資料4
AI for Science 推進委員会
（第5回）令和8年6月29日

AdvanceSoft
http://www.advancesoft.jp/

令和7年度「AI for Scienceの実現に向けた計算基盤等の動向調査」の報告

（令和7年12月～令和8年3月に実施）

令和8年6月29日

アドバンスソフト株式会社
松原聖・高橋邦生・岡崎一行・北野有哉・高原浩志・松尾裕一

Executive Summary

AdvanceSoft
http://www.advancesoft.jp/

研究プロセスの変革

知識体系化
LLMによる膨大な情報統合・知識体系化で、時間的な制約や認識する内容の限界を超える。

シミュレーション加速
物理法則・解析結果などを学習しAIが代替することで、膨大なシミュレーション時間を短縮する。

実験・検証の自律化
AIで実験・検証を制御する自律的な24時間稼働ラボで、時間的な制約や人為的なミスを排除する。

本資料のp.3～4に記載した。

2030年の計算需要推計

500EFLOPS級の演算需要
パラメータ・データ増に対応し、FP16で500EFLOPS、GPU12万基規模のインフラが必要と推計。

640PBのストレージ需要
基盤モデルのための学習データを蓄積するために、640PB程度のストレージが必要と推計。

2030年への技術的展望
実効性能・電力効率を上げるため、混合精度演算やCPU-GPU密結合等の技術の進展。

本資料のp.5～13のデータをもとにp.14～16にまとめた。

戦略的実装への提案

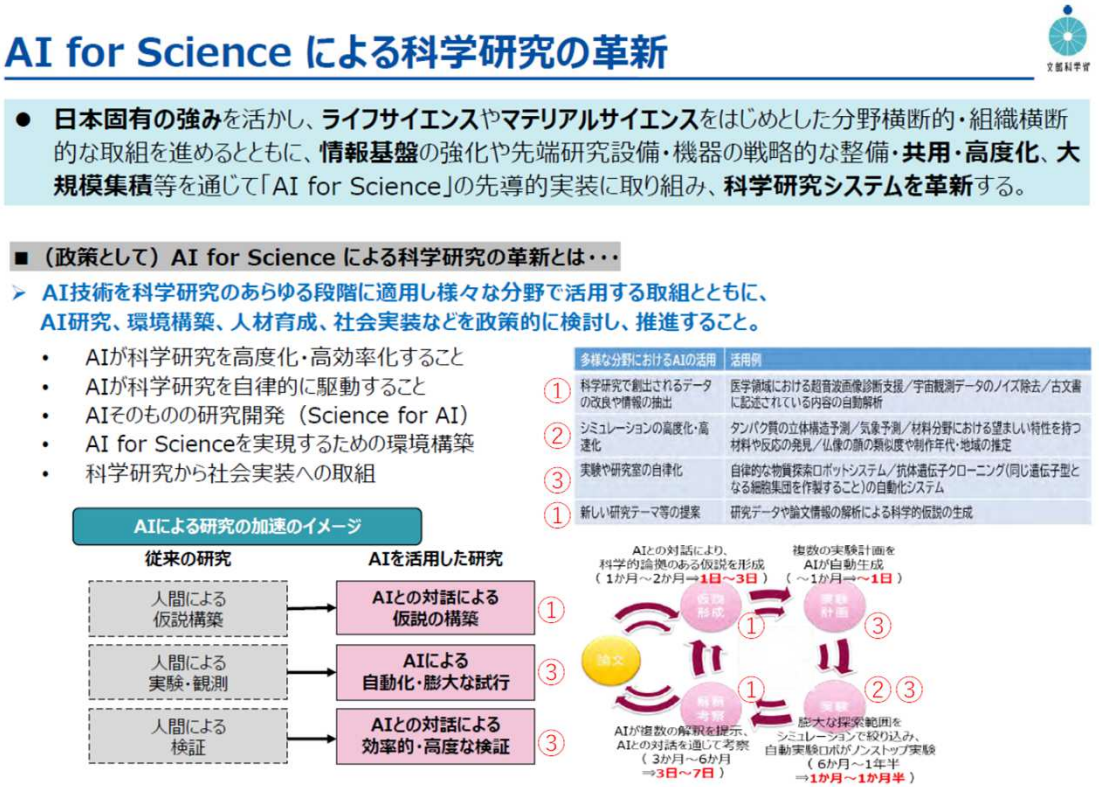
2030年：先導的実装期へ
大規模インフラの構築
演算・データ需要増加への対応
伴走型支援とキャリア確立

将来目指すべき形
自律型研究エコシステムの確立
計算エコシステムの構築
持続可能なグリーン科学基盤

本資料のp.17にまとめ、報告書4.6節に記載した

AI for Scienceによる科学研究の革新

- ✓ 右の文科省様資料に記載された内容に基づき、AI for Scienceの計算基盤と基盤モデルを整理するため、「研究プロセスに着目した基盤モデル」の視点から分類した。
- ✓ 本業務での分類
  - ① 知識体系化と知見抽出
  - ② 現象予測と計算加速
  - ③ 実験・検証の自律化



AI for Science による科学研究の革新

● **日本固有の強み**を活かし、**ライフサイエンス**や**マテリアルサイエンス**をはじめとした分野横断的・組織横断的な取組を進めるとともに、**情報基盤**の強化や先端研究設備・機器の戦略的な整備・**共用**・**高度化**、**大規模集積**等を通じて「AI for Science」の先導的実装に取り組み、**科学研究システムを革新**する。

■（政策として）AI for Science による科学研究の革新とは・・・

➢ AI技術を科学研究のあらゆる段階に適用し様々な分野で活用する取組とともに、AI研究、環境構築、人材育成、社会実装などを政策的に検討し、推進すること。

- AIが科学研究を高度化・高効率化すること
- AIが科学研究を自律的に駆動すること
- AIそのものの研究開発（Science for AI）
- AI for Scienceを実現するための環境構築
- 科学研究から社会実装への取組

	多様な分野におけるAIの活用	活用例
①	科学研究で創出されるデータの改良や情報の抽出	医学領域における超音波画像診断支援／宇宙観測データのノイズ除去／古文書に記述されている内容の自動解析
②	シミュレーションの高度化・高速化	タンパク質の立体構造予測／気象予測／材料分野における望ましい特性を持つ材料や反応の発見／仏像の顔の類似度や制作年代・地域の推定
③	実験や研究室の自律化	自律的な物質探索ロボットシステム／抗体遺伝子クローニング(同じ遺伝子型となる細胞集団を作製すること)の自動化システム
①	新しい研究テーマ等の提案	研究データや論文情報の解析による科学的仮説の生成

AIによる研究の加速のイメージ

従来の研究	AIを活用した研究	
人間による仮説構築	AIとの対話による仮説の構築	①
人間による実験・観測	AIによる自動化・膨大な試行	③
人間による検証	AIとの対話による効率的・高度な検証	③

2

https://www.mext.go.jp/content/20251006-mxt_jyohoka01-000045188_04.pdf
図中の赤い①②③は本資料作成のために追記した。

AI for Scienceによる科学研究の革新

分類	基盤モデルの目的	AIへの期待	事例
① 知識体系化と知見抽出	膨大な論文、特許、実験レポートなどの情報を学習し、把握困難な知見を統合する。キーワード間の潜在的な相関を抽出し、新たな研究仮説の立案や、専門分野を跨いだ知識の再発見を支援する。	時間的な制約や認識する内容の限界を超える。	LLM
② 現象予測と計算加速	物理法則などを学習し、シミュレーションをAIが代替することで処理時間を短縮する。例えば、未知の物質が持つ性質を予測する、目標とする新材料や新薬の構造の探索すべき候補を短時間で絞り込む。	膨大なシミュレーション時間を短縮する。	NNP, PINN サロゲート等
③ 実験・検証の自律化	AIが、実験ロボットや計測機器を制御する。実験結果をリアルタイムで解析し、次に試すべき条件を自律的に再実行する。24時間稼働の「自律型ラボ」を実現することができ、検証全体を加速する。	時間的な制約や人為的なミスを排除する。	実験検証自動化 繰り返し実験

欧米各国・各組織の AI for Science 政策・基盤整備

AdvanceSoft
http://www.advancesoft.jp/

✓ 欧州ではECでデータ戦略、および、EuroHPCでAI Factory, AI Antenna, AI Gigafactory等を進めている。米国連邦政府ではGenesis Mission、各省庁でDOE/SCのScientific AI, NNSAのAI4NS, NSFのNAIRR等を進めている。他の主要国は下記の通りである。

国	政策・基盤整備の特徴
英国	10年間で英国を「AIスーパーパワー」にすることを目指し、公的R&D支出を過去最高レベルに引き上げています。最先端AIの安全性やリスクを研究する「AIセキュリティ研究所」への投資や、政府内での活用を推進する基金に予算を多く配分しています。現在、国内のAI向け計算能力が不足している課題に対し、今後はGPUアクセラレータを備え、国際競争力のある国家レベルのティア1インフラの構築を計画しています。
スイス	シミュレーションとデータサイエンスの強化に向け、フラッグシップ機「Alps」において従来のHPC性能とクラウドの柔軟性を融合したクラウドネイティブ・アーキテクチャを採用しています。「Science as a Service（サービスとしての科学）」の実現を掲げ、研究の初期段階から解決策の提示までの時間を短縮する施策を推進しています。またデータサイエンスセンターの拡張による科学的支援の体制構築を進めています。
イタリア	「ITALIA AI Factory」を軸に、スタートアップ、企業、公共機関などを対象とした計算資源の提供とエコシステム支援を展開しています。利用にあたっては、専門家による個別のアセスメントを通じてニーズに応じた最適なサービスパッケージや支援内容をまとめたレポートを作成する仕組みを整えています。地球物理学のデジタルツイン開発や、AIを用いた人文科学・文化遺産の翻刻など専門的な支援が特徴です。
フランス	「AIクラスター」と呼ばれる9つの卓越した教育・研究拠点を設置し、世界トップクラスのAI人材育成と研究強化を推進しています。また、データセンターを「国家的に重大な関心を持つプロジェクト（PINM）」の対象に含める新法案により、認可手続きの大幅な期間短縮を目指しています。さらに、国立AI評価・セキュリティ研究所（INESIA）を通じて、AIモデルの安全性評価やシステムリスク分析を主導しています。
フィンランド	スーパーコンピュータ「LUMI」を中核とし、HPC、AI、ハイパフォーマンス・データ分析（HPDA）を高次元に融合させた計算環境を提供しています。利用制度として、テスト用のベンチマーク・アクセスから超大規模公募まで習熟度に応じた多様なチャネルを運用しています。また企業向けには、成果を非公開にできる有料利用モデルや、無償で事前に適合性を試せる「Try&Buy」プログラムを用意しています。

