aws FOR HEALTHCARE AND LIFE SCIENCES

CBI学会2023年大会 SS07-03

クラウドと生成系AIを活用した創薬研究：タンパク質構造予測を例に

石尾 千晶

ソリューション アーキテクト
アマゾン ウェブ サービス ジャパン合同会社

自己紹介

石尾 千晶 (Chiaki Ishio)

アマゾンウェブサービスジャパン合同会社
ソリューション アーキテクト

主に製薬企業のお客様を担当し、クラウド活用の
技術支援をおこなっています

創薬領域における生成系 AI のイノベーション

汎用的な生成系 AI を活用した
研究活動の生産性向上

創薬ドメインに特化した
生成系 AI を活用した
解析やドラッグデザインの高度化

創薬領域における生成系 AI のイノベーション

汎用的な生成系 AI を活用した
研究活動の生産性向上

創薬ドメインに特化した
生成系 AI を活用した
解析やドラッグデザインの高度化

このセッションでお伝えしたいこと

- 創薬研究の質をより一層向上させるために、クラウドと生成系 AI を活用できること
- タンパク質構造解析の領域で、用途に応じたツールがあること
- 生成系 AI 活用のために、専任の支援チームがいること

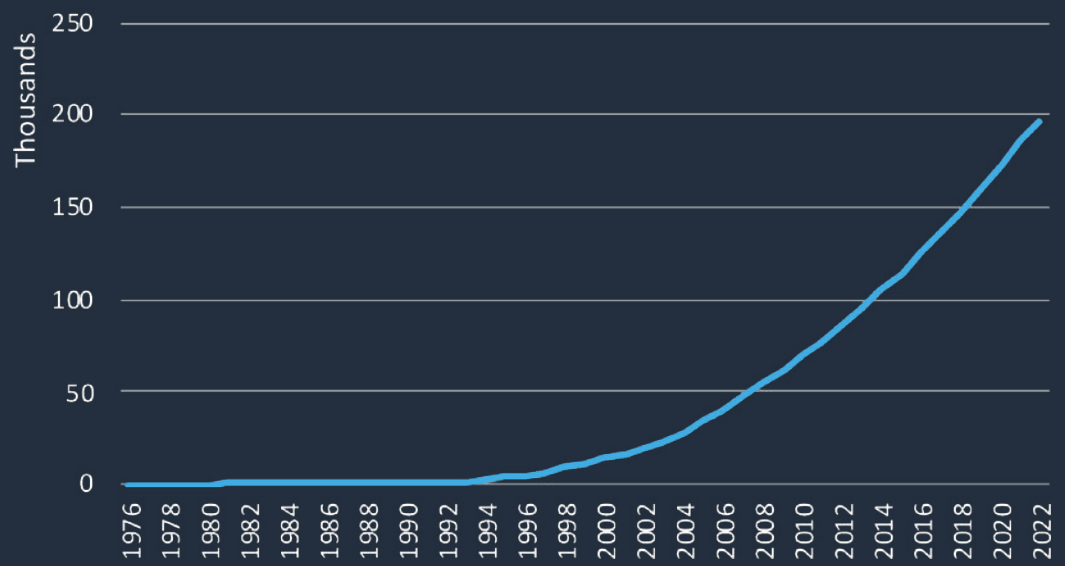
タンパク質構造解析を取り巻く現状

既知のタンパク質のうち、実験的に決定された構造を持つものはわずか0.1%

UniProtKB/TrEMBL への登録数

約2億

PDB への登録数

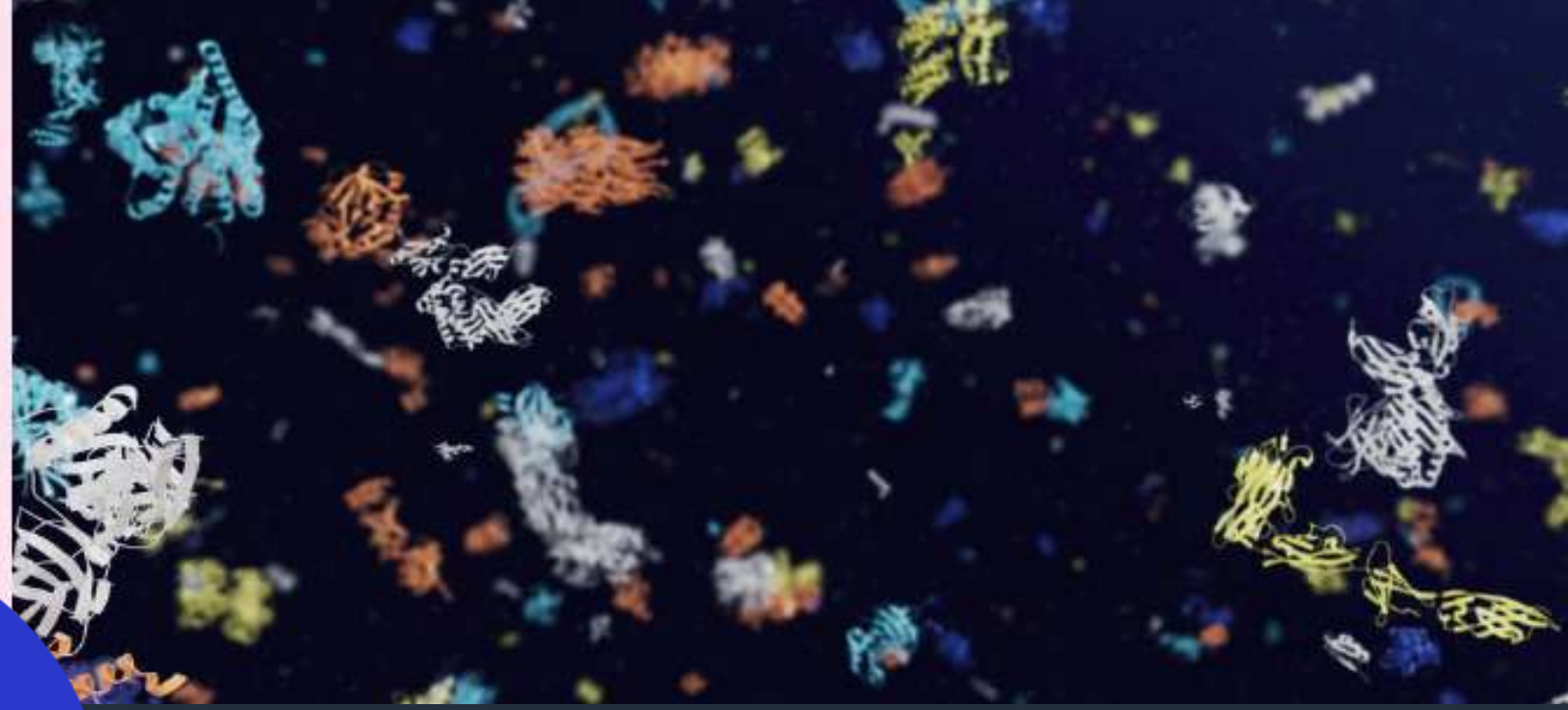


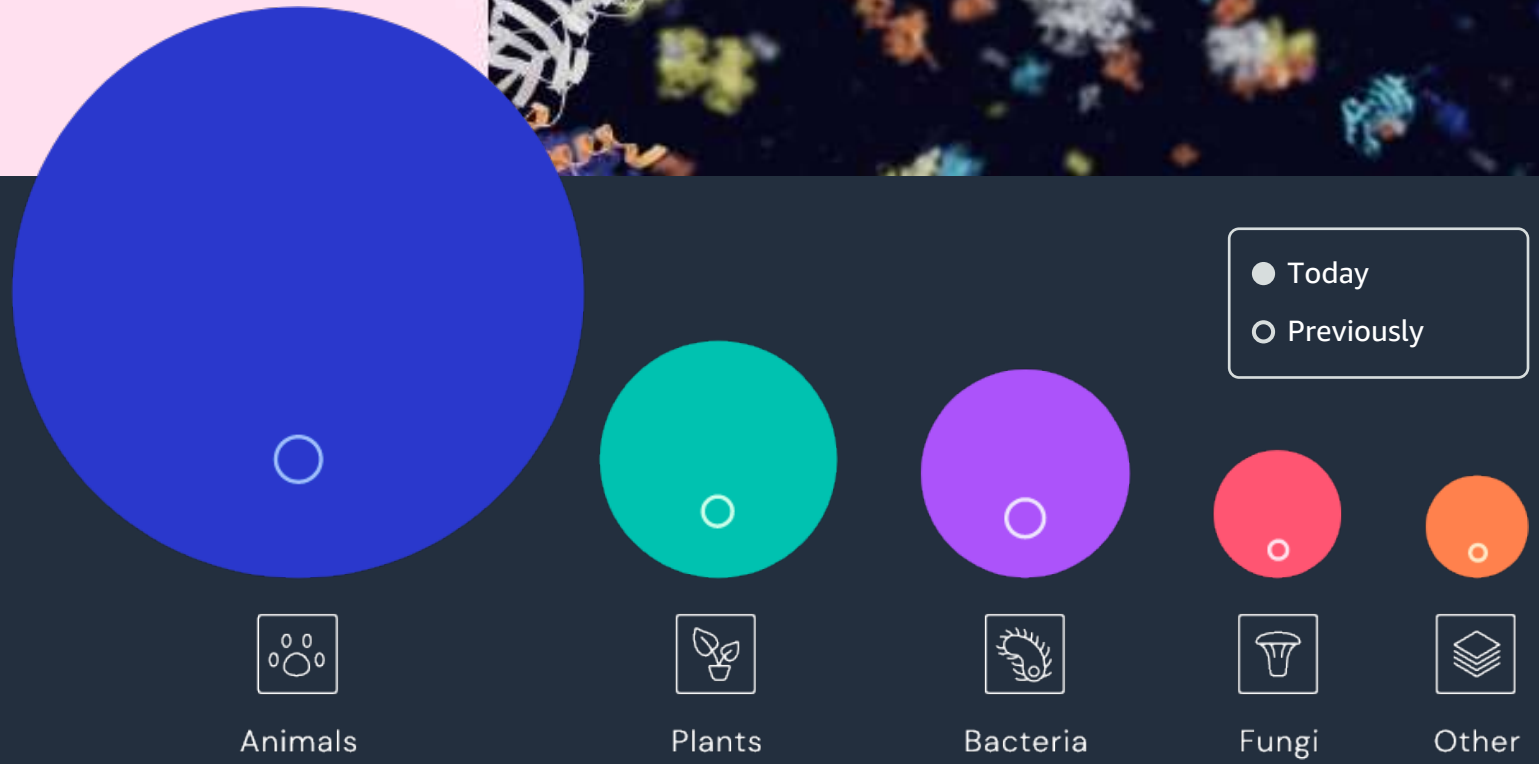
約20万

Research

AlphaFold reveals the structure of the protein universe

July 28, 2022





https://www.deepmind.com/blog/alphafold-reveals-the-structure-of-the-protein-universe

nature methods

Explore content | About the journal | Publish with us

nature > nature methods > brief communications > article

Brief Communication | Open Access | Published: 30 May 2022

ColabFold: making protein folding accessible t

Milot Mirdita, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinni
Steinegger

Nature Methods **19**, 679–682 (2022) | Cite this article

High-resolution *de novo* structure prediction from primary sequence

Ruidong Wu, Fan Ding, Rui Wang, Rui Shen, Xiwen Zhang, Shitong Luo, Chenpeng Su, Zuofan Wu, Qi Xie, Bonnie Berger, Jianzhu Ma, Jian Peng

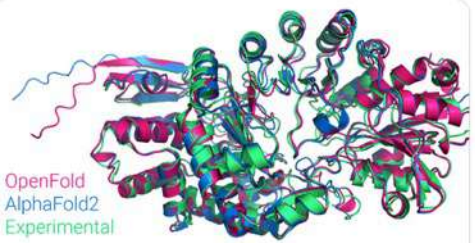
