

戦略プロポーザル

データ科学との連携・融合による 新世代物質・材料設計研究の促進 (マテリアルズ・インフォマティクス)

～物質・材料研究を飛躍的に発展させるための新たなパラダイム～

STRATEGIC PROPOSAL

Materials Informatics 0101 000111 0101 00001
Materials Design by Digital Data Driven Method. 00110

研究開発戦略センター（CRDS）は、国の科学技術イノベーション政策に関する調査、分析、提案を中立的な立場に立って行う公的シンクタンクの一つで、文部科学省を主務省とする独立行政法人科学技術振興機構（JST）に属しています。

CRDS は、科学技術分野全体像の把握（俯瞰）、社会的期待の分析、国内外の動向調査や国際比較を踏まえて、さまざまな分野の専門家や政策立案者との対話を通じて、「戦略プロポーザル」を作成します。

「戦略プロポーザル」は、今後国として重点的に取り組むべき研究開発の戦略や、科学技術イノベーション政策上の重要課題についての提案をまとめたものとして、政策立案者や関連研究者へ配布し、広く公表します。

公的な科学技術研究は、個々の研究領域の振興だけでなく、それらの統合によって社会的な期待に応えることが重要です。「戦略プロポーザル」が国の政策立案に活用され、科学技術イノベーションの実現や社会的な課題の解決に寄与することを期待しています。

さらに詳細は、下記ウェブサイトをご覧ください。

<http://www.jst.go.jp/crds>

エグゼクティブサマリー

本提言は、JST 研究開発戦略センター（CRDS）が、今後の物質・材料分野の研究のあり方として、目標機能の「設計」※¹手法の確立が重要であると考え、そのための手段として、物質・材料研究におけるインフォマティクスの活用を検討した結果を戦略としてまとめたものである。

今後の研究の方向の一つとして、課題解決に貢献する物質・材料の機能「設計」は欠かすことのできない重要な概念である。これは、与えられた元素組成や構造から物質の性質を探る従来の研究手順（順問題）とは異なり、社会・産業が求める特定機能を有する材料を合理的に探索する（逆問題）方法を模索するものである。例えば、目的とする機能を有するリチウムイオン電池材料、熱電変換材料、超伝導材料などに用いられる候補物質が合理的に探索できれば、エネルギー・資源問題等の課題解決に大きく貢献することは言うまでもない。

本提言は、物質・材料の「設計」に必要な新たな方法論として、マテリアルズ・インフォマティクス（materials informatics）を「計算機科学（データ科学※²、計算科学）と物質・材料の物理的・化学的性質に関する多様で膨大なデータとを駆使して、物質・材料科学の諸問題を解明するための科学技術的手法」と定義し、その重要性を提案する。その最大の特徴は、計算機の支援なしでは解明できない事象に対する理論的な捉え方を、質的または量的に進展させることにある。これにより、物質・材料研究に関連する膨大なデータを包括的かつ科学的・系統的に取り扱うことで新法則の発見の促進およびそこから実用材料開発までの時間短縮を目的とする。

先端材料の開発はエネルギー、医療、素材、化学など多くの産業を活気づける源である。政府は、過去 10 数年間にわたって、先端材料を探索・設計するための研究開発に多大の投資を行ってきた。しかしながら、今日でも研究室で新発見された先端物質が実用化され市場に出るまでの過程は、一般的に 10-30 年という長期間を要する状態が続いている。多くの産業がグローバル化し、科学技術が一部の先進国のものではなくなった今日、この過程を加速・短縮することなくしてわが国の科学技術力、産業競争力の維持は難しい。

従来（現在）の実験科学の手法では、経験と勘に基づいて多くの物質を合成し、物性を評価してきた。つまり、構造や物性などに関する複数の実験データを様々な因子で整理し、その相関性をもとに機能発現のメカニズムを推定し、理論モデルを構築するという手法である。この手法はこれまで物質・材料研究に大きく貢献してきた。

また、ここ数年の計算科学の進展や計算機コストの低下により、実験結果の電子論レベルにおける理解・解析や、実験前に電子軌道・構造を推定し、目的の機能を得られるかについての洞察・予測が各研究室で日常的に行われるようになった。

このような背景を踏まえ、最近では、物質・材料研究におけるより効果的な研究の進展を目指して、物質創成、計測・解析、理論・計算の 3 本柱の協調による取組みがようや

く実現し始め、実験科学者と計算科学者の実体的なコミュニケーションが進みつつある。しかしながら、基本的には、これだけでは目標機能を有する物質・材料を「設計」することは不可能であり、経験と勘に基づいて試行錯誤を経て目標機能に達するアプローチの域を出ることは難しい。

これを打破する「設計」促進のための新しい手段として、インフォマティクスを導入し、大量・複雑なデータから知識（規則）を帰納的に獲得するアプローチ（データ駆動型）を構築することを提言する。これによって、理論やモデルに基づく演繹的なアプローチ（原理駆動型）との併用によるシナジー効果が生み出され、今後の物質・材料研究を大きく進展させることにつながる。

これまでに合成された物質の数および物性に関する情報量は膨大であり^{※3}、事実、結晶構造や分子式など従来の整理要素だけでは今後の新物質開発に見通しが立ちにくくなってきている。その上、近年の計測技術や計算技術の進歩によって大量のデータが短時間に生み出されるようになった。今後、データ量はますます増加の一途をたどることは疑いないので、その中から意味のある情報を抽出するためには、膨大な物質データを俯瞰的に可視化し、包括的に整理できる高次のコンセプト（新しい軸）が必要である。

近年、データ（数理・情報）科学の分野においては、大量または複雑なデータの中に潜む情報の構造を抽出する手法である多変量、高次元データ解析が飛躍的な発展をとげており、従来とは質的に異なるデータの科学的分析が実現できる域に達している。

全米科学アカデミー（National Research Council of the National Academies of Sciences）が2008年に発行した“Integrated Computational Materials Engineering”の提言の一節には、「物質・材料研究にかかわる計算手法と情報分析手法を統合すれば、現在10-20年かかっている材料開発サイクルを2-3年に短縮できるだろう」との記載がある。このような提言を受けて、米国では、2012年に国家的な取組みとして、“Materials Genome Initiative”がスタートし、具体的なプログラムとして、NSFでは、2012年から“Designing Materials to Revolutionize and Engineer our Future (DMREF)”が、NISTでは、“Building the Materials Innovation Infrastructure : Data and Standards”等の取組みが始まっている。

翻って日本の状況を見ると、物質・材料研究において、シミュレーションによる解析、予測こそ進展しつつあるものの、データ（群）やその他の計算機能力のうち、計算による網羅性や機械学習等による予測や可視化などをまったく有効に活用できていない。今後早急に、実験、計算科学者に加え、データ科学者の連携が求められる。

想定される主な研究対象と科学的手法は以下のとおりである。

- (1) 物質・材料研究者が解決すべき研究テーマとして、「大量で、多種・多様なデータの分析による構造・物性相関の法則の抽出、複雑な現象等の解明」、「大量データからの物性あるいは構造の予測 (design of experiment)」、「最適化手法などを用いた物質（構造）探索」、「マルチスケール・モデリング」、「高次元データを用いた数理モデルの高度化」、「物質空間あるいは解析データの可視化（方法として俯瞰的可視化、ハイライト的可視化、対象として画像・スペクトルのような一次データ、数値群などの二次デー

タ）」などが挙げられる。

- (2) データ科学的手法として、「機械学習」、「ベイズ推定」、「圧縮センシング（スパースモデリング）」、「データ同化」、「逆問題」、「数理モデリング」、「最適化問題」、その他の種々の内挿・外挿手法などが（1）に適用されるべきである。

このようなデータ駆動型アプローチは、一研究室での実施が困難な場合が多く、多様な専門性をもつ人材・グループが複数関与することが必須である。したがって、上記のような研究テーマを促進するためには、マテリアルズ・インフォマティクスのためのオープンなプラットフォームを構築し、アカデミアから産業界に至るまでマテリアル・イノベーションにかかわるすべての人が容易に全データを使用できる環境の整備があわせて進められる必要がある。

以上の取組みを進めるため必要な政府や研究機関の役割は以下の通りである。

◆政府の役割

わが国においては、研究用の計測機器、計算機だけでなく、シミュレーションソフトやデータベースなどの大半が外国製品であり、国際的にはユーザーとしての位置づけとなっている。このことの弊害は、日本が開発国から高く買っているという経済的な問題だけではなく、開発国の研究者が有利に使えるようにソフトウェアが作られるので、日本の研究者がそれを用いて研究成果を発見するのには大きな不利があるということである。

データベース構築や新しいアルゴリズムを実装したシミュレーションソフトの実用化のような基盤整備は基礎研究の推進や研究成果の実用化を主眼とする競争的資金の性格から外れるため、適切な支援がなされない傾向にある。国としては、短期的な成果を望むのではなく、長期的な視点に立った継続性のある支援が必須であり、ハードインフラの整備・共有だけでなく、「知的基盤（ソフトインフラ）」の整備・共有が火急の責務である。これまで（大型）先端計測研究施設・設備の整備、共用化には力を入れてきた。さらに、ビッグデータに対応する情報基盤技術の整備も始まりつつある。しかし、これらと両輪の存在で同等あるいはそれ以上の価値を有する、物質・材料に関する多種多様なデータを研究者が利用できる環境が整備されておらず、そのような仕組みを早急に構築することが喫緊の課題といえる。

◆研究機関・コミュニティの役割

マテリアルズ・インフォマティクスのオープンプラットフォームを構築するためには、国の財政的な支援の下、大学、公的研究機関、学会などの持続的貢献が不可欠である。大学などは、マテリアルズ・インフォマティクスに関する融合的・包括的な研究の遂行を目指し、物質・材料分野における物質創成（プロセス）、計測、計算の研究者の密接な連携を促進し、データ科学との連携・統合による新しい科学技術研究の潮流を起こすための研究の枠組みを作るべきである。材料開発を加速度的に進めるため、物質・材料研究のコミュニティは、分野を越えてデータセット、モデル化した経験式、より高度化されたアルゴリズムなどの共有を図る必要がある。

最も重要なことは、本提言に掲げるような新興・融合分野をアカデミア自らが発見、奨励し、必要な（意欲のある）人材を積極的に登用し、インセンティブを付与する評価システムを構築することである。

具体的には、下記に関連するプログラムとして、実施することを提言する。

●データ駆動型研究プログラムの創設

データ利用・解析手法の FS（可能性研究）プログラムを実施し、手法の確立・高度化を図るとともに、成果事例を創出する。具体的には、年間 1000 万～2000 万円程度の研究費で、物質・材料科学者とデータ科学者からなる複数名の共同提案を原則とするプログラムを始めるべきである。異分野の共同提案を求めることから、公募期間を通常より長めにとり、学会、ワークショップなどを通じた異分野交流期間を確保することが必要である。

●データ統合・研究開発センターの設立

中核的データ統合・研究開発拠点を整備し、データマネジメントの方針、あるいは収集するデータの種別や範囲の決定やデータ・ツールの管理を行う。サービスセンターとしての機能も有し、複数の研究機関や大学が、LOD（Linked Open Data）の概念を導入し、RDF（Resource Description Framework）／XML などの標準的なデータフォーマットを使い、情報を交換し、材料開発に生かすというマテリアルズ・インフォマティクスのシステム構築を進める。

※1：本提言における「設計（デザイン）」とは、材料に求められる機能や性能などを検討し、どのような構成で実現するのかを仕様として決定することを意味する。工学的な概念であり、科学的な手法としての「アナリシス（分析あるいは解析）」の対極にあり、「シンセシス」と表現されることもある。物質・材料研究において、構成の対象は、組織構造、結晶構造、化学組成、電子構造、磁気構造など多面的であり、これらは各元素のもつ自由度（サイズ、電荷、スピン、軌道など）を考慮した上で決定される。解析的に「設計」することは不可能なので、通常は経験と勘による試行錯誤の作業が必要である。本提言は、従来の経験と勘に基づいた手法から脱して、体系的な設計手法を提案するものである。

※2：本提言における「データ科学」とは、複雑あるいは大量のデータから何らかの特徴的なパターンを発見し、有効な知識を得るための技術の総称であり、工学的手法も含む。データマイニング、機械学習が代表例である。情報学、統計学、応用数学（あるいは情報科学、数理科学）のうちデータ処理にかかわる分野が該当する。

※3：物質・材料分野で取り扱う元素の種類は 80 程度であるが、その組み合わせは 2 元系で約 3 千、3 元系になると約 8 万となる。しかし可能な化合物の数は組成比を考慮すると大きく増える。すなわち化合物 $A_xB_yC_z$ で x, y, z を整数として、 $x + y + z = 10$ と限定しても、その組み合わせは 100 通りあり、3 元系全体では 800 万通りになる。同じ組成であっても結晶多系と呼ばれる様々な構造をもつ場合が多いのでその数はさらに増える。

一方で、世界最大の無機結晶構造データベース ICSD（Inorganic Crystal Structure Database）に収録されている 3 元系までの無機化合物データは現在 76000 件であり、同一の化合物で複数のデータが収録されている重複を除くと、実験的に結晶構造が既知の無機化合物の数は 5 万件を下回る。

Executive Summary

The Center for Research and Development Strategy (CRDS) of the Japan Science and Technology Agency (JST) considers the establishment of “design” *1 methods for materials with target properties and functions to be important for the future of materials research. In this proposal, in order to support this activity, our study results concerning the utilization of informatics for materials research are summarized from a strategic standpoint.

As one of the future courses of materials research, the “design” of materials with functions that contribute to the solutions of problems is an essential and important concept. This involves the development of methods to rationally search for materials which possess the specified properties desired by a society or industry (inverse problem), which is unlike the traditional research procedure, in which the properties of a material are discovered from a given elemental composition and structure (direct problem). For example, if candidate materials with target properties and functions for lithium-ion batteries, thermoelectric conversional, or superconductivity can be rationally discovered, such a method will contribute greatly to the business solution in energy and resource issue and others.

In our proposal, as a new methodology necessary for the “design” of materials, materials informatics is defined as a “scientific and technical method of solving problems in materials science using computer science (data science*2 and computational science) in combination with diverse and massive collections of data regarding the physical and chemical properties of materials,” and the importance of this new study area is proposed. Its most major characteristic is to qualitatively or quantitatively develop the theoretical understanding of problems that cannot be solved without the aid of computers. This allows us to deal with a massive amount of data related to materials research in a comprehensive, scientific, and systematic manner, with the aim of both promoting the discovery of hidden rules, new theories and principle, and shortening the time between theoretical discoveries and the development of practical materials.

The development of advanced materials is essential to activate many industries, such as energy, health-care, materials, and chemical industries. For more than a decade, the Japanese government has invested substantially in research and development for the discovery and design of advanced materials. Nevertheless, a lengthy period of between 10 and 30 years is still generally required before an advanced material that is newly discovered by a research laboratory is put to practical use and entered into the marketplace. Without accelerating and shortening this process, it will be difficult

to maintain the scientific and technological capabilities and industrial competitiveness of Japan in the recent situation where many industries are globalized and science and technology are no longer monopolized by certain advanced countries.

Using the traditional (existing) methods of experimental science, many types of substances have been synthesized based on experience and intuition, and their properties have been evaluated. In such methods, a plurality of experimental data regarding the structure and properties of a material is organized on the basis of various factors, and based on this activity, the functional properties of the material are estimated and theoretical models are constructed. This method has thus far made great contributions to materials research.

Due to recent advancements in computer science as well as computational power, and also reductions in computer costs, many laboratories have been able to perform analyses of experimental results to understand at an electronic theory level on a daily basis. Even before an experiment is undertaken, they are able to often estimate electronic orbits and electronic structure in order to predict whether a target function or property will be obtained.

Based on the above background, a collaborative effort between the three core study areas of materials synthesis/processing, measurement/analysis, and theory/simulation, has recently been initiated in order to advance the effectiveness in materials research, which should be resulted in substantial progress in communication between experimental and computational scientists. Fundamentally, however, such collaborative efforts, by themselves do not make it possible to “design” a material which possesses a target property or function. Accordingly, it remains difficult to progress beyond the standard method of obtaining a target function by a trial and error approach based on experience and intuition.