AI for Science基盤モデルの事例

AdvanceSoft
http://www.advancesoft.jp/

分野	①知識体系化と知見抽出　※1	①知識体系化と知見抽出 ②現象予測と計算加速	③実験・検証の自律化
材料	MatBERT(2022), MatSciBERT(2022), SPT(2022), ChemFormer(2022), MaterialsBERT(2023), HoneyBee(2023), CatBERTa(2023), SolvBERT(2023), SELFormer(2023), MoLXPT(2023), MatChat(2023), ChemDFM(2024), LLaMat / LLaMat-Chat(2024), CrystaLLM(2024), MatterGPT(2024), FlowLLM(2024), LLaMat-CIF(2024), AtomGPT(2024), CSLLM(2024), SynAsk(2025)	**GNoME(2023),** Text+Chem T5(2023), Regression Transformer(2023), MultiMat(2023), DiffCSP(2023), MatterGen(2023), **MatterGen(2023),**ChatMOF(2023), **MACE-MP-0(2024),** HoneyComb(2024), LLMatDesign(2024), MatPilot(2024), ChemCrow(2023), MoL-MoE(2024), nach0(2023), **MatterSim(2024),** DiffCSP++(2024), CrystalFormer(2024), Con-CDVAE(2024), MatAgent(2025), LLM-Fusion(2025), MatterChat(2025), **UMA(2025)**	A-Lab(2023), ChemOS(2020), NIMS-OS(2023), MatPilot(2024), ChemCrow(2023)
ライフ	**BioBERT(2019),** iDNA-ABF(2022), RNABERT(2022), RNA-FM(2022), ProteinBERT(2021), ProtGPT2(2022), ZymCTRL(2022), scBERT(2021), scMVP(2022), **ESM2(2022),** ProGen2(2022), DrugBAN(2022), SpliceBERT(2023), **HyenaDNA(2023), scGPT(2023),** scTranslator(2023), TOSICA(2023), **DNABERT-2(2023),** Nucleotide Transformer(2024), Evo(2024), RNA-MSM(2024), GenerRNA(2024), scFoundation(2024), mvTCR(2024)	MolCLR(2021), OntoProtein(2022), GPN(2022), **GeneFormer(2023),** Mole-BERT(2023), KPGT(2023), DeepMAPS(2023), SiGra(2023), DeepSEED(2023), Bert2Ome(2023), MarsGT(2023), scPROTEIN(2024), **AlphaFold3(2024),** VQDNA(2024), RfamGen(2024), scButterfly(2024), MIDAS(2024)	EVOLVEpro(2025), The Virtual Lab of AI agents (2025)
気象・環境	**Pangu-Weather(2022), FourCastNet(2022),** SimVP(2022), **GraphCast(2023),** FengWu(2023), FuXi(2023), **AtmoRep(2023),** Stormer(2023), **Climax(2023),** GenCast(2023), SFNO(2023), **NeuralGCM(2023),** NowcastNet(2023), **NeuralGCM(2023),** MetNet-3(2023), **Prithvi-EO(2023),** Prediff(2023), DiffCast(2023), **GenCast(2023),** AI-GOMS(2023), **Aurora(2024),** AIFS(2024), AIFS-CRPS(2024), GraphDOP(2024), HEAL-ViT(2024), WeatherGFT(2024), FengWu-W2S(2024), ACE2(2024), ClimODE(2024), DeepPhysiNet(2024), Ola(2024), XiHe(2024), Samudra(2024), **GLONET(2024),** LangYa(2024), **Earth-2 (2024), WV-net(2024),** DLESYM(2024), SamudrACE(2025), WenHai(2025), KIST-Ocean(2025), FuXi-Ocean(2025), ecLand-emul(2025), NoahMP-AI(2025), **WeatherNext 2(2025), TerraMind(2025)**		EarthLink(2025)

※1：①への分類は、Fei Guo et al., “Foundation models in bioinformatics”, National Science Review, Vol. 12, nwaf028, 2025.等における「Language FMs」の分類とした。

別途添付した「表A.1基盤モデル一覧」に詳細を記載した。

AdvanceSoft
http://www.advancesoft.jp/

AI for Science基盤モデルのパラメータ数

■ バイオ

モデル名	開発元	アーキテクチャ	パラメータ数
ESM-3	EvolutionaryScale	Multimodal Transformer	98,000,000,000
HyenaDNA	Stanford Univ.	Hyena Operator	1,600,000
DNABERT-2	UCSD 等	Encoder-only Transformer	117,000,000
scGPT	Toronto Univ.	Generative Transformer	53,000,000
Geneformer	Broad Institute	Encoder-only Transformer	316,000,000
AlphaFold3	Google DeepMind	Diffusion / Pairformer	500,000,000

■ 材料

モデル名	開発元	アーキテクチャ	パラメータ数
GNoME	Google DeepMind	Message Passing GNN	5,000,000
MatterSim	Microsoft	Equivariant GNN	130,000,000
MACE-MP-0	Cambridge Univ.	Higher-order MPNN	4,690,000
MatterGen	Microsoft	Diffusion Equivariant GNN	46,800,000
UMA ※1	Meta (FAIR)	Equiformer V2 / MACE	1,400,000,000

■ 気象等

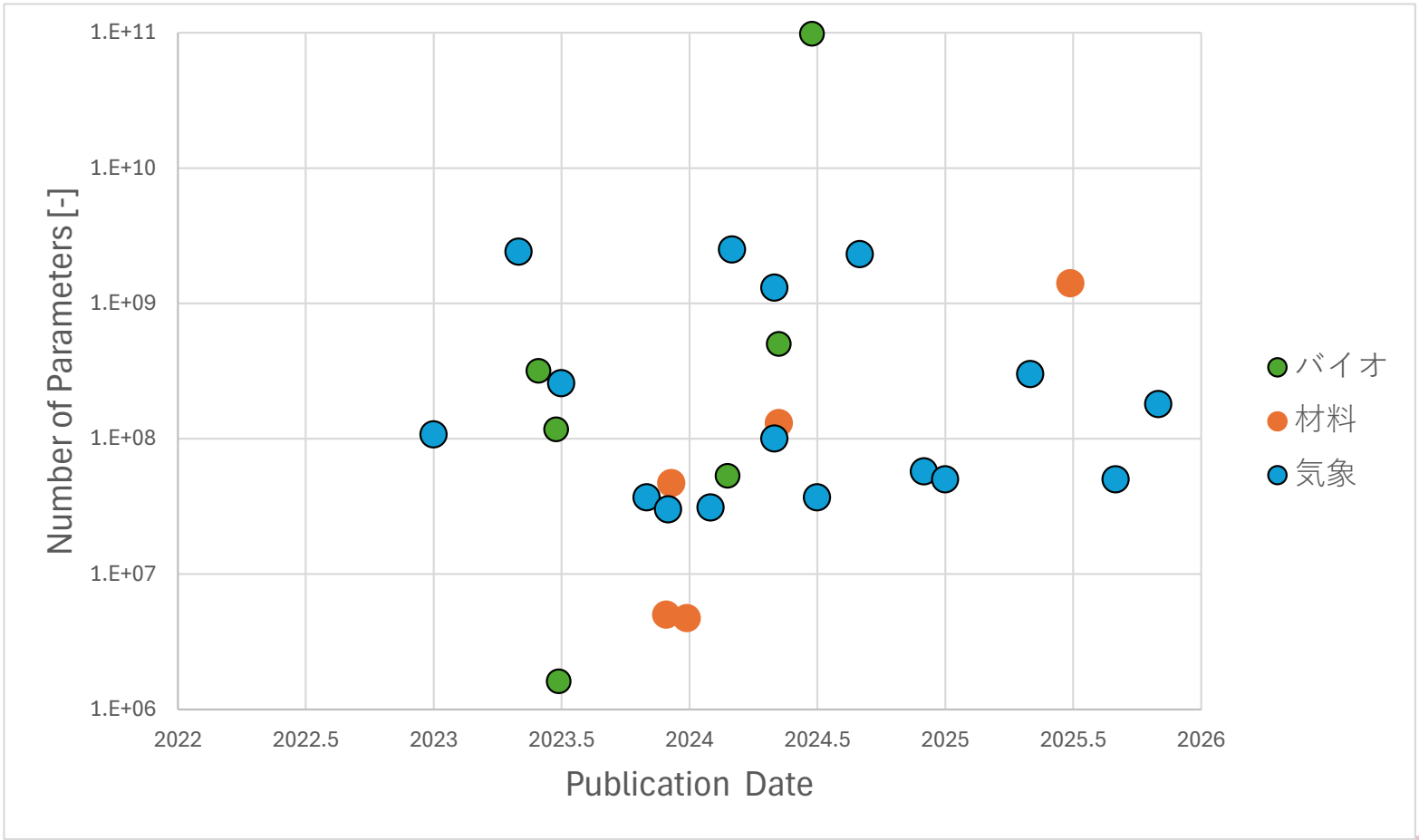
モデル名	開発元	アーキテクチャ	パラメータ数
Earth-2 Medium Range	NVIDIA	Diffusion Transformer	2,500,000,000
FourCastNet3	NVIDIA	AFNO (Fourier/CNN)	300,000,000
WeatherNext2	Google DeepMind	FGN / 生成AIベース	180,000,000
NeuralGCM	Google DeepMind	Hybrid (物理 + GNN/CNN)	36,700,000
GenCast	Google DeepMind	Diffusion / GNN	57,000,000
GraphCast	Google DeepMind	GNN+Multimesh	36,700,000
Aurora	Microsoft	3D Swin Transformer	1,300,000,000
ClimaX	Microsoft	Swin Transformer	107,000,000
AIFS ENS	ECMWF	GNN	100,000,000
AtmoRep	ECMWF + JSC +CERN	Transformer(Multiformer)	2,400,000,000
Prithvi-WxC	IBM + NASA	ViT(MAE)	2,300,000,000
Pangu-Weather	Huawei	Swin Transformer	256,000,000
Granite-Geospatial-Ocean	IBM	ViT(MAE)	50,000,000
GLONET	Mercator Ocean	CNN	30,000,000
WV-Net	Hawaii Univ	ViT(MAE)	31,000,000
TerraMind	IBM + ESA	TiM	50,000,000
Prithvi-EO	IBM + NASA	ViT(MAE)	600,000,000

