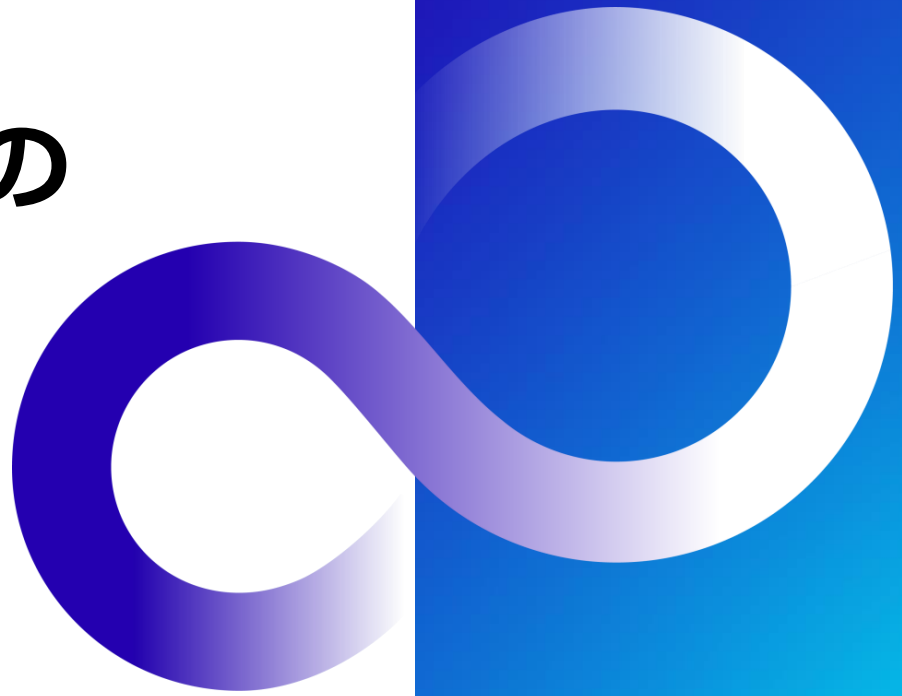
富士通のHPCを活用したAIへの取り組み

2023年12月7日

白幡　晃一
富士通株式会社　富士通研究所
コンピューティング研究所



Our purpose

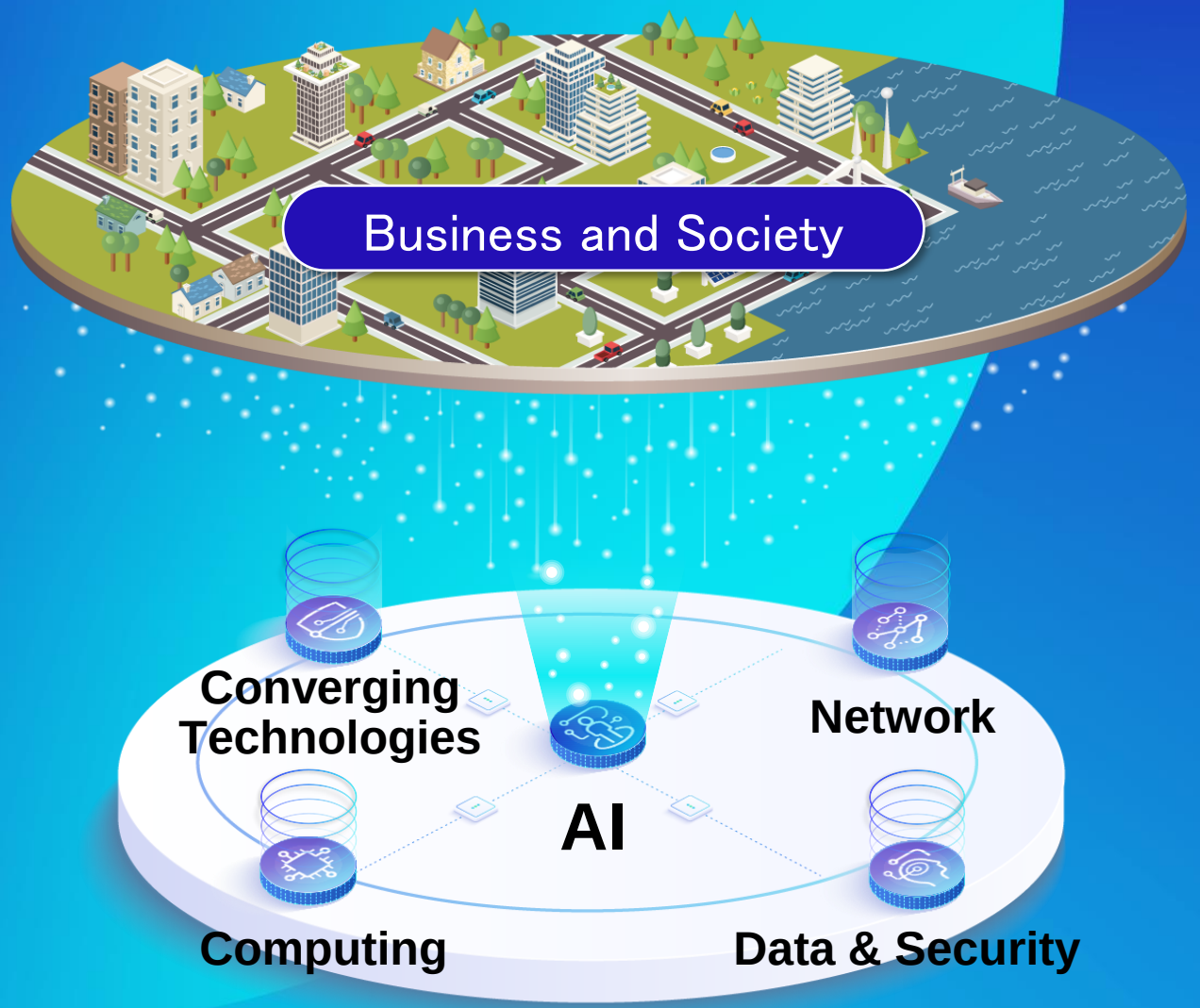
**イノベーションによって社会に
信頼をもたらし、
世界をより持続可能にしていく**

