Symposium (Oral) | Symposium：Developments in materials databases: Accumulating, extracting, and overlooking knowledge

Fri. Mar 14, 2025 1:30 PM - 4:50 PM JST | Fri. Mar 14, 2025 4:30 AM - 7:50 AM UTC K102 (Lecture Hall Bldg.)

[14p-K102-1~6] Developments in materials databases: Accumulating, extracting, and overlooking knowledge

Toyohiro Chikyo(NIMS), Shigetaka Tomiya(NAIST)

1:30 PM - 2:00 PM JST | 4:30 AM - 5:00 AM UTC

[14p-K102-1]

The Future of Research Data: The Fusion of Open Science and Generative AI

○Mikiko Tanifuji[1] (1.NII)

2:00 PM - 2:30 PM JST | 5:00 AM - 5:30 AM UTC

[14p-K102-2]

Advancing Research Data Management and Utilization: Emerging Frontiers in Research Infrastructure with a Focus on Electronic Laboratory Notebooks

○Shogo Takasuka[1] (1.NAIST)

2:30 PM - 3:00 PM JST | 5:30 AM - 6:00 AM UTC

[14p-K102-3]

[The 57th Young Scientist Presentation Award Speech] Learning Semantic of Crystal Structure with Deep Neural Network

○Yuta Suzuki[1] (1.Toyota Motor Corp.)

3:20 PM - 3:50 PM JST | 6:20 AM - 6:50 AM UTC

[14p-K102-4]

Information Extraction in Science: Advancing Knowledge Graphs and Generative AI (Tentative)

○Ken Fukuda[1] (1.AIST)

3:50 PM - 4:20 PM JST | 6:50 AM - 7:20 AM UTC

[14p-K102-5]

Multimodality Data and Graph Representation Learning: Case Studies in the Field of Biology

○Hideyoshi Igata[1] (1.Preferred Networks)

4:20 PM - 4:50 PM JST | 7:20 AM - 7:50 AM UTC

[14p-K102-6]

Material Classification and Feature Extraction of Promising Material Group based on Machine Learning

○Akira Takahashi[1] (1.Science Tokyo)

研究データの未来：オープンサイエンスと生成 AI の融合

The Future of Research Data: The Fusion of Open Science and Generative AI

国立情報学研究所

谷藤 幹子

*E-mail: tanifuji@nii.ac.jp

要旨

1. 研究を取り巻く環境は、ICT の高速化に加えてパブリッククラウドサービスの普及により、実験や計測データの IoT 転送と蓄積環境に、AI アルゴリズムを含む解析環境に繋がり、それらを構造化したデータベース化するアシスト環境など、進展が著しい[1)]。
2. 研究データの世界では、デジタルに情報を扱うデジタルオブジェクト(FAIR Digital Object)を識別子 PID で同定して流通、提供、利用する国プロジェクト FDO One、欧州の EOSC Node のような国や欧州地域間のデータ流通基盤の進展も加速している[2)]。基盤を構築するフェーズから、構築した基盤の上で、どのような利活用や、産業を含むデータ流通が、データ市場を形成しうるか、という実証実験にフェーズが移っている。
3. 日本の研究データ政策は、2023 年の G7 科学技術大臣コミュニケを受けてオープンサイエンスの実践が進み、2024 年、内閣府は日本の学術論文等のオープンアクセス政策を取りまとめ、2025 年の一部の公的研究資金による研究から、査読付き論文および論文根拠データの即時公開を義務化する[3)]。これを受けて全国の大学等機関での DX、すなわち ICT 整備、管理系データベースの連携・統合と、研究データの管理・活用に向けたシステム構築、研究データポリシーの制定が進んだ。また近年に立ち上げが進む研究データエコシステム（東海、中国四国、東北、北陸各地コンソーシアム等）でデータ管理から共有、再利用につなげる研究システムとしての設計、構築、開発ノウハウの共有が進んでいる。
4. 各国で進む研究開発環境 DX を通して、世界で共用可能なデータ記述（スキーマ、モデル）と FAIR な互換性（プロトコール、認証連携等）を担保することの議論が RDA、 CoDATA、WDS 等の研究コミュニティで進んでいる。互いの失敗・成功例を共有することを通じて、世界の研究データスペースが、今や大きなデータ市場の形成に向かって進化している。ここでの市場とは、必ずしも課金のみを意識したものではなく、だれが・何を・どのように利用し・成果を共有し活用に繋げるか、という研究開発サイクルが成立しうるデータ科学の世界を目指していることを意味する。欧米の公的資金が、こうした大規模な動きに大きく投資されることが、産学官が繋がる次世代の研究開発を象徴していると言えるだろう。
5. 米マテリアルゲノム以来「データが多ければ MI は加速＝大きなデータベースが必要」と言われている。質を重視する材料データベースを大きくすることには様々な限界があるが、それを補完する生成 AI を活かす機会が選択肢に入り[4)]、まさに生成 AI の醸成と並走するマテリアルデータベースの新たな道標 2025 年になるのではないか。

