

AI for Scienceを推進する基盤モデル

松原 聖* 高橋 邦生** 北野 有哉*** 岡崎 一行****

Foundation Models Driving AI for Science

Kiyoshi Matsubara*, Kunio Takahashi**, Yuya Kitano*** and Kazuyuki Okazaki****

科学研究の現場では、AI 技術を活用して研究開発を飛躍的に加速させる「AI for Science」が世界的な潮流となっている。アドバンスソフトの取り組みは、当社は長年培ってきた計算科学・シミュレーション技術と、最新の AI 技術を融合させることで、幅広い分野のシミュレーション事業を展開している。本稿は、文部科学省様からの委託業務「AI for Science の実現に向けた計算基盤等の動向調査」（令和 7 年度）としてアドバンスソフトが実施し作成した報告書の内容をもとに、アドバンスソフトが一部をまとめた資料である。本書のもとになった調査報告書は、文部科学省様のサイトで公開されている。

DOI: <https://doi.org/10.69290/j.001234-vol33>

Keywords: AI for Science、基盤モデル、データ、気象、材料、ライフサイエンス

1. はじめに

AI for Science は、単なる解析ツールを超え、研究プロセスのあらゆる段階において 3 つの側面で効率化と深化をもたらしている。第一に、知識の体系化と仮説生成である。膨大な論文や実験データから大規模言語モデルが知見を統合し、人間では把握困難なキーワード間の相関を抽出することで、新たな研究仮説の立案を支援する。第二に、現象予測と計算加速については、物理法則を学習した基盤モデルが、従来のスーパーコンピュータによるシミュレーションを数万倍に加速させ、例えば、材料分野やライフサイエンス分野では、新

材料や新薬の候補を極めて短時間で絞り込むことを可能にしている。第三に、実験の自律化である。ここでは、AI がロボットや計測機器をリアルタイムに制御し、24 時間稼働の自律型ラボを実現することで、検証サイクルから人為的ミスを排除し、発見のスピードを物理的限界まで高めている。本稿では、ライフサイエンス・材料科学・地球科学の 3 分野における AI for Science を支える基盤モデルおよびその基盤モデルを支えるデータベースについて述べる。

2. AI for Science の基盤モデル

基盤モデルとは、膨大なデータセットで学習し、本質的なパターンや構造を深く理解した汎用的な AI モデルの総称である。従来の AI 技術が特定の限定的なデータごとに個別の解析モデルを構築していたのに対し、基盤モデルは大規模な情報を横断的かつ網羅的に学習している。これにより、学習で得た高度な知識の土台を活用し、少量の情報しか存在しない未知の課題などに対しても、柔軟に適応・解析することが可能となる。シミュレーションを高速化するほか、望む特性から新材料や新薬の構造を直接逆設計する機能が期待されている。表 1 と表 2 に基盤モデル一覧を示す。

*アドバンスソフト株式会社 代表取締役社長

President, AdvanceSoft Corporation

**アドバンスソフト株式会社 熱流動エンジニアリングセンター

Thermal Hydraulics Engineering Center, AdvanceSoft Corporation

***アドバンスソフト株式会社 第 3 事業部

3rd Computational Science and Engineering Group, AdvanceSoft Corporation

****アドバンスソフト株式会社 第 1 事業部

1st Computational Science and Engineering Group, AdvanceSoft Corporation

2.1. ライフサイエンス分野

ライフサイエンス分野の基盤モデルは、タンパク質、ゲノム、単一細胞解析を中心に発展している。タンパク質分野では、「ESM-3」や「AlphaFold 3」が、配列や構造、機能を統合的に解析し、DNAを含む複合体の構造予測を可能にした。ゲノム解析では、「HyenaDNA」や「DNABERT 2」が長大な DNA 配列を高精度に処理する。さらに、単一細胞解析では、「scGPT」や「Geneformer」が遺伝子相互作用を学習し、生命科学における高度な予測や新たな知見の創出を支えている。

2.2. 材料科学分野

材料科学分野の基盤モデルは、新材料探索、原子シミュレーション、自律型実験を中心に発展している。Google DeepMind の「GNoME」は、グラフニューラルネットワークを用いて 220 万件の新材料候補を予測した。Microsoft の「MatterSim」は幅広い環境条件下で高精度な物性予測を可能にし、「MatterGen」は指定物性を持つ材料構造を生成する。さらに、AI とロボットを統合した「A-Lab」で自動合成も実現された。これらの技術は、材料開発の高速化と効率化を大きく推進している。