Fujitsu Uvance

**ビジネスを加速し、
社会課題に挑むソリューション**

5 Key Technologies

**5つの重点技術領域の研究開発に
リソースを集中すると共に、幅広い
パートナーとコラボレーション**



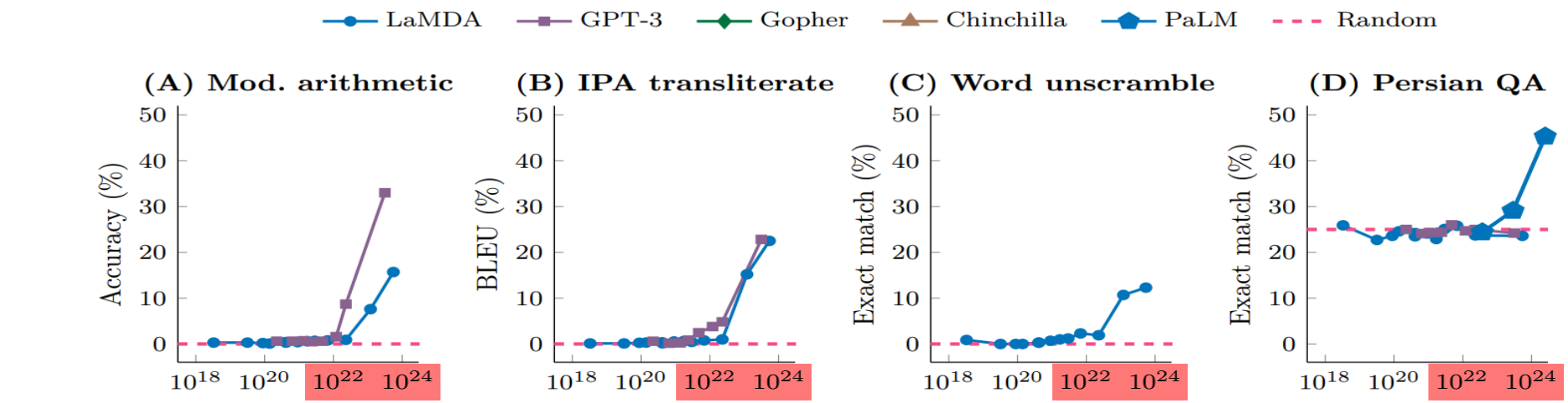
生成AIを支える膨大な計算機リソース

10^{23} FLOPsの計算量の学習で創発性を観測

- これまでできていなかったことが突然できるようになる現象を様々なタスクで観測
- 日本国内の最大のGPUスパコンであるABCIの4000GPUを利用したグランドチャレンジでも達成できない計算量

https://arxiv.org/abs/2206.07682

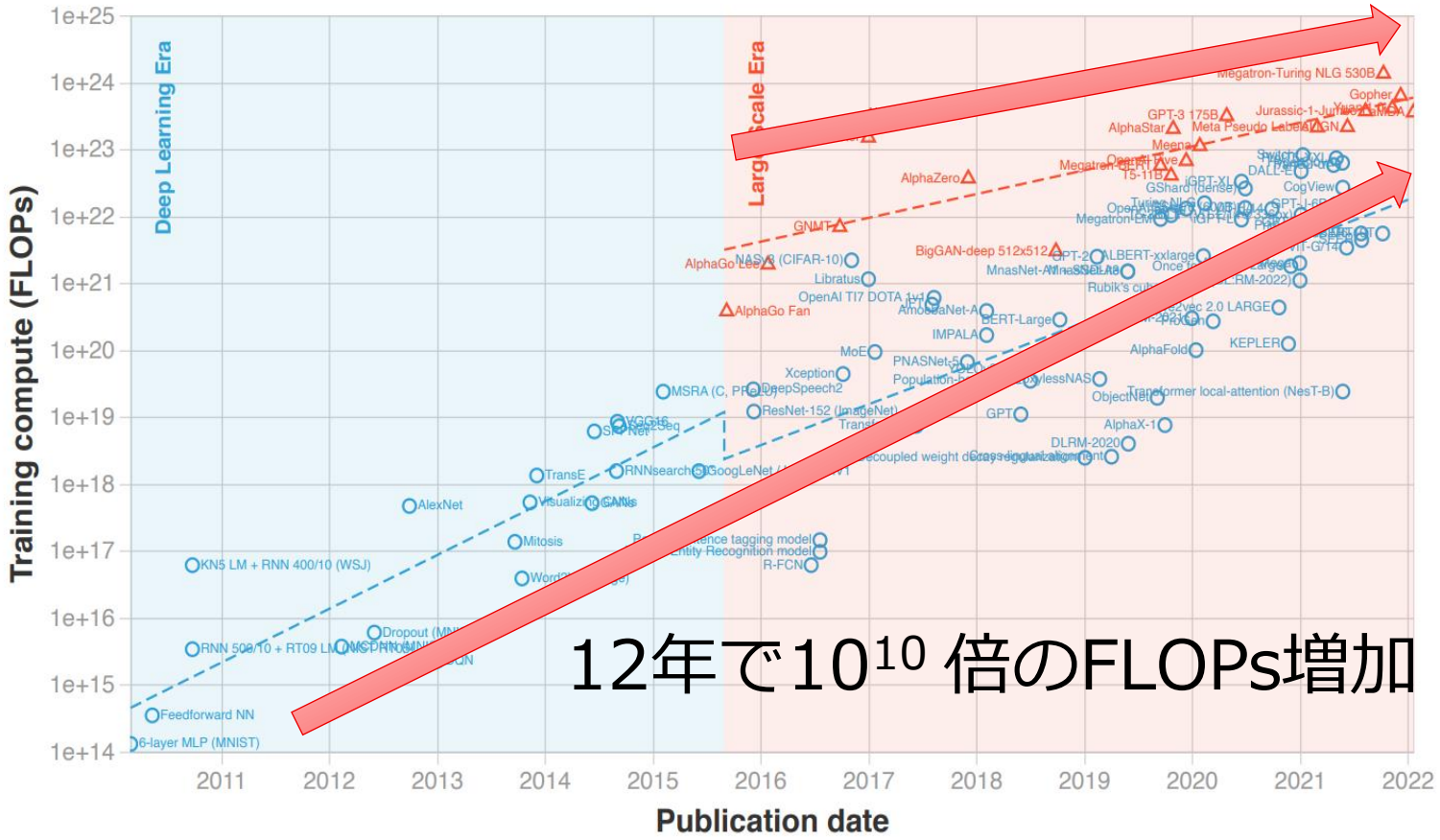
10^{23} FLOPsで精度が大きく改善



縦軸＝精度　横軸＝計算量(FLOPs)

