マテリアルズ・インフォマティクス(MI)の現状と将来展望

徐 一斌（物質・材料研究機構）

１．材料研究における課題

材料研究における課題の全ては、プロセス(P)、構造(S)、物性(P)、いわゆる PSP の関係の中で生じる。すなわち、プロセスによって材料の構造を制御し、材料の構造制御によって望む物性を実現することが材料研究の中心であるといえる。

これまで、材料 PSP 関係の解析、制御に関する研究手法は主に実験であった。19 世紀以降、周期表や結晶構造など材料に関する多くの物理と化学の法則が発見され、量子力学や統計力学に基づいて物性理論の研究手法も発展してきた。近年では、計算科学が著しく進展し、状態図の計算や、第一原理による物性の予測例も大変増えてきている。しかし、材料中で起こる様々な現象は、空間的には原子のサイズからバルクまで、時間的にも電子が関与する現象のフェムト秒レベルから、腐食やクリープのような現象の数十年、数百年レベルまで、幅広い空間・時間スケールで起きていて、現在の物性理論とシミュレーションで解明できるのは、その中のほんの一部に過ぎない。全体的にみると、新材料の開発は、相変わらず trial-and-error の実験が中心となっている。

マテリアルズ・インフォマティクス(MI)は、これまでの材料実験や計算などで蓄積してきた大量のデータを、機械学習などのビッグデータ解析技術を用いて解析し、データに潜んでいる PSP 関係を見つけ出し、材料開発の指針を示す新しい研究手法である。この手法によって、材料研究の伝統的な方法を根本から改革し、材料開発の時間とコストを大幅に短縮するという期待を多くの研究者が抱いている。筆者は、2002 年から材料データの収

集とデータベースの開発に従事し、NIMS の無機材、高分子や鉄鋼材料など、様々なデータベースの開発に参画してきた。さらに、2014 年から機械学習の手法を材料研究に取り込み、2015～2020 年の間、JST イノベーションハブ構築支援事業「情報統合型物質・材料開発イニシアティブ」(MI^{2}I:“Materials research by Information Integration” Initiative)プロジェクトに参画し、データプラットフォームと伝熱制御材料のグループリーダに従事した。これら材料研究の活動において、データ科学の活用に関して、大きな可能性を感じた一方、課題も多く見えてきた。本稿は、筆者のこれまでの経験に基づいて、MI の有効性、課題と将来展望について述べる。

２．材料データの現状

MI は材料データに基づいた技術であるため、まず、その基盤構築の歴史と現状をレビューする。

19 世紀 80 年代から、科学者はデータの重要性を認識し、大規模な材料データ収集を行ってきた。F. K. Beilstein は、1881 年から、有機化合物の特性、スペクトル、合成方法などのデータを文献から収集し、ハンドブックとして纏めた。1989 年に出版された Beilstein ハンドブックの第 4 版は、計 503 巻 440,000 ページが含まれている。Landolt と Börnstein も 1883 年から、無機化合物も含めた幅広い材料対象に関して、データ収集を行い、Landolt-Börnstein ハンドブックとして出版した。現行版の Landolt-Börnstein ハンドブックには、72,000 の化学システム、150,000 の化学物質と 530,000 の物質-特性ペアが掲載されている。その他に、金属合金を対象とした ASM ハンドブックや、熱物性を対象とした TPRC ハンドブックなども出版されている。

Pauling File[1]は、1995年にJSTとスイスMPDSの共同プロジェクトとして立ち上げられたもので、現在、物質･材料研究機構（NIMS）とMPDS社が著作権を共有している無機材料データベースである。1900年から1000種類以上の科学ジャーナルで発表された論文から無機材料の状態図、結晶構造、特性データを網羅的に収集し、有機的にリンクしている。Pauling Fileのデータは、現在12個のデータベース製品、あるいは、ハンドブックとして販売、提供されている。NIMSからは、無償版のAtomWork[2]と有償版のAtomWork-Adv[3]の二つのWebシステムを提供している。2020年12月現在、AtomWork-Advのデータ数は、状態図44,554件、結晶構造334,450件、特性410,401件となる。

近年、計算科学の進歩とスーパーコンピュータの普及により、電子構造や状態図などの計算データが急速に増加している。特に、第一原理計算を用いて生成した電子構造や計算物性のデータベースがMIの重要なデータ源となっている。代表的なものとして、UC Berkelyが中心として開発したThe Materials Project[4]は、約13万無機化合物の計算データが含まれている。Duke Universityが中心として開発したAFLOW[5]は、仮想物質も含めて、3百万以上の化合物データを公開している。FHI Berlinなどヨーロッパの大学と研究機関が共同開発のNOMAD[6]は、世界の研究者から提供した計算データ1億件以上を公開している。

