第１章　マテリアルズ・インフォマティクスの現状と将来展望

徐　一斌*

１　はじめに

材料開発は，これまで主に経験と試行錯誤に支えられてきた。材料科学者たちは，構造・特性・プロセスに関する深い知識と勘に基づいて，新しい材料の発見や改良を進めてきたが，その一方で，膨大な可能性空間を十分に探索するには限界があった。近年，情報科学や計算科学の進展に伴い，こうした経験依存のアプローチを補完・加速する方法として，マテリアルズ・インフォマティクス（Materials Informatics：MI）が注目されている。

MI は，実験や計算から得られる材料データを蓄積・解析し，統計学や機械学習といったデータ科学的手法を用いて，材料の特性予測や新材料設計を効率化する手法である。すでに多くの企業や研究機関が MI に取り組み始めており，材料開発の新たなパラダイムとして社会実装も進みつつある。

MI の最大の魅力は，これまで人間が到達できなかった「未探索の材料空間」に対して，網羅的かつ論理的にアプローチできる点にある。従来の方法では数年かかった開発期間を大幅に短縮することも可能であり，加えて，物性やコスト，環境負荷など多様な設計要件を同時に考慮できる柔軟性も持つ。

しかし一方で，MI は単なるアルゴリズムやデータ処理の問題にとどまらず，「どのような材料データを，どのように用いるべきか」といったデータの質と構造に関わる問題を内包している。特に実用的な応用を目指すには，データの正確性，完全性，多様性を確保し，かつ階層構造をもつ材料情報（元素，化合物，相，材料）をどのように統合して扱うかが重要となる。

本章では，MI の概念的背景とともに，材料データの特性や，データ駆動型研究の成功事例，さらには最近注目を集める大規模言語モデル（LLM）の応用まで，材料科学におけるデータ活用の全体像を展望する。MI が目指す未来像とは何か，それを実現するために今，何が必要かを読み解いていく。

＊　Yibin XU　(国研)物質・材料研究機構　マテリアル基盤研究センター　グループリーダー

2　材料データ

2. 1　高品質材料データの要件

良質な機械学習モデルを構築するには，訓練データとなる材料データの品質が極めて重要である。とりわけ MI においては，正確性・完全性・多様性の三要素が基本的要件とされる。

① 正確性とは，物性値や構造情報が信頼できる精度で測定され，誤差や不確かさが適切に評価されている状態を指す。

② 完全性は，ターゲット特性の理解や再現に必要な背景情報が欠落なく記録されている状態を意味し，材料の化学組成や結晶構造に加え，合成手法や測定条件なども含まれる。

③ 多様性とは，特定の系に偏ることなく，広範な材料空間をカバーすることであり，汎化性能の高いモデル構築に不可欠である。

これら三要素を満たしたデータ基盤の構築こそが，MI の可能性を最大限に引き出す鍵となる。

2. 2　材料データの現状

材料科学におけるデータは，元素 → 化合物 → 物質／相 → 材料という階層構造のもとに整理でき，この構造に従ってシステムの複雑性は増し，利用可能なデータ数は急減する傾向がある。

最も基本的な元素レベルでは，周期表に基づき電気陰性度，イオン半径，融点などの物性が体系的に整理されており，ほぼ全元素の基本特性が入手可能である。

次に化合物は，複数の元素が化学結合で形成する構造体であり，CAS や PubChem などの化合物データベースには，論文などで発表された化合物をほぼ収録していて，その規模は数億件（10^8 件）に達する。

一方で物質／相は，結晶構造や状態（気体，液体，固体など）を具体化した化合物である。無機材料の物質データベースとして，AtomWork-Adv，や ICSD，SpringerMaterials などはよく知られていて，データ数は 100 万件（10^6 件）規模である。

最も応用に近い階層である材料は，製造プロセス，形態，スケール，さらには多相からなる複合構成など，極めて複雑な要素を内包しており，実用性を前提とした高度に調整された存在である。しかしながら，このレベルの材料に関するデータは，ごく限られた件数しか体系的に整備されていないうえ，その多くは企業内部に秘匿された未公開データで占められている。公開されている材料データベースとして，NIMS（物質・材料研究機構）の Kinzoku，AtomWork Battery などがあるが，その規模は，数千件にとどまる。

また，材料の複雑性は含有元素の種類（元素数）によっても表すことができる。一般に，材料中に含まれる元素数が増えるほど，構成可能な化学システムの数は指数関数的に増加する。しかしその一方で，実際にデータベースに収納されている化学システムの割合は急速に低下する。すなわち，多元素系になればなるほど，その化学空間は観測されていない未踏領域となる。図 1