doi: https://doi.org/10.1101/2022.07.21.500999

Uni-Fold: An Open-Source Platform for Devel
Protein Folding Models beyond AlphaFol

Ziyao Li[a,b,*], Xuyang Liu[a,c,*], Weijie Chen[a,b,*], Fan Shen[a], Hangrui Bi[a]
Guolin Ke[a,†] and Linfeng Zhang[a,d,†]
[a]DP Technology
[b]Center for Data Science, Peking University
[c]School of Mathematical Sciences, Peking University
[d]Institute, Beijing

Mohammed AlQuraishi
@MoAlQuraishi · Follow

We have successfully trained OpenFold from scratch, our trainable PyTorch implementation of AlphaFold2. The new OpenFold (OF) (slightly) outperforms AlphaFold2 (AF2). I believe this is the first publicly available reproduction of AF2. We learned a lot. A 1/12



github.com
OpenFold GitHub Repository

HelixFold: An Efficient Implementation of AlphaFold2 using PaddlePaddle

Guoxia Wang, Xiaomin Fang, Zhihua Wu
Yiqun Liu, Yang Xue, Yingfei Xiang
Dianhai Yu, Fan Wang, Yanjun Ma

Baidu Inc.

Accurate prediction of nucleic acid and protein-nucleic acid complexes using RoseTTAFoldNA

Minkyung Baek, Ryan McHugh, Ivan Anishchenko, David Baker, Frank DiMaio

doi: https://doi.org/10.1101/2022.09.09.507333

nature biotechnology

https://doi.org/10.1038/s41587-022-01432-w

Check for updates

Single-sequence protein structure prediction using a language model and deep learning

Ratul Chowdhury[1,8], Nazim Bouatta[1,8] ✉, Surojit Biswas[2,3,8], Christina Floristean[4,8], Anant Kharkare[4], Koushik Roye[4], Charlotte Rochereau[5], Gustaf Ahdritz[6], Joanna Zhang[4], George M. Church[1,2], Peter K. Sorger[1,7] ✉ and Mohammed AlQuraishi[4,6] ✉