※1：UMA量子化学データセットOMol25は60億コア時間（※2）で1億ケースのDFT計算の結果であり、学習データ作成にシミュレーションを利用した。

※2：1コア当たり実効速度12GFLOPSとすると、60億コア時間=7.2e19FLOP/s × 3600s=0.1EFLOPS × 720時間(1ヶ月)である。★

別途添付した「表A.1基盤モデル一覧」に詳細を記載した。

AI for Science基盤モデルのパラメータ数



別途添付した「表A.1基盤モデル一覧」に詳細を記載した。

AdvanceSoft
http://www.advancesoft.jp/

AI for Science基盤モデルの学習データサイズ

■ バイオ

モデル名	学習データセット	学習データ量	データサイズ
ESM-3	タンパク質配列・構造・機能	~27.8億 配列	約 1.0 ～ 2.0 TB
HyenaDNA	ヒトゲノム配列 (HG38)	32億 塩基対	数GB
DNABERT-2	多種ゲノム配列 (RefSeq)	~300億 塩基対	約 5.0 ～ 10 TB
scGPT	単一細胞RNA発現データ	約3,300万細胞 / 7,710億トークン	約 2.0 ～ 4.0 TB
Geneformer	単一細胞RNA発現データ	約9,500万～1億300万細胞	約 6.0 ～ 10 TB
AlphaFold3	PDB 構造データ、リガンド等	17万件超の構造 / 数億件の配列	約 3 TB

■ 材料

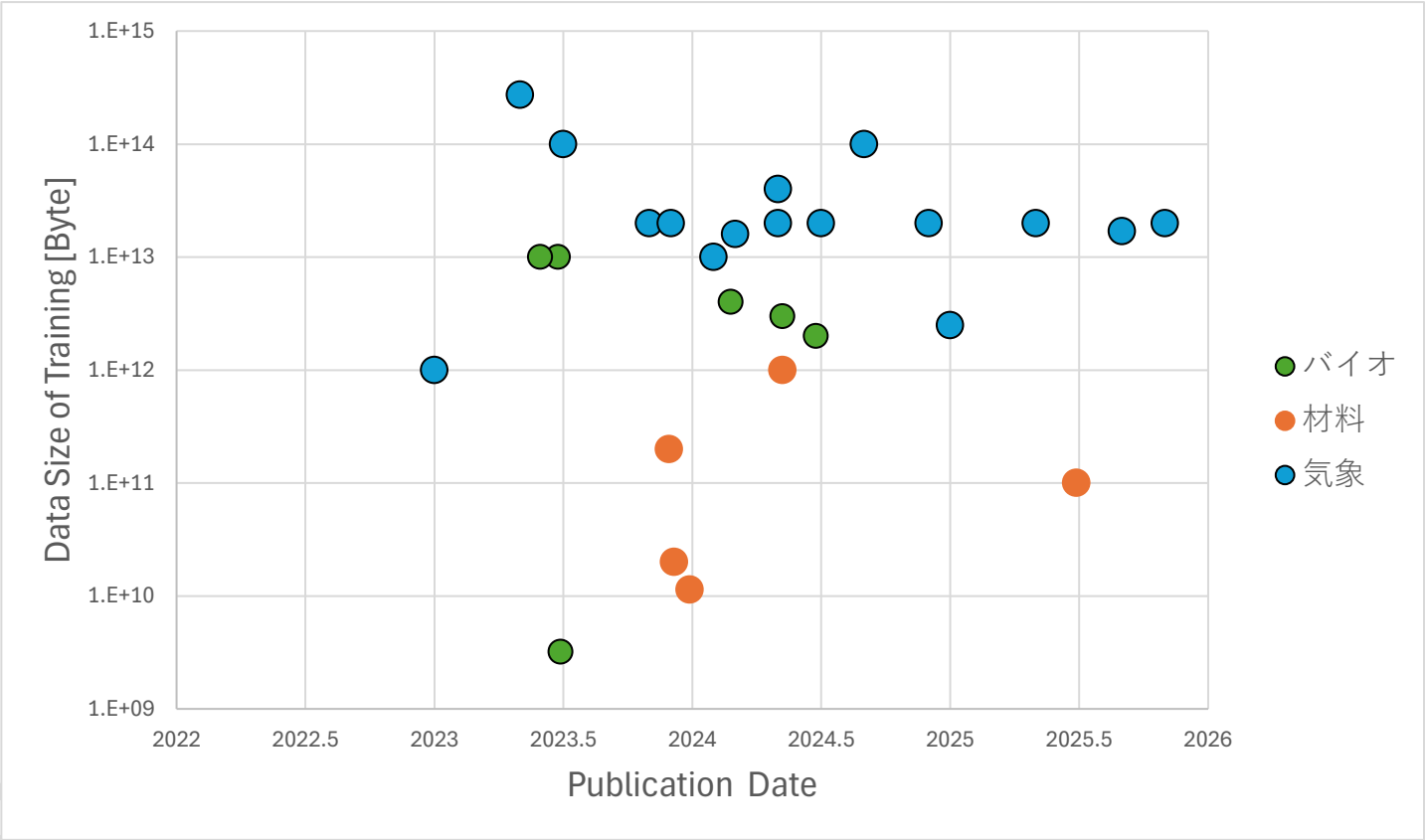
モデル名	学習データセット	学習データ量	学習データ量
GNoME	Materials Project, ICSD, OQMD	2.8万件 / 220万種	約 0.1 ～ 0.2 TB
MatterSim	DFT計算データ	約1,700万件	約 0.5 ～ 1.0 TB
MACE-MP-0	Materials Project Trajectory(MPtrj)	約160万件	11.35GB
MatterGen	Materials Project, Alexandria	約60万8,000件	72GB
UMA	3D構造データ、DFT計算結果	1億件の解析結果	約 0.05 ～ 0.1 TB

■ 気象等

モデル名	学習データセット	学習データ解像度	データサイズ
Earth-2 Medium Range	ERA5 + 高解像度simulation	110,960pnts	16TB
FourCastNet3	ERA5 / 40年分	6時間刻み鉛直13層	数十TBと推測
WeatherNext2	ERA5, HRES-fc0	1時間刻み	数十TBと推測
NeuralGCM	ERA5, HRES-fc0	6時間刻み鉛直32層	数十TBと推測
GenCast	ERA5, HRES-fc0	6時間刻み鉛直13層	数十TBと推測
GraphCast	ERA5, HRES-fc0	6時間刻み鉛直37層	数十TBと推測
Aurora	ERA5, CMIP6, IFS, GFS	6時間刻み鉛直13層	数十TBと推測
ClimaX	CMIP6	6時間刻み鉛直13層	約 1 TB
AIFS ENS	ERA5＋IFS	6時間刻み鉛直13層	約 20 ～ 40 TB
AtmoRep	ERA5	1時間刻み最大鉛直137層	約273 TB
Prithvi-WxC	MERRA-2 (NASA再解析)	1時間刻み鉛直16-25層	数百TB
Pangu-Weather	ERA5	1時間刻み鉛直13層	約 20 ～ 100 TB
Granite-Geospatial-Ocean	Sentinel-3 衛星データ	1日刻み空間300m	2.5 TB
GLONET	GLORYS12	1/12° 1日刻み鉛直50層	2～数十TB
WV-Net	GOAL I, WV Image, ERA5	画像1,000万枚	4～10TB
TerraMind	TerraMesh	気象データ・土地利用	17 TB
Prithvi-EO	HLS V2(30m)	1億	8～10 TB