計算資源の必要性（スケール則）

計算量が大きい領域でのFLOPs増加は緩やか



https://arxiv.org/abs/2202.05924

- ムーアの法則
  - 1970-2020の50年で10^7倍
- 深層学習のスケール則
  - 2010-2022の12年で10^{10}倍
  - この先はムーアの法則に従うので追いつきやすくなる

＊ ここでは **FLOPs** とはアプリケーションの**総演算量** (not 速度) を指す

富岳LLM　「富岳」政策対応枠におけるLLM分散並列学習手法の開発

FUJITSU

LLMで先行するのは、膨大な計算資源を持つ一部の企業のみ

- 日本語特有の課題
- 莫大な計算コストがかかる

富岳を活用した大規模言語モデル分散並列学習手法の開発を開始

- 日本語特有の課題解決、開発した生成AIモデル（富岳LLM）はGitHub等で公開予定
- 業務向けにファインチューニングした特化モデル開発、消費電力効率を考慮した軽量化

富士通　東京工業大学　東北大学
名古屋大学　理化学研究所
サイバーエージェント　Kotoba Technologies

富岳LLM

開 発

公 開　Hugging Face　GitHub

Fujitsu Kozuchi

活 用

培ったノウハウを活用

業務特化型生成AI

・業務特化モデル・省電力

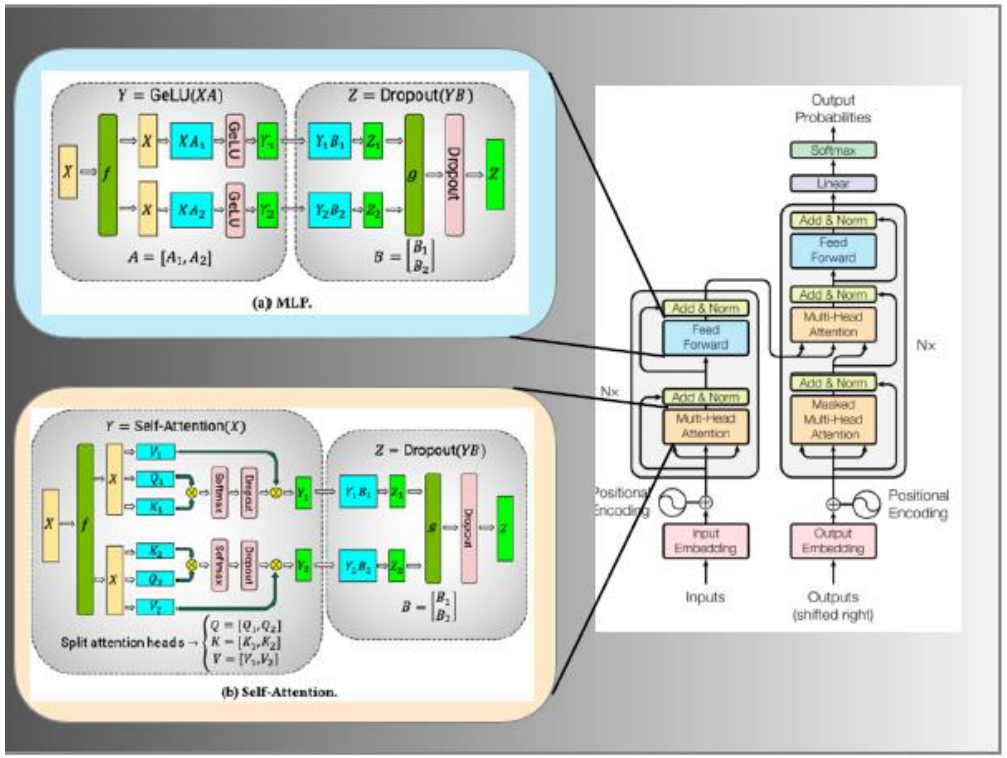
photo credit: RIKEN

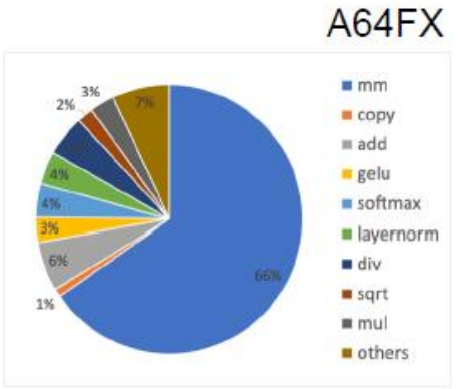
GPTの演算時間の内訳

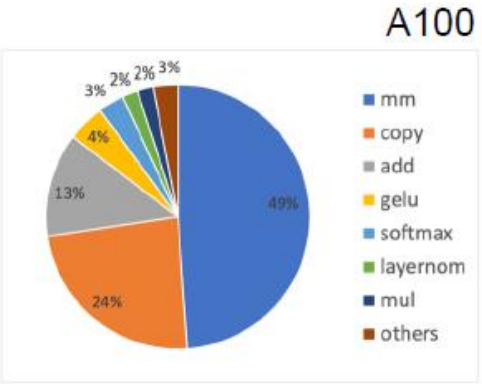
FUJITSU

- 計算の多くは密行列-密行列の積
  - → A64FXでは66%、A100では49%の時間がこの部分に費やされている
  - → 取り組み開始時点で理論ピークの1/3の性能になっており大幅に向上できる可能性がある

ある共通の入力 x に対しそれぞれの変換行列を適用して
$Q=XW_Q$, ,$K=XW_K$, $V=XW_V$ を用意する。
自分自身の要素との注目度合いを抽出する。







2022年度第2回計算科学フォーラム
https://hpcic-kkf.com/forum/2022/kkf_02/data/yokota_kkf2022-02_v2.pdf