医薬品開発の現場では…

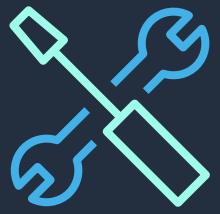
薬価改定による
価格の引き下げ

パテントクリフ

個別化医療への
期待の高まり

研究者の方々のニーズ

構造予測・相同性検索・
ドッキング・タンパク質設計など
次々と開発されるアルゴリズムを
手軽に試したい



使い慣れたインターフェースで
分析を素早く開始したい

独自のデータを分析するための
セキュアな環境

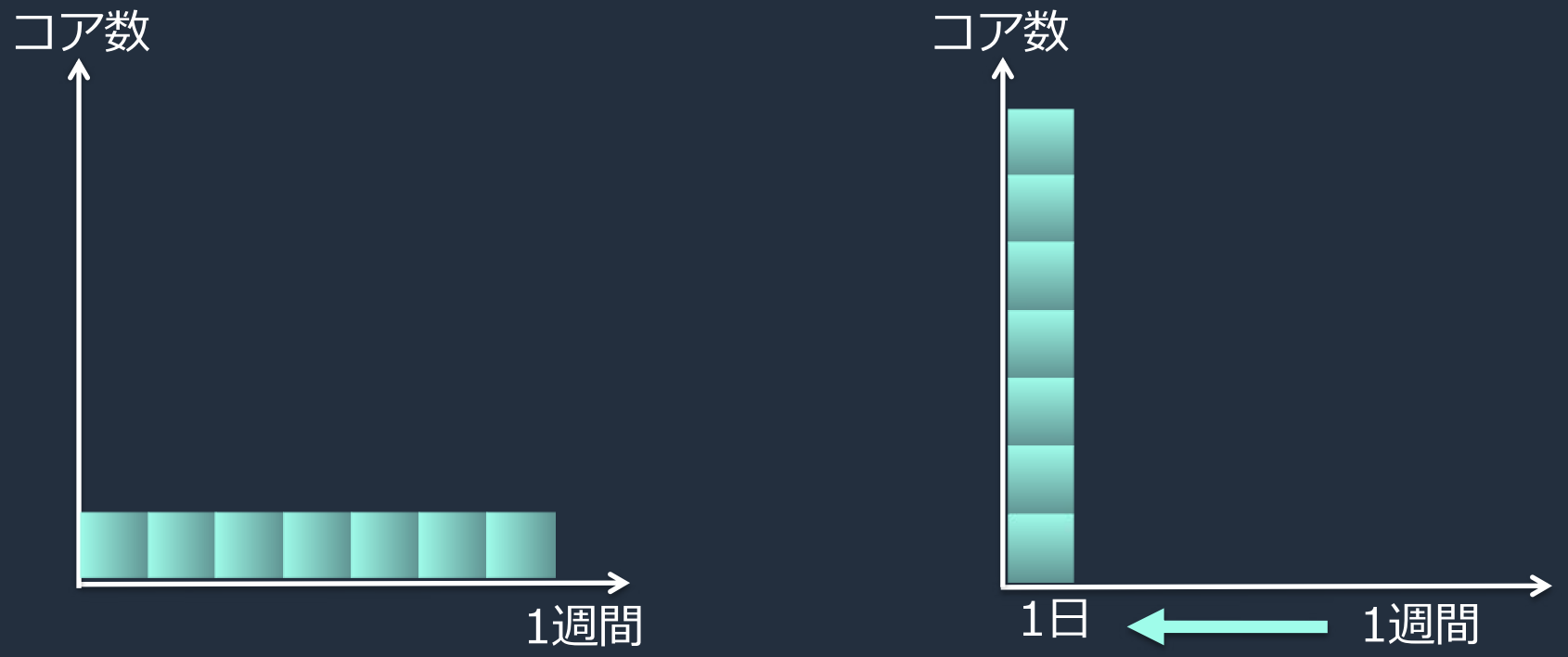
必要に応じて
スケールする環境

大規模なタンパク質ターゲットを
分析するための**強力な計算環境**

幅広い研究チームを
支援するための**コスト最適化**

AWS でタンパク質構造解析を始めるためのツール

クラウド創薬の魅力：
スケーラビリティーをいかして計算時間を短縮



従来は手持ちの限られたリソースで、逐次処理していた計算も
AWSなら必要な台数、インスタンスを起動して、一斉処理。
しかも費用は「時間×台数」なのでどちらも同じ。

使い慣れたインターフェース + スケーラブルな実行環境

ウェブアプリ

Jupyter Notebook

AWS コンソール

AWS CLI

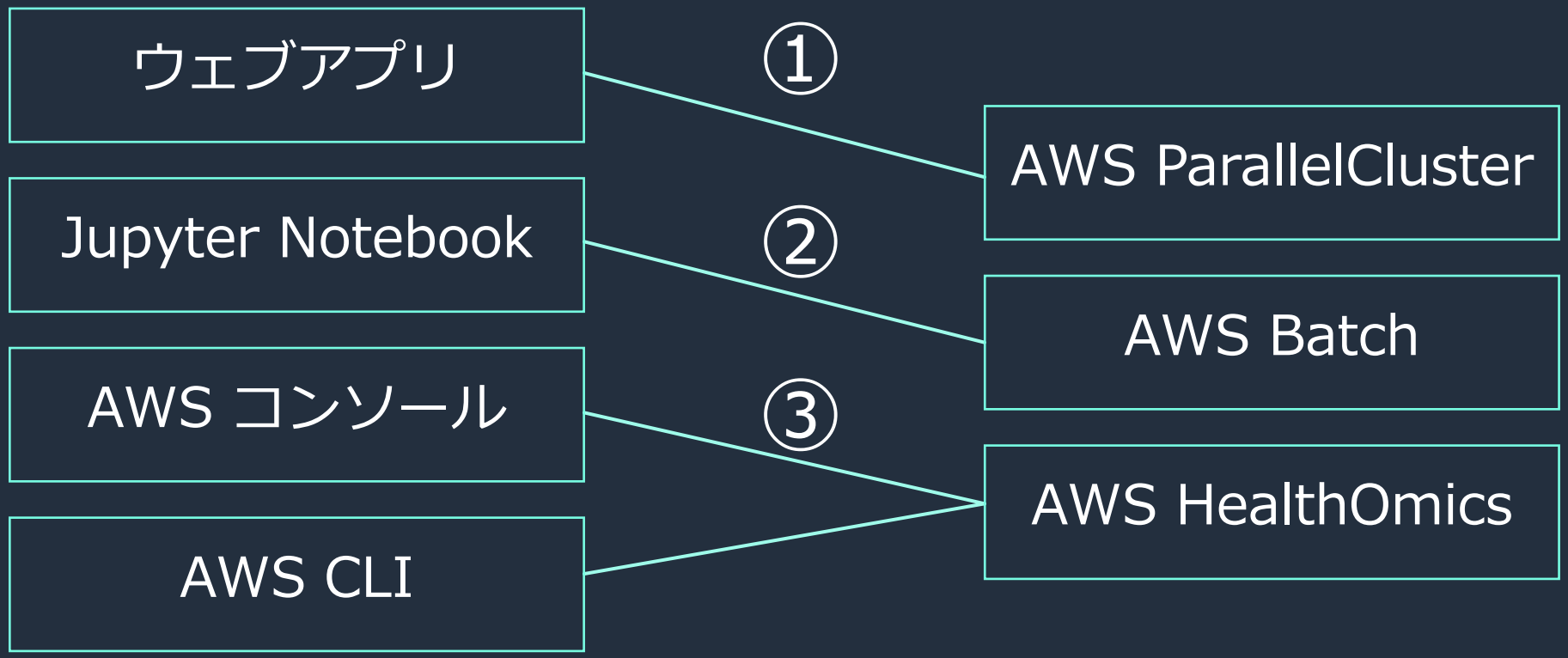
+

AWS ParallelCluster

AWS Batch

AWS HealthOmics

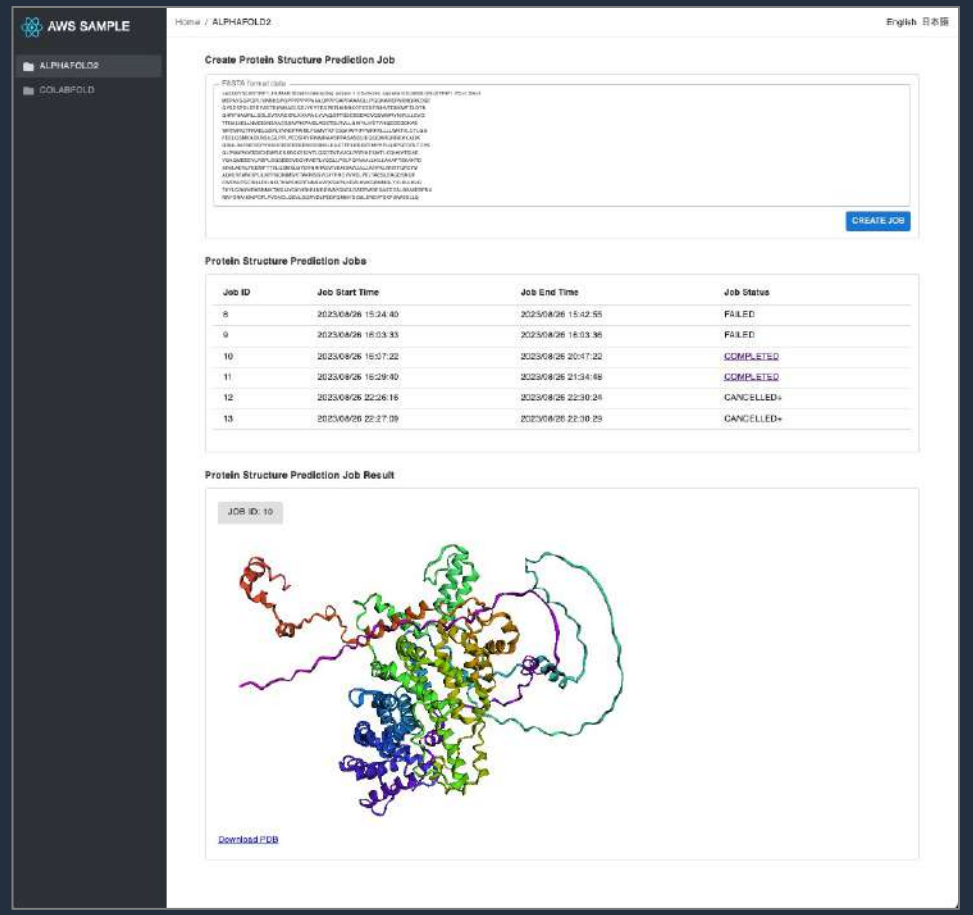
使い慣れたインターフェース + スケーラブルな実行環境



① ウェブアプリ + AWS ParallelCluster

- タンパク質構造予測を手軽に実行できるウェブアプリ
  - オープンソースの実装例として提供
  - AlphaFold2 と ColabFold に対応
  - ジョブの投入・停止、結果の可視化

- 研究者は、AWS コンソールへのアクセス権限を持っていなくても簡単に使える
- IT 管理者は、起動テンプレートを使うことでほぼ自動でウェブアプリを構築可能



- GitHub – AWS Samples ”AlphaFold2 Webapp on AWS”
- AWS ブログ「大規模なタンパク質構造予測を行うための研究者向けウェブアプリを構築するには」