しかし、このようにデータベースなどで収集され、誰もが簡単に入手できるデータは、材料全体のほんの一部に過ぎない。もし全ての材料を一つの材料空間（図１）とすると、この材料空間の中で、既に人類が、「知識を持っている」、「探索したことがある」、「作ったことがある」空間は、灰色で示したような局所的なものであり、その中でさらに、データを入手できる部分は黒い点のようになる。なぜ、このような状況になっているのだろうか。材料の化学組成と構造の組合せは、ほぼ無限の可能性がある。地球上の

元素は全部で約 100 種類ある。2 種類の元素の組合せは約 5 千通り、3 種類では約 16 万、10 種類では約 17 兆通りがある。その膨大な数の組合せに加えて、分子構造、結晶構造、ナノ構造やミクロ組織など、材料の特性に影響を与える多くの要素を考えると、これまでの実験や計算などの探索が至った領域は、ほんのわずかな部分しかないのである。さらに、材料の作製は、何千年以上の長い研究開発の歴史があるにも関わらず、現在入手可能な材料データは、近代に論文などとして発表されているものに限られ、多くても直近 200 年分くらいしかない。AtomWork-Adv.に基づいて算出した全化学システムに対して、データのある化学システムが占める割合を図２に示す。1 元系に関しては、周期表にある 100 種類の元素に対して、ほとんどが網羅されている。２つの元素から構成される２元系になると、全体 5000 種類程度の内の 72%であり、これも概ね網羅していると言える。しかし、３元系に対しては、現在、データがあるのは全体の 16%、４元系に至っては 0.6%、さらに、５元系以上はほぼデータがないという状況になる。

３．世界の MI 研究プロジェクト

2011 年にアメリカの MGI(Materials Genome Initiative)国家プロジェクトが発動して以来、日本、ヨーロッパ、中国など世界中で多くの MI 研究プロジェクトが発足した。MI 研究プロジェクトの目標は、大きく二つある。一つは、優れた機能を有する新しい物質の探索である。アメリカの Materials Project とスイスの MARVEL[7]は、その代表的なプロジェクトである。それらのプロジェクトは、主に科学計算の手法を活用し、既知の物質データを利用して未知の物質の存在可能性と特性を予測する。

MI のもう一つの目標は、材料プロセス条件の最適化と実用材料の性能向上である。それを目標とする代表的なプロジェクトは、日本の SIP 革新構造材料/MI システムと中国の MGE(Materials Genome Engineering)である。それらのプロジェクトは、材料の組成、構造、組織に着目し、実験データとマルチスケール計算を用いて、実用可能な材料の組成、構造とプロセス条件を設計する。

MI^2I は、新物質の探索と実用材料設計の統合を目指すプロジェクトである。そのプロジェクト構成は、図３に示したように三層構造である。下層は、基盤となるデータプラットフォーム、中間層は、データ解析や記述子抽出など要素技術に関する研究、上層は、電池、磁石、伝熱制御材料をターゲットとした実用材料の設計と開発である。MI^2I では、研究チームのほかに、企業を主体としたコンソーシアムも結成した。プロジェクト終了時、88 の企業がコンソーシアムに参加していた。MI は、学術界だけではなく、企業からも多くの関心が集まったことが分かる。

4. MI の有効性

MI は、大量のデータと統計アルゴリズムに基づいて、材料変数の相関性をモデリングする手法である。原則的に、物理と化学のプロセスが分からなくても、材料の特性予測や最適化が可能であるため、物理理論とシミュレーションの限界を超えて、変数の多い複雑な材料現象の研究に特に有用と考えられる。以下に、我々が材料界面を対象として、MI を用いた研究例を紹介する。