提供：東工大 横田研

GPTの並列学習方法

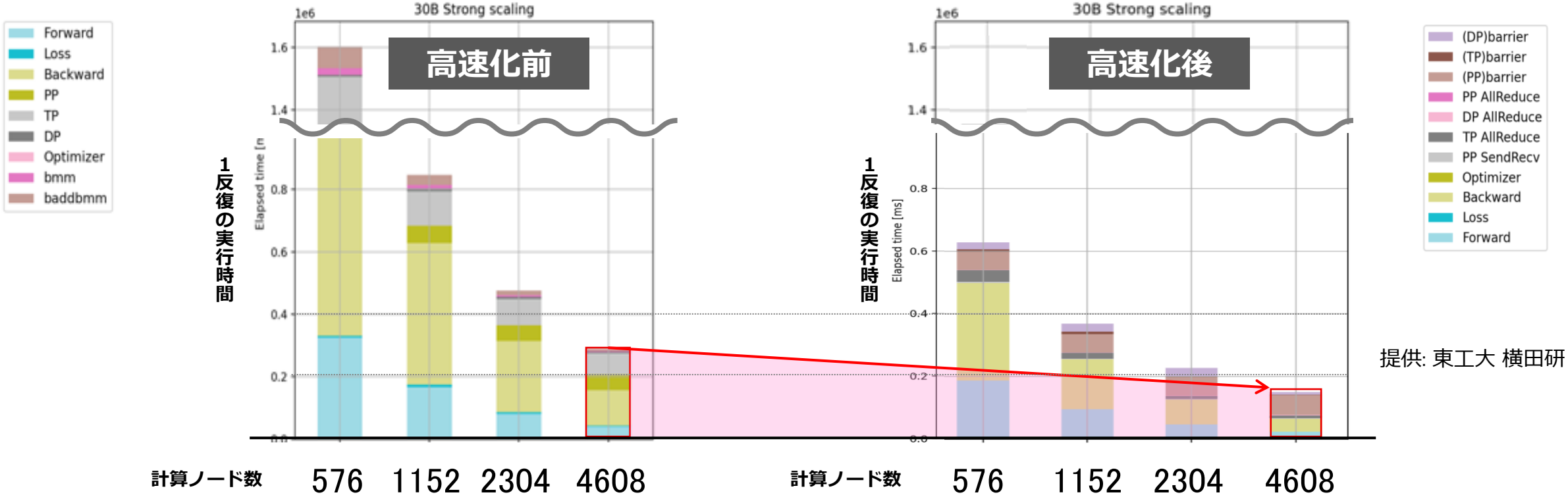
FUJITSU

○ 「富岳」でGPTを高効率に学習するには3種類の並列化を適切に組み合わせることが重要

学習の計算性能

FUJITSU

○計算効率を開発当初の10%から20%まで向上 (576ノード使用時)



