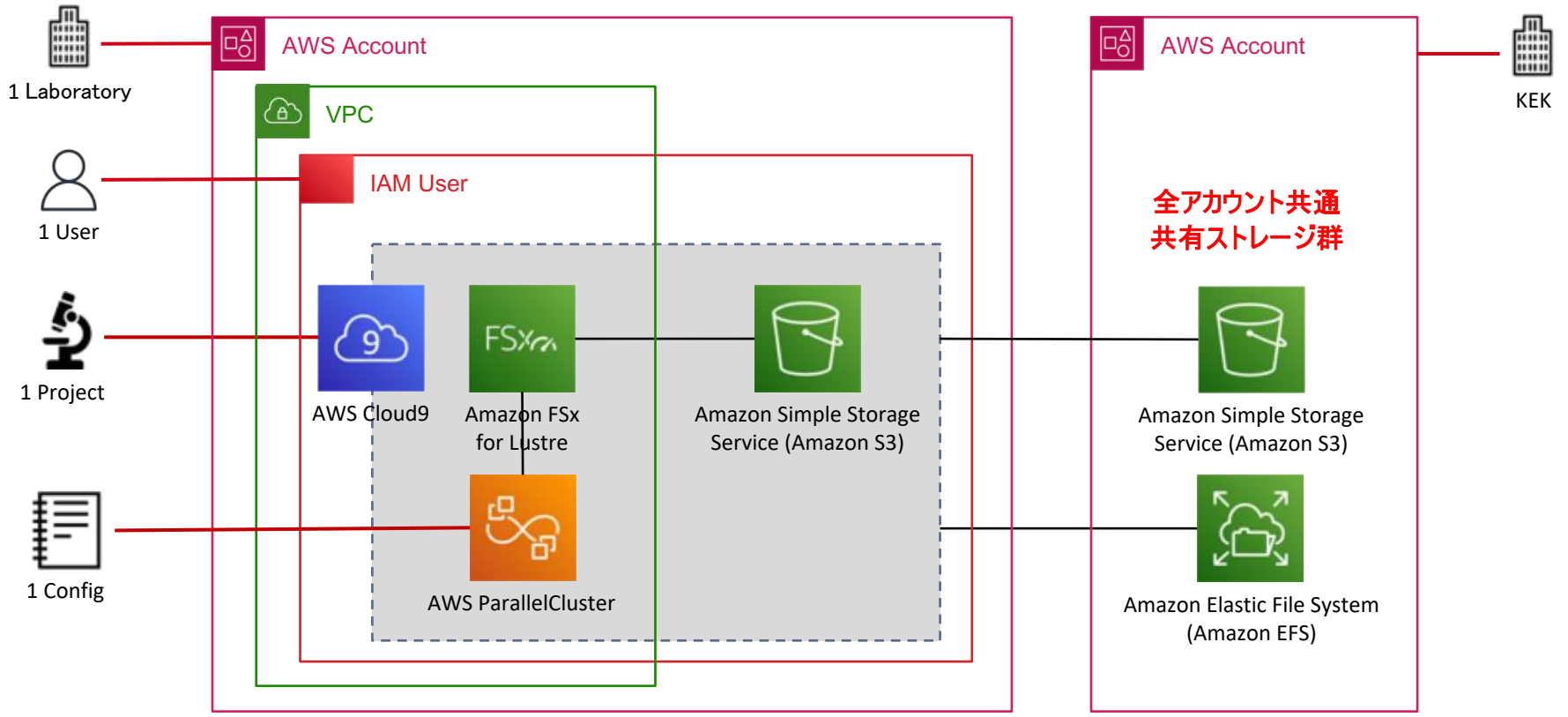
MXデータを自動転送
（担当：山田）

Movieを自動転送
（担当：守屋）

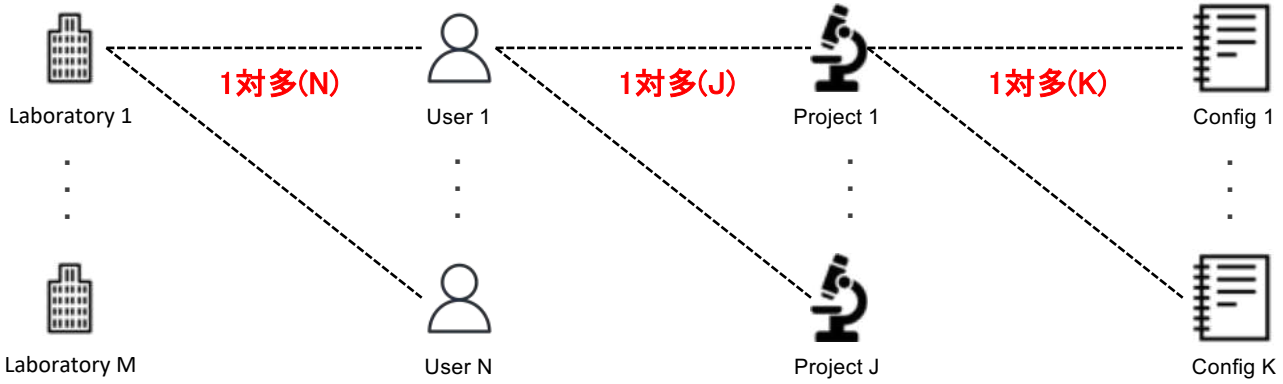
MX-BLs

クライオ電子顕微鏡

GoToCloudシステムマップ



研究室（組織）ごとにAWSアカウントを分けることにより、AWSで既に提供されている堅固なセキュリーティサービスを容易に利用できる！



GoToCloudハンズオンセッション とモニターユーザーテスト

第1回ハンズオン 2021年7月19日
企業ユーザー向け 3社参加
モニター終了 2021年8月31日

第2回ハンズオン 2021年9月15日
アカデミアユーザー向け 3研究室参加
モニター終了 2021年10月31日

GoToCloudのAWS ParallelCluster 作成手順

Step 1: Cloud9上でGoToCloud環境のセットアップ

- GoToCloud共有のAmazon EFSディスクのマウント
- ParallelClusterのインストール
- S3バケットの作成
- EC2のキーペアの作成