デモ動画

aws-sample

https://d5kvhafml2maf.cloudfront.net/alphafold2#11

AWS SAMPLE

Home / ALPHAFOLD2

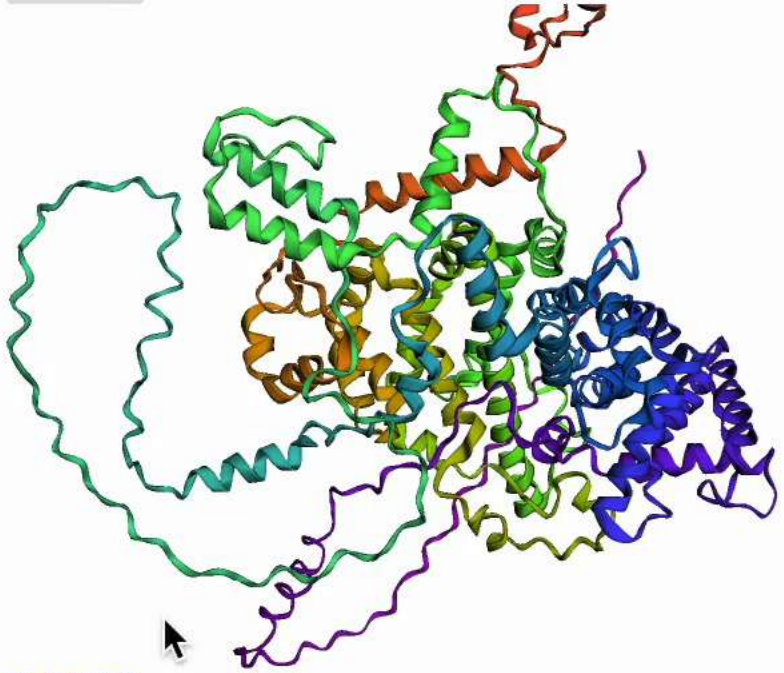
English 日本語

- ALPHAFOLD2
- COLABFOLD

8	2023/08/26 15:24:40	2023/08/26 15:42:55	FAILED	
9	2023/08/26 16:03:33	2023/08/26 16:03:36	FAILED	
10	2023/08/26 16:07:22	2023/08/26 20:47:22	COMPLETED	
11	2023/08/26 16:29:40	2023/08/26 21:34:48	COMPLETED	
12	2023/08/26 22:26:16		RUNNING	TERMINATE JOB
13	2023/08/26 22:27:09		RUNNING	TERMINATE JOB