提供: 東工大 横田研

計算リソースの大量消費という課題

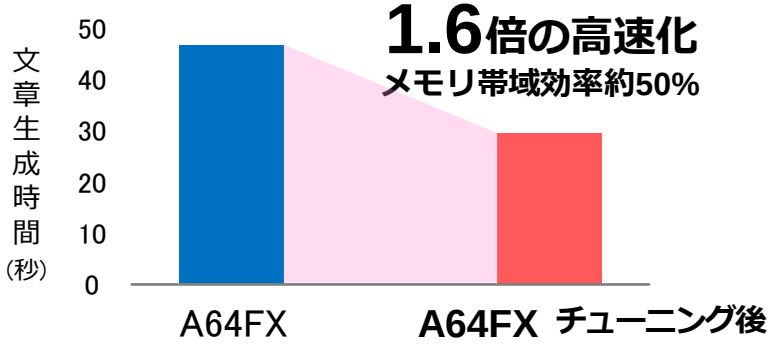
FUJITSU

Transformerベースの巨大DLである、現在のLLMには多くの課題

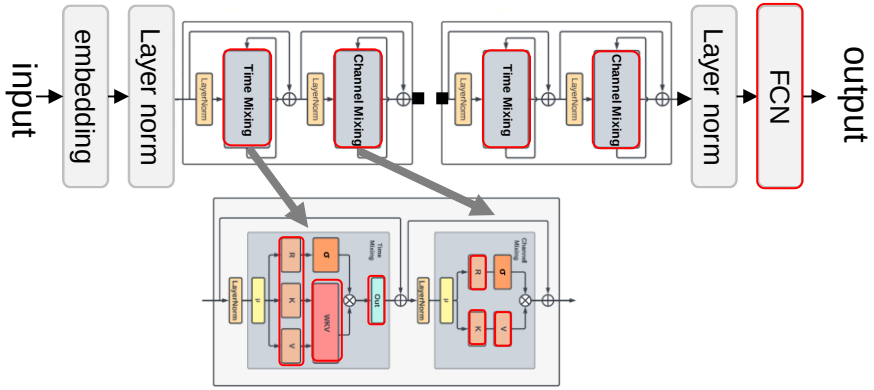
学習時間　推論速度　運用コスト　消費電力

Transformerの次のLLMを探求

「富岳」アーキテクチャの高いメモリ帯域を引き出すことで高速化



独自軽量化技術により省メモリ化と推論高速化を可能に



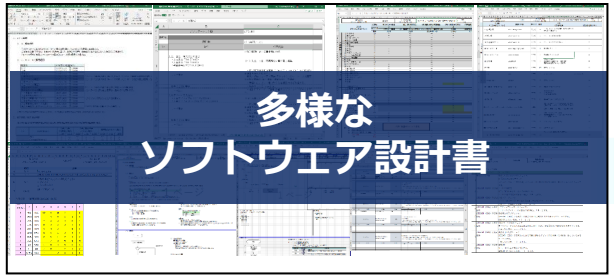
新しいLLM・生成AI・基盤モデルの世界へ！

- 新しいDLアーキテクチャの探求
- 新しいAI学習インフラの提供

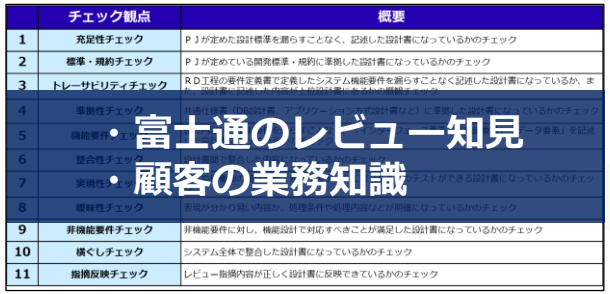
生成AIを活用したシステム開発の効率化

課 題 設計書を生成AI活用するには、データ加工・修正に時間がかかる

技 術 多様な設計書を読み込めることで、ノウハウを活かせる技術



レビュー対象の
設計書



	チェック観点	概要
1	充足性チェック	ＰＪが定めた設計標準を漏らすことなく、記述した設計書になっているかのチェック
2	標準・規約チェック	ＰＪが定めている開発標準・規約に準拠した設計書になっているかのチェック
3	トレーサビリティチェック	ＲＤ工程の要件定義書で定義したシステム機能要件を漏らすことなく記述した設計書になっているか、ま
4	準拠性チェック	
5	機能要件	
6	整合性チェック	
7	実現性	ストができる設計書になっているかのチェック
8	曖昧性チェック	表現が分かり易い内容か、処理条件や処理内容などが明確になっているかのチェック
9	非機能要件チェック	非機能要件に対し、機能設計で対応すべきことが満足した設計書になっているかのチェック
10	横ぐしチェック	システム全体で整合した設計書になっているかのチェック
11	指摘反映チェック	レビュー指摘内容が正しく設計書に反映できているかのチェック