- AWS ParallelClusterのconfigファイルの生成

$ gtc_setup_environment_and_ec2_s3_create.sh

Step 2: AWS ParallelCluster インスタンスの作成

- ParallelClusterインスタンスの作成
- インスタンスへの解析ソフトのインストールとセットアップ

$ gtc_pcluster_create.sh

RELION 3.0, CTFFIND4, (gCTF), UCFS Chimera, UCFS ChimeraX

Step 3: AWS ParallelCluster マスターノードへの接続

- SSH接続

または

- NICE DCV接続

$ gtc_pcluster_ssh.sh

$ gtc_pcluster_dcv_connect.sh

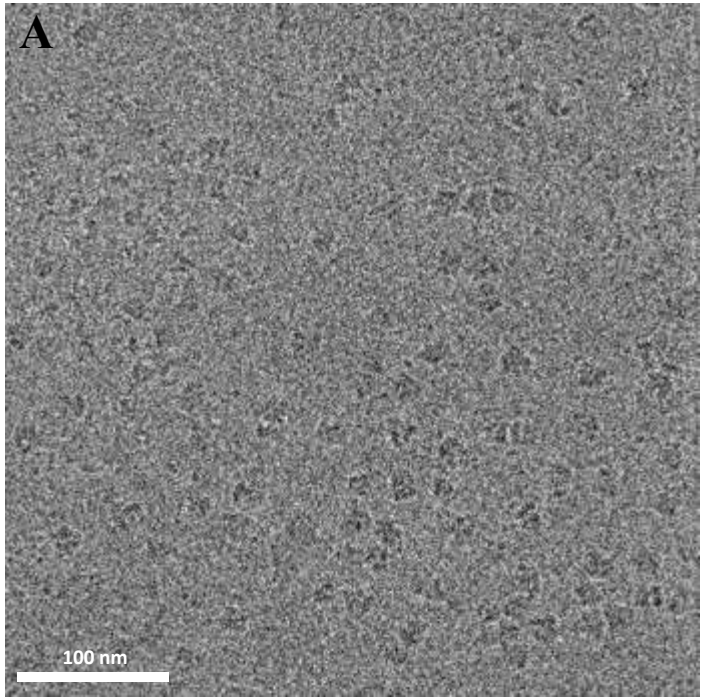
Step 4: AWS ParallelCluster インスタンスの削除

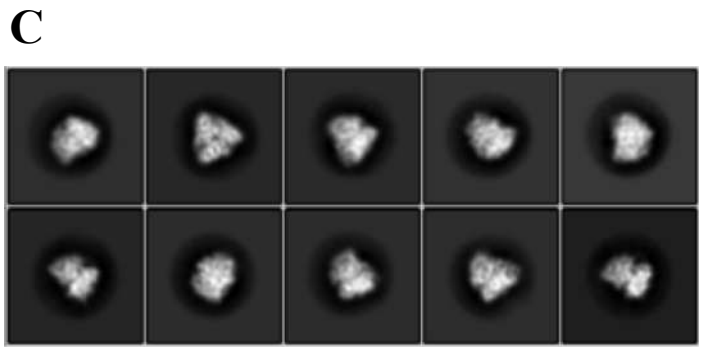
- 解析データのs3バケットへのエクスポート
- ParallelClusterインスタンスの削除