4.1 界面熱抵抗の予測と利用

界面熱抵抗は、熱伝導のキャリアであるフォノンあるいは電子が、材料界面での反射あるいは散乱により生じた熱抵抗のことである。熱が界面を通過する方式として、様々なチャンネルがある。例えば、電子やフォノンの弾性散乱と非弾性散乱、電子―フォノンカップリングなど。それぞれのチャンネルに対して、モデリングやシミュレーションなど、多くの理論研究を行ってきたが、実際の材料界面での熱伝導は、複数のチャンネルが同時に機能する場合が多い。また、界面熱抵抗の影響要因は、界面両側の材料の結晶構造や物性だけではなく、界面の化学と相組成、接触面積、ナノ構造なども考えなければならない。それらの要因を全部取り込む物理モデルは、まだ存在しない。そこで、我々は、MIの手法を導入した。まず、これまでの界面熱抵抗に関する実験と理論研究に基づいて、界面の物理的、化学的、材料的な特徴を表す 12 個の記述子を選定し、論文から収集した界面熱抵抗の実験データと、データベースやハンドブックから収集した材料の結晶構造、比熱、融点などの物性データを合わせて、訓練データセットを作成した。それを利用して、いくつかの機械学習モデルを訓練し、高精度に界面熱抵抗を予測できる機械学習モデルを得た[8-10]（図４）。次に、それらの機械学習モデルを利用して、8 万種類以上の界面から、高熱抵抗を有する可能性の高い 25 界面を絞り出し、その中の Bi/Si 界面に対して、実験検証を行った。Si 基板の上に Bi 薄膜を蒸着し、単一 Bi/Si 界面を形成させ、その熱抵抗を測定した。測定結果は、51.8×10^{-9} m^2KW^{-1} であり、LSBoost という機械学習モデルの予測結果 50.7×10^{-9} m^2KW^{-1} とよく一致した。さらに、コンビナトリアルスパッタ成膜装置（COSCOS）を利用して、分散状態の異なる Bi ナノ結晶を含有したアモルファス Si の複数のナノ複合材料サンプルを全自動で作製し、その熱伝導率も測定した。熱伝導率の最小値は 0.16W/mK となり、これまでに報告された緻密な無機複合材料より小さい熱伝導率を示した[9]。素材である結晶 Bi とアモルファス Si の熱伝導率は、それぞれ約 $7Wm^{-1}K^{-1}$ と 1 $Wm^{-1}K^{-1}$ で

あることを考えると、複合材料の熱伝導率がそれより大きく下回るのは、Bi/Si の高い界面熱抵抗が原因であることは明白であろう。

4.2 粒界 Li^+伝導度の影響要因の解析

全固体電池の基盤材料である固体電解質は、高いイオン伝導率が要求されている。しかし、固体電解質材料は、多結晶やセラミックスが多いので、粒界や界面におけるイオン伝導率は、粒子内部より一桁以上小さく、全体のイオン伝導率のボトルネックとなっている。界面におけるイオン移動のメカニズムも複雑であり、多くの要因に影響されている。特に、材料の作製条件は、イオン伝導率に影響を与えることが分かっているが、定量的な解析モデルが困難であるため、最適化の指針がなかった。我々は、機械学習を利用して、粒子内部と粒界の Li^+伝導度と材料の作製条件や、組織構造、結晶構造などの相関性を解析した[11)]。利用したデータは、論文から収集した 96 Li 固体電解質サンプルの作製条件や粒径サイズ、粒内と粒界のイオン伝導率、および AtomWork-Adv から抽出した結晶構造や原子配位などである。機械学習を用いて、粒内・粒界のイオン伝導率と、構造や作製条件の記述子との相関性の解析により、各記述子の重要性を評価した結果、焼結温度や時間などの焼結条件および粒径サイズは、粒内と粒界の Li^+伝導率の両方に大きな影響を与えることが明らかになった（図 5）。さらに、粒内イオン伝導率に対して、Li の占有率と元素比が重要である一方、粒界イオン伝導率に対して、密度、分極率や添加物が重要であることも分かった。この結果は、新しい固体電解質作製条件の最適化、および、新しい超イオン伝導物質の探索に重要な情報を提供した。

5.　MI の課題

この数年間の MI 研究を通して、データ駆動材料研究と開発には、いくつかの重要な課題も見えてきた。

5.1 材料空間の定義

まずは、材料空間の定義である。理論上の材料空間は、材料の組成、構造や特性など全記述子を含む空間であり、その空間の各点は、それぞれ、一つのユニークな材料に対応しているはずであるが、一つの材料を特定するためには、原子、分子、結晶構造、ミクロ組織、作製方法と条件、各種の特性など、多くの記述子が必要であり、さらに、我々は、どのような記述子が必要なのかを、実質的に把握できていない。また、実験条件などの制限により、成分、構造や特性などを全て解析された材料は、ほぼ皆無であり、材料の全記述子を含む空間は、定義できないのである。実際に、各材料分野は、その分野の専門知識に基づいて、独自のサブ空間、つまり、一部の記述子を定義して対処しているのである。このやり方には、二つの問題がある。第一に、そのサブ空間の定義が間違いあるいは不完全であると、機械学習の結果が誤りである可能性が高い。特に、知識と経験の少ない新しい分野では、記述子の定義が難しいので、機械学習の信頼性を慎重に評価する必要がある。第二に、分野を超えたデータの共有と再利用が困難な点である。なぜなら、データ収集時には、その分野のサブ空間（記述子）の定義に従ってそれらを収集したため、別のサブ空間で利用する時に、記述子の補完が必要となる。しかし、サンプル情報など、後から入手・補完できない重要なデータも沢山ある。