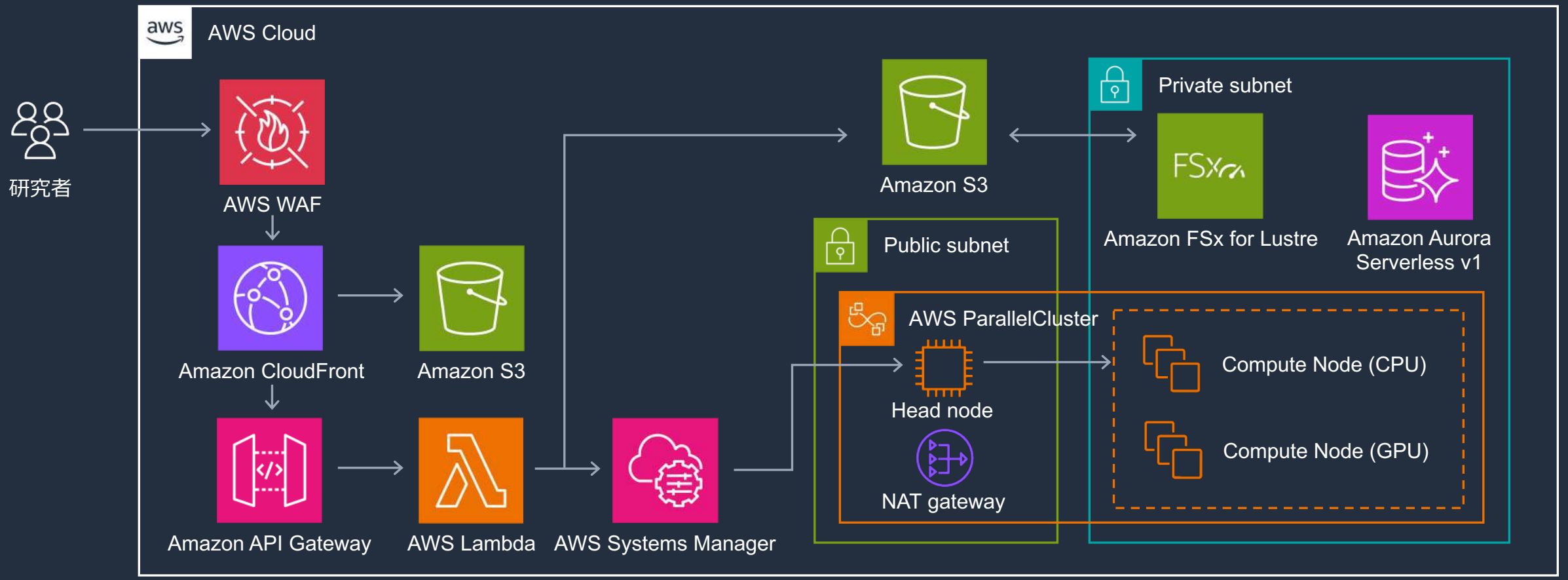
Protein Structure Prediction Job Result

JOB ID: 11

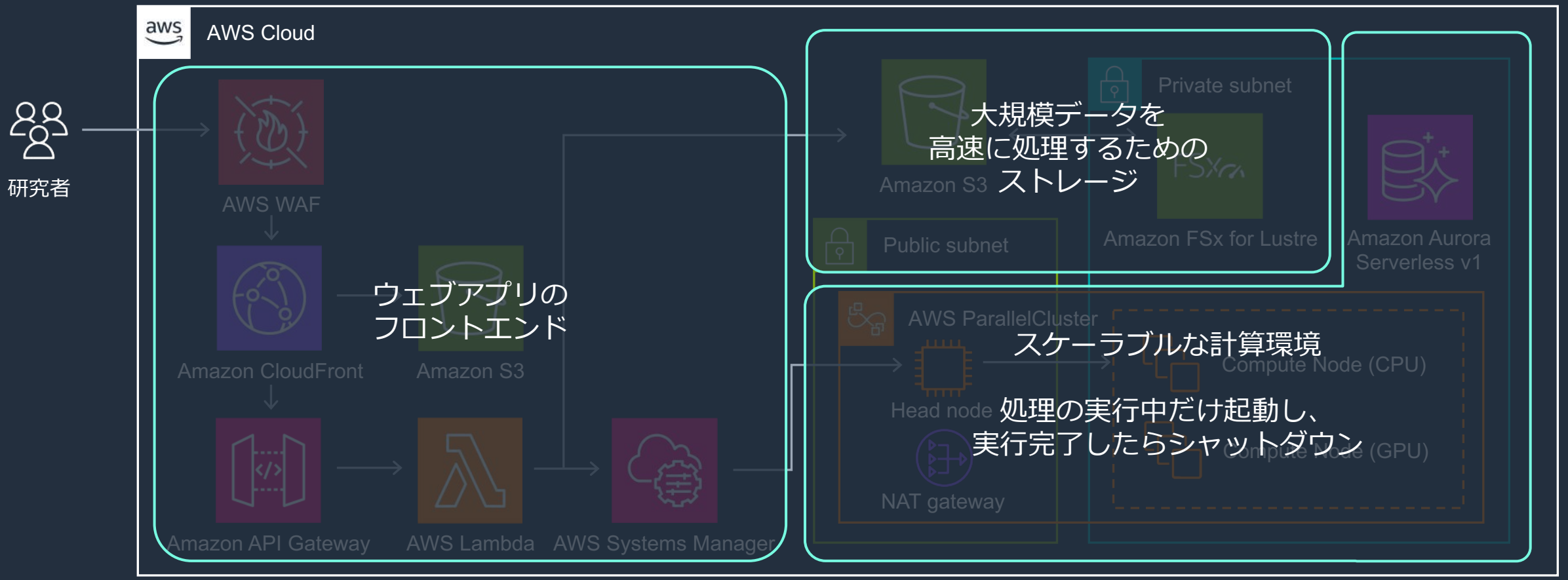


Download PDB

① アーキテクチャ全体像



① アーキテクチャ全体像



スケーラブルな計算環境を実現するために

①

AWS ParallelCluster

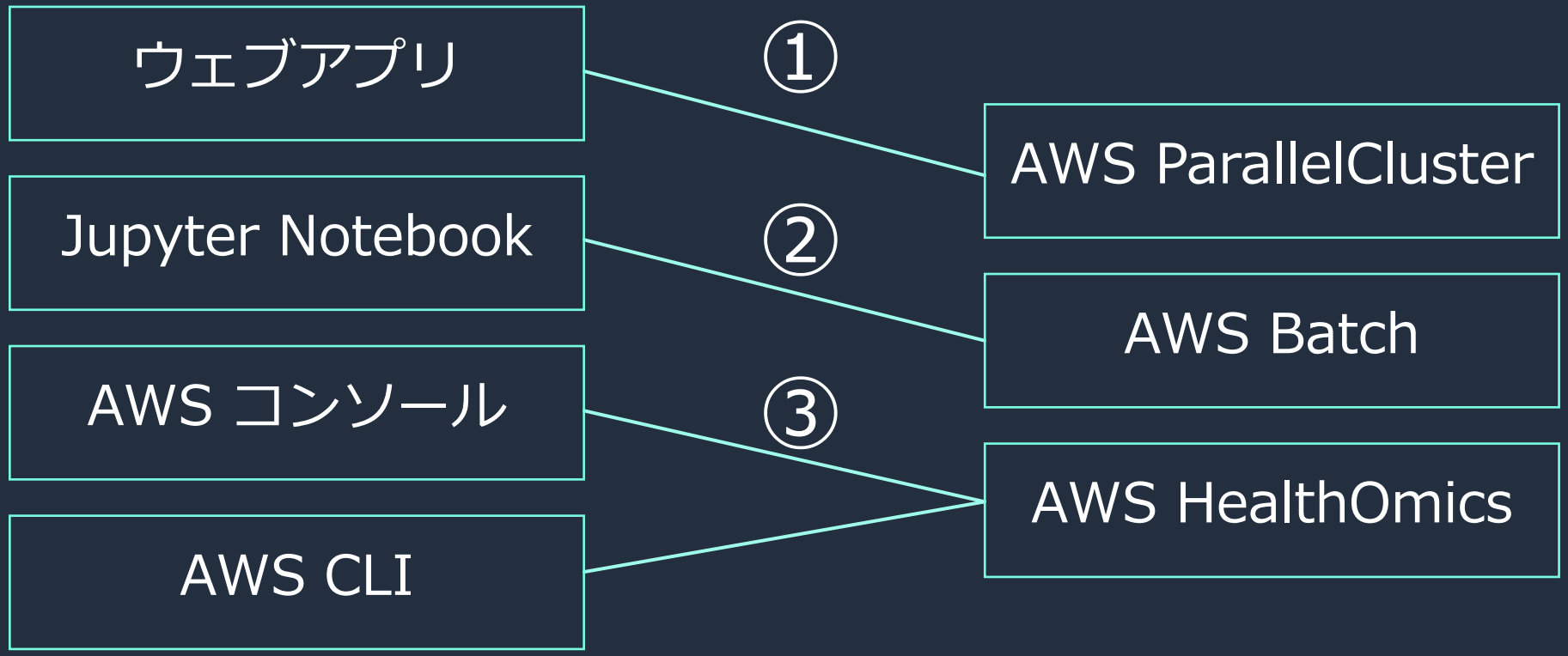
- 自動でスケールする HPC クラスタを AWS 上で構築・管理するためのツール
- HPC クラスタ利用者に馴染みのある **ジョブスケジューラ（Slurm）**を利用可能
- 使用するOSやネットワーク環境、ストレージ構成などをカスタマイズ可能
- 単一のジョブが大量の CPU/GPU core を使用する密結合ワークロードに向いている

②

AWS Batch

- フルマネージド かつ コンテナベースの大規模バッチ処理をおこなうジョブスケジューラ
- AWS Batch がインスタンスの起動や停止を行うため、スケジューラや計算ノードなどの **管理が不要**
- ジョブは **Docker コンテナイメージ** を元に作成し、自動でスケールするコンピューティング環境で実行する

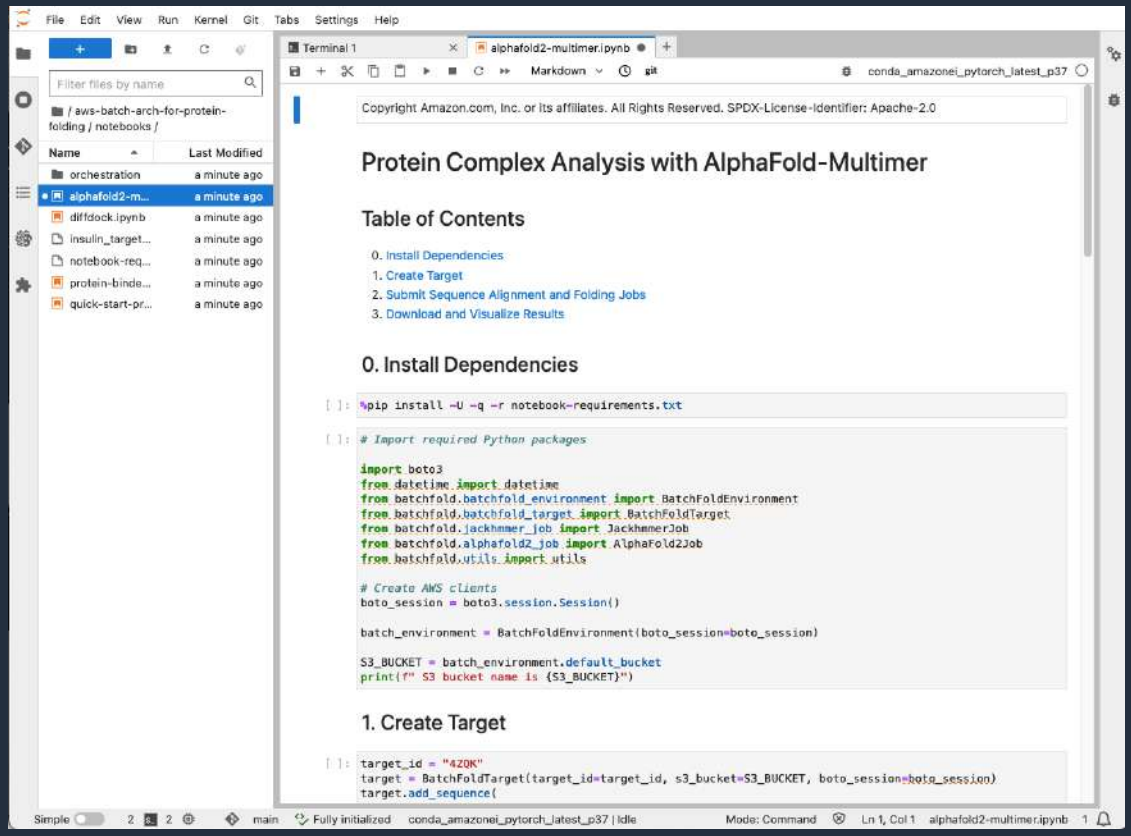
使い慣れたインターフェース + スケーラブルな実行環境（再掲）



② Jupyter Notebook + AWS Batch

- タンパク質構造予測を SageMaker Notebook から実行できる
  - オープンソースの実装例として提供
  - 10個の多様なアルゴリズムに対応（後述）

