Computational Material Discovery

機械学習による新材料探索

膨大に増え続けるデータは、それを活用することで新たな価値を生み出す可能性を秘めています。IBMはこれまで、機械学習によりデータを解析することで世の中で実際に起きる課題に対する解決策を見つけ、新たな価値を提供してきました。機械学習技術の発展とともにその応用領域、適用される産業の幅はますます拡がっています。

既知の材料データを機械学習により解析することで新しい材料を探索する技術「Computational Material Discovery（以下、CMD）」は、実験を中心とするこれまでの材料開発とは全く異なる手法で、材料探索に新たな価値を提供します。

1. 無限に近い新材料発見の可能性

実験や偶然の発見などを通じて存在が知られている材料は、実際に使われたかどうかは別として数十億個とも言われています。しかし、原子の組み合わせの可能性から考えれば、ほぼ無限に近い材料がまだ存在すら知られておらず、そこには新たな材料が見つかる大きな可能性があります（**図1**）。これまでのように実験や発見ではなく、既知の材料データを機械学習で解析することによって、これまでとは全く異なる手法で未知の材料の候補を見つけることができます。多くの産業で実証されているように、材料開発においても機械学習の技術が、新たな可能性とともに破壊的な価値を提供します。

本稿では、機械学習を応用して新しい材料の候補を発見するCMDの全体像を紹介します。

2. 物質データを活用した新材料の探索

これまでに存在が知られている数十億個の材料の多くは、特許や論文、企業も含めた研究機関が保有する実験データなどの記録など、さまざまな形のデータとして存在していると考えられます。個々の材料データには、意図的かそうでないかにかかわらず、経験やノウハウ、知識といったそれまでの蓄積が含まれています。実験による新しい材料の開発では、こういった蓄積を活用してきたことが少なくなく、大きな投資の下に構築された知的財産から新たな価値を生み出してきました。一方で、新

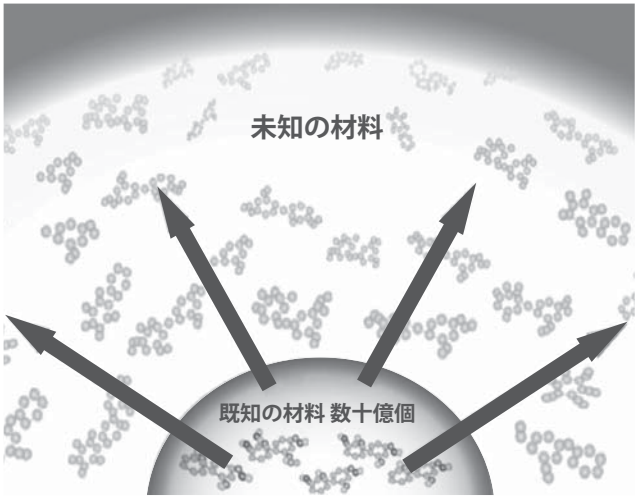

図1. 新物質探索には無限の可能性

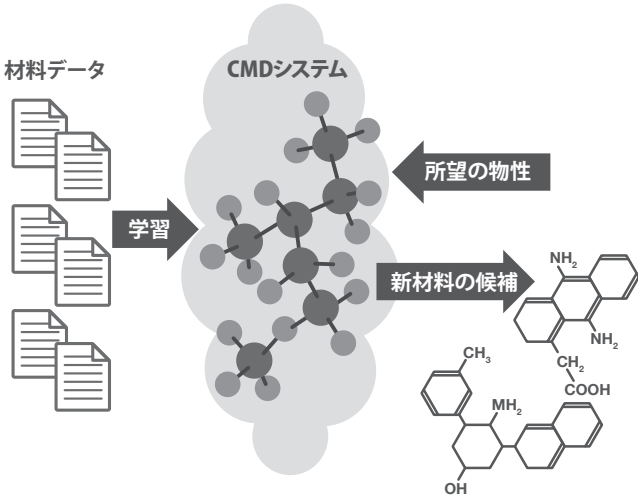

図2. CMDの概念図

たな探索がこれまでのノウハウや経験に束縛される可能性も否定できません。

CMDは、既知の材料データから、材料の構造と物性値の情報を機械学習させることで、構造と物性についての予測モデルを構築することができます（**図2**）。この予測モデルでは、過去の蓄積を取り込みつつもその過度な束縛の影響を抑えることで探索できる新しい材料の可能性を拡げることができます。