As a new means of promoting “design” beyond the existing approach, we propose the introduction of informatics, with the construction of a data-driven model to inductively gain knowledge (rules) from massive, complicated sets of data. This will produce a synergetic effect in parallel with use of a deductive approach (principle-driven) based on theories or models, which may lead to great advancements in future materials research.

The total number of substances that have heretofore been synthesized, and the amount of information on the properties of these synthesized substances are enormous.*3 In fact, it becomes increasingly difficult to obtain prospects for the development of novel materials using only traditional guiding attributes, such as crystal structure and molecular formula. Furthermore, recent advances in

measurement and simulation techniques have resulted in the short-time generation of an enormous volume of data. Because the volume of this data will doubtless continue to increase steadily in the future, in order to extract meaningful information from such massive data, a higher-order concept (new axis) is required to obtain a holistic view of this massive materials data and to organize such data in a comprehensive manner.

In recent years, in the area of data (mathematical or information) science, rapid progress has been made in multivariate, high-dimensional data analysis, in terms of methods to extract the information structure inherent in massive or complicated data. Methods for the scientific analysis of data which are qualitatively different from traditional analysis methods are now at a realizable stage.

“Integration of computational techniques and information analysis techniques that are related to materials research will reduced the materials development cycle from its current 10 to 20 years to 2 to 3 years,” according to a proposal written in “Integrated Computational Materials Engineering,” issued in 2008 by the National Research Council of the National Academies of Sciences. Having accepted this proposition, the American government initiated a national effort called the “Materials Genome Initiative” in 2012. Regarding specific programs, for example, the NSF started “Designing Materials to Revolutionize and Engineer our Future (DMREF)” , while the NIST started “Building the Materials Innovation Infrastructure: Data and Standards” in 2012.

Turning to the present state of affairs in Japan, although simulation analysis and prediction are in progress in materials research, prediction and visualization by exhaustive computation or machine learning, using data or data sets and computational capabilities, are not utilized at all. Moving forward, an urgent collaboration between experimental scientists and computational scientists as well as data scientists is demanded.

Major expected research subjects and scientific methods are as follows:

- (1) The topics to be solved by materials researchers include: “Extraction of rules for structure and physical property correlation and clarification of complicated phenomena, etc. by the analysis of a wide diversity of massive data,” “Prediction of properties and structure using massive data (design of experiments),” “Search for materials (structure) using optimization techniques, etc.,” “Multi-scale modeling,” “Advancement of mathematical models using high-dimensional data,” and “Visualization of material space or analytical data (Method: Holistic visualization or highlight visualization; Target: raw(primary) data, such as images

and spectra, and processed(secondary) data, such as numerical groups in papers).”

- (2) As data scientific methods, “Machine learning,” “Bayesian inference,” “compressed sensing (sparse modeling),” “data assimilation,” “inverse problems,” “mathematical modeling,” “optimization,” and other interpolation or extrapolation methods should be applied to the above (1).

It is frequently difficult for only one laboratory to perform this data-driven approach; thus, the involvement of more than one person and group with a variety of specializations is essential. In order to promote the above-mentioned research subjects, it is therefore also necessary to build an open platform for materials informatics to serve as a common ground in which all of the persons involved in materials innovation, from the academic sector to the industrial sector, can easily use all data.

The roles of the Japanese government and research institutes to advance the above-mentioned efforts are as follows:

◆ Role of the Japanese government

In Japan, most of the measuring equipment and computers used for research, as well as simulation software and databases are foreign products, and the status of Japan is as a user, internationally. The harmful influence of this situation is not limited to an economic problem in which Japan purchases products from countries of manufacture at high prices. Because software is made so that researchers in the countries of manufacture can use it to their advantage, Japanese researchers are at great a disadvantage in the use of such software to obtain the results. For example, consider the situation in which a special analysis may be conducted using certain software, so that special visualization is conducted, and the efficacy of a given medicine can be determined. Now, further consider that this software may be commercialized and exported without revealing this special analysis method to persons other than the software developer and the university researchers involved in the collaborative research. In this case, the Japanese researchers who have to use that standardized software to conduct research will be greatly handicapped, from the outset.

The construction of infrastructure, such as database construction and the commercialization of simulation software equipped with new algorithms, does not tend to receive suitable support because of the nature of a competitive fund that focuses on the promotion of basic research and practical application of research results. The Japanese government should not expect short-term results, but should provide continuous support from a long-term standpoint, and the preparation and sharing of, not only of a hardware infrastructure, but also an “intellectual foundation (software

infrastructure)” are its urgent duty. Thus far, special attention has been given to the development and commoditizing of (large) advanced measurement research facilities and equipment. In addition, the preparation of an information infrastructure technique to cope with big data is at an initial stage. However, there is no environment in which researchers can use a wide variety of data on materials, which is inseparably related to or more valuable than the above-mentioned efforts. Accordingly, the immediate construction of such a system is an urgent task.

◆ Role of research institutes

To build an open platform for materials informatics, continuous contributions from universities and public research institutes are essential, along with financial support from the Japanese government. With the aim of implementing integrated and comprehensive research on materials informatics, they should build the framework for a research system to promote close cooperation among researchers in materials process, measurement, and computation in materials science to create a new current of scientific and technological research via cooperation and integration with data science. To accelerate materials development, the materials research community needs to share data sets, simulated empirical formulae, and advanced algorithms across study fields.

The most important activity is to build an evaluation system to allow the academics themselves to discover and encourage new-risen and integrated areas, as written in our proposal, to positively recruit the necessary motivated persons, and to provide incentives.

Concretely speaking, we propose that the following activities should be conducted as related programs.

● Launch of a data-driven research program

A feasibility study (FS) program for a method for data exploration and analysis will be conducted to establish and advance this method, and result cases will be created. In particular, a program which involves basic collaborative proposals by materials and data scientists should be initiated with an annual research fund of approximately 10 to 20 million yen. Because collaborative proposals involving different research fields are demanded, it will be necessary to provide a longer application period than usual, in order to guarantee a cross-boundary exchange period through academic conferences and workshops.

● Establishment of a data integration/research & development center

A core site for data integration and research & development will be established to determine a data management policy (i.e., the type and range of data to collect) and to

manage data and tools. This site will provide with the functions of a service center, to promote the construction of a materials informatics system in which research institutes and universities introduce the concept of Lined Open Data (LOD) and use standard data formats, such as the Resource Description Framework (RDF)/XML to exchange information for application to materials development.

*1: “Design” in our proposal means the examination of a function or property demanded for a material, and the determination of composition to realize it as a specification. This is an engineering concept, at the opposite end of the spectrum from “analysis” as a scientific means, and is sometimes called “synthesis.” Key factors describing substances in materials R&D are multilateral, such as texture, crystal structure, chemical composition, electronic structure, and magnetic structure, and these are determined in consideration of the degrees of freedom (such as size, electric charge, spin and orbital) of each element. “Design” in an analytical manner is impossible; accordingly, a trial-and-error approach based on experience and intuition is usually required. Our proposal brings forward a systematic design method to surmount the dependence on the traditional method based on experience and intuition.

*2: “Data science” in our proposal is a generic name for techniques to discover characteristic patterns from complicated or massive data to gain effective knowledge, and includes engineering techniques. Typical examples of data science include data mining and machine learning. Among information science, statistics, and applied mathematics (or computer science or mathematical science), areas related to data processing are applicable.

*3: Although the number of elements dealt with in materials science is approximately 80, the number of combinations of these elements is approximately 3,000 in a binary system and approximately 80,000 in a tertiary system. However, the number of possible compounds is greatly increased if composition ratios are considered. For example, in the case of a compound $A_xB_yC_z$, where x , y and z are integers and $x + y + z = 10$, the number of combinations is 100, with 8 million for the entire tertiary system. Even if the composition is the same, crystal polymorphism frequently occurs to yield many types of structure, so the number of combinations increases.

The number of data on up to tertiary inorganic compounds listed in the Inorganic Crystal Structure Database (ICSD) is 76,000 at present, and if duplicated data, which consists of plural data for each compound, are not counted, the number of inorganic compounds with experimentally-known crystal structures is less than 50,000.

目 次

エグゼクティブサマリー（和・英）

1 章 研究開発の内容	1
2 章 提案を実施する意義	
2－1 現状認識および問題点	3
【コラム】物質と材料のギャップ	
ー材料のマルチスケール性と計算・計測手法の発展ー	7
【参考】計算材料科学と数学の協働によるスマート材料デザイン手法の探索	
ー階層構造を解析するー	10
2－2 社会・経済的効果	11
2－3 科学技術上の効果	12
【コラム】バイオインフォマティクスとマテリアルズ・インフォマティクス	14
3 章 具体的な研究開発課題	15
【参考】物質・材料研究における代表的なデータベース	20
4 章 研究開発の推進方法および時間軸	21
【コラム】研究開発プロジェクトを有効に推進するための推進体制	
（領域と役割の連携）	25
付録 1 検討の経緯	26
付録 2 国内外の状況	29
付録 3 参考文献	36
付録 4 専門用語解説	37

1 章 研究開発の内容

本提言におけるマテリアルズ・インフォマティクス（materials informatics）とは、計算機科学（データ科学、計算科学）と物質・材料の物理的・化学的性質に関する多様で膨大なデータを駆使して、物質・材料科学の諸問題を解明するための科学技術的手法である。その最大の特徴は、計算機を用いて人間の能力では処理できない規模のデータから重要な事象を導き出し、理論的な事象の捉え方を質的または量的に進展させることで、物質・材料科学の重要な事象を解明することにある。これを用いることによって、物質・材料に関する設計原理の発見や新物質の発見から材料開発に至るまでの各過程が加速され、全体の時間を短縮することが可能になる。

物質・材料科学の分野では、計測・分析技術の進歩、コンビナトリアル手法などの実験手法の発展、スーパーコンピュータに代表される計算機能力の向上、第一原理計算などの計算科学の進歩などにより、物質・材料に関するデータが短時間で大量に得られる時代になっている。しかし、人間の情報処理能力には限界があり、データ（群）に含まれている多彩な情報の多くは抽出されることなく未利用になっている。

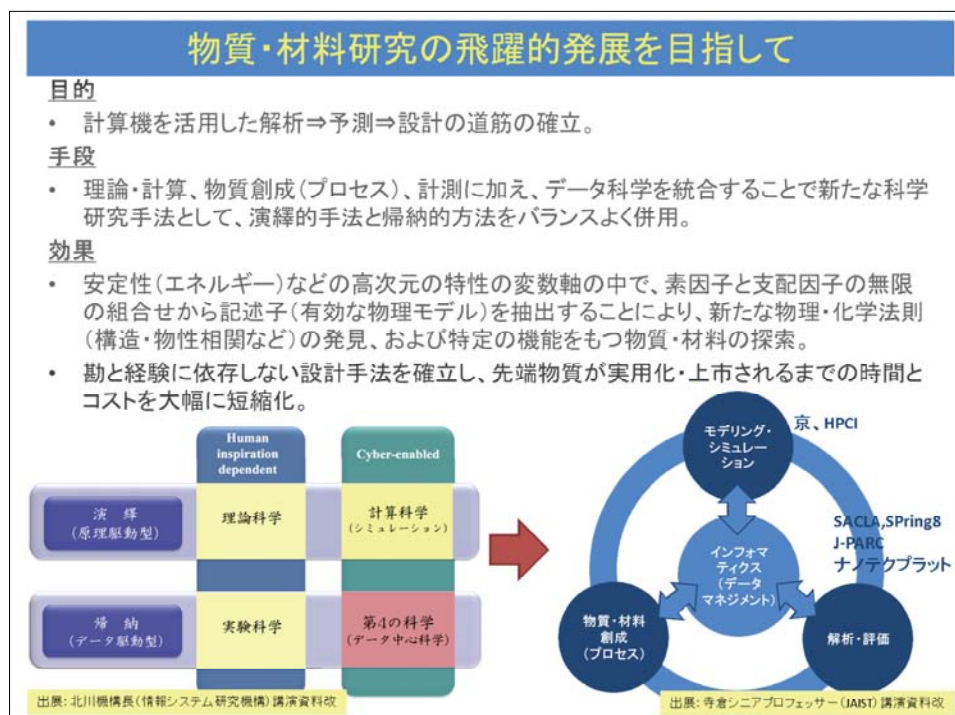
一方で近年、データ科学においては、大量で複雑なデータの中に潜む情報の構造を抽出する手法（多変量、高次元データ解析など）が飛躍的に発展し、従来とは質的に異なる解析を実現できるようになっている。計測や計算から得られるデータや、文献・データベースに収録された大量で複雑なデータから、物質・材料の構造と物性との相関や、実験や計算で探索すべき物質・材料の候補群などに関する所望の情報を得るためには、(1) データの取得・蓄積から解析までを一体として捉え、データ科学で開発された手法を応用・展開することにより、(2) 物質・材料の研究開発におけるそれら手法の有効性を検証するとともに、(3) さらに有効な新たな手法を開発していく、必要がある。以下の例に述べるようなデータ駆動型研究開発の推進を提案する。

- 膨大な物質構造に関する空間を金属、無機、有機、高分子などの別なく俯瞰的に可視化し、機能に基づいて物質群を包括的に構造化・可視化できる高次のコンセプト（新しい軸）を見出すための研究開発。
- 大量のデータから、物質・材料の構造と物性の相関、あるいは目標機能・特性とその他の物性値とを関連付ける法則を抽出するための適切なデータ解析手法（データマイニングなど）の研究開発。
- 実験で探索すべき物質・材料の構造や物性（多次元空間）を予測するための機械学習や最適化などの手法の研究開発。
- マルチスケール現象に関する構造、組織、プロセス間の法則を抽出し、数理モデル化するための、大量データのデジタルアーカイブ化・定量化手法、および適切なデータ解析手法の研究開発。
- 数理モデルの予測精度を向上させるため、当該数理モデルと現象の実測データとの統合によるデータ同化や圧縮センシングを応用する研究開発。
- X線や中性子線などを用いた先端的な計測手法により得られる複雑な画像やスペクトルデータなどから、有用な物性情報を系統的に抽出し、可視化する手法の研究開発。

実験、計測、計算により得られるデータや解析結果、およびそのための計算ツールや解析ツールは、相応の研究開発資源（ヒト・モノ・カネ）の投入によって獲得されたものであり、これらをさらに有効に活用するためのインフラを整備してマテリアルズ・インフォマティクスを広く普及させ、もって材料開発を加速することが国際競争力強化のために不可欠である。データベースは、そのようなインフラの一つとして想定されるが、単にデータや情報の検索機能が高いというだけでは不十分であり、データベース全体がデータ解析の対象になり得る条件を備えていなければならない。また、現存する世界中のデータベースを有効活用するためには、オープンデータ（Open Data）の考え方を考慮したデータの形式やオントロジー（用語の階層・構造化）などによる表現方法の標準化、データの一括ダウンロード手段の担保など、データベースの構造や機能に関する戦略的な設計が必要である。また、データの取得・蓄積・加工・解析、さらには解析結果のデータ取得プロセスへの効率的なフィードバックを実現するため、全自動化システムの開発も必要となる。

以上のように、マテリアルズ・インフォマティクスの研究開発を推進する上では、インフラの整備とそれを実現するための関連技術の研究開発とを並行して実施することが必要である。

本提言におけるマテリアルズ・インフォマティクスの狙いは、物質・材料の研究開発にデータ科学を導入することだけではなく、原理駆動型の演繹的手法とデータ駆動型の帰納的方法とをバランスよく併用する研究開発手法の確立であり、今後の物質・材料研究の飛躍的發展を目指すものである。



2章 提案を実施する意義

2-1 現状認識および問題点

物質の世界は膨大な多様性をもつ。たとえば、多元系化合物には巨大な自由度があり、3元系では、元素組成が1:3:2のように一桁の整数比で表されるものに限っても理論上800万通りあり、結晶多形、不規則相を考えると、その種類はさらに膨大なものになる。この中には材料として有用な特性をもつものが多数存在している可能性がある。