- 研究者は、SageMaker Notebook 経由でジョブの投入をおこなう
- IT 管理者は、起動テンプレートを使うことで簡単に構築可能



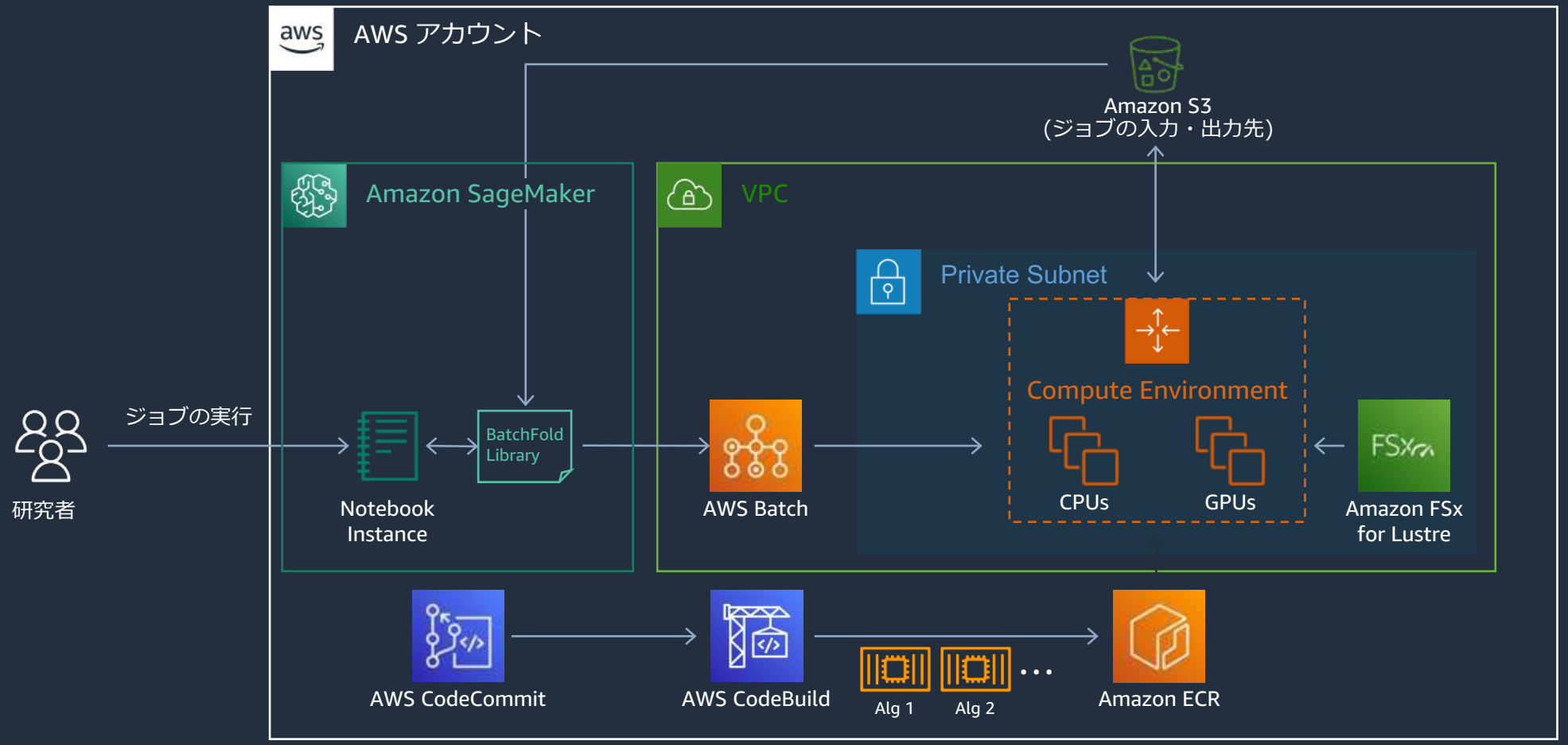
- AWS Solutions Library – “Guidance for Protein Folding on AWS”
- GitHub – “AWS Batch Architecture for Protein Folding and Design”

② 対応している主なアルゴリズム

- **MSA をベースとした構造予測**
  - **AlphaFold 2** (DeepMind)
  - **OpenFold** (Columbia University)
- **タンパク質言語モデル (pLM) をベースとした構造予測**
  - **OmegaFold** (Helixon US)
  - **ESMFold** (Meta Fundamental AI Research)
- ***De novo* タンパク質設計**
  - **RFDiffusion** (University of Washington)
  - **ProteinMPNN** (University of Washington)
- **バーチャルスクリーニング**
  - **DiffDock** (Massachusetts Institute of Technology)

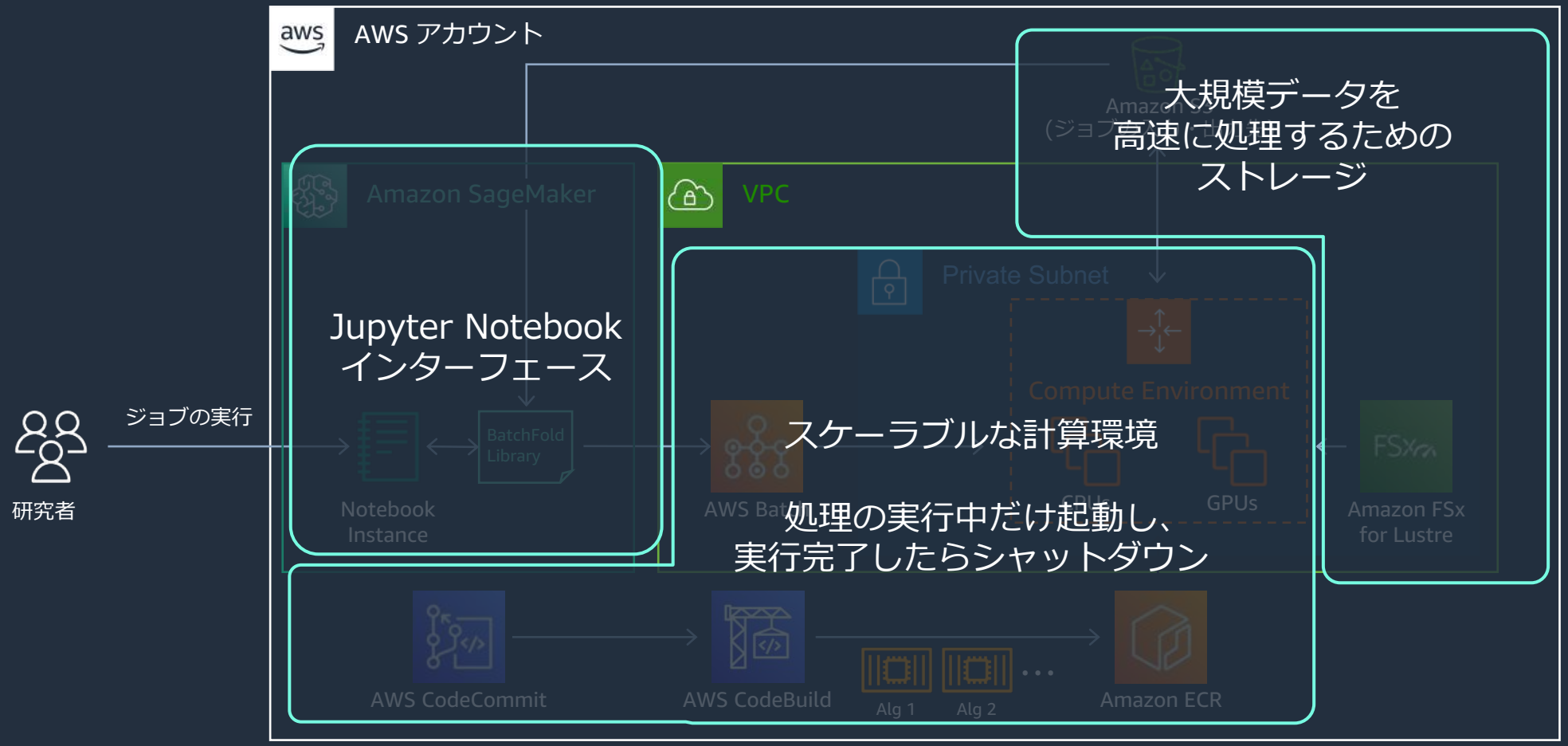
- AWS Solutions Library – “Guidance for Protein Folding on AWS”
- GitHub – “AWS Batch Architecture for Protein Folding and Design”

② アーキテクチャ全体像



- AWS Solutions Library – “Guidance for Protein Folding on AWS”
- GitHub – “AWS Batch Architecture for Protein Folding and Design”