チェックのやり方の
知識を取り込む

設計書

ナレッジ

からプロンプトを
自動生成

AIが解釈
できる形に
データを変換

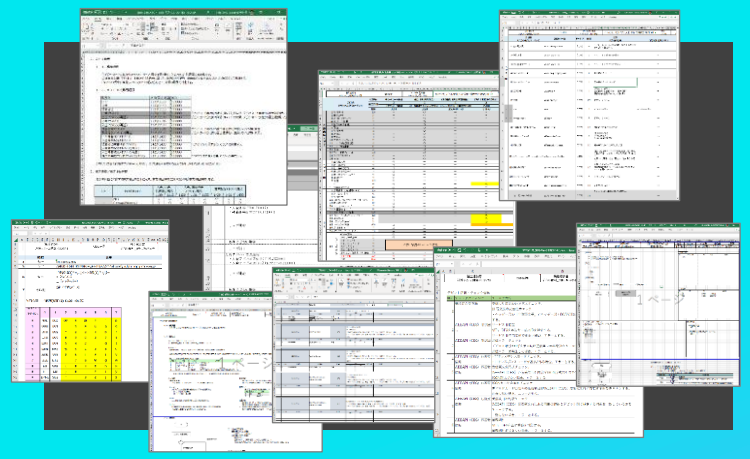
生成AI
LLM

抜け漏れがある設計書の問題点を指摘

FUJITSU

1

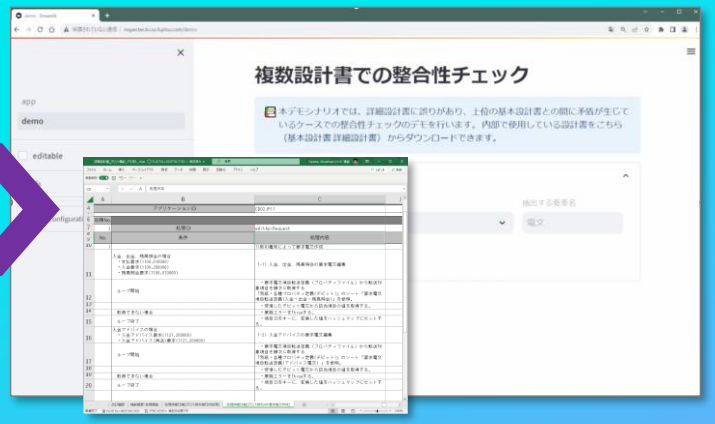
設計書を取り込み



・多種多様な設計書
・チャックのやり方のノウハウ

2

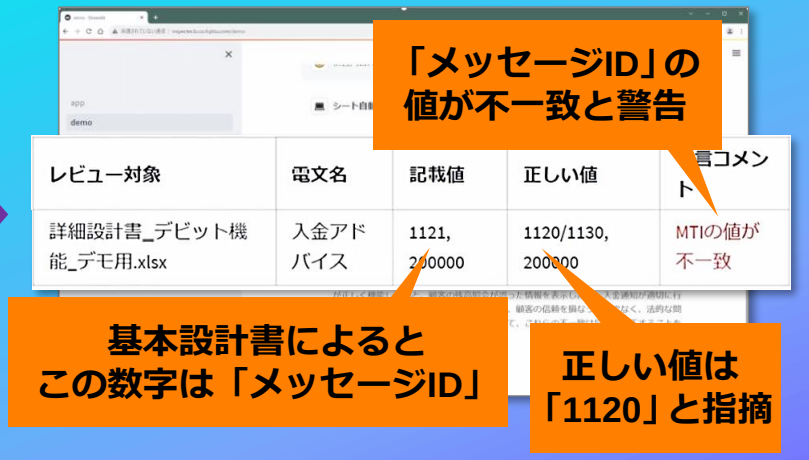
レビュー内容の指定
例：複数設計書の整合性チェック



関連する設計書は規約に
基づいて自動で割り当て

3

生成AIに取り込んで
結果を一覧にして表示



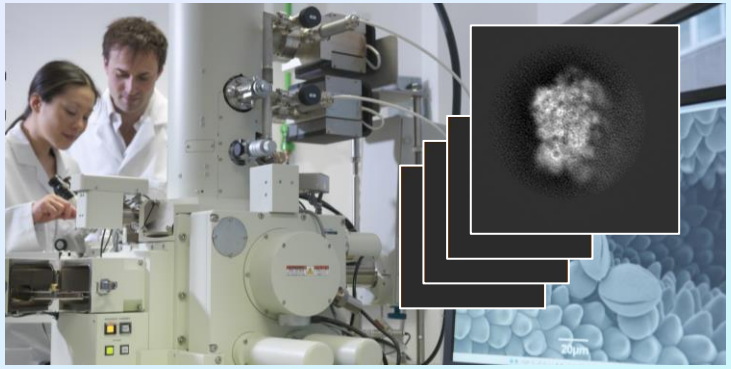
間違っている箇所の
具体的な修正値を提示

従来は専門家にしか指摘ができなかったことが自動で可能に

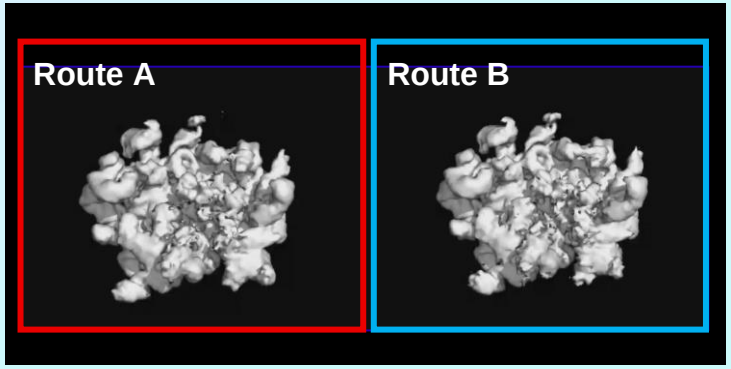
創薬の開発期間や費用を劇的に削減

- 創薬の開発期間10年、費用1200億円、成功確率2.5万分の1
- 細菌やウイルスなどの標的タンパク質の形態や構造変化の把握が重要

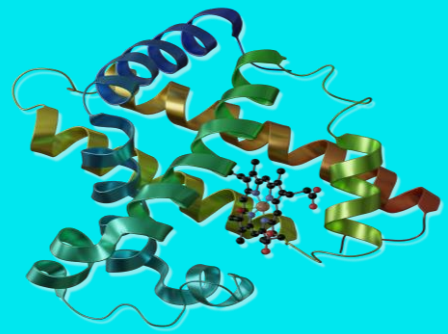
電子顕微鏡で撮影された
大量のタンパク質画像



形態や構造の変化を推定し
複数の反応経路を自動生成



タンパク質の働きを適切に
制御する薬剤を探索



独自の生成AI技術に基づく創薬技術

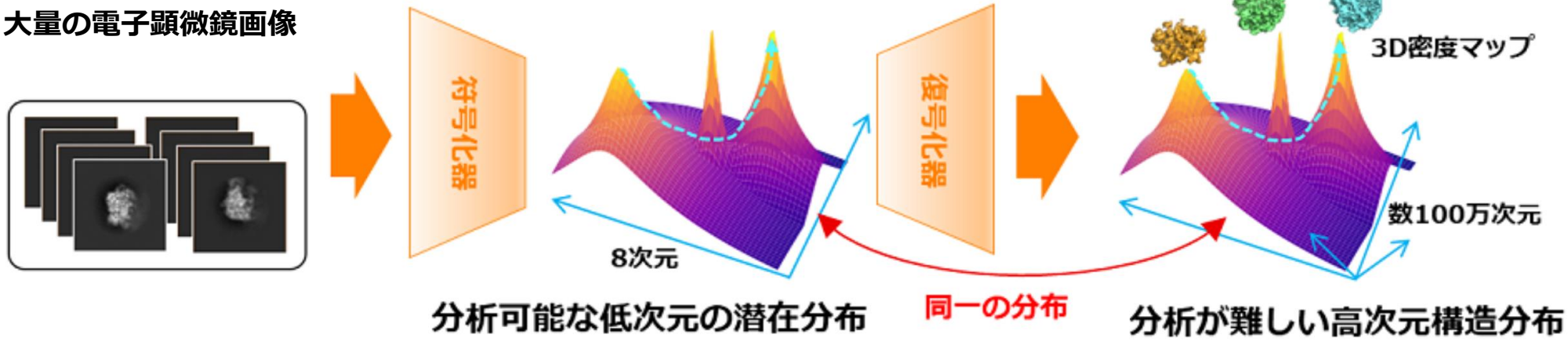
課 題 数万次元の情報をもつ標的タンパク質の構造や動的変化の把握は困難

技 術 タンパク質の形状を低次元で捉えることで立体構造と連続変化を定量的に予測
→構造変化の予測を従来の1日から2時間に短縮

1 低次元空間で形態の割合を見積もり、立体構造を生成

2 実際の高次元空間での形態や構造変化に復号

様々な方位から投影された
大量の電子顕微鏡画像



標的タンパク質に結合する薬剤の設計過程の革新が期待

「富岳」におけるAlphaFold2推論の高速化

FUJITSU

Oyama et al., FlexScience 2023

We accelerate DeepMind's **AlphaFold2 (OpenFold)**† for large-scale folding

MAAHKGAEHHHKAA
EHHEQAAKHHHAAA
EHHEKGEHEQAAHH
ADTAYAHHKHAEEH
AAQAAKHDAEHHAP
KPH

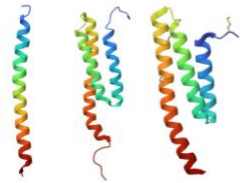
Sequence

Our scope

1. Search from databases

MAAHKGAEHH…
MATHKGAESH…
MATHKGTEHH…

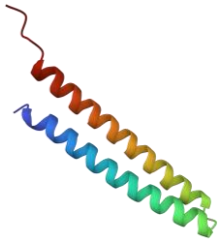
MSA
(Known similar sequences)



Templates‡
(Known structure of similar seqs.)