別途添付した「表A.1基盤モデル一覧」に詳細を記載した。

AI for Science基盤モデルの学習データサイズ



別途添付した「表A.2基盤モデルを支えるデーター覧」にデータの詳細を記載した。

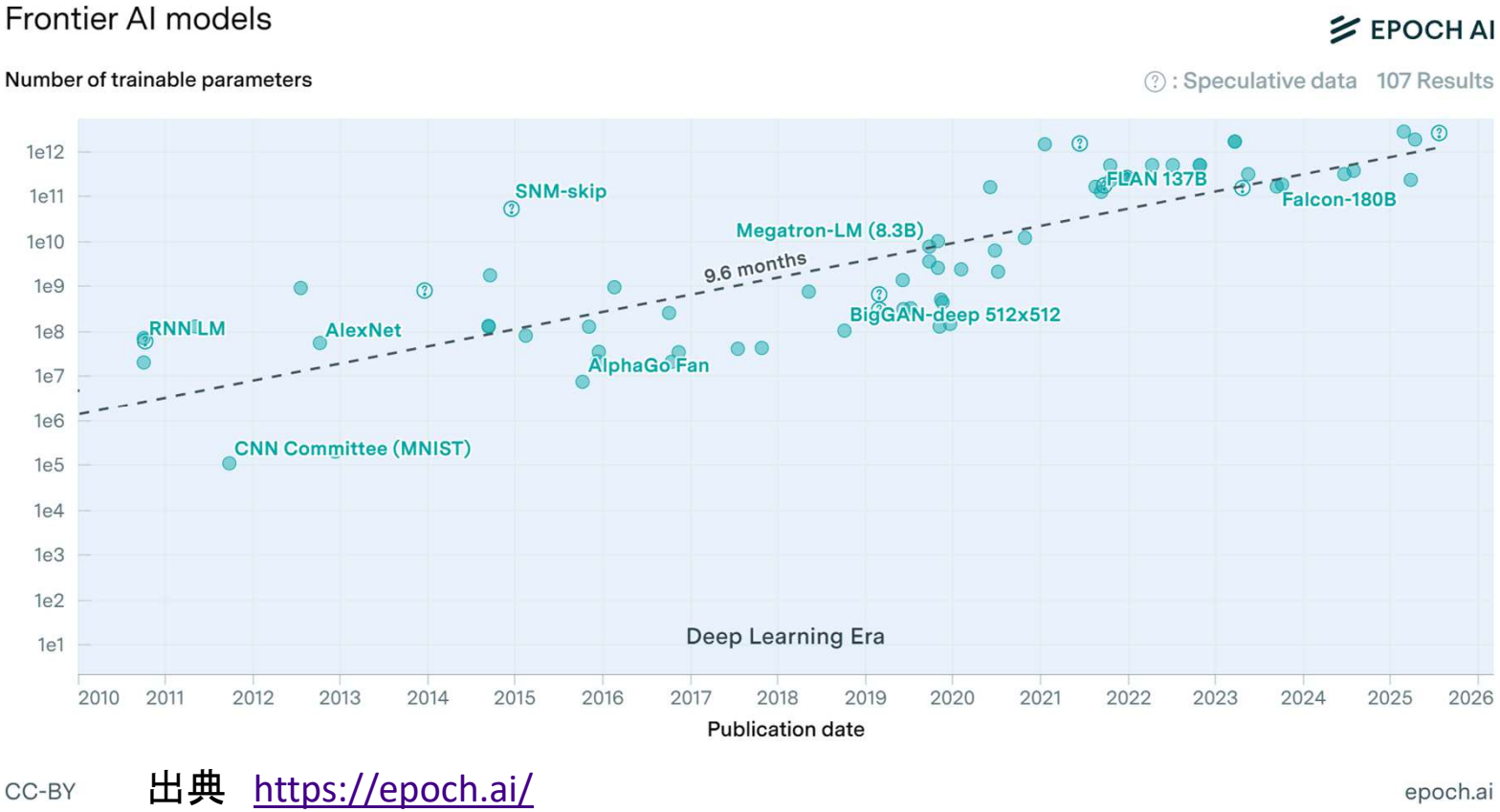
■ 代表的な機関のデータ量

分野	名称	機関	データ量（現在）	年間増分
バイオ	EMBL-EBI	欧州29か国	120 PB	10 PB
	DDBJ	遺伝研	22PB	2～3PB
	SRA,GenBank	NCBI(米国)	100PB	20～30PB
衛星データ	Coperunics	ESA（欧州）	86 PB	20 PB
	Earth Science Data Systems	NASA	178PB	30PB
	G-portal	JAXA	20～45PB	4～6PB
素粒子	CERN Open Data Portal	CERN	6～10PB	1～2PB

※ 参考：現在世界の観測衛星は1000機以上あり、例えば、日本のGOSAT-GW(2026～)は1TB/dayのデータを取得している。

（参考）汎用LLMのパラメータ数

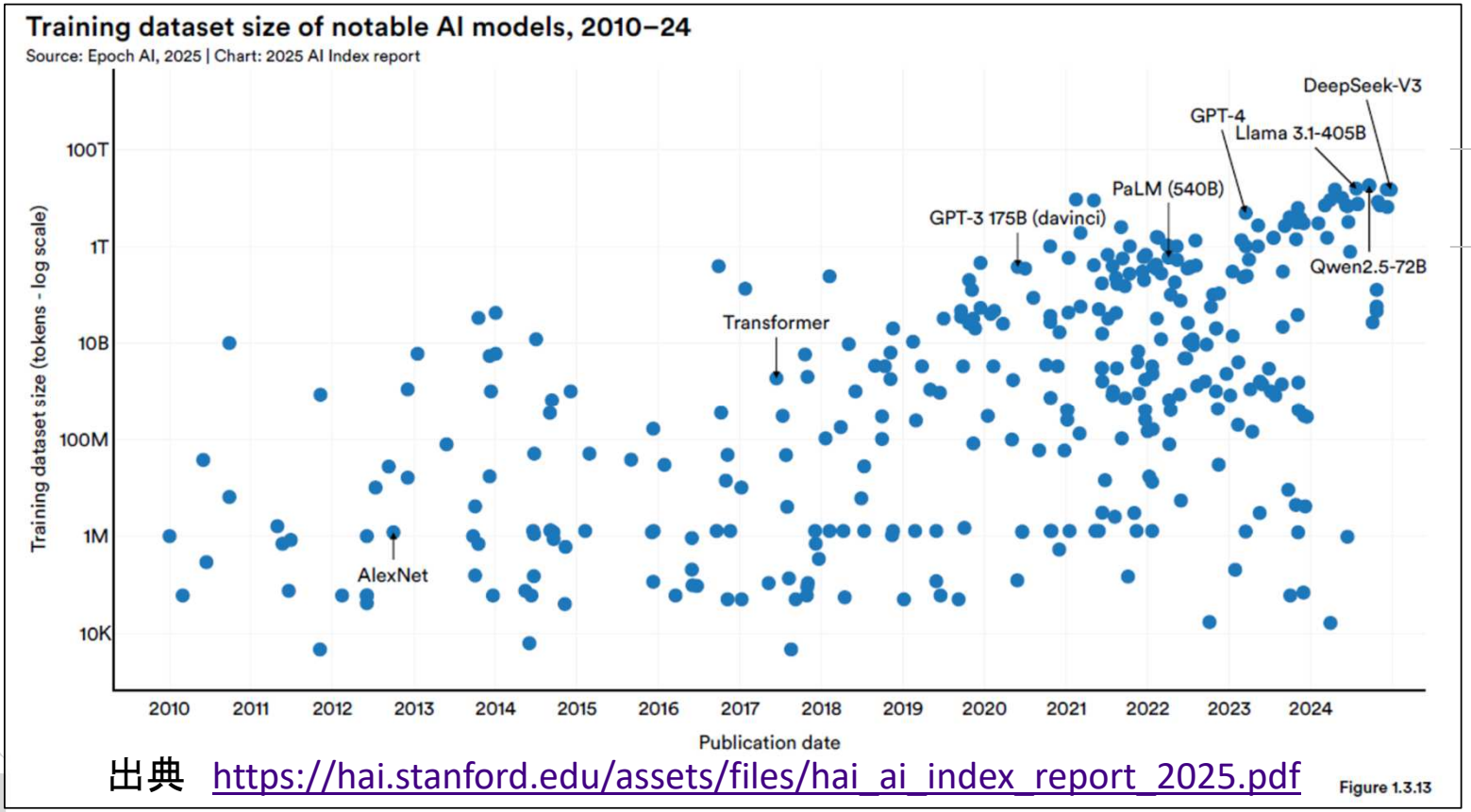
Artificial Intelligence Index Report 2025, Stanford Institute for Human-Centered Artificial Intelligence



- ✓ EPOCH AIのサイトにおいて、データ表示機能を利用して、公開年を横軸とし、パラメータ数を縦軸と指定して表示した図である。
- ✓ また、タイトルに示した文献にもパラメータ数の解説がある。
- ✓ パラメータ数は、9.6ヶ月で倍増する。

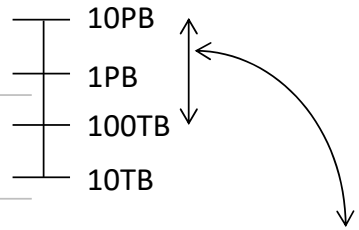
（参考）汎用LLMの学習データ量

Artificial Intelligence Index Report 2025, Stanford Institute for Human-Centered Artificial Intelligence



出典 https://hai.stanford.edu/assets/files/hai_ai_index_report_2025.pdf

1 token = 4 byte で換算
https://platform.openai.com/tokenizer

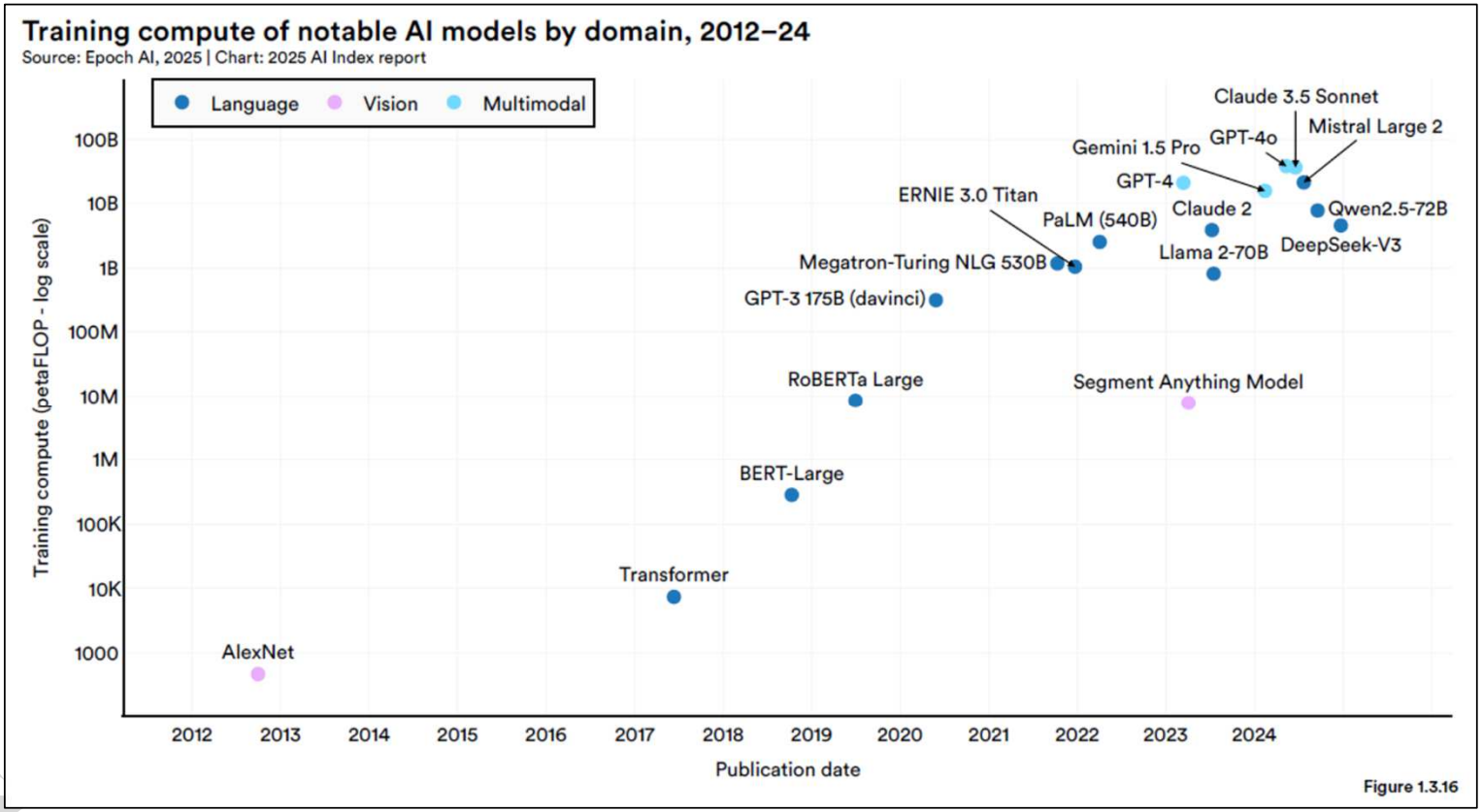


common crawlが200TB～10PBくらい
https://www.mozillafoundation.org/en/research/library/generative-ai-training-data/common-crawl/