2.3. 地球科学分野

地球科学分野の基盤モデルは、高速かつ高精度な予測を実現している。Google DeepMind の「GraphCast」は、10 日先の全球予報を 1 分未満で実行し、一部ベンチマークでスーパーコンピュータベースの予測を上回る精度を達成した。「GenCast」は台風や熱波など極端気象を高精度に予測する。NVIDIA の「Earth-2」や Microsoft の「Aurora」「Climax」も、大規模気象データを統合し高速予測を可能にしている。さらに、「NeuralGCM」や「Prithvi-WxC」など、AI と物理法則を融合したモデルが地球規模の気象理解を加速させている。

3. 基盤モデルを支えるデータ

基盤モデルを構築するために大量のデータ（またはデータベース）が必要となる。このデータと

は、基盤モデルの訓練・改善に不可欠な燃料（欧州のデータ戦略で「fuel」という）となる、大規模かつ高品質なデータコレクションの総称である。単なる受動的なストレージではなく、AI 開発や競争上の優位性を左右する「戦略的資産」と位置付けられる。現実世界の複雑さやデータの多様性をモデルに吸収・学習させるために必要な、圧倒的な量と多様性を提供できる点が特徴である。具体的には、FAIR 原則に基づき高度にキュレートされたゲノム配列、生体高分子の 3D 構造、材料の DFT 計算結果、地球規模の気象再解析データなどが該当する。専門的なキュレーションにより生データと実用的な運用の溝を埋め、高品質なデータの流動性を確保することで、AI による自律的な科学的発見を根底から支える。表 3 と表 4 に基盤モデルを支えるデータの一覧を示す。

3.1. ライフサイエンス分野

ライフサイエンス分野のデータ基盤は、ゲノム、タンパク質、創薬、シングルセル解析を中心に急速に拡大している。TCGA、ArrayExpress、GEO、ENCODE などはゲノム解析の主要基盤であり、UniProt や PDB はタンパク質情報を提供する。創薬分野では ChEMBL、ZINC、PubChem が大規模な化学・生物活性データを蓄積している。また、Single-cell Expression Atlas や Human Cell Atlas は細胞レベルの解析を支えている。これらの高品質データは、AI 基盤モデルの学習と生命科学研究の発展に不可欠である。

3.2. 材料科学分野

材料科学分野のデータ基盤は、無機結晶構造、DFT 計算結果、大規模材料データベースを中心としている。ICSD は世界最大級の無機結晶構造 DB であり、Materials Project や OQMD は DFT 計算に基づく材料特性を公開している。さらに、Meta の「OMat24」や触媒表面データを扱う「OC20」など、AI 学習向けデータセットも拡大している。有機分子分野では QM9 や ZINC-22 が活用されており、膨大な計算・実験データが新材料探索やシミュレーション高速化を支えている。

表 1 AI for Science の基盤モデル (その1)