これまでの新物質探索や材料開発では、多くの場合、研究開発者の経験と勘に基づいて多くの物質・材料を合成し、その構造や物性などの実験データを様々な観点から整理し、そこに見出される規則性をもとに、目的とする特性をもった物質・材料を探索するというプロセスを経ており、実際にこのようなプロセスで開発された物質・材料は産業・社会に様々な恩恵をもたらしてきた。

しかし、新物質の発見から実用化までに要する時間が非常に長いことも事実であり、また、これまでに合成された物質や物性データが巨大な量に膨れ上がったため、従来通りに経験と勘に基づいて研究開発の目的に応じた見通しを立てたり、あるいは機能や製造コスト面で既存の材料を凌駕する新物質を短時間で探索することは極めて難しくなっている。

日本はこれまで物質・材料分野の研究開発において世界をリードしてきた。しかし、韓国、中国の発展を見てもわかるように、産業のみならず科学技術も一部の先進国のものではなくなりつつある。これまでの科学技術的知見の蓄積をもとに、引き続き物質・材料分野において国際的地位を維持・確保するためには、課題解決を志向した材料開発において、材料に求められる機能を特定し、所望の機能を有する物質構造を合理的に「設計」する手段が必要である。

他方、現在の物質・材料科学を俯瞰すると、個別の学術研究が推進される方向、すなわち専門分野の細分化が進行する一方であり、学問の壁を乗り越えて多くの方法や体系を融合する新しい方向、つまり「設計」を指向して、分野としての物質科学、材料科学や手法としての実験科学、計算科学などを統合した全体的理解へのアプローチが必要となる。

「設計」の対象になる物質の「機能」は、それを構成する各元素のもつ自由度（サイズ、電荷、スピン、軌道など）の範囲内で取りうる可能な熱力学的安定・準安定・非平衡な「構造」に対応して決定される。さらに、それら「構造」を実現するための多様な「プロセス」の開発は別途必要であり、プロセスの工夫による材料の微細構造あるいはメソスコピック構造を制御することが肝要である。

材料機能の多くは、様々なスケールでの階層構造の性質が創発的に現れたものが多く、第一原理計算で演繹的に求めることができる物質の一次データ（基礎的なバルク物性）そのものではない。このように、物質の一次データと材料機能の関係は一義的ではないことから、大きな空間スケール（多くの原子が関係する構造、物性、現象）および長い時間スケール（長い時間にわたる現象）も独立して考慮する必要があり、参照すべきデータの数

は膨大なものとなる（P7.【コラム】物質と材料のギャップー材料のマルチスケール性と計算・計測手法の発展ー参照）。

以上のように、機能の「設計」には多様かつ膨大なデータが必要とされるが、近年、そのようなデータの取得は以前よりも格段に容易になっている。理由の一つは、測定技術の進歩によって原子・分子スケールでの構造や物性データが得られ易くなり、また、強力なX線や中性子を利用するSPring-8、SACLA、J-PARCなど大型施設を利用すれば、物性の起源である電子状態やスピン配列などのデータが直接取得可能となったこと。もう一つは、計算機の性能向上や、価格低下に加えてアルゴリズムの急速な発展を背景とする計算科学の普及が著しく、実験による検証が困難もしくは不可能な事象に対しても、背後にある電子論的メカニズムに関するデータを第一原理計算などにより得ることが可能になっているためである。このようにして物質に関する複雑で大量のデータが加速度的に供給される状況にある。

このような状況における一つの問題は、物質・材料に関する膨大なデータがその量の増大に相応して有効に活用されているか否かということであろう。ここでいうデータとは、論文やデータベースとして公表されているものだけでなく、研究開発の現場で日々生成され、蓄積されつつあるものも含む。これらのデータ群の中には未確認の新たな知識が埋もれている可能性があるが、日本ではデータの集合を科学的に整理・分析し物質探索や材料設計につながる有用情報を引き出そうというような、データ科学の観点からの系統的な取り組みが不十分である。つまり、その分、蓄積されたデータは活用されずに放置されている可能性が高い。

データ科学の長所は、大量の複雑なデータセットからデータ間の新しい関係を見出したり、重要な情報の抽出を可能にすることにある。具体的には、生のデータでは見えない背後の現象をデータの統計的な処理のみをベースにして推定し現象の有無を判別すること、また、非常に高次のデータを単純化して因果関係を推論すること、さらには非常に少数のデータから多様な事象を推定することも可能になっている。そればかりではなく、データを研究開発者が理解しやすい形式で提示するためのデータ可視化技術、材料開発で利用される数理モデルの精度を上げるためのデータ同化技術など、データ科学を活用することでデータから材料開発の指針を抽出する局面が多数存在する。

このようなデータ科学を活用する観点から、物質・材料に関わる既存の公開データベースの現状をみると、以下のように改善を要する点が多い。

- 多くのデータベースが研究組織別、材料の種類別、物性・特性別に構築されているが、データ構造や公開方法がデータベース構築者によって異なるため、データ科学の手法を適用するにしても円滑に進めることができない。
- データベースの価値は、情報量と継続的な更新によって決まるため、整備・運用費用の確保が必要である。物質・材料研究機構（NIMS）材料情報ステーションや産業技術総合研究所計測標準研究部門のように、材料データベースの運用維持をミッションとしているところは、一定の予算が確保されており、材料データベース

の公開・運用を継続的に推進している。それでも、先端計測機器や計算機器のハードインフラの整備に比して予算の措置が十分ではなく、新規のデータを追加できない場合も多い。

- ・ 各研究者や研究室はこれまでの研究成果や研究に必要な文献データなどを独自に収集し、データベースを構築して候補物質の探索などに使用している。しかし、データベース構築は研究者の直接的な評価につながらないため、日々データを生み出している研究者には、共用のためのデータを貯蔵・公開するモチベーションが働かないことも、データの整備・公開を難しくしている要因となっている。これは、シミュレーションソフトや本提言に示すようなデータ解析のためのアプリケーションソフトなどの整備・公開においても同様である。
- ・ 国内では日本機械学会や日本材料学会のように、部門や委員会を設置して材料データベースを開発している組織もあるが、委員によるボランティアベースの活動に依存している状況にある。例えばニューガラスフォーラムのように、組織として運用・維持を行うとともに、NEDO プロジェクトの外部資金によってデータベース開発・整備を行っている機関もあるが、プロジェクト終了後は十分なメンテナンスのための人材と費用が断たれるため、開発終了時の状態のまま公開のみ継続せざるを得ないといった例がほとんどである。

一方、以下の例のように、データの利活用のニーズは確実に存在する。

- ・ 実験研究者は、一般に ICSD (Inorganic Crystal Structure Database) などの結晶構造のデータベースや先行論文を参照して候補物質を探索するが、構造既知の物質のデータが限定的であり、かつ電子状態も不明であることが多く、データベースのみから得られる情報には限界がある。そのため、データベース中の結晶構造に加えて、第一原理計算などによる電子状態や観測データを補足して情報をセットで参照し、数理・情報科学的手法で解析できるようにする必要がある。
- ・ メゾ、マクロ領域における材料開発では、計測による実験データとモデリングとの強力な連結が不可欠である。モデリングには、材料の 3 次元の多様な構造、すなわち、単・多結晶相、アモルファス相、粒界、欠陥など多元パラメータによる 3 次元分布の情報が必要である。しかしながら、鉄鋼材などの構造材料においては、プロセスと特性を結びつけるための構造組織のデータがいまだにアナログ情報で、デジタル・データとして整備されていないため、そのままではモデリングにつなげられない事情を抱えている。3 次元微細構造情報の収集や分類方法の取組みは一部で進行中であるが、3 次元データの標準仕様について何らかの合意がなされているわけではない。また、シミュレーションを用いる材料開発においては、パラメータ設定のために基礎的な実験データが必要であるが、現状は個々の研究者が先行文献等を参考にして数値設定している。
- ・ シミュレーションは、計算モデルの中に考慮すべきものが初めからきちんと取り入れられているかどうかによって結果は大きく変わるし、現時点ではモデルが扱いきれない微妙なきわどい性質に関する課題も存在し、十分な信頼を得るまでには至っていない。したがって、実験をきちんと理解していないと誤った結果が導かれるおそれがあり、シミュレーションで得られるデータの質（信頼性）は保証されな

い。計算科学者の立場からすると、計算した結果に関係する実験結果をすぐに参照・比較することができ、それによって計算モデルの妥当性を検証できる仕組みになっている必要がある。また、第一原理計算のようなミクロなシミュレーションを行う際の計算準備段階に用いる擬ポテンシャルのように、元素や化合物に共通のデータとして整備・共有されていないため、多くの研究者が個々に重複して計算を行っている。それに起因した計算結果のバラツキも、当然存在する。

計算機を活用したデータ科学は、網羅的にデータを解析してデータに潜在する規則性を抽出し、それを利用して人間の能力を強力に補強（支援）するものである。物質・材料研究においても今後広く利用される必要がある。しかしながら、現況では、材料に関する多くのデータが整理・構造化されておらず、実験データ、計算データを統合的に活用する枠組みや材料開発に活かす枠組みが存在していない。

データを効果的・効率的に収集・集約し、数理・情報科学の成果を活用して知識発見や新たな価値を創造することの重要性が、国際的に認識されてきているなかで、このような日本の立ち遅れた状況は、将来の物質・材料研究開発における我が国の国際競争力を著しく減じる要因ともなりかねない。

今後は、増大し続ける膨大なデータの積極的な活用という観点から、さらには設計に向けた「知の蓄積・共有・循環」という観点からも、材料開発にデータ科学を適用したデータのハンドリング（インフォマティクス）が不可欠である。物質・材料研究とデータ科学の統合によって付加価値のある情報を抽出し、イノベーションを創出させる研究手法を導入するため、それに適合したインフラの構築と環境の整備が急がれる。

上記のような目的を達成するためには、必要となる研究開発能力、技術的知見、資源などを広く外部から調達して効率的なイノベーションを目指す必要があり、いわゆる「オープンイノベーション」の潮流とも合致するものである。

【コラム】物質と材料のギャップ ～材料のマルチスケール性と計算・計測手法の発展～

材料とは加工されて製品になるものであり、最終消費財を構成する原料として利用される物質群であると定義したとき、物質・材料分野の研究においては、依然として物質（科学）と材料（工学）の間が離れすぎているため、科学の研究成果が実用につながるスピードが遅い。この間を紡ぐ一つの手法がマルチスケールのモデリングであり、データである。

特に鉄鋼材料や高分子材料など構造材料は、空間的にマルチスケール性を有しており、マクロな材料特性の理解・制御が長年の挑戦的課題となっている。実質的に、すべての材料はナノスケールからサブミクロンスケールでの微細構造を有している。このような微細構造は材料製造プロセスによる影響を強く受けて、粒界、二次フェーズ、細孔あるいは転位や空孔のような欠陥を含み、材料に強い不均一性を与える。それらは同時に、材料機能としての強度、磁気特性、誘電特性、劣化などに強い影響を与えるため、微細構造の制御方法の解明なくして材料の設計指針を得ることはできないといえる。また、材料の組織形成過程、劣化、腐食などは空間的マルチスケール性だけでなく、時間発展のマルチスケール性も有している。さらに、材料の構造と特性は熱平衡条件での安定性だけでなく、温度勾配や成長速度、光や電子による励起など、様々な環境の非平衡条件によっても左右される。

したがって、所望の機能を出発点とする材料設計には、基本的な物質としての電子構造からプロセスや環境など外部条件によって決定あるいは変化する組織構造に至るまでの統合的理解が必要であり、究極的にはシステム全体（すべての時空スケール）の定性的かつ定量的理解と階層的連携による理解（システム指向）が求められる。

材料計算科学では、これまでにこうしたマルチスケール性のある材料の挙動を推定・予測するためにさまざまな手法を開発してきた。密度汎関数法に立脚した第一原理計算手法、（第一原理計算に基づいた、あるいは経験的な）原子間相互作用を用いた古典分子動力学法、いくつかの原子や分子をまとめて一つの有効粒子とする粗視化分子動力学法、組織発展や転位組織発展を可能としたフェーズフィールド法、磁性体の磁区分布などを扱うマイクロマグネティクス、結晶塑性理論に基づく有限要素法などの構造材料中の重要な現象が発生する各空間スケールでの計算手法の確立がなされつつある。一方、計算機能力の向上により、それに対応して大きな空間スケール（より多くの原子が関係する構造、物性、現象）および長時間スケール（より長い時間にわたる現象）を取り扱えるようになりつつある。また、材料を構成する個々の物質の基本的性質に関しては、点欠陥、単純な積層欠陥、規則性のある対応粒界、格子振動などを含めて、第一原理電子状態計算によってかなりの高い信頼度をもって記述することが可能になっている。

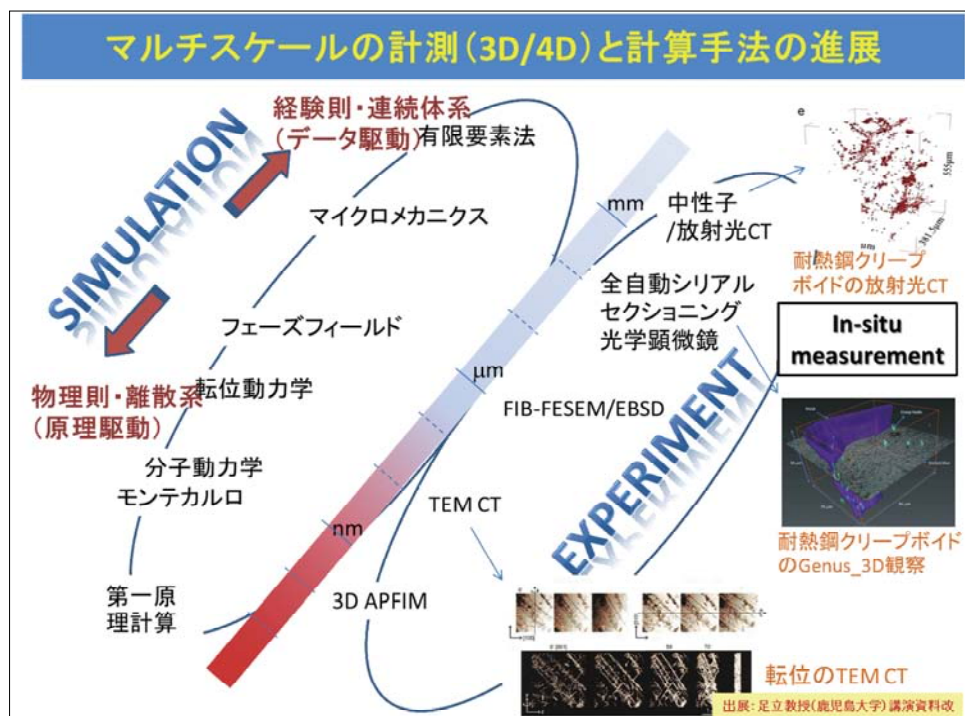
こうした計算科学の手法開発や計算機演算能力の向上は、材料の開発や実用化に確実に貢献してきているものの、その有用性をさらに拡大するためには、解決されるべき課題が多く残されている。以下に、そのいくつかを具体的に挙げる。

- ・ 上記の種々の手法のうち、第一原理計算に基づくもの以外のはほとんどは現象論的なものであり、それらの手法の基礎となる支配方程式が明確でない場合もある。従って、それらの手法で使われるパラメータは経験的に推定されるので、物理的内容が不明瞭である。
- ・ 既に述べたように、実際の材料は諸種の微細構造の集積体であるが、微細構造間には界面が存在する。後述するように、計測手段の急速な進歩により、界面の構造に関しても実験から有用な情報が得られるようになりつつあるものの、複雑な界面の組成分布、原子配置の詳細を得るまでには至っていない。また、シミュレーションにおいても、複雑な界面を扱うことは非常に困難である。そのため、界面に対するこうした不完全な知識のままに、例えばフェーズフィールド法では界面自由エネルギーの空間変化を扱わねばならない現状である。一方で、適切なパラメータを設定すると、実際の問題においてフェーズフィールド法は非常に有用な情報を提供するのも事実である。
- ・ 仮にフェーズフィールド法によって、材料の微細構造が妥当なレベルで予測されたとしても、それに起因する機械強度の予測はまた別の困難な課題である。それぞれの微細構造内部の物質の強度の情報のみならず、界面の形状と機械的性質が情報として必要である。
- ・ 磁石の重要な性質である保磁力は、原子スケールでの結晶磁気異方性だけでなく、