大規模言語モデル（LLM）の学習用データセットのサイズは8カ月ごとに倍増☆
※ 文献のp.54に記載

（参考）汎用LLMにおける学習のための演算量

Artificial Intelligence Index Report 2025, Stanford Institute for Human-Centered Artificial Intelligence



出典 https://hai.stanford.edu/assets/files/hai_ai_index_report_2025.pdf

グラフはFP16でのFLOP

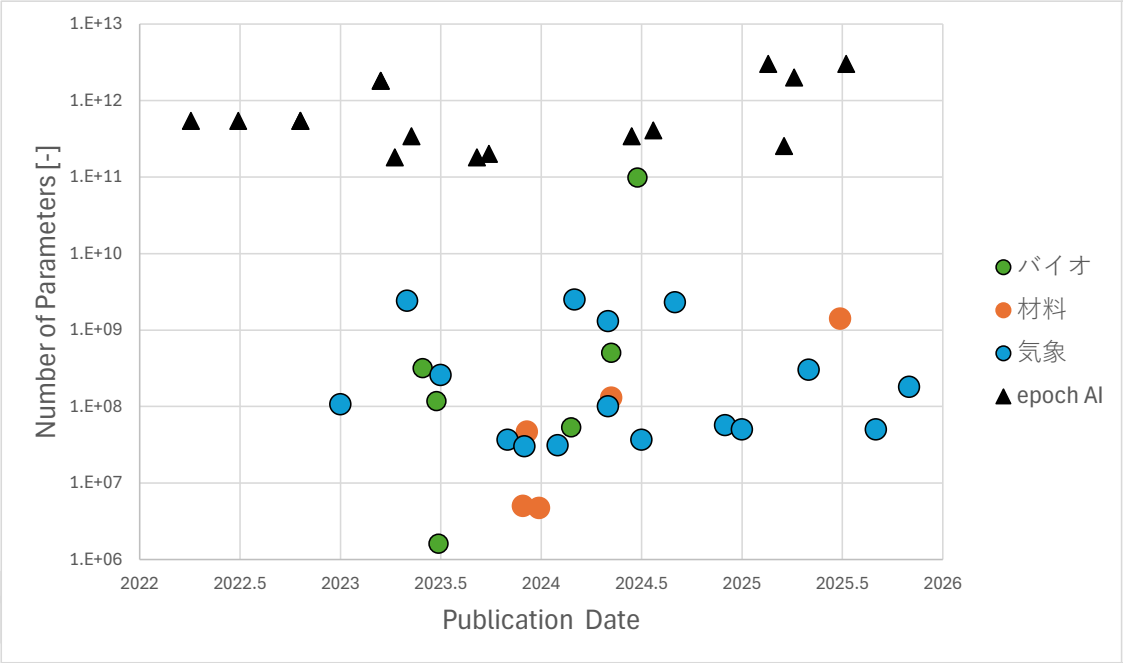
- 1.0 EFLOPS × 1000日
- 1.0 EFLOPS × 100日 ★
- 1.0 EFLOPS × 10日
- 1.0 EFLOPS × 1日

100M petaFLOP
= 100,000 EFLOP
= 1 EFLOPS × 100,000秒
= 1 EFLOPS × 1日

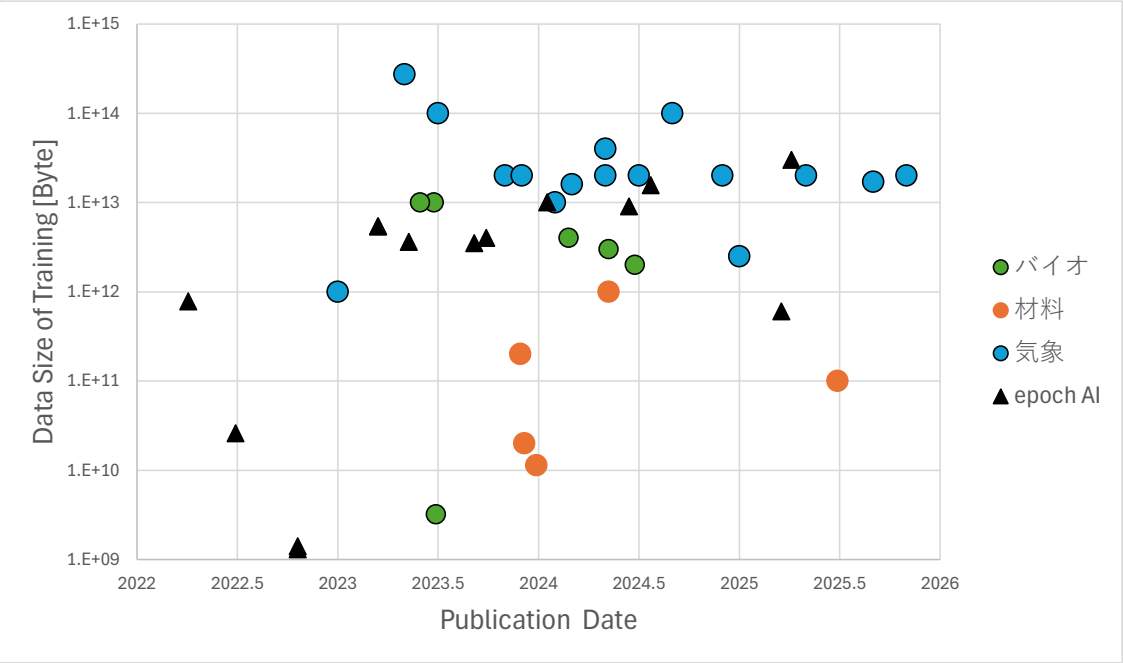
著名なAIモデルの学習に使用される計算資源量は約5カ月ごとに倍増☆
※文献]のp.56に記載

AI4SとLLMのパラメータ数・データ数の比較

- ✓ Epoch AI社の汎用LLMのパラメータ数・データ数と、AI4Sの基盤モデルを比較した。
  図中の▲はEPOCH AI社から公表されているデータである。
- ✓ AI4SはLLMよりもパラメータ数は1～3桁少なく★、学習データ量は同等である★。



縦軸：パラメータ数[個]、横軸：発表年



縦軸：データ数[バイト]、横軸：発表年

2030年に必要とされる計算資源

項目	現状（2026年）	成長トレンドの推定（今後4年間）	2030年予測値（1モデルあたり）	2030年予測値
演算（①知識体系化と知見抽出）	・AI4Sのパラメータは、汎用LLMより1～3桁少ない(p.14★) ・汎用LLMの学習時間はFP16で1EFLOPS×100日程度(p.13★)	学習の演算量は5カ月ごとに倍増（p.13☆） →4年間で750倍に増加	500EFLOPS × 0.15～15日と予測 ※a	500EFlops程度（FP16）※b
演算（②現象予測と計算加速）	・AI4Sのためシミュレーションは最大0.1EPLOPS×1ヶ月(p.7下★)	・演算量増はデータ増と同程度(下記)と推定する	6.4EPLOPS×1ヶ月と予測※c	10EFLOPS程度（FP64）
ストレージ	・AI4Sで学習に利用するデータは100TB～1PB程度（汎用LLMと同程度）(p.14★)	・学習に利用するデータは8カ月ごとに倍増(p.12☆) →4年間で64倍に増加	AI4Sのストレージ需要は6.4～64PBと予測	640PB程度

項目	①知識体系化と知見抽出	②現象予測と計算加速	③実験・検証の自律化
演算性能	GPU: 500EFlops（FP16）	GPU+CPU: 10EFLOPS（FP64）	③≪①②
ストレージ量	640PB		③≪①②

※a：1EFLOPS×100日×（1/10～1/1000）×750
=500EFLOPS×0.15～15日
※b：ユーザ数増の予測とユーザ使用量の分布を仮定し、500EFLOPS×(0.5～1)年と予測した。
※c：0.1EFLOPS×1ヶ月×64=6.4EFLOPS×1ヶ月

①②を発展させながら、
長期的には③の実現を目指す。

（補足）2030年にFP16で500EFlopsは何GPUか？

AdvanceSoft
http://www.advancesoft.jp/

スパコン名	GPU総数（基）	1ノードあたりの構成	FP16 Rpeak (EFlop/s)	搭載アクセラレータの種類	HPL-MxP性能 (EFlop/s)	FP16 Rpeak (EFlop/s)	(参考) FP64 Rpeak
El Capitan	44,544	4基 / ノード	29.1	AMD Instinct MI300A (APU)	16.70	約 25.0 ~ 30.0	2.74 EFlop/s
Aurora	63,744	6基 / ノード	53.4	Intel Data Center GPU Max	11.64	約 20.0 ~ 22.0	1.98 EFlop/s
Frontier	37,632	4基 / ノード	14.5	AMD Instinct MI250X	11.39	約 18.0 ~ 20.0	1.71 EFlop/s
JUPITER	24,000	4基 / ノード	23.7	NVIDIA GH200 (Superchip)	6.25	約 10.0 ~ 12.0	1.25 EFlop/s

世代	代表チップ名	登場時期（予定）	FP16 Rpeak (1基あたり)
Hopper	H200(H100)	2023年	約 1.0 PFlop/s
Blackwell	B200	2024-25年	約 2.2 ~ 2.5 PFlop/s
Rubin	R100	2026年	約 4.0 ~ 5.0 PFlop/s
Feynman	F100 (仮)	2028年	約 8.0 ~ 10.0 PFlop/s

- ✓ 500EFlops=JUPITERの50倍
- ✓ 2030年にGPU性能は10倍
- ✓ したがって、2030年の500EFLOPS計算機のGPU総数は、12万基