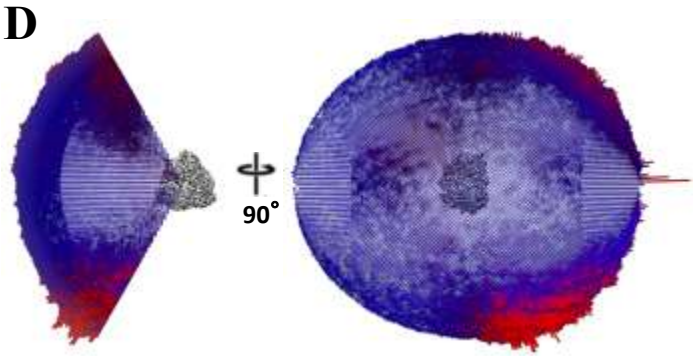
$ gtc_pcluster_delete.sh

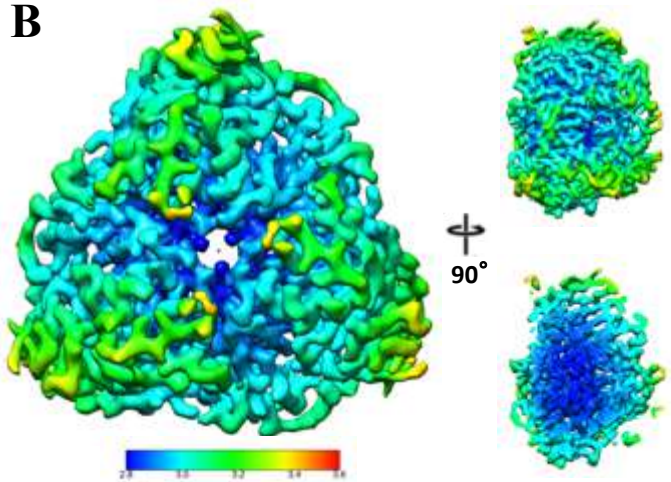
現バージョンではたった3（+1）ステップでデータ解析を開始できる！

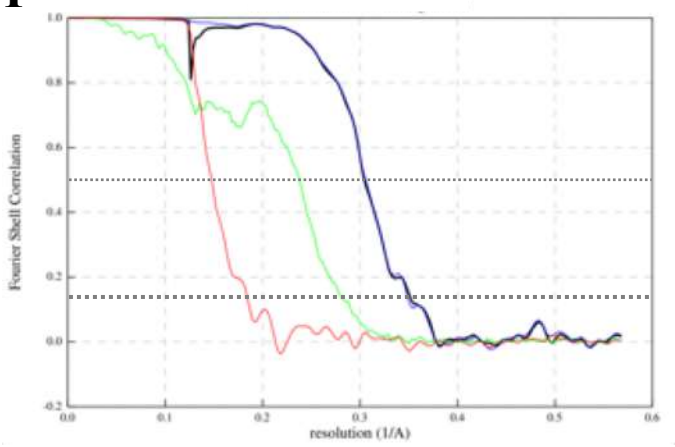
Native Nitrite Reductase (NiR) ベンチマークテスト
110 KDa, C3, 分解能2.85 Å