微細構造やそれらの間の界面などによって大きく変化する。結晶磁気異方性は第一原理電子状態計算で扱えるが、実際の磁石の保磁力を求める筋道はなお不明である。

これらのことから、今後の課題として2つの方向が考えられる。一つは、空間的・時間的スケールが異なる手法を用いて材料の物性や機能の予測を行うための計算科学的課題であり、異なるスケール間をつなぐ筋道をつけることが重要である。具体的には、第一原理計算で扱えるミクロな情報と、現象論であるフェーズフィールド法やマイクロマグネティックスのパラメータをどのようにして関係付けるかという問題である。こうした異なるスケール間の方法の連結方法が明確になると、現象論におけるパラメータの物理的内容が理解され、信頼度の向上につながるだろう。

もう一つ、大量のデータを活用して、各スケールでのシミュレーションに必要なパラメータの設定やモデルの改良を行うことが今後一層重要になる。有効な数理モデルを構築するには、大量のデータを活用してモデルのパラメータを推定する、あるいはモデルそのものの改良につなげることによって予測の信頼度を上げる必要がある。さらには、シミュレーションを経ずして、機械学習などの手法により大量データから直接に所望の機能をもつ物質や材料の予測を行うインフォマティクスの活用が期待される。



物質・材料研究における計測手法の発展も著しい。従来の計測手法ではある断面での観察結果に基づいて解析・評価が行われるため、その観察面でたとえ不連続に第二相組織が観察されたとしても、内部においても不連続であるかどうかについては保証できなかった。また、複相組織が入り組んでいる場合も、それを捉えられる観察方向が限られているため任意断面での観察ではそれを評価することは容易ではない。しかしながら、近年組織を3次元で、かつ動的に測定する計測手法が飛躍的に発達し、組織形成過程（内部を含む）を多次元で理解しようとする試みが精力的に進められており、その対象サイズもナノスケールからセンチメートルオーダーまで広く、様々な手法が開発されている。3次元可視化法は、特性モデルについて定量的な構造情報を引き出すことを可能とし、スケールが大きいほうから、中性子線・X線トモグラフィ法、シリアルセクションング SEM/EBSD 像の3次元再構築法、TEM トモグラフィ法、3次元アトムプローブ法などがある。このような材料組織の3次元情報はデジタルのため、数値的に表現できることから計算手法と相性がよい。

一方、現在の手法の多くは、扱い得る空間的・時間的スケールが一定の範囲に限られており、より広い範囲での空間的・時間的規模での材料挙動や特性を推定する能力に欠けていて、それらが解決すべき課題として挙げられる。たとえば、ピコ秒オーダーでの材料の原子振動の測定は可能であるが、この測定データを活用して材料の経年劣化のような年のオーダーに及ぶ現象を推定することはできない。また、ある製品の材料の結晶構造と物理特性は様々な条件によって製品加工中に変化することがあるが、それは製品の性能に大きな影響を与えることになる。こういった形態上や特性変化をモデル化する能力が得られれば、設計をより迅速により精確に行うことができるはずで、そのためにもモデリングの高度化を支援する計測データの蓄積・共有が必要である。

上述のように理論計算手法と先端計測手法の急速な進展は、これまで実現困難とされていた材料（機能）設計が確実に射程に入りつつあることを示している。物質・材料研究をある程度ベクトルがそろった大きな流れに発展させ、加速するためには、計測研究者と計算研究者が相互に十分な意思疎通を図る必要がある。その上で政府が主体となった計測と計算の融合を図る研究開発の枠組みを構築することが不可欠である。

【参考】計算材料科学と数学の協働によるスマート材料デザイン手法の探索—階層構造を解析する—

東北大学が WPI で実施する原子分子材料科学高等研究機構では、文部科学省委託事業「数学・数理科学と諸科学・産業との協働によるイノベーション創出のための研究促進」の支援を受けて、CMSI とともに表題ワークショップを 2013 年 3 月 13 ～ 15 日に開催した。

- ・ 高性能コンピューティング時代に向かい、機能発現の機構解明と材料設計により、材料科学の新たなフェーズが始まっている
- ・ サイズや時間の階層を超えてミクロとマクロをつなぐにはまだ種々の困難が残されており、その克服には従来にはない発想が必要

という問題意識の下、「階層構造の解析」をキーワードに課題の所在や最新手法をサーベイし、数学・数理科学と計算材料科学とがどのように協働できるか討論を実施した。

以下にワークショップのサマリーを引用する。

◆空間、時間、精度という視点からの研究の現状と課題

- ・ 「空間軸」のチャレンジは順調に進んでおり、第一原理計算を実空間で実装することで数万原子系でも効率的に動作する手法に目処がつきつつある。そこでは、巨大次元の固有値問題から目的の情報をうまく抽出するための数理的工夫などが取り込まれている。
- ・ 「時間軸」の取り扱いは「空間軸」ほど容易ではない。**meta-dynamics** のように位相空間内の極小点から隣の極小点への脱出を加速する手法が考案・適用されてきてはいるが、現状は **ad-hoc** 感が強くまだ発展の余地が大きい。
- ・ 「精度」に関しては、密度汎関数法の適用が困難な系（強相関電子系や励起状態の記述が必要な系など）の計算量はシステムサイズに対してすぐに発散し、「京」コンピュータを用いても扱える系は限られてしまう。

◆新たに明らかになった課題、解決すべきこと

- ・ 計算材料科学は最先端の技術により第一原理計算に基づいて数万原子系の構造と電子状態を議論できるところまで来ており、それ自体、新しい数理科学や計算機科学を取り入れた成果である。
- ・ しかし大規模系に対応する広大な位相空間をうまくサンプリングすることが追いついておらず、材料科学としては不可欠な構造探索・熱力学量の評価・相図の記述が困難なまま残っている。また、この先、数万（あるいは数億）原子系の位相空間サンプリングが実現できたとしても、多次元系の膨大なデータから材料設計の指針となる有用な情報を抽出するのは難しいだろう。
- ・ 一方、**Informatics** の手法を用いて多くの第一原理計算の結果に対するデータマイニングやスクリーニングを行い、材料物質の物性予測やデザインにつなげようという取組みも始まっている。現在は **primitive** な説明変数とバルク物性とを結びつける段階だが、実材料を対象とする以上、粒界やヘテロ界面などの高次構造の取り扱いには不可避である。
- ・ 数学においては、従来微分可能性を仮定して構築されてきた数学の様々な手法や概念を離散化する試みが進展している。ビッグデータと言うが、実際のデータは疎な分布であり、データの構造を見出し、それらが乗っている空間を如何に見つけるかが新しい挑戦であり、数学への期待は高い。

2-2 社会・経済的效果

(1) 新物質・材料の開発から実用化までの時間短縮による産業の活性化

新材料の創出や材料特性の改善が、既存産業の競争力を支えるとともに新たな産業出現の契機や必要条件になることは、すでに世界の共通認識となっており、わが国をはじめ各国政府は、物質・材料の研究開発に多額の投資を行ない、その成果を自国産業の発展に結び付けようとしている。このような状況の中で、物質・材料の研究開発をわが国における産業競争力向上や新産業創出に結実させるためには、他国を凌駕するスピードで新材料の創出、材料特性の改善を行ない、開発された材料が市場に投入されるまでの時間を短縮する必要がある。本提言は、データ科学を物質・材料科学に応用した科学技術としてマテリアルズ・インフォマティクスを確立し、材料の物理的・化学的性質に関する多様で膨大なデータを有効活用して、新たな法則の発見や新物質の発見から材料開発に至るまでの各過程を加速することを目指すものである。

従来の物質・材料研究においては、目的の機能・特性を発現する物質を探索・合成して実用材料の開発に至るまでには多くの試行錯誤が必要であり、これが物質・材料開発に長い時間を要する要因となっていた。たとえば、今日の携帯用電子機器で広く用いられているリチウム・イオン二次電池は、現代の情報技術の状況を一変させている。1970年代中頃に大学の研究室からコンセプト提案がなされ製品として実装されるまでに20年もの歳月を要した。その背景は、実用化には多くの新材料開発が必要であるばかりでなく、二次電池として必要な性能や安全性を確保するために材料機能に改良を加えなければならなかった点にあった。

このようにイノベーションの創出は新物質・材料の開発・改良に要する時間に強く依存しており、半ば常識化している。従って、逆にこの期間を短縮することは、イノベーション創出の加速に直接つながり、産業の活性化に大きく寄与することになる。

(2) インフラ整備・共有による物質・材料研究開発の有効活用・効率化

研究開発の成果を効果的、効率的に社会に還元していくためには、研究成果や研究用材料等の知的資産を体系化し、幅広く研究者の利用に供することができるよう、知的基盤を整備していく必要がある。マテリアルズ・インフォマティクスの普及に必要なデータベース等のインフラが整備されることにより、知の集積と伝播・活用が促進され、科学的発見や社会への実装が加速することになる。また、企業、大学等の研究者は、信頼性あるデータベースやデータ解析ツールが整備されていることにより、高効率な研究活動、無駄な研究コストの削減が可能となる。すなわち、予測に基づく物質・材料の探索・開発を確立することで、直接研究費を削減可能とし、データの公開により、日本全体として、重複研究、重複投資をなくすと同時に、データの利用によって研究開発のスピードを高めることも可能となる。

本提言の実践は、オープンイノベーションを誘起する環境構築に向けて、研究インフラの革新という形で日本社会に貢献する。具体的には、広範囲の科学者・技術者は、データベース、ツールおよびモデルのライブラリーを含むサイバーインフラに自由にアクセスできるようになることで、新材料の迅速な設計と最適化に向かうことが可能となる。最終的

には、2－3章で述べるように科学的発見が促進され、それが革新的な工学的製品として迅速に社会に実装されるものと期待される。

一例として、先端大型研究施設の実験データの公共化が挙げられる。たとえば J-PARC では、一週間の実験のために一組の研究者が必要とするコストは1千万円を超える。得られる実験データには、物質中のスピンやフォノンのダイナミクスに関する膨大で有益な情報が含まれているが、スピンを専門とする超伝導研究者であればスピンに関する情報だけを解析して論文にすることが一般的である。しかしながら、当該データを一定期間後に一般公開することにより、フォノンに興味がある第三者が、同じデータ中に含まれるフォノンに関する情報を解析して、スピンではなくフォノンに起因した別の超伝導メカニズムについて探求するという可能性も考えられる。これは、一種の「オープンデータ」の考え方であり、巨額な国際公共財としての大型実験施設で取得された実験データの幅広い効率的な利活用につながるものである。

2－3 科学技術上の効果

(1) 学術の発展

バイオインフォマティクスが誕生して20年が過ぎたが、それによってライフサイエンス分野の研究手法が変わり、ライフイノベーションにおける重要な局面で大きな役割を果たしている。たとえば、山中伸弥（京都大学）らは、ES細胞の多能性を維持する因子探索のためにゲノムデータベースを利用した *in silico*（コンピュータ上）解析を行い、ブレークスルーを起こしたことは特筆すべきである。物質・材料分野においてもデータに駆動された発見（**a data-driven discovery approach**）というアプローチが期待されている（P14.【コラム】バイオインフォマティクスとマテリアルズ・インフォマティクス参照）。

シミュレーションや理論というのはルール（原理）からデータをつくり出すという演繹的な考え方であるのに対して、経験（実験）や大量データ（データマイニング、機械学習）からルール（知識）を発見するプロセスは帰納的なものである。これら2つの科学的アプローチはまったく異なる考え方に基づいており、帰納と演繹を考え方としてつなぐ仕組みが必要となる。本提言は、この異なる分野の研究者が協力・連携して、演繹および帰納的手法をバランスよく活用することで研究として最大限の効力が発揮されるようにすることを提案している。

帰納と演繹の考え方を繋ぐことにより、学術科学研究に対する新たな方法論の導出が期待される。具体的には、広範囲にわたる膨大な物質系（有機/無機/金属）を俯瞰し、これまで集積された知見を従来の有機、無機、金属といった枠組みにとらわれずに、「機能＝構造×元素」の概念に基づいて、機能発現という視点からの見通しのよい体系にすること（物質科学の因数分解）が可能となる。このような因数分解により、ある機能を実現するための未開拓物質群の構造や構成元素の特定が加速されるだけでなく、データ解析により物質の構造・物性の相関が明らかになれば、新たな物理的・化学的法則の発見が促進される可能性がある。さらに、元素の組み合わせや物質の複合化による機能の発現にある種の周期律が見出されるという期待感がある。

(2) 技術の発展

本提言の研究開発は材料設計の効率化に向けた試みでもある。この推進により、既存データや材料設計の過程で得られたデータが、最終的な材料設計にどう結びついたかを、明示的に記述できるようになるため、材料設計の効率化を系統立てて進めることが可能になる。

今後の材料開発プロセスは計算科学やデータ科学を活用してコンピュータ上に材料の構造、物性を蓄積した上で、所望の機能をもつ（指定した複数の物性を同時に満たす）物質構造の候補について設計し、その中から最低限必要な実験により目的とする材料を選定する、という方向に遷移していくと考えられる。さらに将来的には合成プロセスの設計も大部分がコンピュータ上で実現されることが求められる。

本提言によるインフォマティクスはその遷移を可能にする手段である。長期間を要し、高コストな従来の経験的研究の多くをインフォマティクスで効率化することによって、新物質・材料が社会で有効利用されるまでに要する時間を著しく短縮するとともに、新規に開発される物質・材料の種類も大幅に増加させ得る可能性をもつ。さらに、マルチスケールにおける材料設計原理の解明は、機能要件から製造プロセスまでの一貫したトータルシミュレーション（機能要件→材料設計→分子・素材設計→製造プロセス設計）の開発につながる。

(3) 人材育成

本提言は、全体として物質・材料に関連する知識の構造化・体系化を指向するものである。物質・材料の研究開発に従事するためには、蓄積された膨大なデータを合理的に整理し、そこから得られる知識を習得する必要がある。そのような新しい知識の構造化・体系化によって革新的に進む材料設計の技法を習得することは、人材の質を向上させ、以って人材の有効活用を格段に促進することになる。具体的には、広範囲にわたる膨大な物質系（有機/無機/金属）を俯瞰し、これまで集積された知見を従来の有機、無機、金属といった枠組みにとらわれずに、「機能＝構造×元素」という視点の材料設計研究者の養成につながる。

理論・シミュレーションという演繹的な方法と、実験・インフォマティクスという帰納的な方法の両者を直接つなぐことは一般的には容易なことではなく、現状では科学者自身にも戸惑いがみられる。とくに、日本の教育が演繹に重点を置いているためそういった傾向が強いとの指摘もあり、加えてシステム思考に弱点をもつ日本人の性向を考慮すると、今後、インフォマティクスのような帰納的方法を正當に理解できる人材を中長期的視点で養成する基本教育システムの構築が必須である。例えば、初等中等教育において「情報」あるいは「情報処理」のような科目が必要であるし、高等教育においては物質・材料関係学科の学生のカリキュラムにおける計算機科学、計算科学、データ科学の内容の体系的充実が望まれる。

本提言の実践により、結果として演繹・帰納両方の考え方を身に着けた研究者が育成されることが期待される。このような人材に対して、教育プログラムと産・学・官・社会のプラットフォームの連携を通して人材育成のスコープを共有し、協力して多様なキャリアパスを供給することが求められる。

【コラム】 バイオインフォマティクスとマテリアルズ・インフォマティクス

バイオインフォマティクスの初期の主要な研究対象は核酸塩基配列やタンパク質アミノ酸配列であった。その後、マイクロアレイや二次元電気泳動法などの網羅的な解析手法の普及により大量のデータが得られるようになったことから、ライフサイエンスにおけるさまざまなデータを研究対象とする科学に拡大・発展しつつある。生命現象を生体関連物質間相互作用のネットワークとしてとらえようとするシステムバイオロジーは、バイオインフォマティクスの発展形とみることでもある。

