jh251005

# 材料研究用データプラットフォームの大規模化および深化

華井雅俊（東京大学情報基盤センター）

概要

現代の材料科学研究において、最先端実験施設とスーパーコンピュータ間での効果的かつ大規模なデータ管理は不可欠である。しかし、既存の多くのデータシステムは主に小規模な機関間連携や単一ドメインの運用に焦点を当てており、多様な研究者からの大量データを扱うために必要なスケーラビリティ、効率性、機敏性、学際性が不足している。これら問題を解決するために、本研究では日本全国の材料科学のためのデータプラットフォームを目指す「ARIM-mdx データシステム」を開発している。今年度において、申請時の 1000 ユーザーから、報告時（2026 年 5 月）までで、1700 ユーザーまで拡大する事ができ、目標の 1500 ユーザーを超えることができた。引き続き、先駆的な全国規模の材料データプラットフォームとして、新たな学際研究コミュニティの創出とイノベーション加速への貢献を目指していく。

## 1. 共同研究に関する情報

### (1) 共同利用・共同研究を実施している拠点名

mdx I

### (2) 課題分野

データ科学・データ利活用課題分野

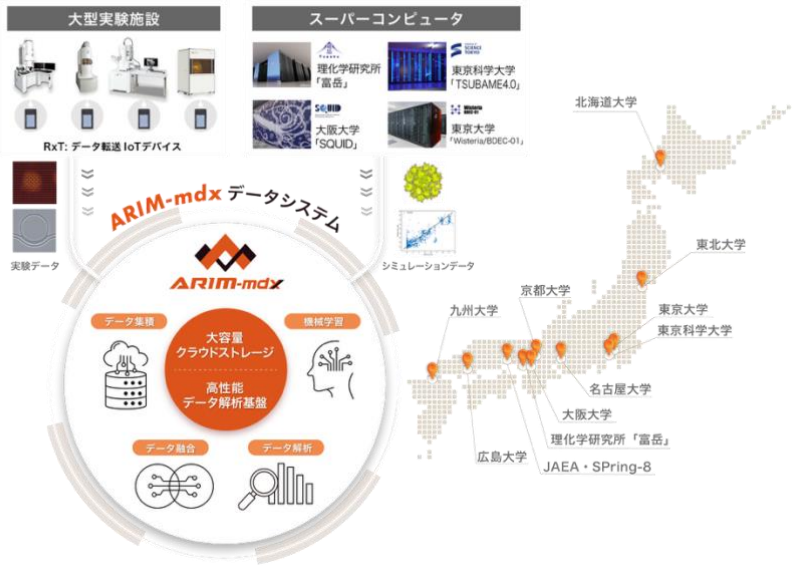
### (3) 参加研究者一覧と役割分担

華井雅俊（研究総括・システム構築実施）
石川亮（実験データユースケース）
田浦 健次朗（システム構築アドバイス）
鈴村豊太郎（システム構築アドバイス）
河村光晶（シミュレーションユースケース）
岡根 哲夫（実験データユースケース）
藤川誠司（実験データユースケース）
松村大樹（実験データユースケース）
大西正人（シミュレーションユースケース）
安永竣（実験データユースケース）
豊倉 敦（実験データユースケース）
村上恭和（他機関連携）
山本剛久（他機関連携）

## 2. 研究の目的と意義

近年、機械学習分野の社会的な盛り上がりが目覚ましく、去年のノーベル物理学賞・化学賞に代表されるように自然科学分野における応用が活発である。機械学習を中心とした今日のデータ駆動型研究の中心は大規模に収集されたデータであり、材料分野においてもデータの重要性がますます強調されている。

材料の研究開発においてデータはおおよそ 2 つに分類され、1 つは理論計算によって生み出されるシミュレーションデータ、もう 1 つは実際の材料実験装置（電子顕微鏡や放射光装置）から得られる実験データである。材料研究開発においてそれらの相互的なデータ解析・データ同化は不可欠であり、個々の研究課題や領域において一般的に行




われている。一方で、材料分野全体、特に実験と理論のデータ融合の文脈において、研究コミュニティに共通した大規模データプラットフォームは未整備であり、各研究グループが個別にシステム整備することを強いられている。その結果、データサイエンス分野などで一般的に行われている研究ツールの共通化や大規模データの共有が限定的であり、分野全体を巻き込んだ共通知の構築が阻害されている。特に実験分野においてその傾向が顕著である。

本研究では、材料分野における大規模な統合プラットフォームの発展及びその深化を目的とする。2024 年度に構築した、大型実験施設・スーパコンピュータ・mdx 連携の材料研究用データプラットフォーム、“ARIM-mdx データシステム”を基盤とし、実際のユースケースを通じたシステムの改良、全国展開および、その深化を実施する。

本研究は、ARIM-mdx データシステムの全国的な本格運用及びその高度化を目標とする。2025 年度末までに 1500 ユーザーを目標（2024 年 12 月現在で 1000 ユーザー）とし、学術組織から産業組織に至るまでの様々な材料研究者を巻き込んだ日本全国規模のデータプラットフォーム展開を目指す。また、ARIM-mdx データシステムを中心とした研究コミュニティの構築、ユースケース研究の追求を目指す。