引用

1) ARIM Japan, 令和 5 年度秀でた利用成果
2) FDO One は前身の FDO から産業利用を含むフォーラムに進化。欧州 EOSC と連携している。
3) 学術論文等の即時オープンアクセスの実現に向けた基本方針(統合イノベーション戦略推進会議)
4) 黒橋禎夫、生成 AI とオープンデータ、情報の科学と技術、74 巻、2024 年、p. 292

研究データの蓄積と活用：電子ラボノートを中心とした新たな研究基盤の新展開

Advancing Research Data Management and Utilization: Emerging Frontiers in Research Infrastructure with a Focus on Electronic Laboratory Notebooks

奈良先端科学技術大学院大学[1]　○高須賀 聖五[1]

Nara Institute of Science and Technology[1] ○Shogo Takasuka[1]

E-mail: takasuka.shogo@ms.naist.jp

デジタル技術の進展により、研究活動におけるデジタル・トランスフォーメーション(研究 DX)が世界的に加速している。中でも、「AI」×「データ」×「リモート化・スマート化」による価値創造 [1]が期待されており、学術分野や産業分野を問わず重要な役割を担っている。特に、リモート化・スマート化によるデータ創出の効率化、AI を用いたデータの高度な利活用の発展が著しい。

これらのプロセスの橋渡しとなる、実験データの統合・管理プラットフォームの開発や実装が十分に進んでいるとは言い難い。さらに、実験データは単に蓄積するだけではなく、FAIR 原則 (Findable, Accessible, Interoperable, Re-usable) [2]に基づいて管理することが重要である。上記を踏まえ、我々は、実験データの統合・管理プラットフォームとして電子ラボノートに着目した。

電子ラボノートの活用により、実験に関する多様な情報(実験(分析)条件、結果、観察記録など)を一元管理し、それを基にデータベースを構築することが可能となるため、データ創出からデータ利活用までを包括的に支援する基盤となり得る。最終的には、多様な実験データを FAIR 原則に沿った形式で蓄積・活用できるプラットフォームの確立を目指す (Figure 1)。

本講演では、研究 DX 推進を目指した電子ラボノートの選定、導入および運用に関する具体的な取り組みを紹介する。例えば、Application Programming Interface (API)の利用に適したテンプレート作成、QR コードによる実験サンプル管理、マテリアルズ・インフォマティクスへの応用および自動実験との統合など、実践的な活用事例について詳述する。