2030年：AI for Scienceの先導的実装期へ

AdvanceSoft
http://www.advancesoft.jp/

大規模インフラの構築

演算基盤
次世代高性能GPUを大量集積した大規模インフラを構築する。

PB級データ共有基盤
AI学習の「燃料」となる高品質データを安全に流通させるプラットフォームを整備する。

最先端アーキテクチャの採用
多精度演算やCPU-GPU密結合などを採用し、国際トレンドに適合させる。

本資料全体を3項目にまとめた

演算・データ需要増加へ対応

FP16で500EFLOPS
2030年の高度なAI4S用基盤モデル開発に必要な演算能力を算出。

12万基のGPU規模
上記の莫大な演算量を実現するため、「AI Gigafactory」級インフラを想定。

最大640PBのストレージ
AI4S用基板モデル構築のために増大する科学データを蓄積するプラットフォームを試算。

本資料のp.15,p.16で試算した内容をまとめた。

伴走型支援とキャリア確立

技術者の育成・確保
GPU等に対応できる高度人材を育成し、そのためのキャリアパスを国内でも確立する。

レガシーコードの最適化
FORTRAN等の膨大な既存コード資産を、次世代機へ移行・最適化する支援プログラムを強化する。

利用者支援の深化
日本固有の強みである手厚い支援体制をAIパラダイムへ進化させ、裾野を広げる。

本資料には掲載していないが、報告書の4.5節にまとめた

Executive Summary（再掲）

研究プロセスの変革

知識体系化
LLMによる膨大な情報統合・知識体系化で、時間的な制約や認識する内容の限界を超える。

シミュレーション加速
物理法則・解析結果などを学習しAIが代替することで、膨大なシミュレーション時間を短縮する。

実験・検証の自律化
AIで実験・検証を制御する自律的な24時間稼働ラボで、時間的な制約や人為的なミスを排除する。

本資料のp.3～4に記載した。

2030年の計算需要推計

500EFLOPS級の演算需要
パラメータ・データ増に対応し、FP16で500EFLOPS、GPU12万基規模のインフラが必要と推計。

640PBのストレージ需要
基盤モデルのための学習データを蓄積するために、640PB程度のストレージが必要と推計。

2030年への技術的展望
実効性能・電力効率を上げるため、混合精度演算やCPU-GPU密結合等の技術の進展。

本資料のp.5～13のデータをもとにp.14～16にまとめた。

戦略的実装への提案

2030年：先導的実装期へ
大規模インフラの構築
演算・データ需要増加への対応
伴走型支援とキャリア確立

将来目指すべき形
自律型研究エコシステムの確立
計算エコシステムの構築
持続可能なグリーン科学基盤

本資料のp.17にまとめ、報告書4.6節に記載した

表A.1 基盤モデル一覧

分野	モデル名	開発元	アーキテクチャ	学習データ	データ量	データサイズ	パラメータ数	公開日	掲載誌・プラットフォーム	ライセンス	主な用途・特徴	アーキテクチャの特徴
ライフ	ESM-3	EvolutionaryScale	Multimodal Transformer	タンパク質配列・構造・機能	~27.8億 配列	約 1.0 〜 2.0 TB	980億	2024年6月	arXiv公開（2025年1月にScience誌掲載）	ESM Open License (※1)	タンパク質の配列・構造・機能を統合的に扱い、狙った役割を果たす最適な配列を設計・生成することでプログラマブル・バイオロジーを実現する。	980億パラメータを持ち約27.8億個のデータを学習。既存の枠組みを超えて人工の蛍光タンパク質設計に成功するなど、圧倒的な生成・予測精度を誇る。
ライフ	HyenaDNA	Stanford Univ.	Hyena Operator	ヒトゲノム配列 (HG38)	32億 塩基対	数GB	（160万）	2023年6月	arXiv公開（ICML 2023 採択）	Apache 2.0 (商用可)	最大100万塩基の長配列を一度に処理し、離れた遺伝子間の相互作用や制御領域、種の分類などをシングルヌクレオチド解像度で精密に解析・予測する。	Transformerの弱点を克服するHyena階層を採用し、計算量が長さに比例。従来の数百倍の文脈を高速に理解し、微細な情報の変化を捉える。
ライフ	DNABERT-2	UCSD 等	Encoder-only Transformer	多種ゲノム配列 (RefSeq)	~300億 塩基対	約 5.0 〜 10 TB	1億1,700万	2023年6月	arXiv公開（ICLR 2024 発表）	Apache 2.0 (商用可)	多種ゲノムを対象にプロモーター特定やメチル化検出、変異の影響予測を行い、ヒトを含む130種以上の生物に共通する進化的なパターンを解析する。	BPEトークン化を採用し、DNAのサブワードを効率的に抽出。1.17億の軽量パラメータで、従来比56倍の高速な学習と最先端の解析能力を両立する。
ライフ	scGPT	Toronto Univ.	Generative Transformer	単一細胞RNA発現データ	約3,300万細胞 / 7,710億トークン	約 2.0 〜 4.0 TB	約5,300万	2024年2月	Nature Methods（オンライン公開）	MIT (商用可)	シングルセル解析で細胞型の自動判別、バッチ補正、遺伝子摂動予測、遺伝子制御ネットワークの推論を一手に引き受け、個別化医療の進展に寄与する。	約3300万個の細胞データを学習した初の生成AI。数万個の遺伝子間の相互作用を捉える5300万パラメータの脳を持ち、生物学的な直感で変化を捉える。
ライフ	Geneformer	Broad Institute	Encoder-only Transformer	単一細胞RNA発現データ	約9,500万〜1億300万細胞	約 6.0 〜 10 TB	3億1600万 (V2 Large)	2023年5月	Nature（オンライン公開）	MIT (商用可)	遺伝子同士の関係性や疾患によるネットワークの乱れを深く理解し、遺伝子操作の影響シミュレーションや疾患に関わる重要因子の特定、転移学習を行う。	世界最大級の約1億個の細胞データを学習。心不全等の研究で有効性が実証されており、少量のデータから高精度な解析が可能な優れたデータ効率を持つ。
ライフ	AlphaFold3	Google DeepMind	Diffusion / Pairformer	PDB 構造データ、リガンド等	17万件超の構造 / 数億件の配列	約 3 TB	5億	2024年5月	Nature（オンライン公開）	非営利研究限定	タンパク質、DNA、RNA、リガンド、イオン等を含む巨大な複合体の3D構造をアミノ酸配列のみから一つのシステムで同時に高精度に予測する。	拡散モデルを導入し、原子の位置をノイズから復元。低分子化合物との結合予測精度を50%以上向上させ、創薬プロセスへの強力な支援能力を発揮する。
分野	モデル名	開発元	アーキテクチャ	学習データ	データ量	データサイズ	パラメータ数	公開日	掲載誌・プラットフォーム	ライセンス	主な用途・特徴	アーキテクチャの特徴
材料	GNoME	Google DeepMind	Message Passing GNN	Materials Project, ICSD, OQMD	2.8万件(基礎) / 220万種(予測)	約 0.1 〜 0.2 TB	数百万規模	2023年11月	Nature（オンライン公開）	Apache 2.0 (商用可)	グラフニューラルネットワークを用い、未知の無機結晶構造の安定性を予測することで、次世代電池等のための新材料をコンピュータ上で高速に発見する。	220万個の新材料を予測し38万個の安定物質を特定。AIの予測と物理計算を循環させるアクティブラーニングで、探索速度を指数関数的に向上させた。
材料	MatterSim	Microsoft	Equivariant GNN	DFT計算データ	約1,700万件	約 0.5 〜 1.0 TB	1億3000万	2024年5月	arXiv公開（Microsoft Research発表）	MIT (商用可)	全元素に対応し、絶対零度から数千度の高温高圧といった極限環境下における物質の挙動や物性を、第一原理計算に匹敵する精度で動的にシミュレートする。	1.82億パラメータの脳を持ち、AIが自信のない領域を物理計算で自ら補強。少量のデータで実験値に近い精度へ微調整可能な「デジタルツイン」を推進。
材料	MACE-MP-0	Cambridge Univ.	Higher-order MPNN	Materials Project Trajectory (MPtrj)	約160万件	11.35GB	469万	2023年12月	arXiv公開	MIT (商用可)	周期表の89元素をカバーし、結晶、液体、ガス等の多様な状態で、原子にかかる力、エネルギー、応力をDFTに近い精度で圧倒的スピードで計算する。	高次の等変メッセージパッシング技術により多体相互作用を正確に記述。160万件のバルク結晶データを学習し、専門領域へは少量の追加学習で即座に適合。
材料	MatterGen	Microsoft	Diffusion Equivariant GNN	Materials Project, Alexandria	約60万8,000件	72GB	4,680万	2023年12月	arXiv公開（Microsoft Research発表）	研究用限定 (Microsoft)	指定された磁性や硬度などの物性から逆算して、それに合致する新しい無機材料の構造をゼロから生成する「逆設計（インバース・デザイン）」を実現する。	拡散モデルを応用し、ノイズから原子の種類や座標を最適化。従来のリストから探す方式に比べ、新規かつ安定な材料をピンポイントで提案する点が画期的。
材料	UMA	Meta (FAIR)	Equiformer V2 / MACE	3D構造データ	50億個以上の原子	約 0.05 〜 0.1 TB	14億	2025年6月	arXiv公開（Meta FAIR発表）	CC-BY-NC 4.0 (非営利限定)	分子、材料、触媒などの化学の複数領域を横断的にカバーし、大規模な原子ポテンシャルの構築や高速な特性予測、複雑な化学現象のシミュレーションを行う。	50億個以上の原子を含む3D構造を学習。14億パラメータの大規模モデルでありながら、Mixture of Linear Experts構造により推論速度が非常に高速。