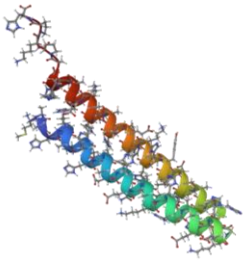
Our scope

2. Inference with AF2 network



Raw structure

3. Relax. with MD



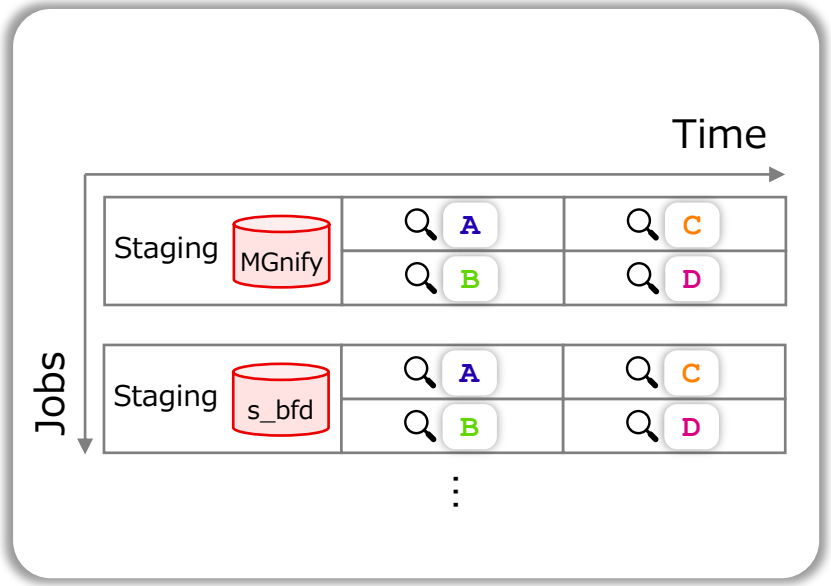
Full structure

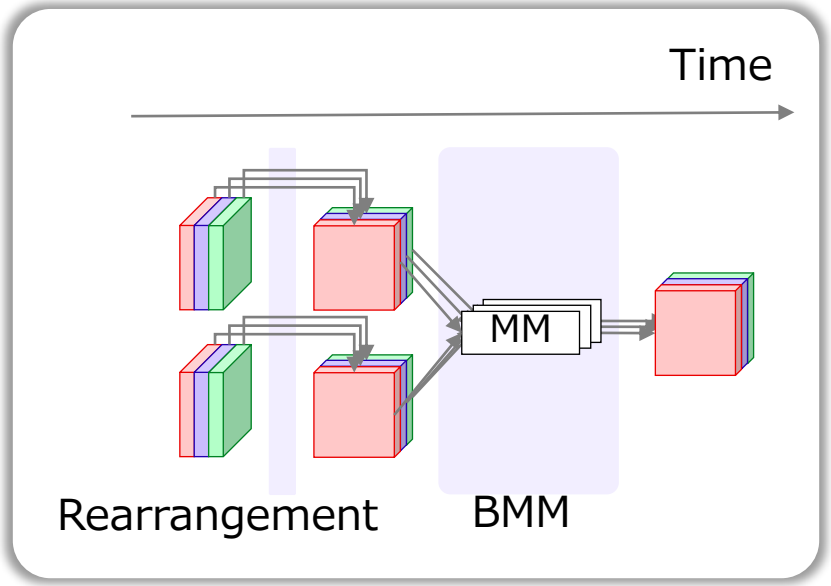
† John Jumper, et al. 2021. Highly accurate protein structure prediction with AlphaFold. Nature 596, 7873 (2021), 583–589.
‡ Images are cited form https://www.rcsb.org/

「富岳」におけるAlphaFold2推論の高速化

FUJITSU

Oyama et al., FlexScience 2023





	Fugaku	ABCI	
		(A)	(V)
# of main processors	158,976	960	4352
Processor throughput [seq./h]	7.6	84.8	55.0
Total throughput [K seq./h]	1,214.6	81.4	239.4

Search phase:
Batching search tasks achieves **8.5x** speedup

Inference phase:
Rearranging input tensors achieves **1.3x** speedup

Fugaku's throughput is estimated to be **3.8x** greater than a GPU supercomputer

HPCとAI技術を活用し、カーボンフリー素材開発を促進

Atmonia社との共同研究

(アンモニア合成触媒を開発するアイスランドのベンチャー企業)

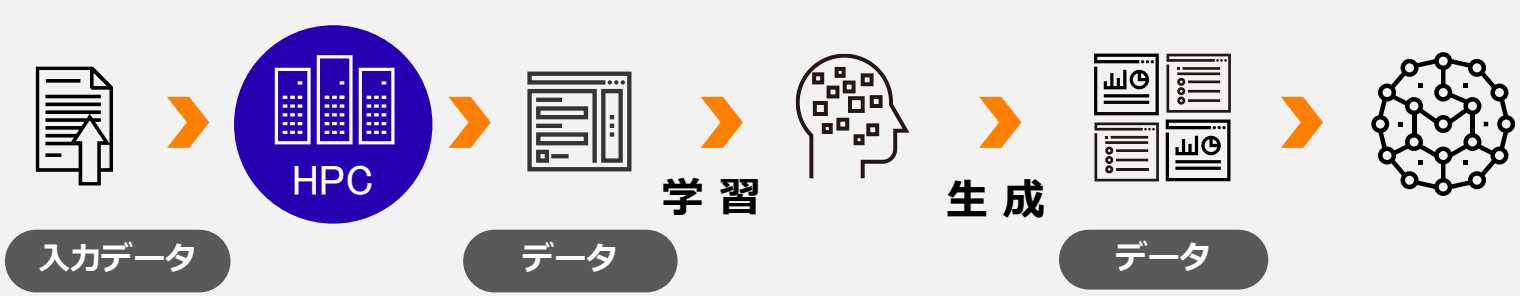
カーボンフリー物質であるアンモニアの触媒探索をHPCとAIで加速

- アンモニアは**CO2の排出が無く**、水素に比べて輸送が容易なため**次世代エネルギー**として注目
- しかし現在主流のアンモニア合成手法では化石燃料から原料の水素を作るため**CO2が大量に排出**されてしまう課題がある
- **富士通のHPC/AI技術**を活用しAtmonia社と　**水と空気（窒素）と電気のみ**でアンモニアを取り出せる革新的な新触媒を研究中

HPC & AI の融合による高速材料探索技術 (現地展示)