5.2 材料空間の分断

MI のもう一つの重要な課題は、機械学習モデルの適応範囲である。機械学習の原理は、既存の材料現象から規則性やパターンを発見し、それを利用して新しい現象を予測することなので、その前提として、新しい現象と既存現象が、同じ規則とパターンに従う必要がある。しかし、材料空間は、様々な材料境界によって分断されていることがある。その一つの例は、相境界である。合金の状態図を思い浮かべると、金属の組み合わせによって、2 つの相が共存する場合もあれば、単一の相が存在する場合もあり、相としても結晶構造の違いや液体、固体といった集合状態の違いもある。材料現象の規則性は、それらの相境界のところで不連続となり、崩れる。これは、この相境界を跨ぐ新材料を探索する場合、厳しい問題となる。相境界の他に、金属/非金属、有機/無機など、様々な材料境界もある。探索したい領域が、既知領域の外になった場合には（図 6）、既知データから見つかった規則性は、どこかで崩壊する可能性があることを常に意識する必要がある。つまり、既知のデータから得られた機械学習モデルは、新しい材料へ適用可能かどうか、慎重に検証する必要がある。

5.3 データ収集

MI の最も大きな問題は、データ不足である。世の中では、データ量が多過ぎてコンピュータが処理しきれないことがよく問題とされるが、材料の場合は、2 章で述べたように、データが多すぎる問題がほとんど無く、データが少ないのが常に問題となる。

データの生成を加速するために、近年、ハイスループット計算とコンビナトリアル実験の技術が、注目されている。第一原理など材料のシミュレーションから得られるデータは、システマティック性と網羅性が良く、ばらつきも比較的小さいので、MI に使いやすい

一方、現実を完全には反映しておらず、未だ実用材料への適用には距離があるという問題点も残っている。一方で、実験データについては、そのコストが「高い」という問題を抱えている。「高い」の意味は、材料を合成する際のコストだけでなく、物性を測定する際のコストも含んでいる。近年では、コンビナトリアル実験によって、自動化、ロボット化することで網羅的に材料データを収集するアプローチもあるが、自動化によって人件費は確かに減らせるが、実際に材料を作るときに使用している原材料費やエネルギーなどに関しては特に削減されていない。また、計測技術の制限により、取得できるデータの種類が限られている問題もあるので、目標を特定してその周辺を探索する方法としてはとても有効であるものの、網羅的な実験データ作成という意味では、限界がある。よって、過去の文献、および、日常の材料実験からのデータ収集と蓄積は、今後も重要なデータ収集手法として強化していく必要があると思われる。ここで気を付けるべきなのは、将来そのデータを様々な分野や目的で利用できるように、材料情報や実験条件などをなるべく共通なフォーマットを用いて、全面的に記述することと、既存の化合物や物質データベースとリンクを付けることである。そして、個々の材料のデータを、物質や化合物のレベルでお互いにリンクすることにより、多様な目的で再編集、再利用することが可能となる。

6. MI の展望

MI は材料データに基づいた技術であることから、今後の展開については、かなりの程度でデータの量と質とに依存するであろう。現在のデータ不足問題の解決方法として、下記に示す幾つかの可能性が考えられる：

（1）インターネットを利用して、内容の関連性のあるデータベースを共通 API（Application Programming Interface）により統合的に検索、利用できるような技術開発が進んでいる。この技術により、入手可能なデータ源が増加し、物質探索などを目的とする大規模、汎用的なデータセットの作成が容易になる。

（2）特定の材料問題と領域に対して、長年のデータ蓄積やコンビナトリアル実験により、小規模ではあるが、システマティックなデータセットを作成し、パラメータの最適化などに利用する。

（3）シミュレーションから得られるデータは、近年、品質も向上し、量も急速に増えて来ている。将来、量子コンピューティング技術の実用化により、さらに現実に近い材料のシミュレーションも可能になるであろう。そうすれば、サイバー空間の中でシミュレーションと機械学習を用いて実用材料を設計する日が来るかもしれない。

MI 活用の核心は、材料問題の設定とデータセットの準備である。実験や計算と違って、MI はゼロからの研究手法ではない。機械学習用データの収集、機械学習の結果に対す理解、および、それに基づいた信頼性の評価には、対象材料分野の専門知識が不可欠である。MI を適切に利用すれば、物理や化学の限界を超えて、新物質の探索、複雑な現象の解析やパラメータの最適化など様々な材料問題に役立つ。現在、様々な機械学習ツールが開発され、プログラミング不要なアプリケーションも提供されていて、普段使っている PC でも手軽に利用できる。将来、材料の研究者とエンジニアたちが、インターネット上のデータベースからデータを取得し、自分のデータと統合して、Excel を使うように機械学習を利用してデータを解析したり、次の実験を設計したりすることが日常の仕事の一部になることを期待したい。

参考文献

1) Pauling File, https://paulingfile.com/

2) 無機材料データベース（AtomWork），https://crystdb.nims.go.jp/

3) 無機材料データベース（AtomWork-Adv.），https://atomwork-adv.nims.go.jp/?lan=jp

4) The Materials Project, https://materialsproject.org/

5) AFLOW Automatic-FLOW for Materials Discovery, http://www.aflowlib.org/

6) NOMAD REPOSITORY & ARCHIVE, https://nomad-lab.eu/prod/rae/gui/search

7) MARVEL, https://nccr-marvel.ch/,

8) T. Zhan, L. Fang, Y. Xu; Scientific Reprots,7, 7109 (2017).