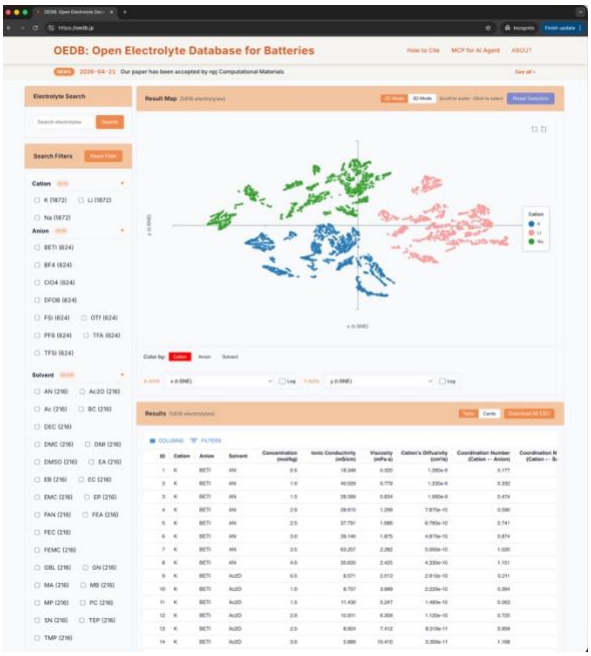
材料の研究開発分野は、Society 5.0 や SDGs に代表される社会指針における重要領域の 1 つであり、そのためのデータインフラ基盤の全国的な整備は学術的・産業的に大いなる意義がある。単なる 1 システムの PoC 開発にとどまらず、幅広い材料研究者が利用可能な本格的な社会実装を実現する。

## 3. 当拠点の公募型共同研究として実施した意義

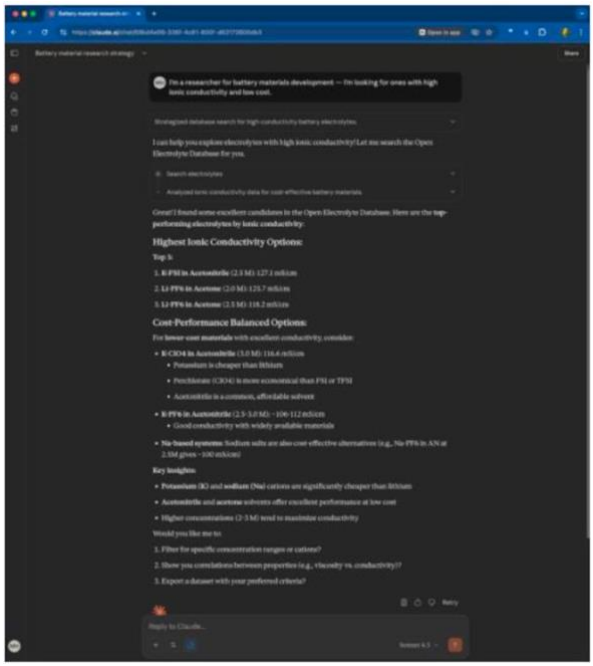
本課題は、物理学・材料科学における実験系分野と理論系分野及び、情報科学における高性能計算分野とデータ科学分野にまたがる学際的課題であり、また、材料実験装置・スーパコンピュータは国内の各大学・研究機関に設置されているため、組織横断的な研究実施体制を必要とする。

## 4. 前年度までに得られた研究成果の概要

2024 年度の成果にて特筆すべきはユーザー数の急増である（1 年間で 168 から 1041）。当初の研究予定で掲げた材料分野におけるデータプラットフォームの強い必要性を確認できた。システム全体に関して、ビックデータに関する国際会議 IEEE BigData 2024 に論文が採択され発表を行った。招待




OEDB: Web UI



OEDB: MCP による AI 連携

講演や利用説明会等での対外発表を積極的に行った。システムを効果的に用いたユースケース研究に関しても徐々に成果創出がなされており、例えば大型電子顕微鏡の実験像に関する新デノイズ手法である。提案のプログラムは ARIM-mdx 上で開発され、一般ユーザーは実行環境整備の手間なく本研究成果を利用可能である。また、大規模なシミュレーションデータの生成・収集を実施し、DFT や MD の計算結果を中心に合計 500TB 集積されている。最後に、自動実験システムとの連携を実施し、実験データのハイスループットな収集に向けた第一歩を踏むことができた。

## 5. 今年度の研究成果の詳細

まずサービス規模に関して、**申請時の 1000 ユーザーから、報告時（2026 年 5 月）までで、1700 ユーザーまで拡大し、着実なサービス拡大を行うことができた。**データ量に関して、計算データ・実験データ合わせて 1PB を超える規模となった。そのうち実験データは 100TB ほどであり、その多くが計算データである。また、接続の大型実験装置に関して、東京大学の武田先端知クリーンルームでの装置を主に拡大し、全体で 70 台規模の運用を実施することになった。