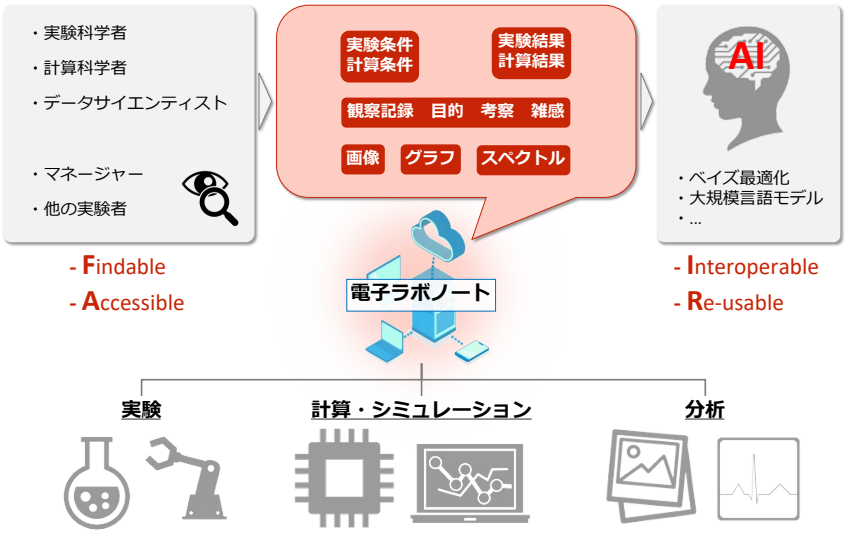

Figure 1. Ideal data flow for research activities.

[1] 文部科学省，総合科学技術・イノベーション会議有識者議員懇談会，2022 年 4 月 28 日. https://www8.cao.go.jp/cstp/gaiyo/yusikisha/20220428/siryo2_1.pdf (2025.1.7)
[2] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

結晶構造の「意味」を学ぶ深層学習技術

Learning Semantic of Crystal Structure with Deep Neural Network

トヨタ自動車[1]，○鈴木雄太[1]

TOYOTA Motor Corp.[1] ○Yuta Suzuki[1]

E-mail: yuta_suzuki_ah@mail.toyota.co.jp

材料開発において、所望の特性を持つ材料を広大な物質空間から効率的に探索することは重要な課題である。近年、深層学習を用いた物質の表現学習が注目を集めており、物質をベクトル空間に埋め込むことで、その類似性を定量的に評価することが可能になった。特に、結晶構造に基づく自己教師あり学習によって得られた埋め込み表現（embedding）は、類似材料の検索や物質空間の可視化に有効であることが示されている[1]。しかしながら、従来の結晶構造のみに基づくembeddingは、その解釈が難しいという課題があった。これは大規模な材料データから特定の機能を持つ材料を探し出す際に、個々の結晶構造や組成式を確認し、その機能に関する知識を別途調べる必要があるという点で、例えるならば地図があっても地名が記されていない状況と言える。材料に関する人類の知識は論文等のテキストとして蓄積されており、これを学習に取り込むことができれば、材料の意味を捉え、かつ人間にも理解しやすい表現が獲得できると考えられる。

この背景を踏まえ、我々の研究では結晶構造と論文の紐づけにより、結晶構造に説明文を付与した約40万件のデータセットを構築した。さらに、テキストと結晶構造の対照学習により両者を共通の埋め込み空間に写像し、テキストの意味と結晶構造の特徴を対応付けることを試みた（Fig.1）。その結果、それぞれのembedding間の距離として、任意のテキスト（例：superconductor）と結晶構造の類似度を直接的に評価できることを示した。これにより、あたかも画像検索のように任意のテキストを用いて目的の材料を検索したり、大規模な物質空間を「地名」と共に俯瞰的に理解したりすることが可能となった。

本講演では、これらの研究成果を改めて紹介するとともに、深層学習技術がデータの背後にある「意味」をどのように捉え、表現することを可能にするのか、コンピュータビジョンや自然言語処理分野における成功例にも触れながら、その仕組みと応用について掘り下げて議論する。