分野	モデル名	開発元	アーキテクチャ	学習データ	データ量	データサイズ	パラメータ数	公開日	掲載誌・プラットフォーム	ライセンス	主な用途・特徴	アーキテクチャの特徴
気象	Earth-2 Medium Range	NVIDIA	Diffusion Transformer	ERA5 + 高解像度シミュレーション	110,960pnts	16TB	25億	2024年3月	NVIDIA Blog / arXiv	非公開	0.25° の空間解像度で全球の地表・上空予測を行い、最長15日先までのアンサンブル予報を数秒で完了させ、意思決定を迅速に支援する。	25億パラメータのDiffusion Transformerを採用。16TBのデータを学習し、既存のスパコンを上回る精度と圧倒的な推論効率を両立している。
気象	FourCastNet3	NVIDIA	AFNO (Fourier/CNN)	ERA5 / 40年分	6時間刻み 鉛直13層	数十TBと推測	3億	2025年5月	NVIDIA Blog / arXiv	Apache 2.0	熱帯低気圧の進路、豪雨、暴風、極端な気温変化といった異常気象の発生を、最長60日先まで高速かつ高解像度で確率的に予測する。	7億パラメータのCNNアーキテクチャを採用。H100 GPU 256基により短期間で学習。確率的な予測を高速に行うAFNO技術を駆使している。
気象	WeatherNext2	Google DeepMind	FGN / 生成AIベース	ERA5, HRES-fc0	1時間刻み	数十TBと推測	1億8,000万	2025年11月	Google Blog / GIGAZINE	MIT	最長15日の全球アンサンブル予報に対応。地域の詳細化から広域予測まで幅広く対応し、TPUインフラを駆使して地球規模の気象理解を加速させる。	1.8億パラメータのFunctional Generative Networkを採用。TPU v5p等490基で学習。1分弱という高い推論効率と精度を誇る。
気象	NeuralGCM	Google DeepMind	Hybrid (物理 + GNN/CNN)	ERA5, HRES-fc0	6時間刻み 鉛直32層	数十TBと推測	3,670万	2024年7月	Nature (2024)	Apache 2.0	物理的な数値予報とCNNを融合させ、気象・気候シミュレーションを従来比3500倍以上高速化し、物理法則に基づいた精緻な全球予測を実行する。	3670万パラメータのハイブリッドモデル。物理的な整合性とAIの学習能力を両立させ、スパコンのリソースを劇的に節約する革新的なモデル。
気象	GenCast	Google DeepMind	Diffusion / GNN	ERA5, HRES-fc0	6時間刻み 鉛直13層	数十TBと推測	5,700万	2024年12月	arXiv (2024)	Apache 2.0	拡散モデルを用いた生成AIにより、台風や熱波などの極端気象や地域風力発電の出力を確率的に高精度で予測し、アンサンブル予報の新境地を開く。	約5700万パラメータを搭載。GNNと拡散モデルを統合したアーキテクチャにより、複雑な気象現象を確率的な分布として捉える能力に優れている。
気象	GraphCast	Google DeepMind	GNN+Multimesh	ERA5, HRES-fc0	6時間刻み 鉛直37層	数十TBと推測	3,670万	2023年11月	Science (2023)	Apache 2.0	10日先の全球予測を1分未満で完了。異常気象の早期警戒、航空運航ルート最適化、電力需給管理など、実運用における決定論的予測を高度化する。	GNNを採用し3670万パラメータで構成。スパコンを凌ぐ精度を低コストで実現し、早期警戒システムなど実用化実績が豊富な先駆的モデル。
気象	Aurora	Microsoft	3D Swin Transformer	ERA5, CMIP6, IFS, GFS	6時間刻み 鉛直13層	数十TBと推測	13億	2024年5月	arXiv / MS Research	MIT	多種多様な大気データを統合し、10日先までの全球予測を決定論的に実行。爆弾低気圧の進路予測や世界規模の大気汚染予報などで高い成果を上げている。	13億パラメータの3D Swin Transformerを採用。ERA5やCMIP6を含む複数の気象・気候データセットを横断的に学習した汎用性の高いモデル。
気象	ClimaX	Microsoft	Swin Transformer	CMIP6	6時間刻み 鉛直13層	約 1 TB	1億700万	2023年1月	ICML 2023	MIT	短期・中期予測から2ヶ月の中長期予測、更には地球全体の長期気候変動予測までをカバーし、広域予報から地域の詳細化まで幅広く対応する。	Vision Transformer（ViT）を基盤とした約1.07億パラメータのモデル。多様な大気データに適応し、気候監視における科学的根拠を提示する。
気象	AIFS ENS	ECMWF	GNN	ERA5＋IFS	6時間刻み 鉛直13層	約 20 〜 40 TB	1億	2024年5月	ECMWF News / arXiv	ECMWF Open	15日先までのアンサンブル予報を行い、気温、風速、降水確率などの気象変数を算出。ECMWFの運用において高効率な予測インフラとして機能する。	1億パラメータのGNNを採用。わずか4基のGPUで3日間という低リソースでの学習を実現。実用的なアンサンブル予測を低コストで提供する。
気象	AtmoRep	ECMWF + JSC + CERN	Transformer(Multiformer)	ERA5	1時間刻み 最大鉛直137層	約273 TB	24億	2023年5月	Science Adv. / NeurIPS	CC-BY-4.0	短期気象予報のほか、降水量補正や時間解像度を超える高精度化を実現し、大気動態を大規模表現学習を通じて精密にモデリングする。	約24億パラメータのTransformer（Multiformer）構造を採用。CERN等と開発され、大気の状態を多角的に表現する確率的確率モデル。
気象	Prithvi-WxC	IBM + NASA	ViT(MAE)	MERRA-2 (NASA再解析)	1時間刻み 鉛直16-25層	数百TB	23億	2024年9月	arXiv (2024)	Apache 2.0	台風進路予測、熱波発生確率推計、CO2上昇に伴う長期気温変化のシミュレーション等、全球と地域の気象現象を高速・高解像度で予測する。	23億パラメータのVision Transformerを採用。NASAとIBMが共同開発。MERRA-2等のデータを用いて気候変動解析に特化している。
気象	Pangu-Weather	Huawei	Swin Transformer	ERA5	1時間刻み 鉛直13層	約 20 〜 100 TB	2億5,600万	2023年7月	Nature / arXiv / ECMWF (運用テスト)	非営利研究限定	台風の進路予測や航空・海運ルートの自動選定、都市部の浸水リスク早期警戒など、決定論的な全球予報を通じて実用的な意思決定を強力に支援する。	約2.56億パラメータのSwin Transformerを採用。最長60日先までの予測が可能。従来比で圧倒的な推論速度と高い精度を世界に示した。
海洋	Granite-Geospatial-Ocean	IBM	ViT(MAE)	海面温度・塩分濃度データ	2.56億 (256M)	2.5 TB	5,000万	2025年1月	Hugging Face (2025)	Apache 2.0	海面水位上昇、生態系解析、海流データに基づく低燃費航路、海氷変動の監視、船舶の安全航行支援を行い、地球規模の海洋環境理解に寄与する。	5000万パラメータのViTを採用。Sentinel-3データを学習。少量のサンプルでも高い精度を維持し、学習効率が極めて高い海洋特化型モデル。
海洋	GLONET	Mercator Ocean	CNN	全球気象データ / GLORYS12	海流、塩分、海水温、海面高度	2〜数十TB	3,000万	2023年12月	KDD 2023 / arXiv	研究用	海流、塩分、海水温、海面高度の予測を高速に行い、海水温の長期変動や海流動向の把握、オンデマンドの海洋予測シミュレーションに利用される。	約3000万パラメータのCNN。Mercator Oceanが開発。10日先の予測を10秒内で行う高い推論速度を備え、実運用レベルの解析を実現する。
海洋	WV-Net	Hawaii Univ	ViT(MAE)	GOAL I, WV Image, ERA5	画像1,000万枚	4〜10TB	3,100万	2024年2月	Remote Sensing (誌)	研究用	衛星画像から直接波高や海面温度を推定。風筋、冷水塊、雨といった物理現象を識別・監視し、海表面付近の動向を決定論的にモニタリングする。	3100万パラメータのViT。1000万枚のSAR衛星画像を用いた自己教師あり学習が特徴。画像から直接、物理量を推定する能力に優れる。
地球	TerraMind	IBM + ESA	TiM	TerraMesh	気象データ・土地利用	17 TB	5,000万	2025年9月	arXiv (2025)	研究用	森林の違法伐採監視、都市開発や農業用地の変化分析、洪水・山火事の被害把握など、地表の環境変化を精緻に捉え最適な土地活用支援に貢献する。	IBMとESAが開発した5000万パラメータのViT。17TBのデータを学習。地球観測データの欠測補完やセグメンテーション、特徴検出を行う。
観測	Prithvi-EO	IBM + NASA	ViT(MAE)	HLS V2(30m)	1億	8〜10 TB	6億	2023年5月	arXiv (2023)	Apache 2.0	自然災害の予見や被害解析のほか、収穫量予測、バイオマス推計、干ばつ監視、インフラモニタリング、避難経路の最適化など、リアルタイム支援を行う。	6億パラメータのVision Transformerを採用。NASAとIBMがHLS V2データを学習。多様な時間軸・地域に対応する汎用地理空間AI。
分野	モデル名	開発元	アーキテクチャ	学習データ	データ量	データサイズ	パラメータ数	公開日	掲載誌・プラットフォーム	ライセンス	主な用途・特徴	アーキテクチャの特徴