3. CMDの概要

CMDは、機械学習モデルの構築に必要な材料データを準備するところから始まります。材料が持つ化学構造の形態的な特徴と、その物性値（ガラス転移点や屈折率など）の対を材料データとして用いることで、予測モデルを構築します。ここで重要なことは、機械学習の対象となるデータは、数値等が表形式に整理された「構造化データ」でなければならないという点です。材料の物性情報はそれ自体が数値の並んだ「構造化データ」を成しますが、化学構造の形態的な情報は数値で表現されていないため、そのままでは機械学習を用いることができません。そこで、化学構造が持つ特徴的なパターン（例えば、芳香環、二重結合、骨格形状など）を特徴ベクトルと呼ばれる数列にエンコードすることで、化学構造の形態的な情報を「構造化データ」に変換することができます（**図3**）。

このようにして構造化された材料データを機械学習にかけることで、新材料探索のための予測モデルを構築します。ここでいう予測モデルとは、材料の構造に関する情報を入力すると、その材料が持つ物性値を予測するモデルのことです。その概要を**図4**に示します。材料の構造は、通常は数百から数千次元（種類）からなる特徴ベクトルにより表現されますが、ここでは図示できるよう2種類の特徴ベクトル（特徴Aと特徴B）で表現できると仮定しています。また、ターゲットとなる物性値としてXとY（例えば、融点と沸点など）を仮定します。このとき、既知の材料データの分布をプロットすると、構造と物性値のおおよその相関を知ることができます。この相関をより定量的に得るために、物性値XとYに対して、それぞれのデータ分布にフィットするような曲面を描きます。このフィッティングの過程を「学習」といい、でき上がった曲面を「予測モデル」と呼んでいます。

予測モデルの種類は、扱う材料や物性値の種類や数に応じて、線形回帰やサポートベクター回帰、ランダムフォレスト、ディープ・ニューラル・ネットワークなどの中から適切なものを選択します。表現力の高いモデルを選択すると、データの微小な変化を捉えたより複雑なモデルを作ることができますが、既存の材料データに過剰に適合するように学習してしまい、新しい未知のデータに対して正しい予測ができないモデルになるケースもあるため、モデルの選択には注意が必要です。

予測モデルを構築したら、次は新構造を発見するプロセスです。XとYそれぞれに関する予測モデルの曲面上で、所望の物性値を満たすような構造的特徴AとBの組を見いだします。**図4**に示した「所望の特徴領域」の「積集合」をとることで、目指す物性値を満たす構造的特徴の集合が得られ