モデル	主な用途・特徴	開発元	アーキテクチャ	パラメータ数	公開
ESM-3	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。	Evolutionary Scale	Multimodal Transformer	980 億	2024 年 6 月
HyenaDNA	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。	Stanford Univ.	Hyena Operator	(160 万)	2023 年 6 月
DNABERT-2	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。	UCSD 等	Encoder-only Transformer	1 億 1,700 万	2023 年 6 月
scGPT	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。	Toronto Univ.	Generative Transformer	約 5,300 万	2024 年 2 月
Geneformer	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。	Broad Institute	Encoder-only Transformer	3 億 1600 万 (V2 Large)	2023 年 5 月
AlphaFold3	タンパク質だけでなく、DNA、RNA、低分子化合物を含む複合体構造まで高精度に予測できる点が特徴である。拡散モデル技術を導入し、分子間相互作用を精密に再現することで、創薬や生命科学研究の高速化に貢献。	Google DeepMind	Diffusion / Pairformer	5 億 (推定)	2024 年 5 月
GNoME	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。	Google DeepMind	Message Passing GNN	数百万規模	2023 年 11 月
MatterSim	全元素に対応し、絶対零度から数千度の高温高压といった極限環境下における物質の挙動や物性を、第一原理計算に近い精度で動的にシミュレートする。	Microsoft	Equivariant GNN	1 億 3000 万	2024 年 5 月
MACE-MP-0	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なスピードで計算する。	Cambridge Univ.	Higher-order MPNN	469 万	2023 年 12 月
MatterGen	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。	Microsoft	Diffusion Equivariant GNN	4,680 万	2023 年 12 月
UMA	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。	Meta (FAIR)	Equiformer V2 / MACE	14 億	2025 年 6 月
FourCastNet	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。	NVIDIA	AFNO(Fourier/CNN)	3 億	2025 年 5 月
WeatherNetx2	最長 15 日の全球アンサンブル予測に対応。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。	Google DeepMind	FGN / 生成 AI ベース	1 億 8,000 万	2025 年 11 月
NeuralGCM	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。	Google DeepMind	Hybrid (物理 + NN/CNN)	3,670 万	2024 年 7 月
GenCast	拡散モデルを用いた生成 AI により、台風や熱帯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。	Google DeepMind	Diffusion / GNN	5,700 万	2024 年 12 月
GraphCast	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。	Google DeepMind	GNN+Multimesh	3,670 万	2023 年 11 月
Aurora	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。	Microsoft	3D Swin Transformer	13 億	2024 年 5 月
ClimaX	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。	Microsoft	Swin Transformer	1 億 700 万	2023 年 1 月
AIFS ENS	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。	ECMWF	GNN	1 億	2024 年 5 月
AtmoRep	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。	ECMWF + JSC + CERN	Transformer (Multiformer)	24 億	2023 年 5 月
Prithvi-WxC	台風進路予測、熱帯発生確率推計、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。	IBM + NASA	ViT(MAE)	23 億	2024 年 9 月
Pangu-Weather	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。	Huawei	Swin Transformer	2 億 5,600 万	2023 年 7 月
Granite-Geospatial-Ocean	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、海水変動の監視、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。	IBM	ViT(MAE)	5,000 万	2025 年 1 月
GLONET	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。	Mercator Ocean	CNN	3,000 万	2023 年 12 月
WV-Net	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。	Hawaii Univ	ViT(MAE)	3,100 万	2024 年 2 月
TerraMind	森林の違法伐採監視、都市開発や農用地の変化分析、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 ESA +	TiM	5,000 万	2025 年 9 月
Prithvi-EO	自然災害の予見や被害解析のほか、収穫量予測、パイオマス推計、干ばつ監視、インフラモニタリング、避難経路の最適化等リアルタイム支援を行う。	IBM NASA +	ViT(MAE)	6 億	2023 年 5 月

表 2 AI for Science の基盤モデル (その2)