9) Y. J. Wu, M. Sasaki, M. Goto, L. Fang, Y. Xu, ACS Appl. Nano Mater. 1, (7), 3355 (2018)

10) Y. J. Wu, L. Fang, Y. Xu, npj Comput. Mater. 5, 56 (2019).

11) Y. J. Wu, T. Tanaka, T. Komori, M. Fujii, H. Mizuno, S. Itoh, T. Takada, E. Fujita and Y. Xu, Sci. Technol. Adv. Mater, 21:1, 712 (2020)

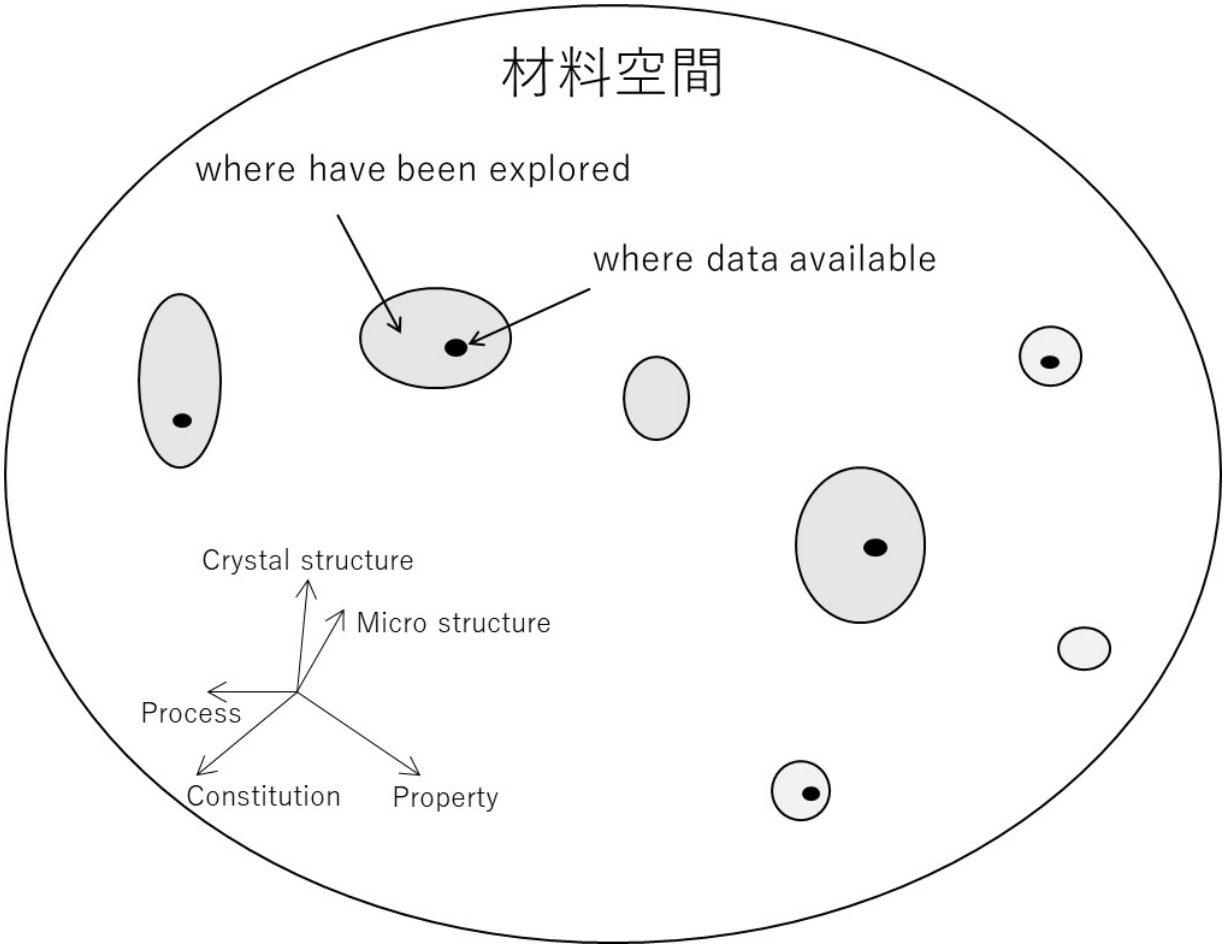


図１　材料空間におけるデータの分布

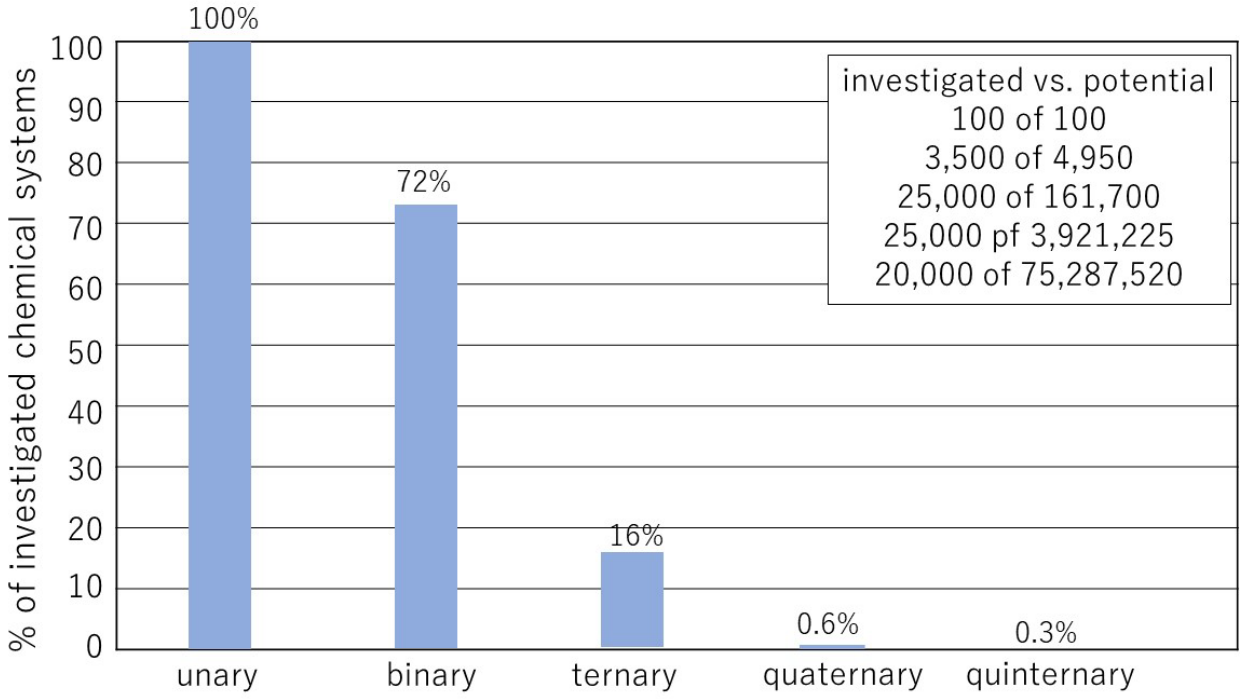
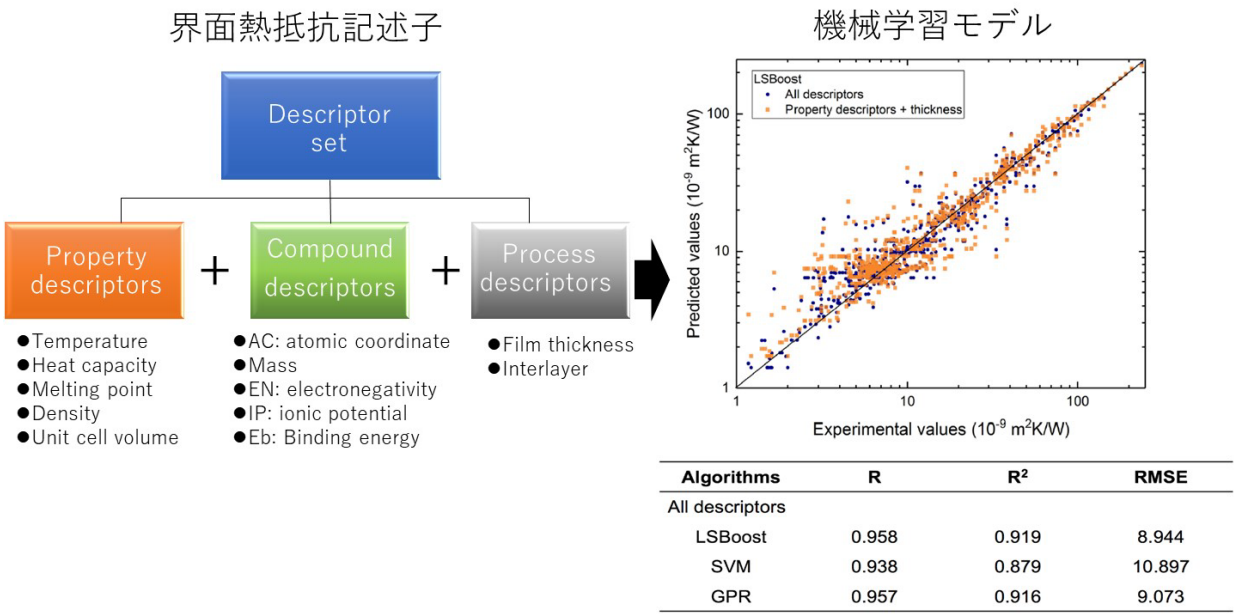