茨城大学
Ibaraki University

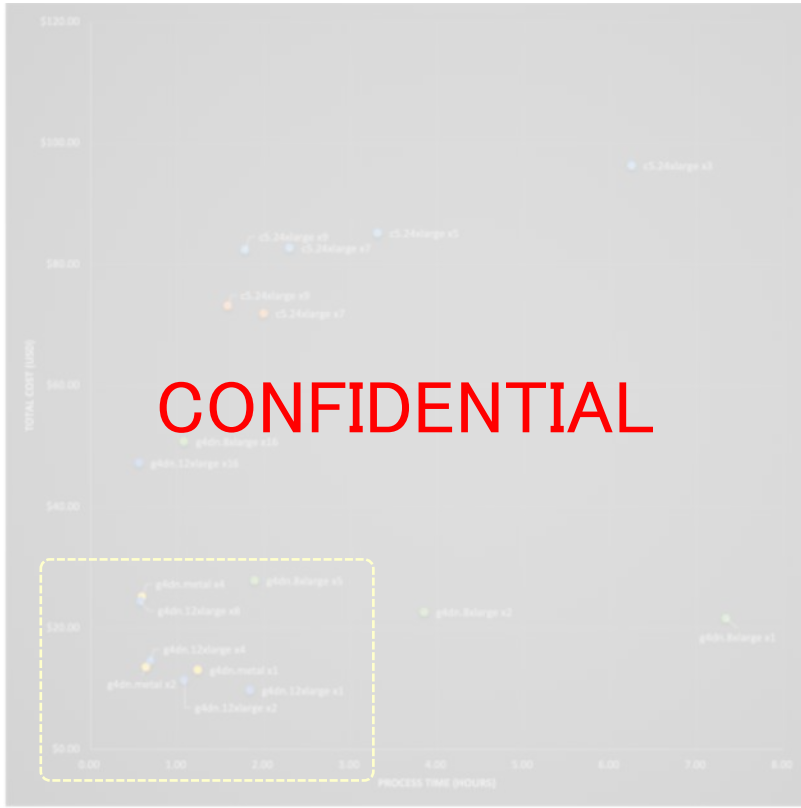
Dr. Kozuma
高妻 孝光 先生

Dr. Yamaguchi
山口 峻英 先生

Refine3Dベンチマーク

EMPIAR-10581
CryoEM map and model of Nitrite Reductase at pH 8.1

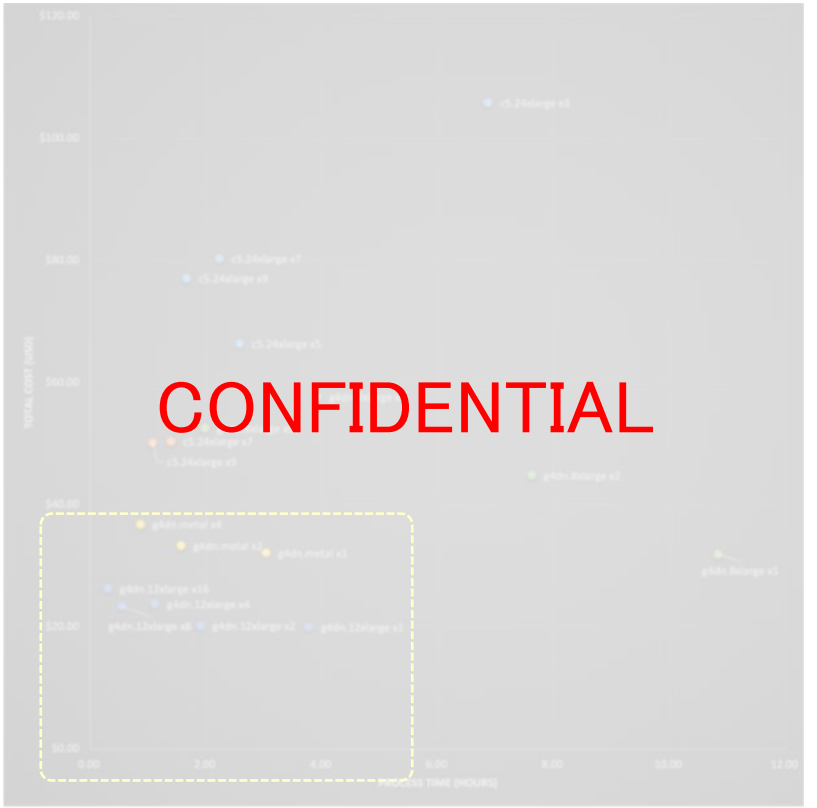
Rescaled Box [Pix]	: 352	Particles	: 129,298
Rescaled Apix [A/Pix]	: 1.17	Symmetry	: C3
Original Box [Pix]	: 468	Mask Diameter [A]	: 164
Original Apix [A/Pix]	: 0.88		



Class2Dベンチマーク

EMPIAR-10581
CryoEM map and model of Nitrite Reductase at pH 8.1

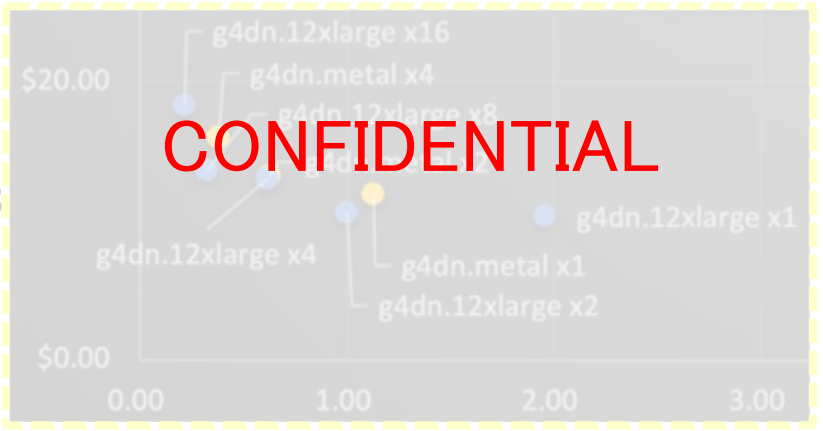
Rescaled Box [Pix]	: 64	Particles	: 176,256
Rescaled Apix [A/Pix]	: 3.3	Classes	: 200
Original Box [Pix]	: 240	Use fast subsets	: No
Original Apix [A/Pix]	: 0.88	Mask Diameter [A]	: 120



Calss3Dベンチマーク

EMPIAR-10581
CryoEM map and model of Nitrite Reductase at pH 8.1

Rescaled Box [Pix]	: 64	Particles	: 129,298
Rescaled Apix [A/Pix]	: 3.3	Symmetry	: C1
Original Box [Pix]	: 240	Classes	: 4
Original Apix [A/Pix]	: 0.88	User fast subsets?	: No
		Mask Diameter [A]	: 120