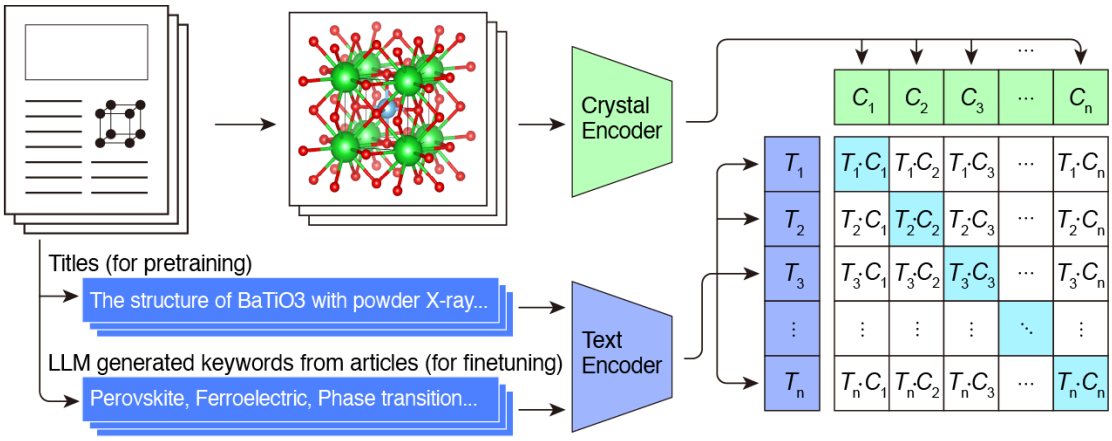

Fig. 1 Overview of our strategy, Contrastive Language-Structure Pre-training (CLaSP).

References

1. Yuta Suzuki *et al* 2022 *Mach. Learn.: Sci. Technol.* **3** 045034

マルチモダリティデータとグラフ表現学習 ～生物学分野における例～

Multimodality Data and Graph Representation Learning: Case Studies in the Field of Biology

株式会社 Preferred Networks　井形 秀吉

Preferred Networks, Inc., Hideyoshi Igata

E-mail: igata@preferred.jp

複数の種類の情報(モダリティ)を統合的に処理することが様々な分野で関心を集めている。本発表では、生物学のマルチモーダルなデータに Graph Neural Network (GNN) を適用した研究を紹介する。

本研究では、生物学の中心的な概念である遺伝情報の伝達・変換 (セントラルドグマ) を扱う。シングルセル解析技術の進歩により、同じ細胞から複数の情報 (モダリティ) を一度に取得可能となっているが、これら全てのデータを取得するのは費用がかかり、技術的にもまだ発展途上である。そこで、手元にある単一のモダリティデータから他のモダリティを予測する手法が重要となっている。

ここでは、生物学的プロセスを理解し、これをグラフ構造で表現することにより、モダリティ間の関係性の予測を行う。具体的には、生物学的プロセスを表現する生体分子グラフと、深層学習パラダイムである GNN を組み合わせる。GNN は、グラフの隣接するノードの情報を元に各ノードの特徴量を更新し、最終的にはグラフ全体の特徴表現を生成する。これを実際の細胞に由来する生物データに適用し、ヘテログラフとして捉えることで、複数の種類のノードやエッジから派生する情報を利用する。

その結果、学習データの数が少ない場合や使用できる特徴量が少ない場合に、予測対象モダリティの制御に関与する特定のグラフ構造を用いることで予測性能の改善が見られた。特に予測対象となるタンパク質の発現量制御における生物学的プロセスとの関連性を捉えることが、予測性能向上に寄与したと考えられる。一方で、サンプル数や特徴量が十分なケースでは本モデルを用いることで飛躍的な予測性能の改善は認められなかった。

予測対象の制御に関与する生物学的プロセスをグラフ構造として考慮したモデルでは「少数サンプルかつ少数特徴量」の条件下で、より頑健な推定ができる可能性が示唆された。比較的限られた条件だが、臨床検査値などの検体数が少なく遺伝子の測定値も網羅的でないデータへの応用が期待される。