ライフサイエンスにおいてインフォマティクスが発展した背景には、生命現象の複雑さがあり、実験から得られるデータから生命現象の理解に資する情報を引き出すためには、計算機を活用した数理・情報科学の援用が不可避であったためである。また、ライフサイエンスには、物質・材料科学における量子力学のように、現象を原理的に演繹できる理論はなく、実験や観察から得られるファクトから規則性を見出す帰納的なアプローチこそが、この分野の発展を支えてきたという歴史的経緯もある。

一方、物質・材料科学では、化学においてケモインフォマティクスが進んだほかは、インフォマティクスを積極的に取り込もうとする動きは乏しく、21世紀に入ってからようやく米国を中心に目立った動きが見られるようになってきた。マテリアルズ・インフォマティクスのスタートがライフサイエンスや化学に遅れた背景の一つには、物質・材料科学では扱う現象の空間的スケールに応じて物性などを演繹できる一定の有効性をもった理論やモデルの存在がある。このような理論やモデルは、計算機の進歩、理論やモデル自体の精緻化、計算手法の改良などにより演繹できる現象の範囲を広げてきたこともあり、その可能性への期待からインフォマティクスのような帰納的アプローチへの関心が薄かった。

しかし、材料開発の視点に立ったとき、多くの材料機能は物質の基本的情報から演繹することができないことも事実であり、無機化合物を中心とした物質・材料科学にも、材料機能の定量的データと物質の基本的情報からなるデータセットにインフォマティクスを適用し、材料機能を予測し得るルールを抽出しようとする試みがみられるようになってきている。このようなマテリアルズ・インフォマティクスでは、実験的に得られたデータに加え、第一原理計算から定量予測可能な物質の基本的情報もデータ解析の対象とすることが可能である。この点は、主に実験データを解析の対象とするバイオインフォマティクスとは異なる点であり、マテリアルズ・インフォマティクスの特徴のひとつといえよう。

マテリアルズ・インフォマティクスにおいても、解析対象に利用できる実験データが豊富にあることが重要であることに変わりはなく、インフォマティクス適用を前提としたデータの取得・蓄積にも配慮が必要である。無機化合物では、化学組成が同一であっても、結晶構造が異なれば物性は異なり、ひとつの物質が多面的な性質をもつ。たとえば、ゲート絶縁膜材料としての ZrO_2 は、できるだけ酸素空孔を少なくして絶縁特性を向上させることが求められるため、絶縁特性を向上させるための添加材料や絶縁特性などが中心的データとして蓄積されている。一方、イオン伝導体やセンサーとしての ZrO_2 は、逆に空孔濃度を上げて伝導性を確保することが求められるため、電気抵抗を下げるための作製条件と伝導性を高める不純物材料が中心的データとして蓄積されている。同じ物質でありながら、目的とする用途が異なると、プロセスおよび構造が異なり、実験的に全く異なるデータを蓄積している。すなわち、ひとつひとつの物質・材料をそれぞれ異なったプロセスで合成し、特定の特性だけを評価するという従来の実験データの蓄積方法では、マテリアルズ・インフォマティクスには不十分である。酸素分圧や電気抵抗といった材料特性を自動計測し、これらをデータセットとして蓄積することができれば、一部は高誘電体ゲート絶縁膜情報として利用でき、別の一部はイオン伝導体情報として利用することができる。このようなデータセットの収集と蓄積がマテリアルズ・インフォマティクス構築の第一歩となる。

3章 具体的な研究開発課題

3-1. データ科学の手法を活用した研究開発の推進

物質・材料の研究開発にデータ科学の手法を活用する場合、原因究明か予測かのいずれを主目的とするかによって、大きくは2つの方向性がある。両者に共通するのは、注目する機能を表す目的変数とその機能が依存すると思われる説明変数の相関を適切な関数（モデル）として表すことである。

原因究明とは、膨大な化合物の構造や物性のデータから目的変数が強く依存する説明変数（記述子）を探ること、ひいては構造物性相関（物理・化学法則）を探ることを目的とするものであり、いくつかのモデルを用いてこれを試み、目的変数が最も少ない記述子で表わされるように、モデルと記述子を決めることにより、より意味のある法則が得られることになる。

予測とは、既知のデータから学習によりモデルを導き出し、それによって所望の機能を有する物質・材料を推定することを目的とする。たとえば比較的少数のデータから多くの系の情報を予測するために、限定的な組み合わせの第一原理計算や実験データから、機械学習などに基づいて目的変数と説明変数のモデルを抽出し、その規則性を利用して他の膨大な組み合わせの結果を推測することができる。

物質・材料研究においては、説明変数として、温度、圧力、化合物の組成などの外部条件や原子番号やイオン化エネルギー、原子量のような元素についての根源的な値のほかに、第一原理計算によって獲得した物質の一次情報を利用することが考えられる。

これらの研究を行う際、重要な手順として、(i) 探索空間の設計、(ii) 目的に応じた構造、特性などのデータセットの作成（計算や計測によるデータベースの活用、必要に応じて系統的に実験や計算を行う）、(iii) 目的に応じたデータ科学的手法を用いて、構造と物性などの相関を探る原因究明または予測モデルの構築、そしてそれらに依るスクリーニング、となる。

データ科学の多様な解析手法を物質・材料の研究開発のさまざまな局面に適用するため、以下に例示する新たな研究開発の方法論、これを用いた研究開発の取組みが必要となる。

1) 物質・材料研究における定量的構造・物性相関解明法の確立とそれを用いた物質探索・設計

計算や計測から得られた大量で多種多様なデータ群から、物質のある特性（目的変数）とその特性に関与すると思われる複数の説明変数としての構造に関する因子や物理的パラメータとの関係を統計学的に比較検討し、定量的な相関式を導き出すことを目的とした研究開発を行う。いくつかのモデルを用いてこれを試み、目的変数が最も少ない記述子で表わされるように、モデルと記述子を決めることで、物理的に意味のある規則性が得られる。また、機能・特性を表す目的変数そのものが明確でない場合は、最適な目的変数の探索手法の研究開発も必要となる。このような研究開発は現象の根底にある問題の発見にもつながるものであり、新たな学問領域を開くものと期待される。

2) 既知のデータセット（基本的な物性値など）に基づく、未知物質の複雑特性の予測手法の高度化とそれを用いた物質・材料探索、設計（design of experiment）

膨大な空間の中から最適な構造（化合物）を探索する手法の研究開発を行う。新しい物質の物性を予測する際に、ベイズ推定を用いて、既存の物質群についての大量データ（計算の結果など）からその物性をできるだけ正しく予測できる関数を割り出し、その関数の最適値をできるだけ少ない観測で、より高精度に発見するアプローチや多くの類似の系を探索するアプローチが考えられる。このような予測を目的としたデータ解析では、予測の結果に基づいて行った実験や計算から得られたデータを加えて、データ解析を繰り返すことが肝要であり、特に第一原理計算と組み合わせて行うことが有効である。

この他にも、組成を与えて、ランダムに結晶構造をいくつか発生させて第一原理計算を行い、構造最適化計算と局所構造緩和をすべて計算し、得られた構造を基に次の子供を発生させ進化論的アルゴリズムを使って、200 世代程度まで生存競争させながら標的物質に近づいていくなどの方法が考えられる。

このような方法でピンポイントに予測された組成および近辺の組成について、コンビナトリアル実験を使って合成・評価することによって、優れた特性をもつ物質の探索をすることができる。その結果をフィードバックし、また計算する。逆に、計算する上でモデルを作る際にも、実験のデータが多いほど精度が高くなる。そこで、コンビナトリアル実験から始め、得られた情報に基づいてモデルを作り、計算機により新しく予測し、またコンビナトリアル実験を行うなどのシステムが考えられる。

3) データおよび物理法則に基づいたモデルの構築・検証、およびスケール間の連結による材料設計指針の確立（マルチスケール・マテリアルデザイン）

数理モデルの予測精度を高めるために、データ同化や圧縮センシングを用い、現象の実測データとの統合により、数理モデルの予測精度を向上させるための研究開発を行う。

特に、鉄鋼材料や高分子材料など構造材料のマクロな材料特性は、非平衡、非線形性が強く、短時間・微小空間で生じたことが、長時間・材料全体の挙動に大きな影響を与えるマルチスケール性を有している。このような複数の現象が相互に影響しあうマルチスケール性に由来する現象には、有効な数理モデルがない場合も多い。時空間スケールを横断する実験（計測）と計算の連携、データの蓄積・共有により、巨視的挙動に及ぼす微視的挙動の影響を最適に評価し、現象の制御に対してより有効な物理モデルを構築する研究開発（マルチスケール・モデリングの研究）を行う。これは、広範な条件で適用できる材料のプロセス-構造-特性相関を確立すること、と言い換えることができる。

具体的には以下のような取組みが想定される。

- ・ 材料構造（欠陥含む）と材料特性の間の関係を定量化する手法
- ・ 複数のモデルを互いに連結する手法の開発
- ・ データ同化の物質・材料研究への応用
- ・ 物理モデリングと圧縮センシングの融合による機能発現のための法則の抽出
データ同化や圧縮センシングはすでに気象や地震などの分野で活用されているが、物質・材料分野の研究開発への適用を検討する必要がある。
- ・ モデルの不確実性が全体にどのように影響するかを理解する手法

マルチスケールにおけるモデルの統合を成功させるためには、各スケールで情報の抽出

が必要である。また、データのモデルと実験への取り込みを簡素化し、研究者がお互いのデータを容易にかつ追加的にモデルに組み入れられるようにすることが必要である。このためにも、材料研究コミュニティはデータを公開し、共有することが求められる。

4) 大量または／および複雑データの可視化・解析手法の高度化（多変量解析などの物質・材料研究への応用）

非常に大きなデータセットや解析が困難な画像データを人間が理解できる形式で表すための研究開発である。コンピュータは、10億もの自由度をもつシミュレーションの結果を表示することができるかもしれないが、研究者がデータを理解し利用するには、可視化技術による支援が必要である。このような可視化技術の開発には2つの方向性がある。一つは大量のデータを俯瞰的な視野で構造化し、新しい軸の発見を促す手法、もう一つは必要なスペクトル情報を恣意的に特定の場所から抽出するだけでなく、データ全体から含まれている有用な情報を余すことなく抽出し、その結果の中の重要な特徴をハイライトする手法の開発である。

このために、計測から得られるデータに機械学習、多変量解析などの手法を適用し、新たな可視化の手法を開発する。

データ取得の方法としては、既存のデータベースの利用の他、シミュレーションやコンビナトリアル合成・評価によって取得する方法があり、このようにして得たデータ群と、パブリックなデータベースを統合して解析できることが理想である。

3-2. データ基盤の整備

データ駆動型の研究開発を促進するためには、研究コミュニティがアクセスできる共通のオープンアクセスのデータベースが重要であり、アクセス可能なデータの数を増やしていく必要がある。このような基盤の整備では、計算結果などのデータの蓄積に留まるのではなく、「蓄積されたデータを共有し、それをもとにどのような情報が得られるか」を十分勘案することが肝要である。蓄積されたデータは二次利用可能であることが重要であり、データの標準化、利用ガイドラインの策定等が必要である。

2-1章で述べたとおり、いずれの領域の研究者にとってもデータが非常に重要な役割を果たし、今後の物質・材料設計加速のキー・ドライバーとなることは確実である。

データベース作成には、いくつかの方法がある。(i) 文献からの情報をアーカイブしていく方法、(ii) 電子構造や物性などを第一原理計算等により大量にアーカイブしていく方法、(iii) コンビナトリアル的な実験的手法により大量合成・評価したデータをアーカイブする方法などである。これらには、それぞれ下記のような課題を考慮した取組みが必要となる。

- ① については、人手により公開文献から必要な情報を抽出して共通仕様に落とし込んでデータベース化するため、データ収集のコストが高くなり、数をこなすことが難しい。また人材の確保も課題となっている。
- ② については、計算機シミュレーションによるデータベースの場合は、専用計算機を導入することにより、数百万のオーダーで蓄積することが可能となる。しかし、計算の

精度には限界があるため、データの信頼性の問題があり、実験データと突合させる必要がある。

- ③ については、物質・材料分野においてコンビナトリアル手法を用いた研究に取り組んでいる大学等公的研究機関が少なく、高効率にデータをアーカイブする方法が確立されていない。ただし、合成化学・製薬分野では実績があり、当該技術を取り込んでいくことが求められる。

これらを踏まえたうえで、以下のような取り組みが必要である。

データの取得・蓄積手法

1) 第一原理計算等を介したハイスループットかつ均質な計算データの生成・蓄積

たとえば、構造既知の物質の網羅的な計算、構造未知でも周期表の元素組み合わせによる網羅的な計算、様々な物性値の網羅的な計算による結果の蓄積が考えられる。

この場合において、計算条件（系）、最適結晶構造、電子状態（全エネルギー、バンド構造、状態密度）、これらを利用して得られる二次情報である物性（電気伝導度、熱電能、熱伝導度、磁性、弾性定数、光反射スペクトル等）とともに蓄積することが求められる。

2) 実験（計測）による構造、物性に関する多面的なデータの蓄積

たとえば、ハイスループットな実験手法（たとえばコンビナトリアル手法）によるデータの生成・蓄積手法の構築や計測機器等における多様（不均質）な構造のデータを取り扱う技術などが挙げられる。

データの統合手法（解析）

3) 計算および実験（計測）によって得られる物質・材料の電子・結晶構造、物性などデータの統合

米国では“Materials Project”や“aflowlib.org”のデータベースが既に公開されていることから（付録2参照）、日本では独自に工夫したデータベースを作成する必要がある。データの信頼性を担保するため、使用ソフト+擬ポテンシャル+計算物性の3点セットを開示することによる計算の再現性が求められる。

4) RDF（Resource Description Framework）など Linked Open Data に基づくデータベースの整備やオントロジーの活用

データ解析に比して、データを取得するほうが問題になりがちである。データ解析はいろいろな説明変数（記述子）を試行し、結果を見て再度試行するという試行錯誤の繰り返しとなる。生産性を上げるためには自身が実験して得たデータと、パブリックに公開されているデータを統合して解析することが望まれる。多数の説明変数（記述子）を、WEBを通して瞬時に用意できることが理想であり、そのためにはデータ様式の標準化が望まれる。

マテリアルズ・インフォマティクスには、材料を説明し、分類する、広く同意された分類学（オントロジー）の確立が重要となる。これによってデータの応用・統合の取り組みが可能となり、専門家が重要な情報を見つけ出し、関連づけることが可能になる。

以上により、研究者がデータのインデキシング、検索および比較できるデータ交換システムが稼働できるようになる。

その他、既存の物質・材料データベースと新しいデータベースのインタフェースやデータベース管理機関によるデータの生成、収集と同時にユーザによるデータ登録機能なども考慮されるべきである。

【参考】物質・材料研究における代表的なデータベース

● ICSD (FIZ Karlsruhe-NIST)

無機結晶構造データベース。鉱物・セラミックス・金属間化合物など、元素と無機化合物の結晶構造データを収録。約 15 万エントリ。結晶構造表示や粉末パターンシミュレーション等のプログラムも含む。有料。

ICSD に収録されている 3 元系までの無機化合物データは 2012 年末時点で、76000 件であり、同一の化合物で複数のデータが収録されている重複を除けば、実験的に結晶構造が既知の無機化合物の数は 5 万件を下回る。このうち生成エネルギーなどの熱力学データが既知である化合物は数千件であり、物性値が既知である化合物の数はさらに少ない。構造既知な化合物は単純 2 元系であり、それ以外は僅か。さらに物性既知な化合物は極めて少ない。

● Landolt-Boernstein (Springer)

ランドルト・ベルンシュタインは、物理学・化学・工学分野において、厳選された体系的かつ包括的な物質情報を提供する世界最大のデータベース。自然科学・工学全分野の研究雑誌の原著論文を元とし、その中から厳しい評価を経て選択されたファクト・データ集。有料。契約金額が高く、日本の大学における契約数は多くない。