CONFIDENTIAL

CONFIDENTIAL

CONFIDENTIAL

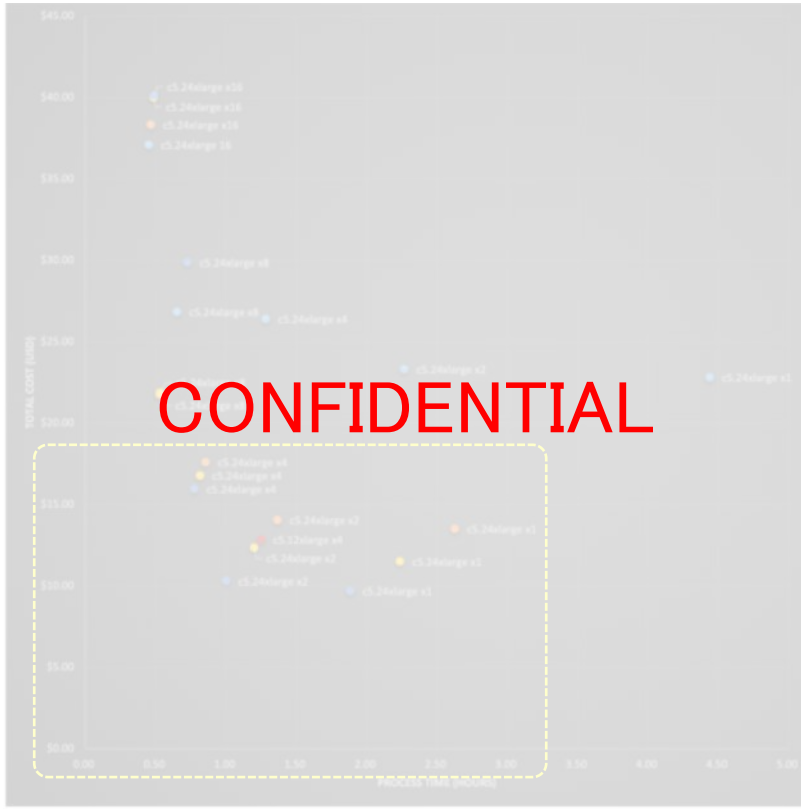
Polishベンチマーク

EMPIAR-10581
CryoEM map and model of Nitrite Reductase at pH 8.1

Rescaled Box [Pix]	: 352	Particles	: 129,298
Rescaled Apix [A/Pix]	: 1.17		
Original Box [Pix]	: 468		
Original Apix [A/Pix]	: 0.88		

CONFIDENTIAL

CONFIDENTIAL


CONFIDENTIAL

コストダウンの方策～処理時間に影響する要因

撮影枚数	倍率（ピクセルサイズ）、撮影サイズ、フレーム数はいつも同じにする。最大撮影枚数を制限する（e.g. 1,000-2,000）
ピック粒子数	ピック粒子数を制限する？（e.g. 100K - 200K）
ボックスサイズ	粒子サイズが大きい場合（直径200[A]~350[A]以上）は、それに合わせて目標分解能を調整して、ボックスサイズをいつも同じにする？ 最大デフォーカス値がいつも同じにする？（e.g. 2.0 or 3.0 um）
対称性	C1処理をいつも行い、最後だけ対称性を使うシナリで固定料金？ それとも対称性に合わせて変動料金設定？
構造状態の数	最も高い分解能が得られるはずのメジャーな状態の構造一つだけを再構築。
目標分解能	サイズに合わせて目標分解能を設定？通常は3.0-3.5[A]で良い？

今後の計画～GoToCloud環境の拡張

KEK
IMSS PF SBRC


山田悠介
(KEK)


山本美里
(KEK)


守屋俊夫
(KEK)