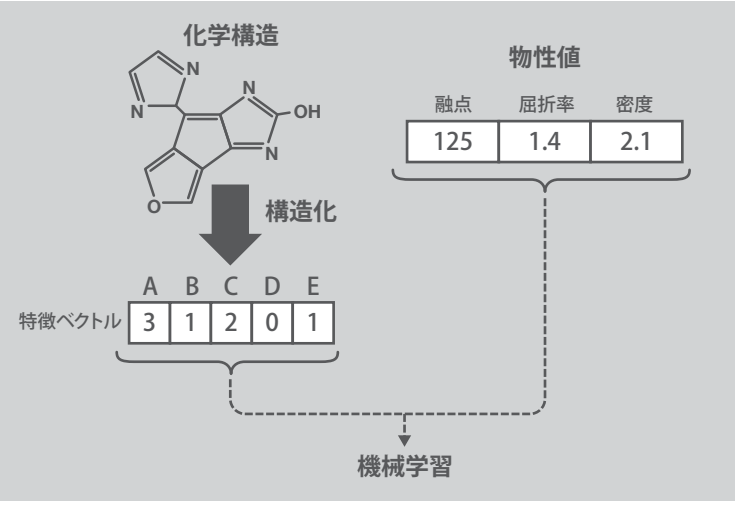

図3. 材料データの構造化

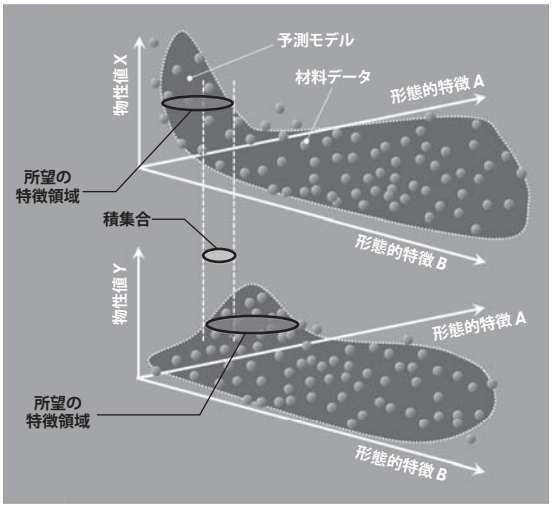

図4. 新材料探索のための予測モデル構築

ます。このような解集合を求める手法には、特徴空間を網羅的に調べ上げるグリッドサーチや、複数の仮想的な粒子が相互に通信しながら特徴空間内を探索する粒子群最適化など、さまざまな解探索のアルゴリズムが用いられます。

最後は、得られた特徴ベクトルの候補を実際の化学構造にデコードする、構造生成プロセスです。ここでは、特徴ベクトル中に埋め込まれた構造情報を頼りに、化学構造を自動的に組み上げていくボトムアップの生成アルゴリズムを用います。ただし、特徴ベクトルには化学的な安定性や合成可能性に関する情報が入っていないため、生成される構造が必ずしも化学的に意味のあるものとは限りません。そこで、生成の過程で化学的な安定性や材料としての有用性に関する、構造上の知見を取り入れた最低限のスクリーニング・プロセスを複数設けることで、最終的には有意な化合物のみを複数列挙することができます。ここで取り入れる知見は多くの場合、材料ごとの分野依存性が大きいため、お客様と打ち合わせを繰り返すことで、分野に応じたノウハウや専門知識を徐々に取り込んでいきます。ただし、専門知識を過剰に取り込むことで化合物を専門的常識の範囲内に束縛することがないよう、適度な生成の自由度を担保しておくことが重要です。

4. CMDの実例

CMDの手法により新しい化合物を発見する実例を紹介します。本実例の目標は、3種類の物性について、新しい組み合わせの値を持つ化合物を発見することです。まず、180個の有機化合物の材料データを準備します。ここでは、物性値としてLog P(分配係数)、TPSA(極性表面積)、MolWt(分子量)の3種類を選びます。**図5(a)**に、それぞれの物性値のヒストグラムを示しています。予測モデルは基本的には内挿の範囲でのみ精度が保証されていますので、ターゲットとなる各物性値の各々は、データセットの範囲内である必要があります。ヒストグラム内に、ターゲット物性値:(Log P, TPSA, MolWt)=(4.5, 85, 380)を矢印で示しています。

180個の物質それぞれを特徴ベクトルにエンコードし、それぞれの物性値に対する予測モデルを構築します。このモデルを用いてターゲット物性値を満たす特徴ベクトルの候補を列挙すると、100種類程度の候補が得られます。これらの候補の中から構造の生成に適した特徴ベクトルを選択し、ボトムアップで構造を生成すると、約700種類の新しい化合物が得られました。検証として、これらの化合物の物性値を原子団寄与法により計算すると、いずれもターゲット物性値と非常に近い値を持っていることが確認されました。得られた化合物の一例を**図5(b)**に図示しています。