表A.2 基盤モデルを支えるデータ

カテゴリ		データベース名	中心的な機関		件数（目安）	データサイズ	年間増加量 (目安)	根拠リンク
ライフサイエンス	ゲノミクス	TCGA	NIH (NCI/NHGRI)	33種のがんを対象とした大規模なゲノム解析データ。変異や発現、臨床情報を統合した、がん研究に欠かせない包括的リソースである。	33がん種、2万サンプル	2.5 PB	プロジェクト終了	https://www.cancer.gov/ccg/research/genome-sequencing/tcga
ライフサイエンス	機能的ゲノミクス	ArrayExpress (BioStudiesに統合)	EMBL-EBI	次世代シーケンス等の遺伝子発現データの公共リポジトリ。機能ゲノミクス実験の詳細なメタデータと生データを、欧州を拠点に提供する。	実験: 77,585件 アッセイ: 2,429,810件	56.68 TB	–	https://www.ebi.ac.uk/arrayexpress/
ライフサイエンス	統合アーカイブ	BioStudies	EMBL-EBI	論文の全データを集約するデータベース。ArrayExpressを統合し、マルチモーダルな研究データの一括管理とリンクで再現性を支える。 (データ量は電子顕微鏡データEMPIREが支配的)	26.464.332 files; 10.369.537 links; 2.398.055	8PB以上	1PB以上	https://www.ebi.ac.uk/biostudies/
ライフサイエンス	機能的ゲノミクス	GEO	NCBI (NIH)	NCBIが運営する世界最大の機能ゲノミクスレポジトリ。公開されたマイクロアレイやNGSの生データを収集し、検索や再利用を可能にする。	Samples: 8,388,256件 Series: 280,160件	数百 TB	15%	https://www.ncbi.nlm.nih.gov/geo/
ライフサイエンス	機能的ゲノミクス	ENCODE	NHGRI (NIH)	タンパク質、RNA、遺伝子発現や細胞活動を制御する調節要素などヒトゲノムの必須要素を網羅した、包括的な機能ゲノミクスデータベース	実験： 23,439件 ファイル： 1,171,460件	1 PB以上	不明	https://www.encodeproject.org/
ライフサイエンス	プロテオミクス	UniProt	UniProt Consortium (EMBL-EBI,SIB,PIR)	タンパク質の配列情報と機能注釈を統合した世界標準のリソース。高精度なキュレーションデータにより、生物学研究の基盤を広く支える。	エントリ203,130,941件 (アミノ酸:約757億個)	1TB以下(2KB/エントリと仮定)	2%	https://www.uniprot.org
ライフサイエンス	プロテオミクス	PDB	wwPDB	タンパク質や核酸など、生体高分子の3次元立体構造データを集約。新薬開発や生命現象の解明など、世界の生命科学研究を支える必須の基盤。	250,496件の登録	366 TB	87 TB	https://www.rcsb.org/stats/data_storage_growth
ライフサイエンス	創薬	ChEMBL	EMBL-EBI	医薬品候補やバイオアクティブ分子の活性情報を収録。標的タンパク質との相互作用データを整理し、データ駆動型の創薬研究を強力に支援。	化合物: 2,425,720件 活性データ: 20,405,100件	10〜20GB (ChEMBL 34)	10%程度	https://www.ebi.ac.uk/chembl/
ライフサイエンス	創薬	ZINC	UCSF医薬化学科	バーチャルスクリーニングに特化した市販化合物の巨大DB。ドッキング計算に即時利用可能な3D構造や物理化学的特性データを公開。	370億 化合物	数十 TB	10%程度	https://zinc22.docking.org/
ライフサイエンス	創薬	PubChem	NCBI (NIH)	化学物質の名称、構造、生物活性を網羅した世界最大級のDB。化合物とタンパク質の関係や毒性情報を統合し、オープンに提供している。	化合物： 123,475,690 件 物質： 342,870,557 件	約1.4TB以上	約5〜10%	https://pubchem.ncbi.nlm.nih.gov/statistics/
ライフサイエンス	シングルセル	Single-Cell Exp. Atlas	EMBL-EBI	1,000万超の細胞を対象とした、シングルセル遺伝子発現の包括的DB。異なる組織や疾患間での発現差異を細胞レベルで精密に解析可能。	研究数: 383件 細胞数: 17,846,119個	数 TB以下	10%程度以上	https://www.ebi.ac.uk/gxa/sc/home
ライフサイエンス	シングルセル	Human Cell Atlas	HCA Consortium(Broad	ヒトの全身の細胞をマッピングする国際プロジェクトの成果。細胞の種類や状態、位置情報を網羅した、生命科学のデジタル地図を提供。	細胞数：1億1,100万個以上 プロジェクト数：約400件	830 TB	約 100 TB	https://data.humancellatlas.org/
カテゴリ		データベース名	中心的な機関		件数（目安）	データサイズ	年間増加量 (目安)	根拠リンク
材料	無機結晶構造	ICSD	FIZ Karlsruhe	世界最大の無機結晶構造データベース。1913年からの実験データを収録し、単位格子定数や原子座標、空間群などの詳細な情報を提供。	約 327,833 件	数百GB以下 (想定50KB/	約 1.2万件 (4%程度)	https://icsd.products.fiz-karlsruhe.de/
材料	結晶固体・材料特性	Materials Project (MP)	LBNL (米国)	第一原理計算により、既知および未知の無機結晶材料の構造や熱力学的安定性、電子状態などの物性データを網羅的に提供し、材料探索を支援する	20万件以上の材料	14 GB	10%程度	https://matbench-discovery.materialsproject.org/data
材料	材料特性・DFT	OQMD	Northwestern Univ.	130万件超の材料に関するハイスループットDFT計算結果を収録。熱力学的・構造的特性を網羅し、新材料探索や機械学習に活用。	1,407,395 件	10 TB以下 (数MB/件を想定)	10%程度	http://oqmd.org/
材料	データ管理	NOMAD	NOMAD Lab. (Max Planck等)	計算材料科学データを標準化し共有・解析する世界最大級の基盤。膨大な計算結果を公開し、研究の効率化やAIによる新材料探索を支援	19,295,396件	114.4 TB	不明	https://nomad-lab.eu/
材料	大規模材料DB	OMat24	Meta FAIR	1億1千万件以上のDFT計算結果を収録する、材料特性予測モデル訓練用の大規模な無機材料オープンソースデータセット。	約 1億1,800万 構造	185.67 GB	不明	https://hyper.ai/en/datasets/35287
材料	合成材料	SNUMAT	SNU	約1万件の合成材料とそのDFT計算特性を収録したDB。APIを通じてデータアクセスが可能であり、合成可能性の評価等に利用。	15,280件	100GB (数MB/件を想定)	不明	http://snumat.com/
材料	触媒・原子軌跡	OC20	Meta FAIR/ CMU	触媒表面の吸着エネルギーや原子軌跡を対象とした大規模データセット。原子間ポテンシャルの開発や複雑な反応シミュレーションに貢献。	DFT計算結果：2億6000万件以上	1.2 TB（ダウンロード時）	不定期にupdate	https://opencatalystproject.org/
材料	トラジェクトリ	MPTrj	Materials Project (LBNL, UCB)	aterials Projectの構造緩和過程を集約した158万件のデータセット。CHGNet等の汎用的な機械学習原子間ポテンシャルの学習に活用	1,580,395 件（構造データ数）	約11.35 GB	不定期にupdate	https://figshare.com/articles/dataset/CHGNet_MPTrj_Dataset/23457473
材料	分子構造 (DFT)	Alexandria	ICAMS	結晶構造DB。DFT計算による580万件超の3D/2D/1D材料データを集積。AIによる高速な材料探索と物性予測の基盤として世界最大級	約 580万 構造	数TBと推定(学習済で72GB)	1回のupdateで130万件	https://alexandria.icams.rub.de/
化学	有機分子	QM9	Prof. Raghunathan Ramakrishnan	最大9個の重原子を含む約13.4万個の小有機分子の量子化学的性質（幾何構造、エネルギー等）を網羅した機械学習用ベンチマークデータセット	133,885 件	172.68 MB	（固定）	https://hyper.ai/en/datasets/38089
カテゴリ		データベース名	中心的な機関		件数（目安）	データサイズ	年間増加量 (目安)	根拠リンク
気象・気候	気候再解析	ERA5	ECMWF	1940年から現在までの全球気象再解析データ。過去の状況を最新モデルで再現した標準セットとして、気候学研究で多用。	数兆レコード	10 PB	約 500 TB 〜 1 PB	https://www.ecmwf.int/en/forecasts/dataset/ecmwf-reanalysis-v5
気象・気候	気象予報	HRES	ECMWF	ECMWFによる最高解像度9kmの全球気象予報モデル。10日先までの詳細な予測を日々提供し、極端気象への早期警戒を支援。	(日次運用)	50 TB	(運用データのため膨大)	https://www.ecmwf.int/en/forecasts/datasets/set-i
気象・気候	気候予測	CMIP6	WCRP / PCMDI	IPCC報告書の根拠となる世界共通の気候予測。多様な将来シナリオに基づく地球システムモデルの結果を集約し対策策定を支える。	数百万 データセット	40 〜 80 PB	(フェーズ毎に数PB増加)	https://pcmdi.llnl.gov/CMIP6/
気象・気候	気候再解析	MERRA-2	NASA (GMAO)	NASAによる衛星データ同化に強みを持つ全球再解析。特にエアロゾルやオゾン等の大気組成の変化を精密に捉え、理解を深める。	数百万 ファイル	800 TB	約 15 〜 20 TB	https://gmao.gsfc.nasa.gov/reanalysis/MERRA-2/
気象・気候	長期再解析	JRA-3Q	気象庁 (JMA)	日本気象庁作成の1947年以降の最新全球再解析。台風やアジアのモンスーン、梅雨等の地域特性を正確に再現し気候監視に貢献。	75年分の一貫データ	473.33 TB	約 20 TB	https://jra.kishou.go.jp/JRA-3Q/index_ja.html
気象・気候	地上観測	GHCN	NOAA (NCEI)	全国の観測所から収集された地上気温等の歴史的記録。100年以上の変化を分析する基盤データであり、温暖化の証拠を示す。	10万以上の地点	100 GB	約 1 GB	https://www.ncei.noaa.gov/products/land-based-station/global-historical-climatology-network
カテゴリ		データベース名	中心的な機関		件数（目安）	データサイズ	年間増加量 (目安)	根拠リンク
衛星	地表観測	HLS	NASA GSFC + USGS	LandsatとSentinel-2を30m解像度で統合。雲除去等を施した一貫性のある地表反射率データにより詳細な土地被覆解析を実現。	約3,090万件以上 NASA Earth Searchより	8.7 PB	10TB/day	https://hls.gsfc.nasa.gov/
衛星	大気成分	Sentinel-5P	ESA (欧州宇宙機関)	対流圏のNO2やO3等を毎日全球規模で観測。大気質のモニタリングや排出源特定、気候変動研究に不可欠なデータを提供。	約 420万件	1.7 PB (2024年1月)	1.5TB/day	https://sentinel.esa.int/web/sentinel/missions/sentinel-5p
衛星	気象衛星	ひまわり8/9号	気象庁 (JMA)	日本周辺を10分毎に多バンド観測する静止気象衛星。高頻度画像は台風の監視や集中豪雨の予測、防災業務において極めて重要。	約7,000万件以上 2万件/day	1.5 PB	150 TB	https://www.data.jma.go.jp/mscweb/ja/index.html
衛星	降水観測	GPM IMERG	NASA GSFC + JAXA	複数の衛星観測と地上計を統合した全球降水マップ。0.1度解像度で30分毎の降水量を提供し洪水予測や水資源管理に活用。	約120万件以上	1.6 PB	200〜300 TB	https://gpm.nasa.gov/data/imerg
衛星	広域観測	MODIS	NASA	NASAの衛星による25年以上の環境観測データ。大気、海洋、陸域の状態を全球で捉え、生態系変化や温暖化の長期分析を支える。	数億件以上	約 5 〜 10 PB以上	約 200 〜 300 TB	https://modis.gsfc.nasa.gov/
衛星	大気組成	CAMS	ECMWF	大気汚染物質や温室効果ガスの監視・予測を行う。衛星データとモデルを融合した再解析を提供し、大気環境の健全性評価を支援。	約6,700億件 （気象フィールド数）	22PB以上	175 TB	https://atmosphere.copernicus.eu/
衛星	降水データ	CHIRPS	UCSB CHC	衛星と地上観測を融合した1981年からの高解像度降水データ。農業支援や干ばつ監視に特化し食料安保のリスク評価に利用。	1981年〜現在	数百GB〜数TB	不明	https://www.chc.ucsb.edu/data/chirp , https://swat.tamu.edu/data/chirps-chirts/
カテゴリ		データベース名	中心的な機関		件数（目安）	データサイズ	年間増加量 (目安)	根拠リンク