図２　AtomWork-Adv のデータに基づいて算出した全化学システムに対するデータのある化学システムの割合

図３　MI^2I プロジェクトの三層構成



Algorithms	R	R^2	RMSE
All descriptors			
LSBoost	0.958	0.919	8.944
SVM	0.938	0.879	10.897
GPR	0.957	0.916	9.073

図４　界面熱抵抗の記述子と訓練した機械学習モデル

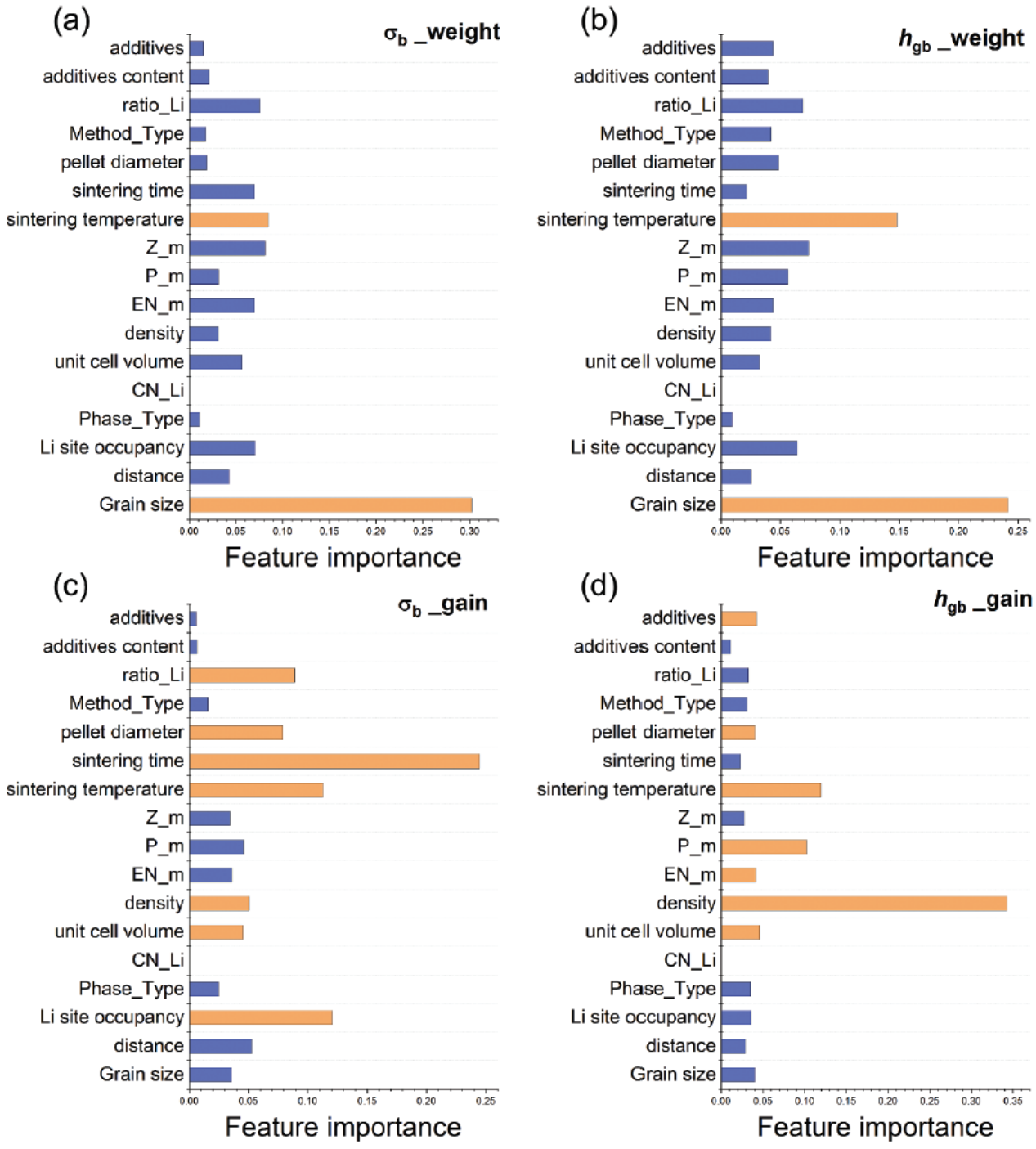


図５　結晶粒内の Li^+伝導率と粒界 Li^+イオン伝導度に対する記述子の重要度。(a) 粒内の Li^+伝導率に対する weight タイプ重要度、(b) 粒界 Li^+イオン伝導度に対する weight タイプ重要度、 (c) 粒内の Li^+伝導率に対する gain タイプ重要度　(b) 粒界 Li^+イオン伝導度に対する gain タイプ重要度