### (i) 計算データユースケース

まず、計算・シミュレーションデータに関しては大きく 2 つのテーマで大規模データセットの作成を実施した。

1 つは、電池開発における電解液データベースの構築である（https://oedb.jp/）。電解液の構成である、Anion, Cation, Solvent の組み合わせ合計 5000 に対して、大規模 MD 計算を実施しイオン伝導度など 5 種類の電池性能に関する物性を計算した。前年度に引き続き mdx の CPU ノードを 100 VM 単位で確保し Slurm 等バッチシステムを構築、ARIM-mdx とのシームレスな連携により中間データを含めた数百 TB 単位のデータを ARIM-mdx に保存している。また、実験による実データ作成も実施し計算・実験データの融合分析を実施中である。本研究の計算データ生成手法に関して論文が出版された。

現在計算データのうち、中間データを省いた物性結果は Web だけでなく AI Interface

(MCP)での公開を実施した。Web UI では物性の相関を 2D / 3D プロットし、インタラクティブに選択可能にし、所望の物性値から対象の組み合わせ材料を効率的に探索できるようにした。また、MCP による AI Interface では、Claude や ChatGPT 等、商用の AI サービスとの連携を可能にし、例えば、Web から最新研究論文をさがし、それらと DB 内のデータを突き合わせなどが簡単に実施できるようになった。

2 つ目のシミュレーションユースケースは固体材料の熱伝導データの大規模データセットである（ https://phonix-db.org/ )。熱伝導の DFT による理論計算は 他物性の中でも非常に計算コストがかかることが知られ (非調和フォノン計算を含むため) 既存のオープンデータでは 100 材料ほ どしかなかったが、本研究ではそれらを 6000 材料にまで拡大した。本 JHPCN 課題に加え、HPCI 等各種スパコンに申請し、これまでの数 年間で約 1 億 CPU コア時間をかけ大規模計算 を実施、ARIM-mdx に計算結果を保存している。本研究に関しては、論文を npj Computational Materials に採択、出版された。

（ii）実験データユースケース

実験系ユースケースに関して、上の 電池材料の実験データ収集に加え、電子顕微鏡データのデコンボリューションアルゴリズム（2D イメージからの 3D 構造の同定）の開発およびシミュレーションとの位置同定研究を実施した。

また、SPring-8 での利用に関して、ARIM-mdx 上で SPring-8 からのデータを処理するためにこれまで Fortran ベース書かれた ローカルプログラムを Python に移植し Jupyter 環境で簡単に利用できるような基盤を整備した。

最後に制度面・運用面に関して、

（i）利用規約の正式な設定

（ii）機関間での契約書の定義

（iii）ユーザー管理システム等、運用管理機能の充実

などを行い今後の利用拡大に備えた。

## 6. 進捗状況の自己評価と今後の展望

進歩は概ね良好である（自己評価 80%)。

着実なユーザー数拡大に加え、今年度はデータ量としても 1PB を超えることになり、国内でも有数のデータ量が保存される材料系データシステムとなった。

制度面の整備が進んだ結果、

(i) MoonSHOT 10 目標 10 “超次元状態エンジニアリングによる未来予測型デジタルシステム ” 星 PM

(ii) K Program “次世代半導体微細加工プロ

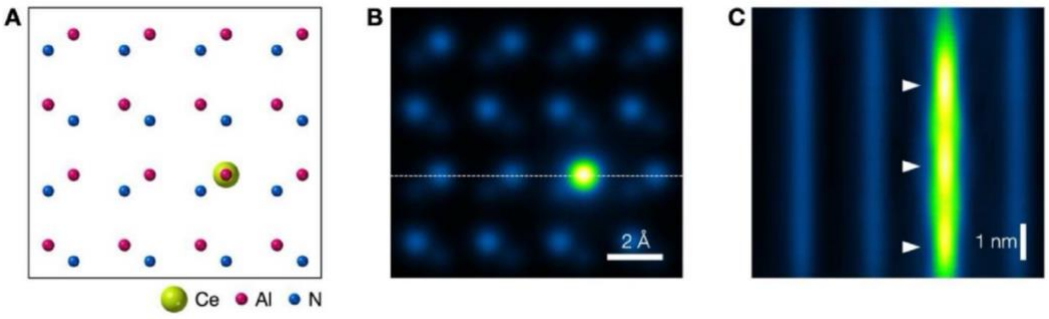



**Fig. 1.** (A) Point defect structure model of Ce-doped *w*-AlN viewed along the [11–20] direction, where the specimen thickness is 74.7 Å. The yellow sphere represents substitutional point defects of Ce at Al site, with the depth locations of 21.0, 42.8, 64.5 Å, respectively. (B) Projection of simulated depth-dectioning image of Ce-doped *w*-AlN. (C) Cross sectional view along the dashed line in (B). Three triangles denote the locations of Ce dopants.

セス技術” 理研 緑川 PM

(iii) NanoTerasu 東京大学物性研究所 松田研グループ

などでの導入が実施できた。今後、さまざまな大型プロジェクトの誘致や機関での大規模利用を着実に進めるとともに、2026 年度から始まっている AI for Science の各種プロジェクトに関しても着実にサービス拡大を進めていきたい。