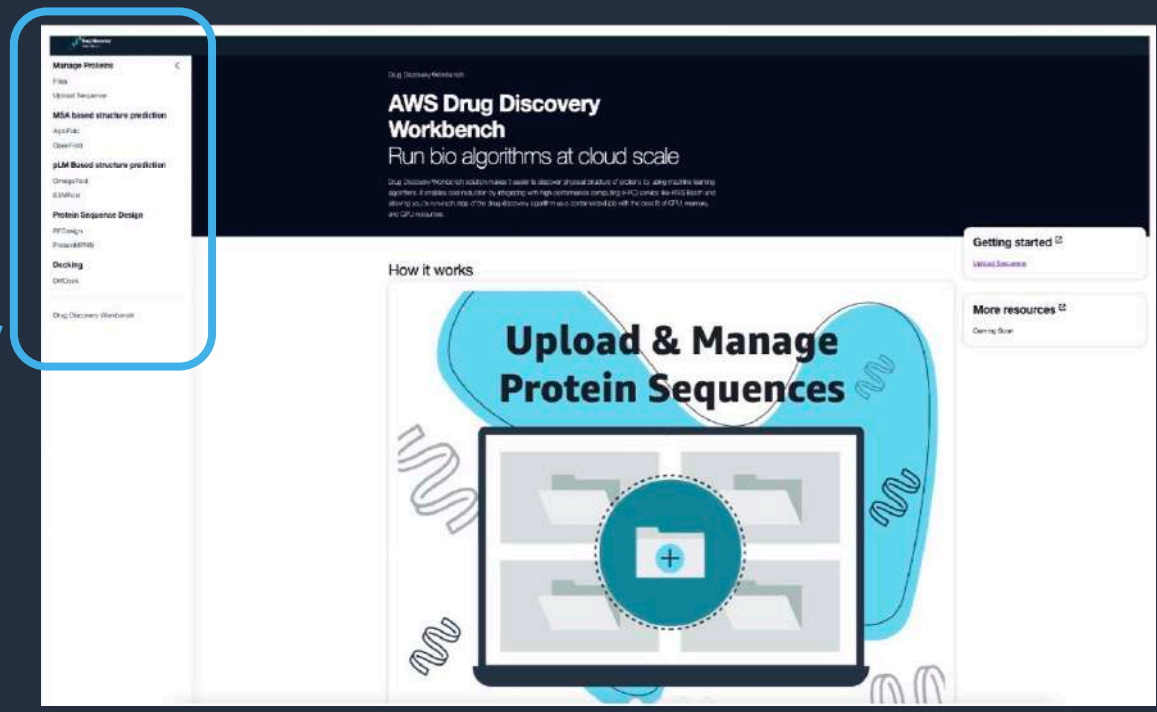
② アーキテクチャ全体像



- AWS Solutions Library – “Guidance for Protein Folding on AWS”
- GitHub – “AWS Batch Architecture for Protein Folding and Design”

②’ AWS Drug Discovery Workbench（プレビュー中）

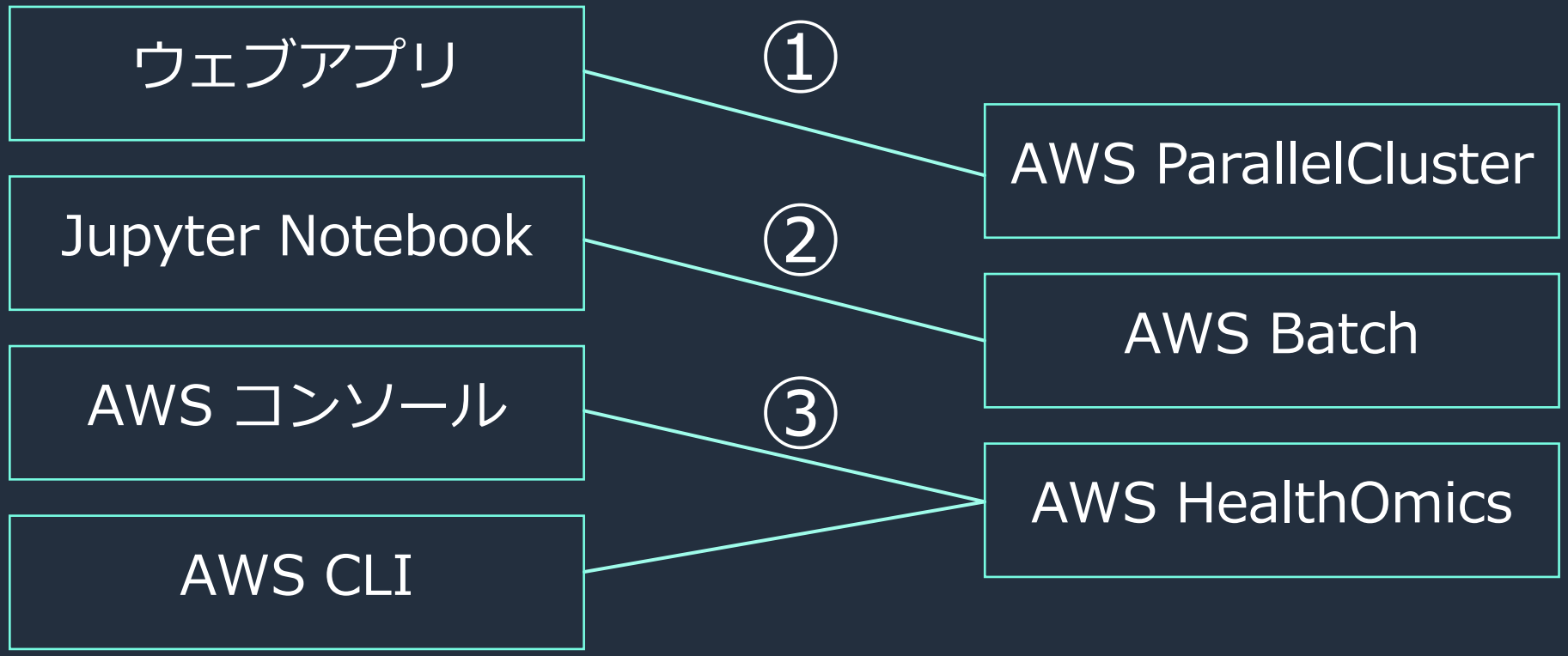
- ② の構成をさらに拡張し、ウェブアプリのフロントエンドが追加されたオープンソースの実装例
- 現時点での対応アルゴリズム
  - AlphaFold2, OpenFold, OmegaFold, ESMFold, RFDiffusion, ProteinMPNN, DiffDock, ImmuneBuilder
- 主な機能
  - 入力ファイルのアップロード
  - 出力ファイル（PDB）の可視化
  - メタデータの付与、検索



具体的な提供時期は未定ですが、
ご興味ある方はお問い合わせください

- AWS ブログ「AWS ライフサイエンスエグゼクティブシンポジウム 2023 のハイライト：ML およびジェネレーティブ AI で創薬を加速する」

使い慣れたインターフェース + スケーラブルな実行環境（再掲）



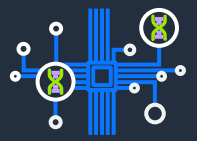
③ AWS サービスの機能を活用して、すぐに分析を開始するなら

AWS HealthOmics

本番利用可能なオミクス解析環境を
フルマネージドで素早く提供する

マルチオミックスとマルチモーダル分析

集団ゲノム解析レベルの規模に対応



フルマネージドなバイオインフォマティクス計算環境

セキュリティ、プライバシー、コンプライアンス機能搭載

※ 2023年10月現在、AWS HealthOmics は東京リージョンで未提供

③ AWS HealthOmics：3つの機能

ストレージ

store raw sequencing data

GCTTGGTCA
GATAATGCA *fasta*
→

Reference Store

ref. sequence


fastq
bam
cram
→

Sequence Store

raw sequence

オミクスデータの保存

ワークフロー

run analysis workflows

Workflow Definition Language


Nextflow

CWL

→

Omics Workflow

オミクスデータの 2 次解析

分析

store and query variant data

GCTTGGACA
GATCATGCA *vcf*
gvcf
→

Variant Store

variants


vcf
gff
tsv
→

Annotation Store

annotations

オミクスデータの 3 次解析と
マルチモーダル解析

③ より手軽にワークフローを実行する

Ready2Run ワークフロー

特定のユースケースに特化した
事前構築済みワークフローのセット

オープンソースパイプライン

gatk from Broad Institute

scRNAseq from **nf-core**

AlphaFold from DeepMind

ESMFold from Meta Research

業界各社によるワークフロー

Element Biosciences

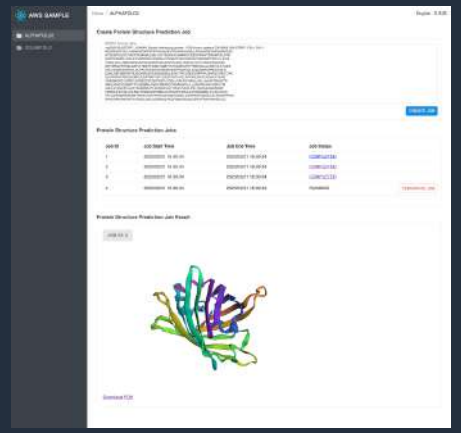
Sentieon

NVIDIA

ここまでのまとめ：タンパク質構造解析を始めるためのツール

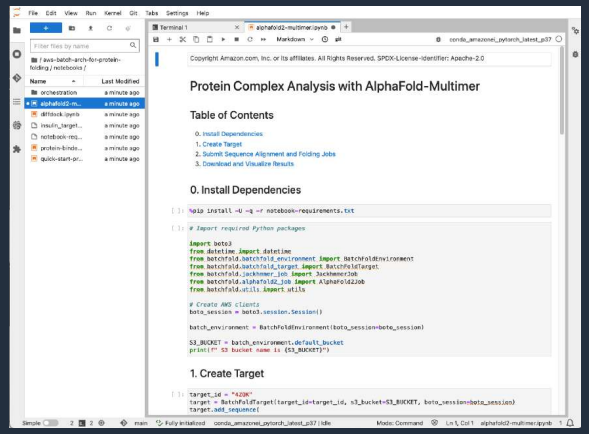
- スケーラブルな実行環境
- 用途に応じたインタフェースと組み合わせられるのも AWS の良さ