AMED
TCEF

ファイルサーバー、外付け
USB HDD等

クライオ電子顕微鏡

自動的な
データ圧縮と
バックアップ
データ転送

測定データを
自動転送

[1] オン・ザ・フライ処理@KEK

フレーム間
ドリフト補正
↓
CTFパラメータ
推定
↓
粒子ピッキング
↓
２次元平均化

初期３次元構造
再構築
↓
３次元構造
クラス分け
↓
３次元精密化

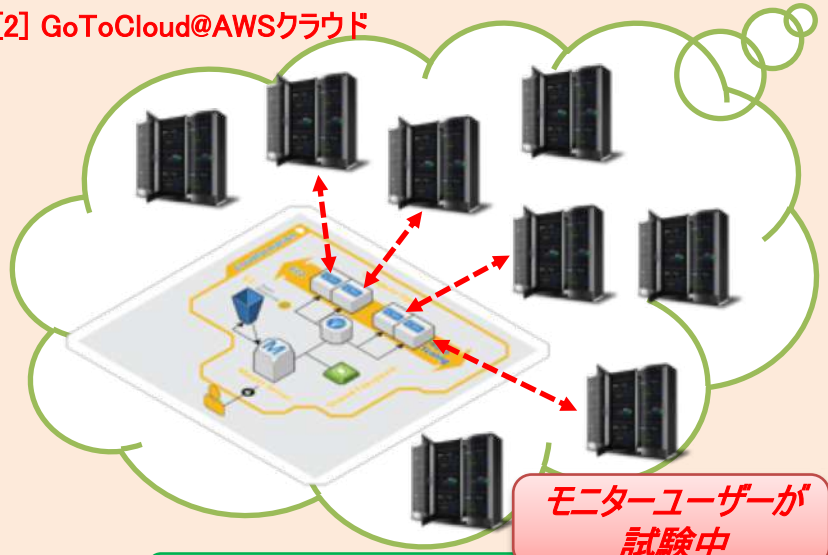
電顕像毎処理とサブセット処理

オン・ザ・フライ全自動処理で計測終了と
ほぼ同時に低～中分解能でコンセンサス
マップまでを再構築する。

入力データ及び解析結果
を自動転送

データ転送速度
KEK→AWS(Tokyo Region)
凡そ80～100MB/sec
(1TBで3時間弱)

[2] GoToCloud@AWSクラウド



***モニターユーザーが
試験中***

［Phase1］従来通りに人が解析

[*]「未知の構造」のデータセットの高分解能の構造解析は自動化が難しいのでは当面は人が解析計算を行う必要がある。

［Phase2］化合物スクリーニング
全自動解析処理

[*]阻害剤が結合しているホロ体構造解析をAWSで全自動解析！
[*]ターゲットタンパク質のアポ体構造が既知であることが前提！
[*]アポ体構造の高分解能化は人が行う。