FUJITSU

限界を突破する技術



高速量子化学シミュレーション

コンピューティング技術で
10倍高速化しデータ生成

AIシミュレーションモデル

AI活用で精度を維持しつつ
計算量を3000分の1に削減

因果発見技術の適用

試行錯誤結果の因果関係を
分析し、未知の特性を発見

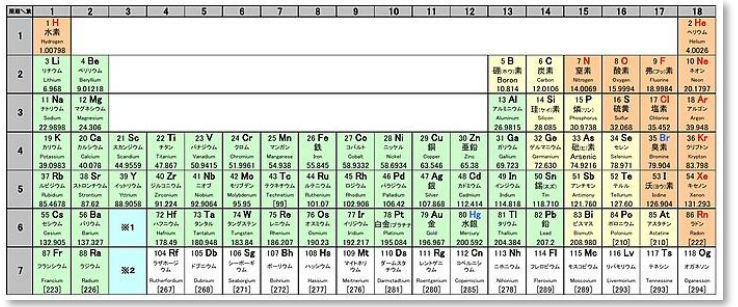
AIとコンピューティングで少ない精度劣化で計算量を削減

Atmonia社との実証

- アンモニア触媒開発をターゲット
- 大量にエネルギー計算を試行し、未知の物性傾向を発見

実証による新たな発見

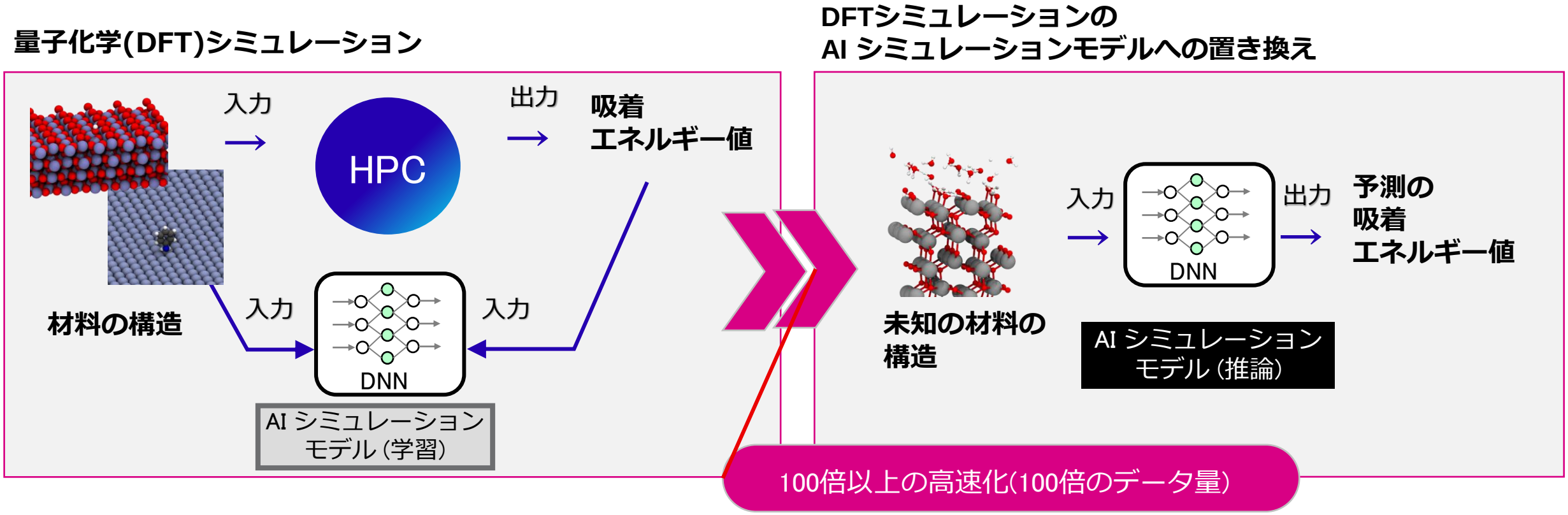
クリーンなアンモニア合成に向けて、
高価な貴金属以外の触媒材料候補を発見



※2023年10月プレスリリース

AI シミュレーションモデル

- HPCで大量に実行したDFTシミュレーションデータをAIシミュレーションモデルに学習させ、未知の材料構造の吸着エネルギー値を瞬速で推論
  → **約3000倍の高速化**
- 大量データ生成により、探索範囲の拡大、因果発見技術の精度向上が図れる

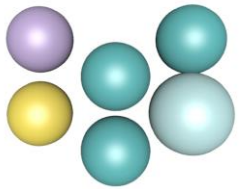


材料探索向けNNの予測精度を改善する技術

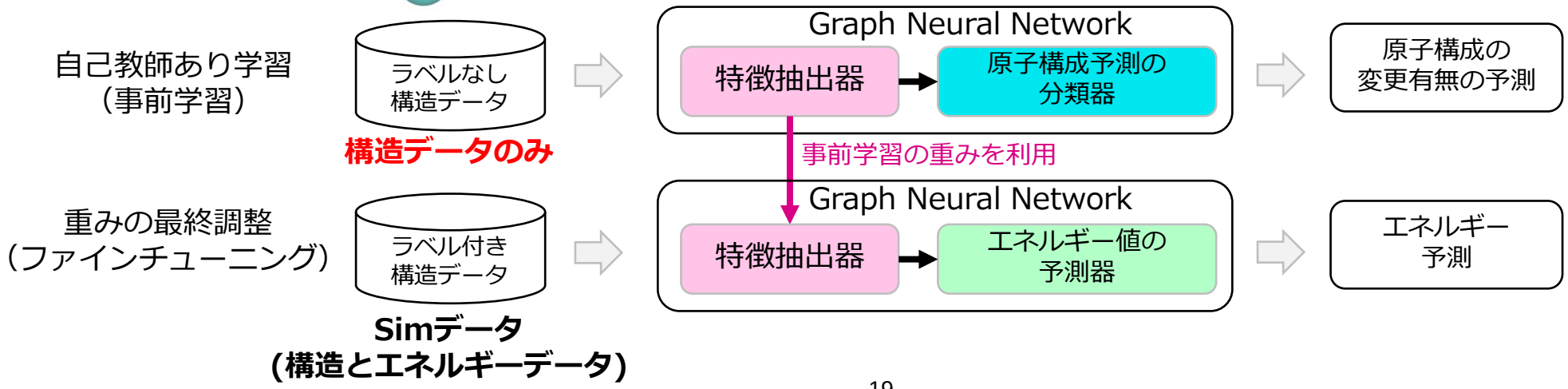
FUJITSU

Sakai et al., IEEE IJCNN 2023 Workshop

- 背景・従来の課題
  - エネルギー予測精度の高いNNの実現に必要なデータを用意するには数十～数千時間かかる
- 開発技術
  - より少ないエネルギーデータでNNを高精度化する技術(触媒探索向け自己教師あり学習)を開発
    - 同じ精度に到達するまでに必要なデータを用意するコストを半減
    - 同じデータ量でNNが高精度化 → より高精度に材料を探索できる



エネルギーデータ無しでNNを訓練する方法を開発
（材料の構成を変更し、NNに変更の有無を予測させる訓練方法）



まとめ

FUJITSU

- 富士通のHPCを活用したAIへの取り組みを紹介

- 「富岳」上での大規模言語モデルの分散並列学習手法の開発
- ドメイン特化生成AIの事例
  - システム開発の効率化
  - たんぱく質の構造変化予測
- HPCとAIを活用した科学シミュレーション高速化
  - 「富岳」におけるAlphaFold2推論の高速化
  - HPCとAIを用いた材料探索高速化

Thank you