● CALPHAD 法を用いた平衡状態図のデータベース

CALPHAD 法とは、熱力学モデルを基に、様々な実験データや理論計算データを解析し、種々の状態変数の関数として各相のギブズエネルギーを決定し、コンピュータにより状態図を計算する手法。熱力学モデルの仮定が満足される範囲内であれば、多元系や高温・低温、高圧・低圧領域への外挿精度も高く、さらに、ここ数年では第一原理計算を援用することで、これまでは実験データを得ることができなかった準安定相を含めた熱力学解析が行われるようになり、得られるギブズエネルギー関数の精度が格段に向上している。その有用性から、現在ではこれに基づく数多くの熱力学データベースが市販・公開され、研究や材料開発に広く用いられている。

● MatNavi (物質・材料研究機構)

高分子、無機材料、金属材料、超伝導材料など、11 のデータベース、4 つのアプリケーションシステム、6 種の構造材料データシートで構成される。公表されている膨大な学術文献から有用な論文を抽出し、数値データを採取、データベース化しているため信頼性の高いデータベースと言える。無機系は 2003 年以降 10 年間更新されていない。上記のデータベースと異なり無料である。

方法論、概念の確立。

- ステップ 1' 新しい研究手法、方法論のための基盤作り
新しい科学技術を社会実装する際の基盤となる装置、ソフトウェア、データベース等の試作。
- ステップ 2 新しい手法、方法論の様々な現象・ケースへの適用
ステップ 1 で創出された方法論、研究手法、実験手法などの様々な現象や事例への適用可能性の拡大、イノベーションに向けた多様なシーズの創出が期待される。同時に、新しい方法論、研究手法などに触発された研究者の参入による研究者・技術者コミュニティの拡大が期待される。また、研究手法、リサーチ・ツールの普及や将来の標準化に向けた活動への取組等を開始する。企業からの研究者・技術者も参画・協力し、研究ノウハウ等を共有する。
- ステップ 3 シーズの展開、技術の実証、産業化への橋渡し
これまでのステップで創出された多様なシーズと産学官の研究者の相互刺激が更なる研究開発や共同研究等を誘発し、シーズの発展や産学官連携による技術の創出と実証、産業化への橋渡しが期待される。
- ステップ 4 技術の効果の検証、上位ステップから、下位ステップへのフィードバック、新たな基礎的研究課題

これまで、研究者に新興・融合科学技術への取組みを躊躇させてきた要因の一つは、融合的な新しい研究をしても、成果を発表し、評価される場が少ないという問題であった。加えて大学や研究機関には、論文数など形式的な定量指標よりも、論文の内容やインパクトに重きを置いた評価が求められている現状がある。また、新興・融合科学技術の研究者のための新たな学会を立ち上げたとしても、学会の縦割り構造を細分化するだけになることが懸念される。このような現状を打破するためには、学会等は「異」や新しいものを受容する文化を持ち、新興・融合科学技術に取り組む研究者を積極的に登用し、自ら変わり、学術を発展させようという意識をもたなくてはならない。同時に、国は、そのようなインセンティブを与える仕組みやファンディング制度を構築するべきである。

1. データ科学の手法を活用した研究開発プログラムの創設

ステップ 1 として、データ駆動型の FS（可能性研究）プログラムを実施し、手法の確立・高度化を図るとともに、成果事例を創出することを目指す。

具体例としては、さきがけ程度の研究費（年間 1000 ～ 2000 万円の直接経費）で、原則、物質・材料科学者とデータ（数理・情報）科学者 2 名以上の異分野共同提案とする。異分野の共同提案となることから、公募期間を通常より長めにとり、学会やワークショップなどを通じた異分野交流のための期間をとるなどの考慮が求められる。

2 ないしは 3 年間募集し、各年 10 件程度採択する。研究期間は 5 年を原則とするが、3 年目でステージゲートを設けることもありうる。プログラム中には、異分野の研究者が互いに研究をベンチマークし、新たなアイデアを触発する仕組みの導入が期待される。

ステップ 1 と同時並行して、ステップ 1' として、データベースの整備を行う必要があるが、これについては、2. 基盤整備プログラムの項で記載する。

2. データ基盤整備プログラムの実施

設計型の物質・材料研究を促進するための中核となる物質・材料データ研究拠点の設置が求められる。データやソフトウェアの蓄積や維持を個別の機関や時限プロジェクトに委ねる方法は過去に成功例が無く、うまくいかないことが経験的に証明されている。これは国が責任をもって継続的に推進すべき案件である。

具体的な整備の対象は、物質・材料の化学組成、組織構造、結晶構造、電子構造、磁気構造や熱力学的情報、光学的情報など様々な構造・機能が想定される。整理の方法としても酸化物や窒化物、超伝導や磁性材料、ワイドギャップやトポロジカル絶縁体など様々な切り口が考えられる。

本研究拠点が、長期的視点に立ってデータマネジメントの方針、収集するデータの範囲を決め、データやツールの管理を行う。サービスセンターとしての機能も有し、複数の研究機関や大学が、RDF / XML などの標準的なデータフォーマットを使い、情報を交換して、次の材料開発に生かすというマテリアルズ・インフォマティクスのシステム構築を進める。

さらに、データベース構築専用のスーパーコンピュータ、パラメータセットのための実験データを補完する計測設備も有し、ユーザーのニーズを踏まえた新規のデータ整備も行なうことが求められる。その他具体的には下記のような取組みが求められる。

- ・ データ様式の標準化
- ・ 蓄積・維持すべきデータの範囲の定義
- ・ 必要なソフト（ツール）の管理と利用体制の整備
- ・ 第三者のデータを受け入れる仕組みの構築
- ・ 新しく出現するデータ様式への拡張
- ・ ユーザビリティの保証

これは、言い換えれば、JST に設置されたバイオサイエンスデータベースセンター（NBDC）が実施している取組みとほぼ等価である。物質・材料分野においても同様の取組みを可及的速やかに行うことが必要であり、拠点の運営に当たっては、先行する NBDC の知見を有効に活用することが必要不可欠かつ効率的である。

具体的には、物質・材料研究機構の運営する MatNavi を中核として、JST 関係部門が協力して知的基盤の整備に当たるとするのが最も現実的である。将来的には、理研の放射光センターや J-PARC などの大型計測施設の物質データと連携することが必要となる。

また、このセンターにおいて、1. データ活用研究開発プログラムの創設を実施することがガバナンスやデータ・ツールの整備を考えたとき最も効果的である。

さらに政策的には少なくとも 4 つの検討すべき課題がある。

- 1) 物質・材料研究機構（NIMS）のようなデータベースを作成することを任務としている機関のデータベースの発展・改良。
- 2) 学術振興会や鉄鋼協会が取組む「鉄鋼ゲノムの解明」のようにデータのアーカイブを進めたいと考えているコミュニティの支援や研究室が独自に作成したデータベースの運営・公開支援。

- 3) 放射光や中性子など物質・材料科学にかかわる大型研究施設のデータ（非常にコストがかかっている）の公共化と活用。
- 4) 「元素戦略」や「計算物質科学イニシアティブ（CMSI）」、「ナノテクノロジープラットフォーム」などの国家プログラムにおけるデータの蓄積・共有を図り、プロジェクトの終了後はデータを公的研究機関に引き継ぐなどの仕組みの構築。

なお、データのオープン化については、産業競争力の維持という視点も入れつつ、戦略的なデータの整備・公開を検討していく必要がある。

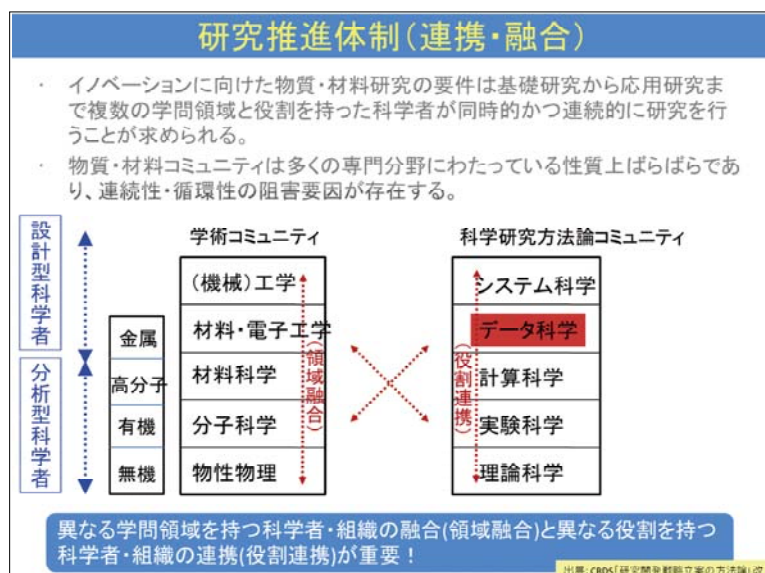
【コラム】研究開発プロジェクトを有効に推進するための推進体制（領域と役割の連携）

イノベーションに向けた物質・材料研究の要件は基礎研究から応用研究まで複数の学問領域と役割をもった科学者が同時的かつ連続的に研究を行うことが求められるが、物質・材料研究コミュニティは多岐の専門分野にわたっている性質上ばらばらであり、下記のような連続性・循環性の阻害要因が存在する。

1. 理論・物性物理（分析型）科学者の変化に対する現実観察の不足
2. 理論・物性物理科学者の変化に対する全体性観察の欠如
3. 理論・物性物理科学者と材料（設計型）科学者との連携の不十分性
4. 材料科学者の方法的未熟成
5. 材料科学者と産業・社会の間の交流の不足

具体的な例として、現状を計算科学の人材の面から見ると、材料の変形挙動を対象としてモデリングおよびシミュレーションを実施できる人材が明らかに不足している。力学的な考察については機械系の研究者が高い研究能力を有しているが、その一方で原子・電子レベルの知見に基づく有効な解析ができる人材が少ない。材料系の研究者は、材料組織や組織と変形形態との関連について豊富な知見を有しているが、力学的観点からの解析や考察ができる人材が少ない。物理、化学の分野においても、原子・電子レベルの解析には高い研究能力を有しているが、力学的観点からの研究についてはあまり行われていない。実験を主とした研究者も計算を利用して大幅に試行実験の回数を減らしたり、実験結果の解釈に応用することが可能になっている。しかしながら、実験と計算の間にある敷居はまだ高く、有効に利用できていないとも言えない。

物質・材料科学とデータ科学の知識をバランス良く身につけた研究者はほとんどおらず、マテリアルズ・インフォマティクスの未来を担う新しい人材の育成が急務となる。



したがって、基礎研究から開発に至る科学技術・学術領域の融合と、理論からシステムに至る色々な役割をもつ研究の間の役割連携の実現こそが重要であり、「異」を受容し、研究者、特に若手研究者が新興・融合科学技術に挑戦しやすい環境を構築することが必要である。

日本の強みは、「HPCI 戦略プログラムの CMSI」、「ナノテクノロジープラットフォーム」など基盤となるインフラストラクチャーやコミュニティがまとまる場があること、「元素戦略（拠点形成型）」など理論と実験が連携するプロジェクトが存在することである。これらのすでにある仕組みを活用することが最大の実効性を発揮するための鍵となる。

付録 1 検討の経緯

- 平成 24 年 3 月に JST 研究開発戦略センター戦略スコープ 2012 策定委員会により、承認された戦略スコープに基づき、CRDS 内にチームを編成し、提言の作成を開始することとなった。
- 4 月より活動を開始し、スコープの焦点を絞るとともに、ワークショップの設計、提言の方向性を検討するため、以下の有識者（敬称略）へのインタビュー・意見交換を実施した（ワークショップ参加者は別途次頁に表記）。
 - 物質・材料科学（実験・計測、計算）
井下猛、大野隆央、山崎政義（以上、物質・材料研究機構）、大友明、神谷利夫（以上、東京工業大学）、岸本浩通（住友ゴム工業）、久保百司、塚田捷、一杉太郎（東北大学）、榊茂好、谷口貴志、山本量一（以上、京都大学）、澤田英明（新日鐵住金）、高木英典、永長直人、溝口照康、山口由岐夫（以上、東京大学）、中山将伸（名古屋工業大学）、西川宜孝（みずほ情報総研）、兵頭志明（兵庫県立大学）、吉田博（大阪大学）、武藤俊介（名古屋大学）
 - 数理科学・データ科学
儀我美一、杉原正顕、山西健司、山本昌宏（以上、東京大学）、佐伯修、福本康秀、若山正人（以上、九州大学）、西浦廉政（東北大学）、杉山将（東京工業大学）
 - ライフサイエンス（バイオインフォマティクス等）
豊田哲郎（理化学研究所）
- インタビューなどを踏まえて、ワークショップの設計を行った。寺倉清之教授（産業技術総合研究所、北陸先端科学技術大学院大学）にオーガナイザーとして多大なご協力をいただいた。
- 同時並行して、産業競争力懇談会（COCN）の 2012 年度推進テーマにおける研究会「シミュレーション応用によるものづくり連携システムおよび新材料設計手法」にオブザーバ参加した。
- 平成 25 年 2 月 11 日および 6 月 1 日に「データを活用した機能に基づく新物質・材料の設計（マテリアルズ・インフォマティクス）」ワークショップを開催し、関連する取組みを共有するとともに、日本における今後の展望について議論を行った（次頁プログラム参照）。の方々に御礼申し上げます。
- 上記ワークショップにおける話題提供、議論を収めた報告書を平成 25 年 8 月に発行した（CRDS-FY2013-WR-03）。
- WEB や文献の調査、インタビュー、ワークショップの結果から総合的な検討を重ね、本プロポーザルを執筆した。

この場を借りて、お忙しいところ貴重な時間を割いてインタビュー・相談に乗っていただき、有益なアドバイスをくださった有識者の皆様、COCN の佐々木直哉氏（日立製作所）およびメンバーとして各企業から参加された皆様、ワークショップにご参加いただき貴重なご意見をくださった有識者の皆様、ワークショップの設計のために何度もアドバイス・意見交換いただいた寺倉教授に感謝・御礼申し上げます。

第1回ワークショッププログラム（敬称略、五十音順）

開催日時：平成25年2月11日（月）10時00分～17時45分

開催場所：科学技術振興機構 東京本部別館1階ホール（東京都千代田区五番町7）

オーガナイザー：

寺倉清之（北陸先端科学技術大学院大学、産業技術総合研究所）

田中一宜（JST 研究開発戦略センター）

10:00 ～ 10:30	オーガナイザー挨拶 趣旨説明	田中 一宜 寺倉 清之
10:30 ～ 11:00	「機能に基づく材料の設計～理論、実験、計算、データの統合への期待～」 細野 秀雄（東京工業大学）	
11:00 ～ 11:30	「データ科学による予測と原因究明」 津田 宏治（産業技術総合研究所）	
11:30 ～ 12:00	『『京』を用いたインシリコ創薬』 奥野 恭史（京都大学）	
休憩		
12:40 ～ 13:10	「第一原理計算を用いたマテリアルズ・インフォマティクス」 田中 功（京都大学）	
13:10 ～ 13:50	「Combinatorial and High-Throughput Screening of Materials Libraries、および MGI 等米国の動向」 竹内 一郎（メリーランド大学）	
13:50 ～ 14:10	「コンビナトリアルツールの貢献と課題」 知京 豊裕（物質・材料研究機構）	
質疑・休憩		
14:40 ～ 15:10	『『鉄鋼ゲノムの解明』について』 足立 吉隆（鹿児島大学）	
15:10 ～ 15:40	「データ同化によるモデルの高度化～物質・材料研究への応用可能性～」 樋口 知之（統計数理研究所）	
15:40 ～ 16:10	「企業における事例紹介および課題とアカデミアへの期待」 射場 英紀（トヨタ自動車）	
16:10 ～ 17:40	論点に沿って議論、コメンテーター等によるコメント	
17:40 ～ 17:45	閉会 田中 一宜	

コメンテーター：新井 正敏（J-PARC）、伊藤 聡（理化学研究所）、小谷 元子（東北大学）、
曾根 純一（NIMS、CRDS 特任フェロー）、常行 真司（東京大学、CMSI）、福山 秀敏（東
京理科大学）、船津 公人（東京大学）、松宮 徹（新日鐵住金）、鷲尾 隆（大阪大学）