［Phase3］プロティンダイナミクス
全自動解析処理

[*] 構造状態分離をAWSで全自動解析。
[*] 高分解能のコンセンサスマップの再構築までは当面は人が行う必要がある必要がある。

目的に適したデータ解析の全自動化をしたい！

GoToCloud環境で化合物スクリーニング

ファイルサーバー、外付け
USB HDD等

クライオ電子顕微鏡

自動的な
データ圧縮と
バックアップ
データ転送

測定データを
自動転送

[1] オン・ザ・フライ処理@KEK

フレーム間
ドリフト補正

CTFパラメータ
推定

粒子ピッキング

２次元平均化

初期３次元構造
再構築

３次元構造
クラス分け

３次元精密化

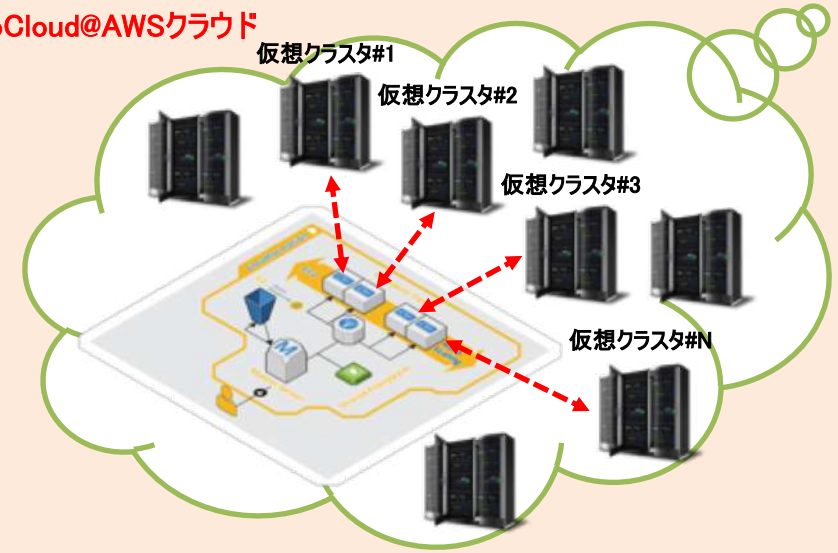
電顕像毎処理とサブセット処理

オン・ザ・フライ全自動処理で計測終了と
ほぼ同時に低～中分解能でコンセンサス
マップまでを再構築する。

入力データ及び解析結果
を自動転送

データ転送速度
KEK→AWS(Tokyo Region)
凡そ80～100MB/sec
（1TBで3時間弱）

[2] GoToCloud@AWSクラウド

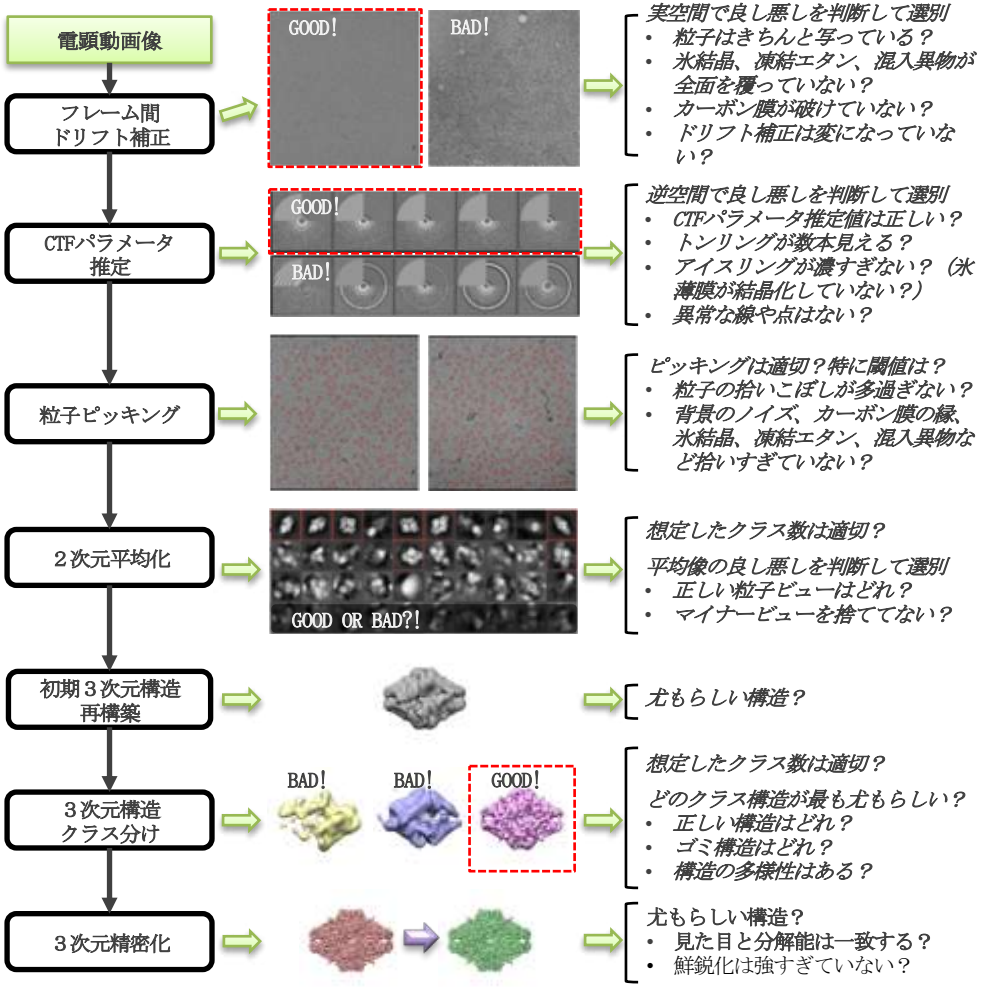


仮想クラスタ#1	[標的タンパク質 + 低分子化合物”A”」の解析計算
仮想クラスタ#2	[標的タンパク質 + 低分子化合物”B”]の解析計算
仮想クラスタ#3	[標的タンパク質 + 低分子化合物”C”]の解析計算
・・・	
仮想クラスタ#N	[標的タンパク質 + 低分子化合物”X”]の解析計算

全てのデータセットを並列処理で同時に実行！

創薬産業向け構造ベースドラックデザイン（SBDD）工程の効率化！

IoT化の有効活用による全自動化

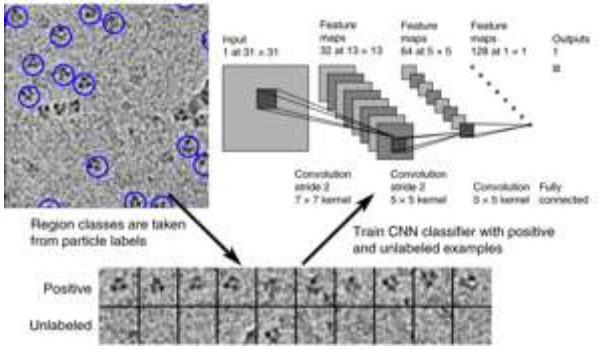


目標

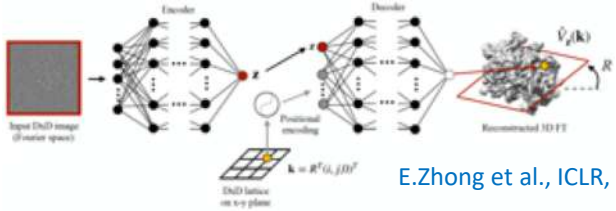
自動化を妨げている人の介入を完全に排除したい！

解決策

解析エキスパートの選別方法をAIによる“学習”でアルゴリズム化、これによって単粒子解析計算の完全自動化！ IoT化によって蓄積されるエキスパートが行った解析結果を教師データとして進化アルゴリズム的な競争による高度AIを育成し、解析技術の高度化を図る。“経験”が増えれば増えるほど、高度なAIの育成が可能となる。