この実例で注目すべきは、学習に用いた材料データでは、log PとTPSAにほぼトレードオフの関係があるに

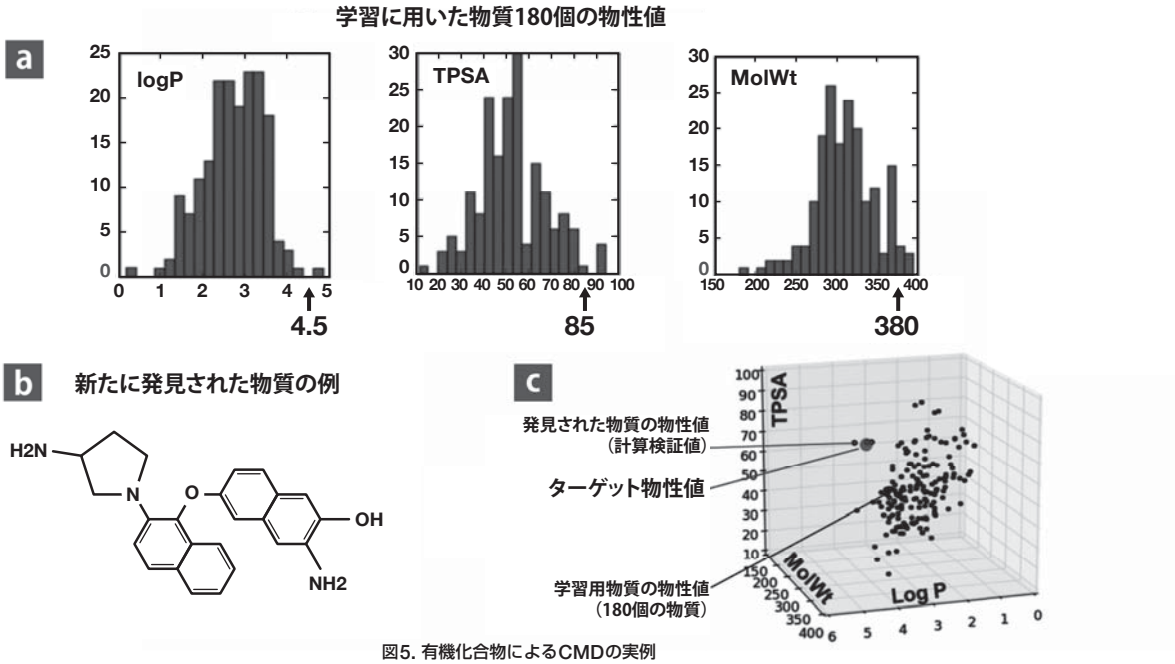

図5. 有機化合物によるCMDの実例

もかかわらず、両者ともに高い値を持つ物質を発見できたということです（**図5(c)**）。硬さと軽さを両立させるなど、従来の常識的なトレードオフを超えた新材料を発見することは、あらゆる材料開発における普遍的な要求です。CMDはこのような要求に対して、ソリューションを提供することができます。

5. おわりに

東京基礎研究所の技術であるCMDは、IBM Research Frontiers Institute（RFI）[1]のAccelerated Material Discovery（AMD）の中核技術の一つであるアナリティクス・アプローチとして、RFI参画企業に利用可能なプラットフォームとして提供されると同時に、JDA（共同開発契約）を通じて、お客様個別の課題に対する材料発見のソリューションとして提供されています。

CMDは、材料開発を通じて製造業全体に破壊的なイノベーションをもたらすことを目指しています（**図6**）。すなわち、自動車や航空機などの製品メーカーに、新材料開発をも含めた製品設計のソリューションを提供することで、設計の自由度を大幅に増大し、従来にないスケールで製品の差別化を図ることができます。また、そのために必要な材料を提供する材料メーカーには、新材料開発のソリューションを提供することで、圧倒的な材料の差別化が可能になります。このように、製品メーカー、材料メーカー、IBMの技術力が一体となった強固なエコシステムを構築することで、製造業におけるイノベーションの質とスピードを高め、製造業をめぐる世界の産業構造を次なるステージへと推進します。