モデル	学習データ	データ量	データサイズ	ライセンス	アーキテクチャの特徴
ESM-3	タンパク質配列/構造/機能	~27.8 億 配列	約 1.0 ~ 2.0 TB	ESM Open License	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。
HyenaDNA	ヒトゲノム配列 (HG38)	32 億 塩基対	数 GB	Apache 2.0 (商用可)	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。
DNABERT-2	多種ゲノム配列 (RefSeq)	~300 億 塩基対	約 5.0 ~ 10 TB	Apache 2.0 (商用可)	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。
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	MIT (商用可)	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。
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	MIT (商用可)	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。
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	非営利研究限定	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。
GNoME	Materials Project, ICSD, OQMD	2.8 万件(基礎) / 220 万種(予測)	約 0.1 ~ 0.2 TB	Apache 2.0 (商用可)	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。
MatterSim	DFT 計算データ	約 1,700 万件	約 0.5 ~ 1.0 TB	MIT (商用可)	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。
MACE-MP-0	Materials Project Trajectory	約 160 万件	11.35GB	MIT (商用可)	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。
MatterGen	Materials Project, Alexandria	約 60 万 8,000 件	72GB	研究用限定	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。
UMA	3D 構造データ	50 億個以上の原子	約 0.05 ~ 0.1 TB	CC-BY-NC 4.0 (非営利限定)	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。
FourCastNet	ERA5 / 40 年分	6 時間刻み 鉛直 13 層	数十 TB と推測	Apache 2.0	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 技術を利用。
WeatherNet2	ERA5, HRES-fc0	1 時間刻み	数十 TB と推測	MIT	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。
NeuralGCN	ERA5, HRES-fc0	6 時間刻み 鉛直 32 層	数十 TB と推測	Apache 2.0	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たスパコンリソースを節約する革新的なモデル
GenCast	ERA5, HRES-fc0	6 時間刻み 鉛直 13 層	数十 TB と推測	Apache 2.0	約 5700 万パラメータを搭載。GNN と拡散モデルを統合し、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。
GraphCast	ERA5, HRES-fc0	6 時間刻み 鉛直 37 層	数十 TB と推測	Apache 2.0	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の先駆的モデル。
Aurora	ERA5, CMIP6, IFS, GFS	6 時間刻み 鉛直 13 層	数十 TB と推測	MIT	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。
ClimaX	CMIP6	6 時間刻み 鉛直 13 層	約 1 TB	MIT	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。
AIFS ENS	ERA5+IFS	6 時間刻み 鉛直 13 層	約 20 ~ 40 TB	ECMWF Open	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提供する。
AtmoRep	ERA5	1 時間刻み 最大鉛直 137 層	約 273 TB	CC-BY-4.0	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モデル。
Prithvi-WxC	MERRA-2 (NASA 再解析)	1 時間刻み 鉛直 16-25 層	数百 TB	Apache 2.0	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。
Pangu-Weather	ERA5	1 時間刻み 鉛直 13 層	約 20 ~ 100 TB	非営利研究限定	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。
Granite-Geospatial-Ocean	海面温度・塩分濃度データ	2.56 億 (256M)	2.5 TB	Apache 2.0	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。
GLONET	全球気象データ / GLORYS12	海流、塩分、海水温、海面高度	2~数十 TB	研究用	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。
WV-Net	GOAL I, WV Image, ERA5	画像 1,000 万枚	4~10TB	研究用	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。
TerraMind	TerraMesh	気象データ・土地利用	17 TB	研究用	IBM と ESA が開発した 5000 万パラメータ 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。
Prithvi-EO	HLS V2(30m)	1 億	8~10 TB	Apache 2.0	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。

表 3 基盤モデルを支えるデータ（その1）

データ名	分野	内容
TCGA	ゲノミクス	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。
ArrayExpress (BioStudies に統合)	機能的ゲノミクス	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。
BioStudies	統合アーカイブ	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)
GEO	機能的ゲノミクス	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。
ENCODE	機能的ゲノミクス	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
UniProt	プロテオミクス	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。
PDB	プロテオミクス	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。
ChEMBL	創薬	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。
ZINC	創薬	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。
PubChem	創薬	米国国立衛生研究所 (NIH) が提供する世界最大級の化学情報データベースであり、化合物の化学構造、物性、生物活性、毒性、薬理作用など多様な情報を公開している。数千万件規模の化学物質データを収録し、創薬、材料科学、生命科学など幅広い研究分野で利用されている。また、AI を活用した創薬や分子設計における学習用データ基盤としても重要な役割を担っている。
Single-Cell Exp. Atlas	シングルセル	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。
Human Cell Atlas	シングルセル	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。
ICSD	無機結晶構造	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。
Materials Project (MP)	結晶固体・材料特性	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
OQMD	材料特性・DFT	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。
NOMAD	データ管理	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
OMat24	大規模材料 DB	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。
SNUMAT	合成材料	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。
OC20	触媒・原子軌跡	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。
MPTraj	トラジェクトリ	Materials 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
Alexandria	分子構造 (DFT)	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
QM9	有機分子	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
ERA5	気候再解析	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。
HRES	気象予報	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。
CMIP6	気候予測	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。
MERRA-2	気候再解析	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。
JRA-3Q	長期再解析	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。
GHCN	地上観測	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。
HLS	地表観測	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。
Sentinel-5P	大気成分	対流圏の NO ₂ や O ₃ 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。
ひまわり 8/9 号	気象衛星	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。
GPM IMERG	降水観測	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。
MODIS	広域観測	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。
CAMS	大気組成	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。
CHIRPS	降水データ	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た食料安保のリスク評価に利用。

表 4 基盤モデルを支えるデータ (その2)