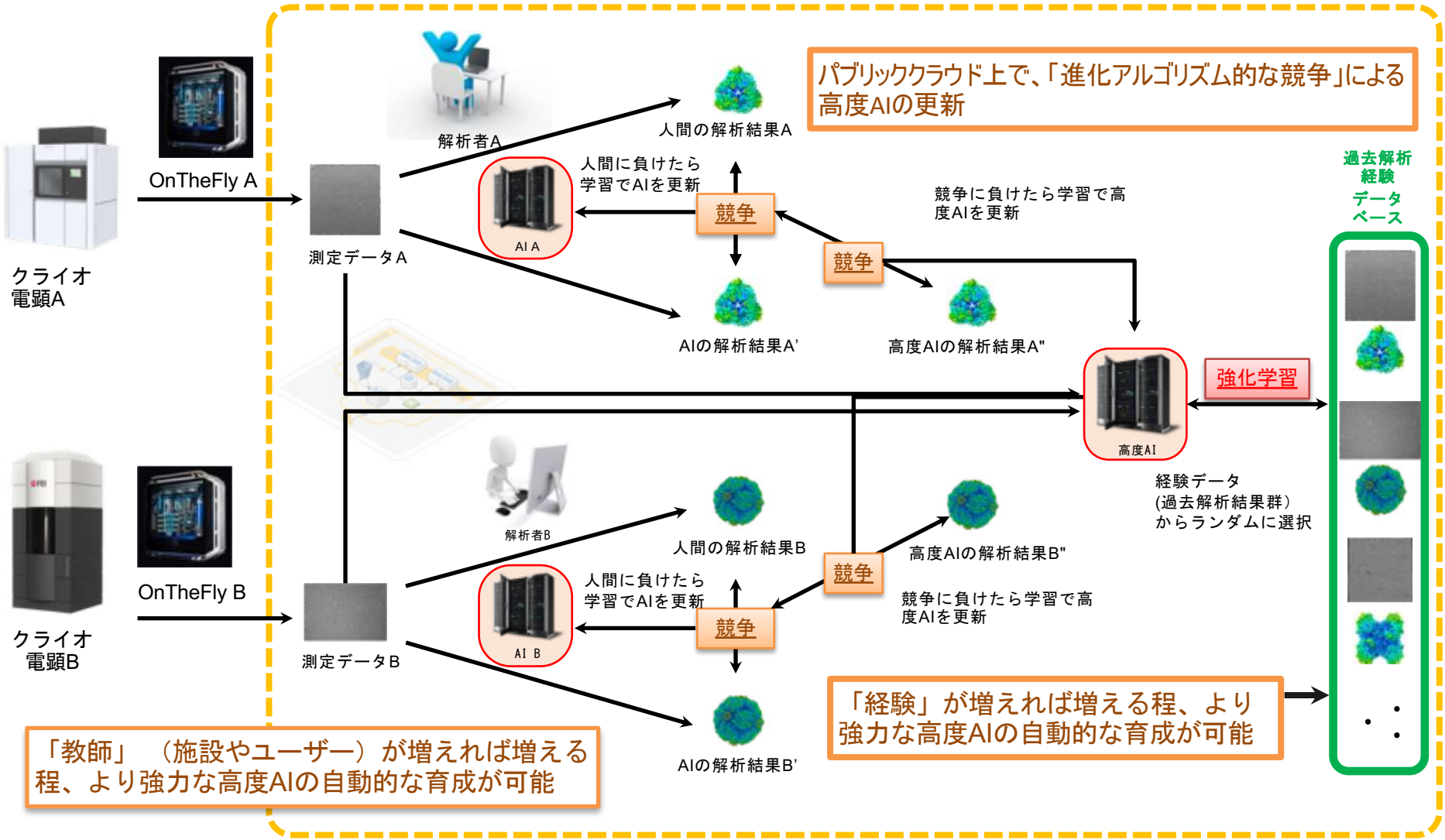
T. Wagner et al., Commun Biol, 2019



E.Zhong et al., ICLR, 2020

全自動化により創薬とバイオ分野の発展を加速！

IoTを超えて



IoT化、深層学習、進化アルゴリズム的な競争を組合せて高度AIを育成！
データ解析の永続的かつ自律的な高度化の仕組みを確立！

まとめ

- 単粒子クライオ電子顕微鏡法の概要
  - 画像処理ヘビーなタンパク質立体構造決定手法
- GoToCloudプロジェクト@KEK
  - クライオ電顕ネットワークのIoT化
  - AWS ParallelClusterをIoT化のハブとして利用
  - ベンチマークでベストコストパフォーマンスを模索
- 今後のGoToCloud環境の拡張計画
  - GoToCloud環境で化合物スクリーニングを全自動化

謝辞（敬称略）

高エネルギー加速器研究機構（KEK）
物質構造科学研究所（IMSS）
構造生物学研究センター（SBRC）

アマゾンウェブサービスジャパン（株）
田代 皓嗣
宮本 大輔
片岡 勇人

センター長
千田 俊哉

GoToCloudプロジェクト
山田 悠介
山本 美里
モニターユーザーの皆様

クライオ電顕チーム
安達 成彦
川崎 政人
池田 聡人
荒牧 慎二（TVIPS）

解析計算環境整備
篠田 晃
久保田 孝幸

秘書
増田 千穂
鮭川 理恵子

SBRCメンバの皆様