① ウェブアプリ + AWS ParallelCluster



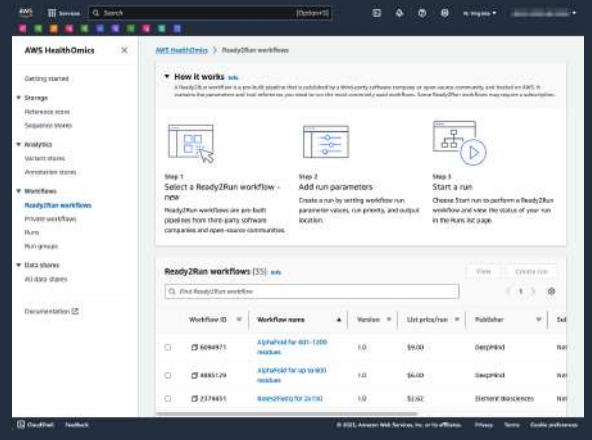
[AlphaFold2 Webapp on AWS]

② SageMaker Notebook + AWS Batch



[AWS Batch Architecture for Protein Folding and Design]

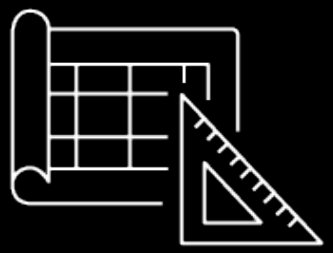
③ AWS HealthOmics (Ready2Run Workflows)



[AWS HealthOmics の新機能紹介ブログ]

生成系 AI 活用をさらに一歩先へ：Generative AI Innovation Center

Generative AI Innovation Center のご紹介



Design

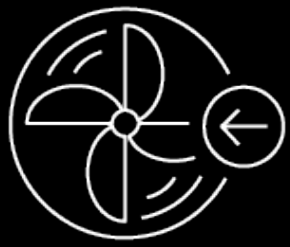
設計ガイダンス:

- 最もビジネスインパクトのある生成系AIユースケースの選定
- 生成系AIの開発、学習、本番環境へのデプロイまでの計画策定

Deploy

推奨ソリューションの実用化:

- ビジネス目的を満たす生成系AIソリューションの開発とファインチューニングを行い、実現可能性を実証



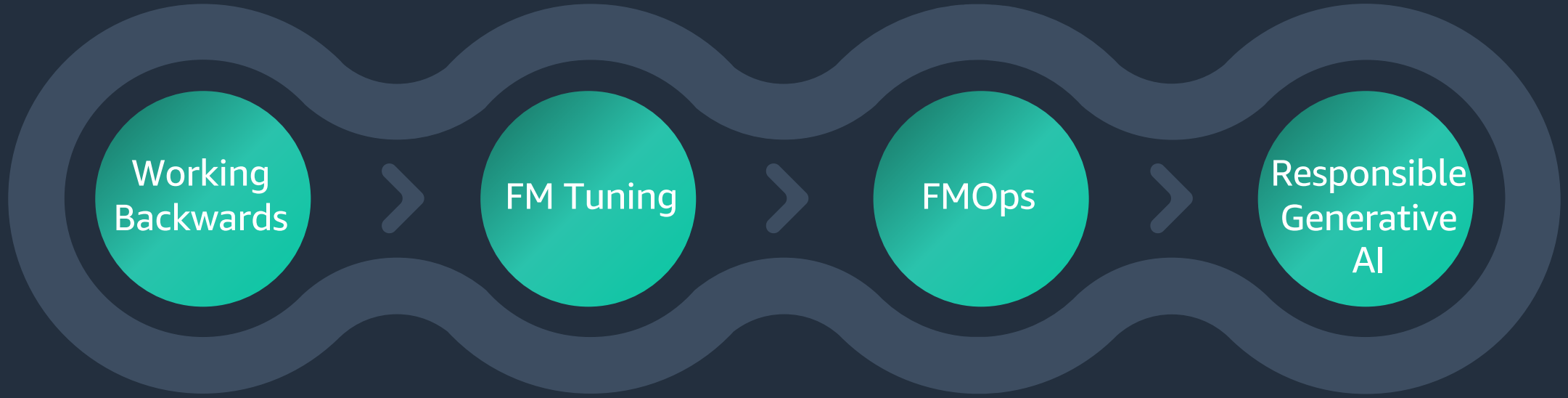
Drive

実利用の加速:

- 生成系AIソリューションをアプリケーションに組み込み、定着率と導入を促進

- AWS What's new 「AWS、生成系AIイノベーションセンターを発表」

生成系 AI 活用のロードマップ



ビジネスバリュー

生成系AI をビジネスバリューに活用するユースケース機会を特定

モデル探索

カスタムおよびドメイン固有のFM*チューニング、ユースケースに合わせたFM の学習と構築のための高度な実装支援

実用化への道筋

- 継続的なFMチューニングとモデル圧縮
- FM の知識とプロンプトの精緻化
- 学習データの自動ラベリング

生成系AIガイダンス

原則論から実践に至るまで、信頼できる生成系AIの製品とソリューションを構築およびローンチするためのアプローチを指南

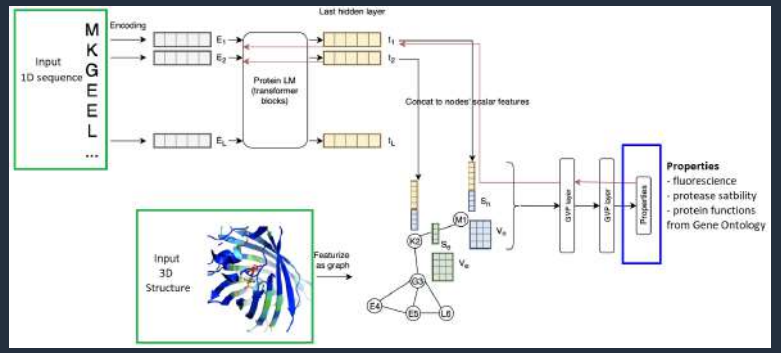
* FM = Foundation Model (基盤モデル)

創薬研究領域での GenAI Innovation Center 支援事例

Janssen Biotherapeutics 様での事例

LM-GVP: an extensible sequence and structure informed deep learning framework for protein property prediction

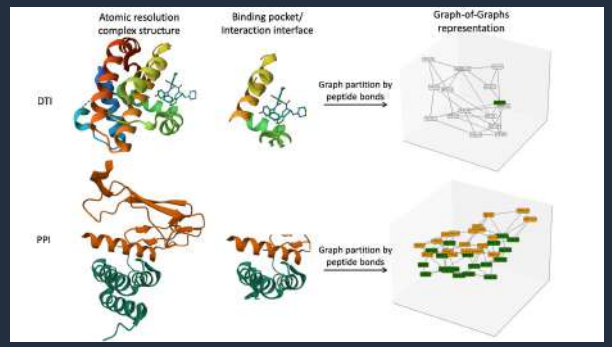
- LM-GVP* という深層学習フレームワークを構築し、タンパク質の性質を予測した
- フレームワークは pLM と GNN からなる



*Language Model – Geometric Vector Perceptrons, https://www.nature.com/articles/s41598-022-10775-y

EGGNet, a generalizable geometric deep learning framework for protein complex pose scoring

- EGGNet* という深層学習フレームワークを構築し、タンパク質複合体のドッキングポーズのスコアリングを算出
- タンパク質の安定性予測を 12 %改善



*Equivariant Graph of Graphs neural Networks, https://www.biorxiv.org/content/10.1101/2023.03.22.533800v1

このセッションのまとめ

- 創薬研究の質をより一層向上させるために、クラウドと生成系 AI を活用できる
- タンパク質構造解析の領域で、用途に応じたツールがある
- 生成系 AI 活用のために、専任の支援チームがいる

AWS の展示ブースにてデモを
ご覧いただけます！
ぜひお気軽にお立ち寄りください。

aws FOR HEALTHCARE AND LIFE SCIENCES

Thank you!

Chiaki Ishio
ciishio@amazon.co.jp