データ名	中心的な機関	件数	データサイズ	年間増加量	リンク
TCGA	NIH (NCI/NHGRI)	33 がん種、2 万サンプル	2.5 PB	プロジェクト終了	[1]
ArrayExpress (BioStudies に統合)	EMBL-EBI	実験: 77,585 件 アクセス: 2,429,810 件	56.68 TB	—	[2]
BioStudies	EMBL-EBI	26,464,332 files; 10,369,537 links; 2,398,055 studies	8PB 以上	1PB 以上	[3]
GEO	NCBI (NIH)	Samples: 8,388,256 件、Series: 280,160 件、 Platforms: 28,191 件、DataSets: 4,348 件	数百 TB	15%	[4]
ENCODE	NHGRI (NIH)	実験: 23,439 件 ファイル: 1,171,460 件	1 PB 以上	不明	[5]
UniProt	UniProt Consortium (EMBL-EBI, SIB, PIR)	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%	[6]
PDB	wwPDB	250,496 件の登録	366 TB	87 TB	[7]
ChEMBL	EMBL-EBI	化合物: 2.4 百万件, 活性データ: 20 百万件, 標的: 15,622 件, 論文: 88,421 件	10~20GB (ChEMBL 34)	10%程度	[8]
ZINC	UCSF 医薬化学科	370 億 化合物	数十 TB	10%程度	[9]
PubChem	NCBI (NIH)	化合物: 123,475,690 件, 物質: 342,870,557 件, 生物学的活性試験: 1,910,133 件, 生物活性データポイント: 299,438,868 件, 特許: 83,167,183 件	約 1.4TB 以上	約 5~10%	[10]
Single-Cell Exp. Atlas	EMBL-EBI	研究数: 383 件, 細胞数: 17,846,119 個, 対象生物種: 21 種	数 TB 以下	10%程度以上	[11]
Human Cell Atlas	HCA Consortium(Broad Institute, EMBL-EBI 等)	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	[12]
ICSD	FIZ Karlsruhe	約 327,833 件	数百 GB 以下 (想定 50KB/件)	約 1.2 万件 (4%程度)	[13]
Materials Project (MP)	LBNL (米国)	20 万件以上の材料	14 GB	10%程度	[14]
OQMD	Northwestern Univ.	1,407,395 件	10 TB 以下 (数 MB/ 件を想定)	10%程度	[15]
NOMAD	NOMAD Lab. (Max Planck 等)	19,295,396 件	114.4 TB	不明	[16]
OMat24	Meta FAIR	約 1 億 1,800 万 構造	185.67 GB	不明	[17]
SNUMAT	SNU	15,280 件	100GB (数 MB/件を想定)	不明	[18]
OC20	Meta FAIR/ CMU	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	[19]
MPTj	Materials Project (LBNL, UCB)	1,580,395 件 (構造データ数)	約 11.35 GB	不定期に update	[20]
Alexandria	ICAMS	約 580 万 構造	数 TB と推定(学習済 で 72GB)	1 回の update で 130 万件	[21]
QM9	Prof. Raghunathan (Univ. of Basel)	133,885 件	172.68 MB	(固定)	[22]
ERA5	ECMWF	数兆レコード	10 PB	約 500 TB ~ 1 PB	[23]
HRES	ECMWF	(日次運用)	50 TB	(運用データのため膨大)	[24]
CMIP6	WCRP / PCMDI	数百万 データセット	40 ~ 80 PB	(フェーズ毎に数 PB 増加)	[25]
MERRA-2	NASA (GMAO)	数百万 ファイル	800 TB	約 15 ~ 20 TB	[26]
JRA-3Q	気象庁 (JMA)	75 年分の一貫データ	473.33 TB	約 20 TB	[27]
GHCN	NOAA (NCEI)	10 万以上の地点	100 GB	約 1 GB	[28]
HLS	NASA GSFC + USGS	約 3,090 万件以上 NASA Earth Search より	8.7 PB	10TB/day	[29]
Sentinel-5P	ESA (欧州宇宙機関)	約 420 万件	1.7 PB (2024 年 1 月)	1.5TB/day	[30]
ひまわり 8/9 号	気象庁 (JMA)	約 7,000 万件以上 2 万件/day	1.5 PB	150 TB	[31]
GPM IMERG	NASA GSFC + JAXA	約 120 万件以上	1.6 PB	200~300 TB	[32]
MODIS	NASA	数億件以上	約 5 ~ 10 PB 以上	約 200 ~ 300 TB	[33]
CAMS	ECMWF	約 6,700 億件 (気象フィールド数)	22PB 以上	175 TB	[34]
CHIRPS	UCSB CHC	1981 年~現在	数百 GB~数 TB	不明	[35]