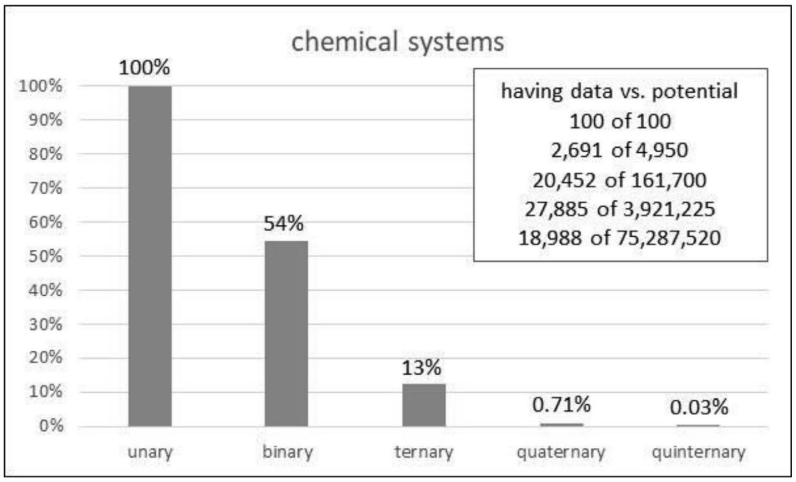

図 1　一元系から五元系までの全組成可能な化学システム数に対する，実際にデータが存在するシステムの割合（AtomWork Adv[1]のデータによる）

は，AtomWork-Adv に収録されている約 17 万件の物質データを対象に，一元系から五元系までの全組成可能な化学システム数に対する，実際にデータが存在するシステムの割合を示している。結果からは，四元系以上の化学システムの大多数が未観測または未記録であることが明らかであり，多元素材料探索におけるデータの希薄性と限界を如実に示している。

2. 3　マテリアルズ・インフォマティクスに必要なデータ

材料の特性やパフォーマンスは，化学組成や結晶構造，微細組織，界面状態など多様な因子の相互作用によって決定される。そのため，MI においては，元素レベルから化合物，物質，そして最終的な材料に至るまで，階層横断的かつ多様なデータの統合的活用が不可欠である。しかし現状では，多くの材料データベースは特定階層に特化しており，単一のデータベースだけで材料設計に必要な全情報を網羅することは困難である。したがって，複数の異なるスケール・性質のデータベースを有機的に接続・統合する仕組みが求められている。

その一例が，NIMS によって開発された電池材料データネットワーク[2]である。このシステムは，電池材料に関する異なる階層・スケールの情報を横断的に統合することを目的として構築されている。中核となるのは，文献から収集した電池材料の特性と電池パフォーマンスに関するデータベース AtomWork Battery である。AtomWork Battery は物質レベルで，無機材料の結晶構造や物性，状態図データを収録する AtomWork-Adv，さらに電子状態に関するデータベース CompES-X と連携されており，これらをリンクさせることで，化学組成，結晶構造，電子構造，プロセス条件，物性値，電池性能といった異種の情報を一元的に取得することが可能となっている（図 2）。このように，材料システムを構成する要素を階層横断的に統合する枠組みは，

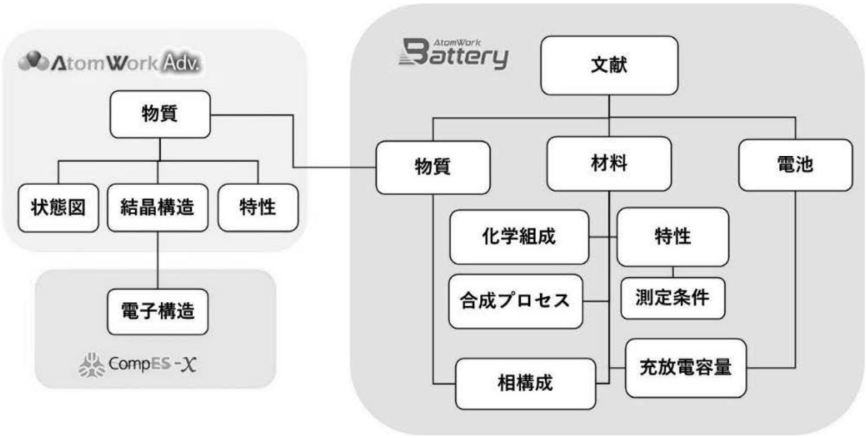

図2　NIMSの電池材料データネットワーク

複雑な電池材料設計において，データ駆動型アプローチを実現するための有力な基盤となっている。

3　MIの研究事例

MIの主な目的として，知見の獲得，特性予測，材料設計の三つが挙げられる。以下に，それぞれの目的に対応した代表的な事例を紹介する。

3. 1　知見の獲得

まず知見の獲得においては，データの統計解析や機械学習モデルを用いて，既存データから未知の構造-物性関係や因果的なパターンを抽出する試みがある。例えば，アモルファス材料は構造が不規則であるため，構造と物性の関係を明確にすることが困難であった。NIMSと東北大学の研究チームは，熱伝導率が異なるアモルファスゲルマニウムの構造的差異を，トポロジカルデータ解析（TDA）と主成分分析（PCA）というデータ科学的手法を用いて明らかにした[3)]（図3）。従来の構造解析では見分けがつかなかった試料間の違いを，原子鎖のリングサイズの分布として定量化し，熱伝導率の違いが原子鎖の長さとネットワーク構造に起因することを突き止めた。この研究は，物理的要因が不明だった準安定相材料に対し，データ駆動で新たな構造-物性相関の知見を導き出した好例である。

3. 2　特性予測

データ駆動特性予測とは，計算や実験から得られたデータをもとに構築した機械学習モデルを