参考文献：

東一織, 井形秀吉, 遺伝子に関するグラフを利用したモデルの開発

https://tech.preferred.jp/ja/blog/model-learning-using-gene-graph/

機械学習を用いた物質分類と有望物質群の特徴抽出

Material Classification and Feature Extraction of Promising Material Group based on Machine Learning

東京科学大学総合研究院フロンティア材料研究所[1]　○高橋亮[1]

Materials and Structures Laboratory, Institute of Integrated Research, Institute of Science Tokyo[1]

○Akira Takahashi[1]

E-mail: takahashi.a.f9db@m.isct.ac.jp

近年、文献収集や実験・計算シミュレーションのハイスループット化により非常に大規模な材料データベースを構築・入手することが可能となっている。我々はこうした材料データベースから所望の特性を実現するための構成元素・原子配列の特徴を抽出することを目的とし、機械学習に基づいて対象物性が類似し、かつ構成元素・原子配列にも共通な特徴を持つような物質分類を行う手法を開発した[1]。本講演では本手法の解説と大規模第一原理計算材料データベースに対する適用例の紹介を行う。

結果の一例として、本概要では Materials project database[2]に含まれる酸素を含む 7,981 物質についてバンドギャップを基に分類した結果を示す。対象物性についてランダムフォレスト回帰[3]を行い、得られたモデルに基づいて各物質を *z*-vector に変換し、その z-vector に対してクラスタリングを行うことで物質分類を行った。回帰分析に用いた特徴量に関しては、Matminer[4]に実装されている化学式・結晶構造から求まる 696 個の記述子から、Recursive feature elimination[5]を用いて選択した。

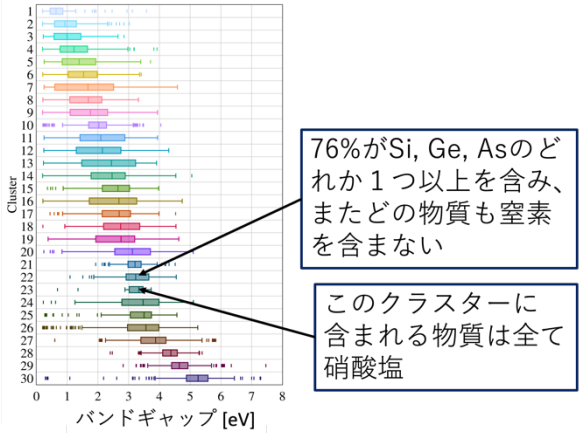

図 1　バンドギャップを考慮し、酸素を含む 7,981 物質を 30 物質群に分類した結果。文献 [1] (CC BY 4.0) から一部改変して引用。

クラスタリングを用いて 30 物質群に分類した結果を図 1 に示す。各物質群に含まれる物質のバンドギャップは狙い通り概ね類似しており、また異なる特徴をもつ物質は物質群としても異なるものに分類されていることがわかった。

さらに講演では複数物性を同時に考慮する場合やスペクトル物性への拡張についても解説を行う予定である。

本研究は科学大（旧東工大）の佐藤暢哉氏・清原慎氏（現東北大）・大場史康氏、東北大の熊谷悠氏、横浜市大の寺山慧氏、NIMS の田村亮氏との共同研究である。(所属は研究当時)

参考文献

[1] N. Sato, A. Takahashi, S. Kiyohara, K. Terayama, R. Tamura, and F. Oba, *Adv. Intel. Sys.* **6**, 2400253 (2024)
[2] A. Jain et al., *APL Mater*. **1**, 011002 (2013)
[3] L. Breiman, *Mach. Learn.* **45**, 5 (2001)
[4] L. Ward et al., *Comput. Mater. Sci.* **152**, 60 (2018)
[5] I. Guyon et al., *Mach. Learn.* **46**, 389 (2002)