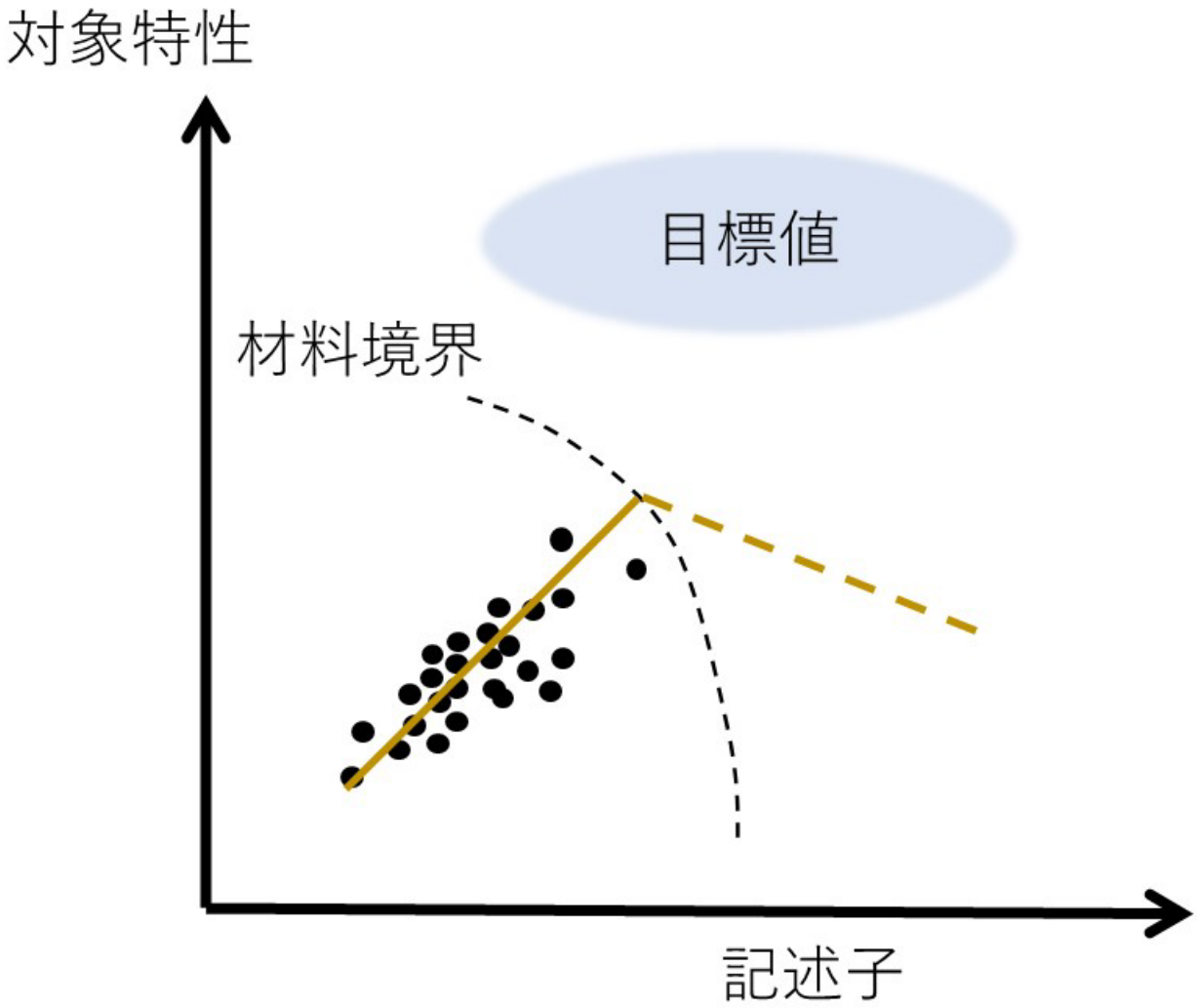


図６　未知な材料領域を探索するために、材料現象の連続性が崩れる材料境界線を超える可能性がある。構築した機械学習モデルの適用範囲を検証する必要がある。