第2回ワークショッププログラム（敬称略）

開催日時：平成25年6月1日（土）13時00分～17時45分

開催場所：科学技術振興機構 東京本部別館2階セミナー室（東京都千代田区五番町7）

参加者

CMSI: 寺倉 清之（産総研）、常行 真司（東京大学）、田中 功（京都大学）、伊藤 聡（理研）

元素戦略：細野 秀雄（東京工業大学）、田中 功

NIMS 材料情報ステーション：曾根 純一、緒形 俊夫（NIMS）

東北大学 WPI：小谷 元子（東北大学）

数学イノベーション：中川 淳一（新日鐵住金）、小谷 元子

データ・統計：井手 剛（IBM）、鹿野 豊（分子研）

文部科学省（ナノテクノロジー・材料開発推進室）

I 挨拶と趣旨説明（13時00分～13時30分）

13:00～13:10 田中 一宜（JST 研究開発戦略センター）および寺倉 清之（産総研）
ご挨拶

13:10～13:30 「第一回 WS の振り返り」
島津 博基（JST 研究開発戦略センター）

II 話題提供（背景・意義、手法、事例、課題、展望）（13時30分～15時30分）

13:30～14:00 「ベイズ推論と物性科学」
岡田 真人（東京大学）

14:00～14:30 「材料設計とデータベース」
及川 勝成（東北大学）

14:30～15:00 「オープンデータの潮流とデータの統合利用～ライフサイエンスの例～」
山口 敦子（ライフサイエンス統合データベースセンター）

15:00～15:30 「材料データベースの国際動向と今後の展望」
芦野 俊宏（東洋大学）

小休止

III 提言案について討論（15時45分～17時45分）

15:45～16:10 「提言案の説明」
島津 博基（JST 研究開発戦略センター）

16:10～17:45 討議

付録 2 国内外の状況

1. 海外

◆ 米国 “Materials Genome Initiative”

米国では、全米研究会議（National Research Council）や米国エネルギー省（DOE）の報告書において、計算科学や情報科学を活用して材料開発を効率化すべきであるとの主張がなされてきた。また、計算機を使った研究活動について、数千件におよぶヒアリングが行われ、大規模数値計算に対する期待よりも、大量データをどう使うかという問題の解決に対する期待の方がはるかに大きいという結果が得られている。これらを受けて、製造業の国際競争力強化に向けた施策の一つとして、国家科学技術会議（NSTC）から 2011 年に “Materials Genome Initiative” が公表され、2012 年には年間 1 億ドルの予算が措置された。このイニシアティブでは、材料開発に要する期間を 2 分の 1 に短縮するという目標を掲げ、「計算ツール」、「実験ツール」、「デジタル・データ」を材料イノベーション基盤として整備すべきであるとしている。それぞれの内容は以下の通りである。

「計算ツール」

- ・ 材料挙動の精確なモデルを作成し、理論および実験データによりモデルのバリデーションを行う。
- ・ 材料イノベーションと実用化に関係する人々がソフトウェアを容易に使用・維持できるオープンプラットフォームの枠組みを実現する。
- ・ 幅広いユーザー・コミュニティが便益を享受できるようモジュール性をもち、利用しやすいソフトウェアを作成する。

「実験ツール」

- ・ 合成・加工過程や実使用条件・環境下で、諸特性を効率的、定量的に評価できる方法、技術を実現する。
- ・ *in situ* 評価技術を計算ツールと組み合わせ、材料、反応条件、加工条件の高速スクリーニングを広範な空間・時間スケールで可能にする。
- ・ モデルパラメータの決定、予測検証、モデルの有効性・信頼性の向上・補完に必要な実験結果を提供する。

「デジタル・データ」

- ・ 材料研究者、開発者のためのデータ・ストレージシステム、データ転送システムを確立する。
- ・ モデルや実験ツールから得られるデータの精確性と検証可能性に重点を置くとともに、インフォマティクス研究を支援し、材料データの最も有効な検索と解析を可能にする。
- ・ 材料開発の一連のプロセスにおいて高度なデータ共有を行い、本イニシアティブを駆動する。

➤ 米国国立科学財団（NSF）” Designing Materials to Revolutionize and Engineer our Future (DMREF)”

DMREF は” Materials Genome Initiative (MGI)” の目標達成を支援するため、NSF の部局を横断した研究ファンディングの枠組みであり、特定の部局が担当する独立したファンディングプログラムではない。第1回目の公募が2012年1月15日～2月15日に行われ、同年10月に22の助成対象（14のプロジェクト）に対し約12百万ドルの助成を行うことが公表された。第2回目の公募は2013年1月15日～2月15日に行われている。

第1回目、第2回目とも、所望の特性や機能をもった物質・材料を設計するために必要な知識基盤の構築を通じて、物質・材料の発見・開発を促進する提案、マルチスケールで物質・材料の基礎的理解を深化し、微細構造、表面構造・加工が物質・材料の特性・性能に及ぼす影響を明らかにするような提案など、MGI の目標達成に資する提案を求めている点は共通している。提案される研究は、計算が実験・理論を先導し、実験・理論が計算の発展を促すようなプロセスが繰り返されるように、計算・実験・理論の連携が図られていなければならないとしている。

第2回目の公募では、さらに、新しいデータ解析手法や統計アルゴリズム、機械学習・データマイニング・スパースモデリングなどを活用した予測モデル構築、アクセス性や拡張性やもち継続的に維持されるデータインフラ、さらに膨大かつ複雑で多様データを管理する共同施設、の必要性を指摘した上で、提案には、透明性確保・データ共有、オープンソースソフトウェア供給に向けたデータ・マネジメントの強化計画を含むべきであるとしている。また、第1回目の公募では所管部局が Directorate for Mathematical & Physical Science (MPS) と Directorate for Engineering の2部局の3つの Division であったのに対し、第2回目の公募では Directorate for Computer & Information Science & Engineering が新たに加わり、Division も12に増え、多くのファンディングプログラムが DMREF に対応することになる。助成規模も要求段階で35百万ドルに拡大している。

MPS の Division のうち、第2回目の公募から DMREF に加わった Division of Chemistry (CHE) と Division of Mathematical Sciences (DMS) は個別に DMREF について “Dear Colleague Letter” を発表している。その中で、CHE は、化学構造と反応性や、所望の性質をもち、多機能性を有する新物質の発見・設計を可能とする分子間相互作用の基礎的理解に資する研究に関心があるとしている。また、DMS は、複雑で高次元の状態空間における設計最適化、微細構造とバルクの性質や組成・加工とバルクの性質との間にある関係を見出すデータマイニング手法、物質理解のための「第一原理」、マルチスケール問題やデータ同化といった計算科学の課題への対処に関心あるトピックとして例示している。

➤ 米国国立標準技術研究所（NIST） “Building the Materials Innovation Infrastructure: Data and Standards” ワークショップ

NIST 主催によるワークショップ “Building the Materials Innovation Infrastructure: Data and Standards” が2012年5月に開催された。本ワークショップでは、MGI において整備すべきとされた「計算ツール」、「実験ツール」、「デジタル・データ」の3つの材料イノベーション基盤のうち「デジタル・データ」を取り上げ、短期的（5年未満）・

長期的（5年以上）課題の抽出と MGI の材料開発期間の短縮という目標に照らした優先順位付けが行われた。

125 名を超える産学官からの専門家の参加を得たワークショップでは、長さスケール（マクロから原子スケール）と技術応用分野（電気化学的エネルギー貯蔵、耐熱合金、触媒、軽量構造材料）の 2 つの観点から課題抽出と優先順位付けが行われ、いずれの観点にも共通する優先度の高い課題として以下がワークショップ報告書に記載されている。

コミュニティの取組みをまとめる強いリーダーシップ：データの生成、解析、共有の各レベルでコミュニティを引っ張るリーダーシップが必要である。物質・材料コミュニティには多くの分野や応用があり、それぞれの利害・関心にも独自のものがある。そのため、コミュニティ全体を考えた取組みを支援・擁護するようなリーダーシップは限定的なものしかみられない。データ共有に向けてコミュニティ文化を変えていくためには、コミュニティ全体の支援と理解をとりつけるためのリーダーシップが求められる。このようなリーダーシップの問題は今後も課題として残り続けると考えられるが、マテリアルズイノベーション基盤が成果をあげるために解決が不可欠である。

データ共有：データ共有を促進するための仕組みとインセンティブが必要である。データの共有・流通に対する個々の研究機関のニーズとより広いコミュニティのニーズとの間のバランスをとるメカニズムが現在欠けている。発見、知的財産、差異化要因の先取権・所有権を保護しながら、ニーズ間のバランスをとる仕組みが必要である。共有されたデータセットに対し、一定のレベルでのデータ保護やデータの帰属確保と一体となった報酬システムを構築できれば、データを共有・流通させるためのインセンティブとなる可能性がある。Digital Object Identifiers (DOI) がデータ共有の仕組みのヒントを与えるものかもしれない。DOI はコンテンツ識別のための枠組みを提供するもので、知的コンテンツのメタデータを管理し、ユーザーをコンテンツ・ソースにリンクさせ、コンテンツ媒体の自動的管理ができるようにしている。

計算のバリデーション：計算モデルのバリデーションは長さスケール、応用技術分野、双方の観点から優先度の高い課題であるとされている。バリデーション手法は開発に長い時間を要するが、バリデーション手法の欠如は材料イノベーション基盤成功への全体的障害となる。バリデーションに向けた系統的で適切なデータ取得プロセスの評価が重要な課題である。また、バリデーションを経た計算手法が新物質・材料の設計をスピードアップできるか、実際の製造物に適用できるか、物質・材料の試験に必要な時間とコストを削減できるか、を比較・評価することも重要である。

集中化されたデータレポジトリ：集中化され利用しやすいデータレポジトリに対するニーズへの対処も高い優先度の課題である。このようなレポジトリは現在限られたものしかない。また、規格化されたフォーマットがないため、コミュニティ全般によるデータ利用を難しくしている。データ管理は一般にコストがかかり、データの収集・保存・メンテナンスにも相応の資源が必要である。集中化されたデータレポジトリに必要とされる一般的要件には、

- ・規格化されたフォーマット
- ・蓄積し、メンテナンスすべきデータの定義
- ・アップロードされるデータのデータ規格
- ・新しいデータタイプに対応できる拡張性
- ・非専門家でも利用できるような配慮
- ・長期的なデータメンテナンスのための方策

などがあげられる。ハブシステムを利用して多数のユーザーに対応するのも一つの方策である。この場合、ユーザーは単にレポジトリにアクセスできるだけでなく、データの処理についてもツール提供や支援が受けられるのが望ましい。

大規模データセット：大規模なデータセットを適切に保存・伝送・解析する方法の欠落が技術応用分野の観点から指摘されている。データ量は指数関数的に増加しており、この傾向は今後も続くものと予想される。大量のデータを適切に管理し、有効な情報を引き出すための新しい方法が必要である。

インタフェースと相互運用性：オープンソースのインタフェースやデータベース間の相互運用性の構築が必要である。データベース間のメタデータに関するインタフェースの欠如は、実験データと計算データの統合、異なるソースコードによる計算データの統合ともに、大きな課題である。異なるアプリケーションソフト間のフォーマット変換も解決の難しい問題である。柔軟性をもつ共通のデータスキーマ（データベース構造）とインタフェースの開発がこれらの課題のうちいくつかを解決するために必要である。複数のデータベースを単一の検索可能な環境に統合することも求められる。

データ基準：計算データ、実験データの双方に対し、厳格な基準が必要である。基準にはデータキュレーション、試験条件に基づくデータの品質を記載した品質指標が含まれる。データベースに収録されるデータについて、データフォーマットやメタデータに対する規格の策定も優先度の高い課題である。このような取組みは、データ共有やデータの品質および取得状況に対する理解を促進するものである。

◆ その他関連する動向

マサチューセッツ工科大学（MIT）とローレンス・バークレー国立研究所による“Materials Project”では、ICSDに収録された結晶構造データをもとに網羅的に第一原理計算を行った結果とNISTの熱力学データベースをリンクさせて、状態図や構造予測、化学反応予測などを収録したデータベースを公開している。さらに、これらのデータを活用できる各種アプリケーションソフトウェアも公開している。同様なデータベースはDuke大学のCenter for Materials Genomicsが主導するコンソーシアム“aflowlib.org”からも公開されており、結晶構造、電子特性、熱電特性などを見ることができる。米国ではこの2グループが第一原理計算から得られたデータの高度な共有を牽引している。

DOEが2009年に設立した46のEnergy Frontier Research Center（EFRC）の一つであるCenter for Inverse Design（2009～2014年）では、年間4百万ドルの予算で、

太陽電池材料の探索が行われており、遺伝的アルゴリズムなどデータ科学の手法を活用して、材料の網羅的探索を行った結果、これまで未知であった化合物 Ta-Co-Sn が安定であり、太陽電池として使用できるバンドギャップを有することを示している。

米 国 National Nanotechnology Initiative (NNI) は、2012 年 に NNI Signature Initiative（重点領域）の一つとして、これまでの太陽エネルギーの利用、ナノ製造、ナノエレクトロニクスの 3 つに加え、「ナノテクノロジー知識基盤（Nanotechnology Knowledge Infrastructure）」の構築を進めることを明らかにした。ナノテクノロジーは世界規模の課題解決への貢献が期待されながら、情報量が膨大なうえ未整理であるため、ナノテクノロジーに関する情報の収集・立証・保存・共有・抽出・分析・モデル化・応用を効率化するナノインフォマティクスの発展が望まれているとし、ナノインフォマティクスの発展は、再現性向上、実験データ共有の進展、データを情報や応用へと転換・活用するためのツールやモデルの開発と実証の促進など、ナノテクノロジーの研究から開発・応用に至る全過程に影響を及ぼすと予想している。このような知識基盤の一例として NSF には“Network for Computational Technology”を通じて施行されているプログラム“nanoHUB”がある。このプログラムでは、無償でダウンロード可能なモデリングとシミュレーション・アプリケーションが提供されており、ナノテクノロジー研究での計算ツールの使用をサポートするものである。研究者は最先端のモデリング・アルゴリズムにアクセスすることができ、またウェブサイト経由で共同研究が可能である。

物質・材料科学分野の代表的な学会である Materials Research Society (MRS) でも、2012 年の秋季大会では、マテリアルズ・インフォマティクスのセッションが設けられ、50 件近い発表が行われており、マテリアルズ・インフォマティクスの重要性が強く認識されていることが伺える。第一原理計算と X 線回折実験結果と遺伝的アルゴリズムを組み合わせて未知の結晶構造を探索する試みや、第一原理計算結果でトレーニングしたニューラルネットワークによりポテンシャルエネルギー曲面を構築する試みなどが新しい展開を予期させるものであった。

欧州では、CECAM (Centre Européen de Calcul Atomique et Moléculaire) の 2011 年のカンファレンスでマテリアルズ・インフォマティクスがテーマとして取り上げられたほか、ワークショップも頻繁に開催されている。DPG (Deutsche Physikalische Gesellschaft) の春季大会 (2013 年 3 月) でも、マテリアルズ・インフォマティクスの研究者による招待講演が複数の集中セッションやシンポジウムで行われている。

その他、2013 年 2 月には豪州でマテリアルズ・インフォマティクスの国際会議が開催された。韓国では、韓国科学技術研究院 (KIST) により MIDAS という第一原理計算結果のデータベースが構築され、これを利用して材料設計を行おうとする動きがある。

2. 国内

物質・材料研究にデータ科学、インフォマティクスを活用するという取組は世界的にも緒についたばかりであり、日本においても未だコミュニティは形成されておらず、一部の材料研究者が試験的な試みとして研究をはじめた段階である。今後、物性物理、数理・情報学などのコミュニティを巻き込んだ発展が望まれる。母体となりうる取組として下記の2つが挙げられる。

➤ 元素戦略＜研究拠点形成型＞（平成24年度～）

我が国の産業競争力に直結する4つの材料領域（磁石材料、触媒・電池材料、電子材料、構造材料）を対象に、物質の機能を支配する元素の役割の理論的な解明から新材料の創製、特性評価までを、拠点を中核として形成する共同研究組織の密接な連携・協働によって一体的に推進。