3.3. 地球科学分野

地球科学分野のデータ基盤は、再解析データ、気候予測、地上観測など多岐にわたる。「ERA5」は 1940 年以降の全球気象データを収録した世界標準の再解析 DB であり、「CMIP6」は IPCC 報告書を支える気候予測シナリオを集約している。日本気象庁の「JRA-3Q」は台風や梅雨の再現に優れ、NASA の「MERRA-2」は大気組成変化の解析に活用される。また、「GHCN」は世界各地の気温記録を蓄積しており、これらのデータ基盤が高精度予報や長期的な気候監視を支えている。

4. まとめ

AI for Science の実現は、単なる技術的進歩ではなく、我が国の研究開発能力を根本から規定する国家戦略上の最優先課題である。これまでの HPCI の運用で培われた「利用者支援のノウハウ」や「産業利用の裾野」といった日本固有の強みは、AI for Science という新パラダイムにおいても盤石な基礎となる。この既存政策の延長線上に、「富岳 NEXT」を核とする次世代インフラを大胆に構築し、データ・アルゴリズム・人材を有機的に結合させることこそが、2030 年以降の日本の科学技術の地位を不動のものにする。

参考文献

[1] <https://www.cancer.gov/ccg/research/genome-sequencing/tcga>
 [2] <https://www.ebi.ac.uk/arrayexpress/>
 [3] <https://www.ebi.ac.uk/biostudies/>
 [4] <https://www.ncbi.nlm.nih.gov/geo/>
 [5] <https://www.encodeproject.org/>
 [6] <https://www.uniprot.org>
 [7] https://www.rcsb.org/stats/data_storage_growth
 [8] <https://www.ebi.ac.uk/chembl/>
 [9] <https://zinc22.docking.org/>
 [10] <https://pubchem.ncbi.nlm.nih.gov/statistics/>
 [11] <https://www.ebi.ac.uk/gxa/sc/home>
 [12] <https://data.humancellatlas.org/>

[13] <https://icsd.products.fiz-karlsruhe.de/>
 [14] <https://matbench-discovery.materialsproject.org/data>
 [15] <http://oqmd.org/>
 [16] <https://nomad-lab.eu/>
 [17] <https://hyper.ai/en/datasets/35287>
 [18] <http://snumat.com/>
 [19] <https://opencatalystproject.org/>
 [20] https://figshare.com/articles/dataset/CHG_Net_MPTrj_Dataset/23457473
 [21] <https://alexandria.icams.rub.de/>
 [22] <https://hyper.ai/en/datasets/38089>
 [23] <https://www.ecmwf.int/en/forecasts/dataset/ecmwf-reanalysis-v5>
 [24] <https://www.ecmwf.int/en/forecasts/dataset/set-i>
 [25] <https://pcmdi.llnl.gov/CMIP6/>
 [26] <https://gmao.gsfc.nasa.gov/reanalysis/MERRA-2/>
 [27] https://jra.kishou.go.jp/JRA-3Q/index_ja.html
 [28] <https://www.ncei.noaa.gov/products/land-based-station/global-historical-climatology-network>
 [29] <https://hls.gsfc.nasa.gov/>
 [30] <https://sentinel.esa.int/web/sentinel/missions/sentinel-5p>
 [31] <https://www.data.jma.go.jp/mscweb/ja/index.html>
 [32] <https://gpm.nasa.gov/data/imerg>
 [33] <https://modis.gsfc.nasa.gov/>
 [34] <https://atmosphere.copernicus.eu/>
 [35] <https://swat.tamu.edu/data/chirps-chirts/>

※ 技術情報誌アドバンスシミュレーションは、それぞれの文献タイトルの下に記載した DOI から、PDF ファイル（カラー版）がダウンロードできます。また、本雑誌に記載された文献は、発行後に、JDREAMIII（日本最大級の科学技術文献情報データベース）に登録されます。