[参考文献]
[1] IBM : Research Frontiers Institute, http://www.research.ibm.com/frontiers/

日本アイ・ビー・エム株式会社
事業開発. 事業開発部
シニア・テクニカル・スタッフ・メンバー

中川 茂
Shigeru Nakagawa

IT企業で青色発光半導体レーザーなどの研究に7年半従事、その後USのスタートアップ企業で事業立ち上げ、さらに日本のスタートアップ企業でChief Designerとして製品開発を担当。2005年にIBM東京基礎研究所に入社。サーバー用光インターコネクト、Computational Material Discovery、脳を模した機械学習する計算チップを開発するグループを担当。2011年、林巌夫賞を受賞。

日本アイ・ビー・エム株式会社
東京基礎研究所
サイエンス&テクノロジー
スタッフ・リサーチャー

武田 征士
Seiji Takeda

慶應義塾大学特任助教を経て、2012年IBM 東京基礎研究所に入社。光インターコネクトの技術開発を経て、CMDおよびIBM Research Frontiers Institute（RFI）Accelerated Material Discovery（AMD）の東京チームの技術戦略リーダーを務める（22ページを参照）。また、脳を模した機械学習デバイスに関する東京大学との共同研究にも従事。

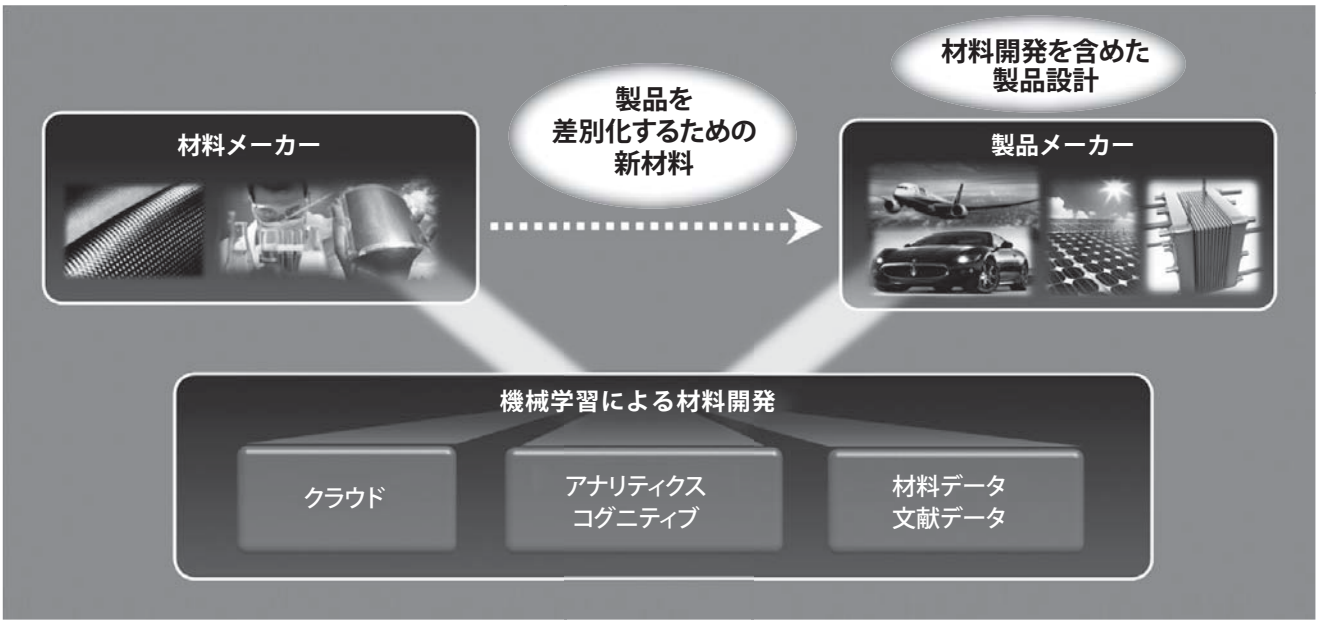

図6. CMDの提供する価値