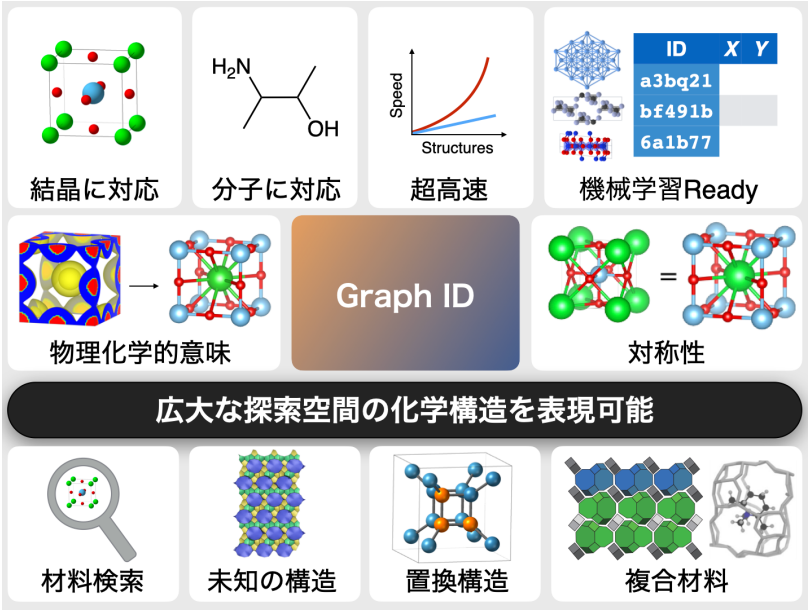
PRESS RELEASE

東京大学
科学技術振興機構（JST）

化学構造の「共通 ID」を開発
――材料データベース統合で探索や機械学習を加速――

発表のポイント

◆ 結晶や分子の原子配列に対して、重複なく固有の識別子（ID）を付与する新アルゴリズム「Graph ID」を開発。
◆ 化学構造を数学的なネットワーク（グラフ）として捉えることで、従来の自動命名手法では困難だった微細な構造の違いを正確に識別可能に。
◆ 世界中に分散する材料データベースの統合や、機械学習、材料探索のための高速な検索エンジンの構築など、材料科学のデジタルトランスフォーメーション（DX）への応用が期待される。



Graph ID

発表内容

東京大学大学院工学系研究科の中山 哲 教授と、村岡 恒輝 准教授らによる研究チームは、世界中に分散する膨大な材料データベースの統合を可能にする画期的な識別子「Graph ID」を開発しました。本研究成果は、蓄電池、触媒、半導体などの新材料探索において、世界規模でのデータ統合と重複排除を可能にし、開発スピードを劇的に向上させることが期待されます。

近年、ハイスループット計算技術（注 1）の普及により、未知の材料を含む膨大な構造データが日々生成されています。これらのデータは「Materials Project」や「AFLOW」といった国際的なデータベースに蓄積されていますが、それぞれが独自の管理体系を持っているため、「あるデータベースに登録された材料が、別のデータベースのどの材料と同じか」を即座に判断することは困難でした。

これに対し、従来は専門家による手作業での命名が行われてきましたが、数百万件を超えるビッグデータを処理することは現実的に不可能です。また、既存の自動命名手法では、数値的な誤差や座標系の取り方の違いにより、同一の構造を別物と判定したり、逆に似て非なる構造を混同してしまったりという問題がありました。

本研究チームが開発した「Graph ID」は、化学構造を数学的なグラフ（注 2）として捉えることで、これらの問題を解決しました。Graph ID では各原子の周囲の環境を反復的に解析し、その構造固有の「指紋」となるハッシュ文字列（注 3）を生成します。

検証の結果、Graph ID は、高い精度、高速性、汎用性という優れた特性を示しました。従来の対称性に基づく手法では判別が難しかった、複雑な結晶構造や吸着分子を含む表面構造も正確に識別可能です。また、データベース内の新規構造の照合にかかる計算コストはごく小さく、従来のペア比較法に比べて大幅な高速化を実現しました。本成果は、結晶に限らず、分子や表面構造など、幅広い化学構造に適用可能です。

さらに、Graph ID を用いて世界最大級の 3 つの材料データベース（Materials Project、AFLOW、OQMD）を統合的に解析し、異なるデータベース間で共通している材料を特定することに成功しました。これにより、複数のデータベースを横断した統合データセットを構築することが可能となりました。

研究チームは、本技術を科学コミュニティの共通基盤とするため、Graph ID を生成するプログラムコードをオープンソースとして公開しました。また、15 万件以上の既知構造に ID を付与したデータベースも併せて公開しています。今後、この共通 ID が「材料のマイナンバー」のように普及することで、AI を用いた新材料予測や、世界中の研究者が知見を共有するプラットフォームの構築が加速すると期待されます。

発表者・研究者等情報

東京大学大学院工学系研究科
化学システム工学専攻
中山　哲　教授

附属総合研究機構
村岡　恒輝　准教授
兼：同研究科化学システム工学専攻

論文情報

雑誌名：Nature Communications
題　名：Universal graph-based identifiers of chemical structures for linking large material databases
著者名：Koki Muraoka, Taku Tanimoto, Tsubasa Munekata, Akira Nakayama
DOI：10.1038/s41467-026-74536-5
URL：https://www.doi.org/10.1038/s41467-026-74536-5

研究助成

本研究は、JST 戦略的創造研究推進事業 さきがけ（課題番号：JPMJPR2378）、JSPS 科研費（課題番号：JP21K20551、JP22K14751）の支援により実施されました。

用語解説

（注1） ハイスループット計算技術：コンピュータを用いて、膨大な数の候補材料の特性を自動的かつ高速にシミュレーションする技術。

（注2） グラフ：点（ノード）とそれらを結ぶ線（エッジ）で構成される「ネットワーク」のこと。

（注3） ハッシュ文字列：元のデータから一定の計算規則に従って生成される、固定長の短い文字列。データの「指紋」のように機能し、データの同一性確認に利用される。

問合せ先

<研究内容について>
東京大学大学院工学系研究科
教授　中山　哲（なかやま　あきら）

<機関窓口>
東京大学大学院工学系研究科 広報室

科学技術振興機構 広報課

<JST 事業について>
科学技術振興機構 戦略研究推進部 グリーンイノベーショングループ
安藤　裕輔（あんどう　ゆうすけ）