プロセス（材料創成）、計算（理論・電子論）、計測（特性評価）の3つのコミュニティが協力して応募をすることを初めて義務付けた画期的な試みといえる。平成24年度にスタートした文部科学省の10年事業である。

➤ 次世代スパコン戦略プログラム「計算物質科学イニシアティブ（CMSI）」（平成22～27年度）

「次世代スーパーコンピュータ戦略プログラム」の分野2＜新物質・エネルギー創成＞を進める研究ネットワーク拠点である。計算科学に限定されているものの、物性科学・分子科学・材料科学の3つのコミュニティが、分野の壁を越えて組織のネットワークを形成し、シンポジウム、ワークショップ、実験研究者との連携などを通じて、戦力となる研究者、明日を担う若手のネットワークを構築している。また、研究成果だけでなく、人材、ソフトウェア、データも広報誌やホームページを通じて紹介し、世界に発信する拠点づくりを目指している。

社会の要請を重視すると、シミュレーションも解析⇒予測⇒設計と進化していくことが求められる。日本でも計算科学を材料探索・設計に活用するアイデアは多くの研究者が考えているが、計算結果を利用して材料設計にうまくつなげるための方法論はまだ確立していない。

なお、知的基盤整備に関する関連法令として、下記のものが挙げられる。

①第4期科学技術基本計画（平成23年8月19日閣議決定）

研究開発活動を効果的、効率的に推進していくためには、研究成果や研究用材料等の知的資産を体系化し、幅広く研究者の利用に供することができるよう、知的基盤（注）を整備していく必要がある。

（注）研究用材料、計量標準、計測・分析・試験・評価方法及びそれらに係る先端的機器、関連データベース等

②研究開発力強化法（平成23年6月11日法律第63号）（研究開発システムの改革の推進等による研究開発能力の強化及び研究開発等の効率的推進等に関する法律）

国は、研究材料、計量の標準、科学技術に関する情報その他の研究開発の推進のための

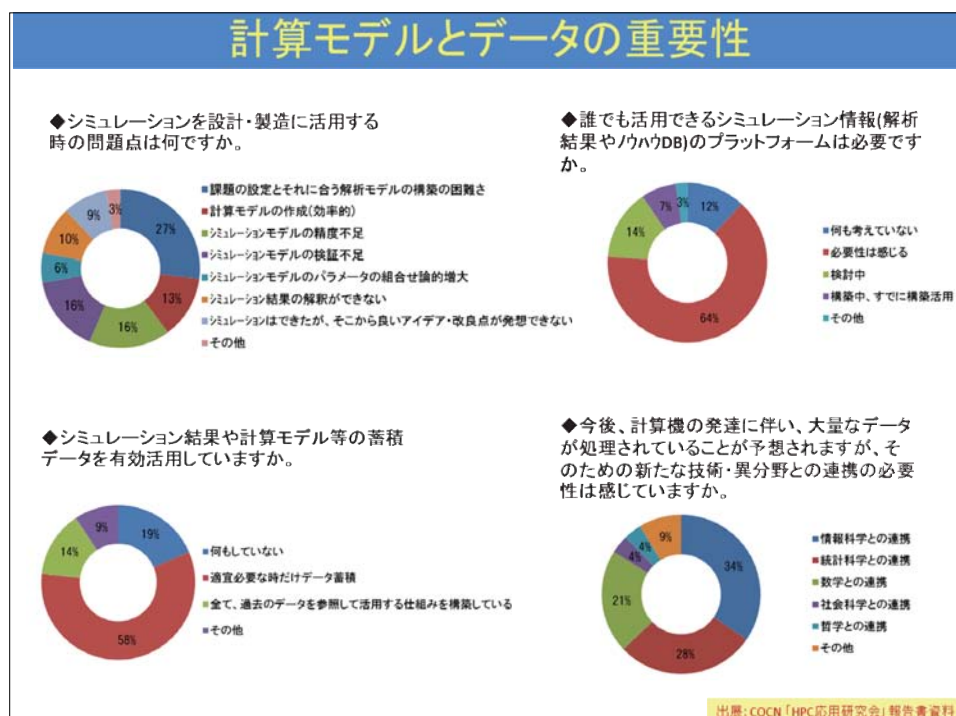
知的基盤をなすもの（以下この条において「知的基盤」という。）の供用の促進を図るため、国、研究開発法人及び国立大学法人等が保有する知的基盤のうち研究者等の利用に供するものについて、研究者等が知的基盤を利用するために必要な情報の提供その他の知的基盤を広く研究者等の利用に供するために必要な施策を講ずるものとする。

➤ 産業界の課題 ～COCN「HPCの応用」より～

シミュレーションや解析情報データと思考（設計）をうまく繋げたものづくりへの活用、すなわち、シミュレーションをうまく活用するための「知識、知恵の方法論（解析妥当性や目利き、創造性）」がまだ確立されていない。

産業界におけるシミュレーション活用の課題として以下の事項が抽出されている。

- ・ オープンソフト活用、解析情報データベースの有効活用、新しい方法論の構築、異分野連携手法
- ・ 解析情報の有効活用、活用データベース、複雑系モデリング理解
- ・ 産学連携解析知識共有データ化、知的活用構造化手法
- ・ シミュレーション解釈学の体系化（分野、業界ごとに考慮）



付録 3 参考文献

- NISTEP「無機材料研究におけるマテリアルインフォマティクスの動向」2006年1月
- NISTEP「忘れられた数学」2006年5月
- NEDO、東洋大学「材料データベースの連携プラットフォームの構築に関する調査研究」2008年3月
- JST-CRDS「新興・融合科学技術の推進方策に関する戦略提言 社会的課題の解決と科学技術のフロンティアの開拓を目指して」2009年11月
- JST-CRDS「研究開発戦略立案の方法論」2010年6月
- 戦略的創造研究推進事業（CREST）「理論と実験の連携強化を目指したシンポジウム」資料 2012年1月
- 産業競争力懇談会（COCN）「HPC（High Performance Computing）応用研究会」2012年3月
- 情報機構「第一原理計算」2012年3月
- 文部科学省「元素戦略シンポジウム資料」2012年7月
- 文部科学省 アカデミッククラウドに関する検討会「ビッグデータ時代におけるアカデミアの挑戦 ―アカデミッククラウドに関する検討会 提言―」平成24年7月
- 文部科学省 科学技術・学術審議会 先端研究基盤部会「数学イノベーション戦略（中間報告）」平成24年8月
- 文部科学省 科学技術・学術審議会 先端研究基盤部会「科学技術イノベーションを牽引する研究基盤戦略について―研究開発プラットフォームによる研究開発力強化策―」平成24年8月
- 経済産業省 産業構造審議会・日本工業標準調査会合同会議 知的基盤整備特別委員会「中間報告 ―知的基盤整備・利用促進プログラム―」平成24年8月
- 産業競争力懇談会（COCN）「シミュレーション応用によるものづくり連携システム及び新材料設計手法」2013年3月

- 米国 National Research Council「Integrated Computational Materials Engineering」2008年
- 米国 OSTP「Material Genome Initiative」2011年
- 米国 NSF「Nanoinformatics 2020 Roadmap」2011年4月
- 米国 NIST「Building the Materials Innovation Infrastructure: Data and Standards」2012年11月

付録4 専門用語解説

第一原理計算

原則として実験データや経験パラメータを使わないで、物質中の電子の量子力学的運動を記述するシュレディンガー方程式を解くことによって、物性発現の舞台となる物質の電子状態を数値的に解く方法。物質が原子から構成されていることを利用して、構成原子の種類と位置を指定して計算する。物質の静的な特性を精度良く知ることができるだけでなく、時間発展する電子状態を計算することで、系の時間変化を追うこともできる。

第一原理分子動力学法

物質中の電子と原子核のセットに対して、原子の動きの度毎に密度汎関数理論（多体電子系のエネルギーなどの物性を電子密度から計算することを可能にした理論）に基づく第一原理計算を実行することで原子間に働く力を計算しながら、大規模な系の電子状態と原子運動を同時に求め、安定な原子配置（構造）や原子の運動の様子を高精度に調べる手法。古典的には、原子間に働く力を実験データや計算結果を利用して適切な「経験的」関数で与え、原子の運動方程式を解いて物理量を算出していたが、この手法の出現により、系の構造の分子動力学計算や最適化が電子状態計算と同時に高速に行えるようになった。

擬ポテンシャル

多くの物性は価電子の状態で決まっていることから、内核電子の効果と原子核からのクーロンポテンシャルを合わせて、あたかも内殻電子がないとした、価電子が実効的に受けるポテンシャル。擬ポテンシャルを用いたバンド計算において、最初は、経験的な擬ポテンシャルが使われていたが、現在では第一原理（非経験的）擬ポテンシャルの使用が主流となっている。

モンテカルロ法

平衡状態にある系の任意の物理量に関する統計平均を求めるための、ランダムな数を使った統計熱力学的な計算手法。相転移の振舞いや安定構造の探索等に用いられる。時間変化を模す運動学的モンテカルロ法は、反応や結晶成長のシミュレーションなどにも使われる。現象の全体的理解に適している。この名前はモナコにあるカジノの町（Monte Carlo）に由来している。

フェーズフィールド法

メソスケールにおける相変態を連続体モデルで扱うことにより、微細構造等の組織形成を現象論的にシミュレーションする手法。微細構造間の境界のダイナミクスは、それぞれの微細構造の相の自由エネルギー、弾性エネルギーなどにより決定され、境界が進行する速度、境界の曲率の間に成立する関係を満たしながら時間発展していく。組織の形態を濃度や規則度等の複数の変数で表現し、その時間・空間変化を発展方程式に基づいて計算することにより、複雑な材料組織形態を形成する界面の移動過程を解析できる。

機械学習

コンピュータに学習データを読み込ませることで、ヒトの情報処理能力を超えて、自動的に今あるデータを分類・予測したり、隠れた規則や法則を発見する手法。一定数以上の訓練データを基にして、様々なアルゴリズムによる規則性と法則性の学習を行った後、正解が明示されているテストデータを用いて、学習結果から導き出される解答とその正解を比較して、精度や汎用性などを評価する。

ベイズ推定

ある証拠に基づいて、その原因となった事象を推定するための確率論的方法。「ベイズの定理」を用いて、新しい証拠に照らして命題の尤も（もっとも）らしさを改訂していく方法である。

最尤法

ある分布から観測値が取り出されたとして、“そうなる確率（尤度）”が最も高くなるように分布の具体的な形状を決める方法。尤度とは、ある前提条件に従って結果が出現する場合に、逆に観察結果からみて前提条件が「何々であった」と推測する尤もらしさを表す数値を、「何々」を変数とする関数として捉えたものである。

観測者は、ある観測対象に対して、Aという事象があることを認識しているとする。その事象が起こる確率を $P(A)$ とする。これを、事前確率と呼び、観測者が観測以前にもっている信念を確率分布関数で表現したもの。

多変量解析

ある対象から得られた、お互いに関連のある多種類のデータを「多変量データ」といい、そのようなデータを総合的に解析する手法。あるデータが、結果として他のデータに影響を与えている時、影響を与えているデータを「説明変数（記述子と同じ）」と呼び、影響を与えられているデータを「目的変数（外的基準）」と呼ぶ。

外的基準がある場合→予測・判別の問題。

外的基準がない場合→データの内部構造の分析。

予測・判別の問題は、目的変数が説明変数からどのように影響を受けているのかをデータに基づいて分析し、その内容を検討したり、あるいは説明変数が特定の値だった時に目的変数がどのような値になるかを予測したりするもの。データの内部構造の分析は、項目と項目がどのようにからみ合い、どのように影響を及ぼし合っているかを分析したり、それに基づいて効率的な要約値を求めたりするもの。

- 主成分分析

因子分析は変数相互の関連を分析し、たくさんの変数を数少ない変数（因子という）に集約していく。目的変数はなく、各変数相互の関係を分析し、数多くある変数を少数の因子に集約する。

- クラスター分析

サンプルの中で似たものを集めて、グループ化する。クラスター分析は量的データに基づいて似たサンプルを集め、サンプルを任意のグループにまとめる。また、似た量的データを集め、データをグループにまとめる。

- ・ 回帰分析

ある結果（量的データ）に対して、その原因（量的データ）を1次式で結ぶもの。目的変数に影響を及ぼしているいくつかの変数との関係を分析し、目的とその要因となる変数の解明。目的変数の値の今後の予測。

データ同化

モデルから複雑現象を再現する演繹アプローチと複雑現象の観測結果からモデルを推測する帰納アプローチが融合した次世代のシミュレーション技術。数値モデルと観測データを組み合わせて、現実的なデータセットを作成することであり、シミュレーションモデルの変数と実際の観測値をつなぐ観測モデル（一般状態空間モデル）を用意することが重要である。複雑な現象を科学的に理解し、精度よく予測したいという要求にこたえる。

スパースモデリング

高次元データから背後に潜むルールや法則の抽出という問題に対して「高次元データを説明するための変数がデータの次元数と比べて少数である」というスパース性を仮定し、説明変数になるべく少なくなることとデータとモデルの適合とを同時に要請し、自動的な説明変数の選択を行うためのデータ解析法。

オントロジー

文書などデータの内容を説明する意味情報（メタデータ）を各文書などに付加し、メタデータを記述する用語を定義する構造を構築すること。オントロジーを導入することにより、検索対象となるデータが単なる単語や情報の集まりとしてではなく、データ全体で意味を持ったデータとして扱われ、各データについて統一的な付加情報をもたせることができる。これにより、必要な情報を的確に検索することが可能となる。

■戦略プロポーザル作成メンバー■

田中 一宜	上席フェロー	(ナノテクノロジー・材料ユニット)
島津 博基	フェロー	(ナノテクノロジー・材料ユニット)
中本 信也	フェロー	(ナノテクノロジー・材料ユニット)
的場 正憲	フェロー	(電子情報通信ユニット)
永野 智己	フェロー	(ナノテクノロジー・材料ユニット)
宮下 哲	主査	(戦略研究推進部)
荒岡 礼	主査	(戦略研究推進部) ※ 24 年 11 月まで
及川 智博	主査	(産学基礎基盤推進部) ※ 24 年 9 月まで
石原 聡	特任フェロー	(ナノテクノロジー・材料ユニット) ※ 25 年 3 月まで

※お問い合わせは、下記までお願い致します。

CRDS-FY2013-SP-01

戦略プロポーザル

「データ科学との連携・融合による新世代物質・材料設計研究の促進 (マテリアルズ・インフォマティクス)」

～物質・材料研究を飛躍的に発展させるための新たなパラダイム～

STRATEGIC PROPOSAL

Materials Informatics

Materials Design by Digital Data Driven Method.

平成 25 年 8 月 August, 2013

独立行政法人科学技術振興機構 研究開発戦略センター ナノテクノロジー・材料ユニット
Nanotechnology/Materials Unit, Center for Research and Development Strategy
Japan Science and Technology Agency

〒102-0076 東京都千代田区五番町 7 K's 五番町

電 話 03-5214-7481

ファックス 03-5214-7385

<http://www.jst.go.jp/crds>

© 2013 JST/CRDS

許可無く複写／複製することを禁じます。

引用を行う際は、必ず出典を記述願います。

No part of this publication may be reproduced, copied, transmitted or translated without written permission.

Application should be sent to crds@jst.go.jp. Any quotations must be appropriately acknowledged.

ATTAATC A AAGA C CTAAC T CTCAGACC
CT CTCGCC AATTAATA
TAA TAATC
TTGCAATTGGA CCCC
AATTCC AAAA GGCCTTAA CCTAC
ATAAGA CTCTAACT CTCGCC
AA TAATC
AAT A TCTATAAGA CTCTAACT CTAAT A TCTAT
CTCGCC AATTAATA
ATTAATC A AAGA C CTAAC T CTCAGACC
AAT A TCTATAAGA CTCTAACT
CTCGCC AATTAATA
TTAATC A AAGA C CTAAC T CTCAGACC
AAT A TCTATAAGA CTCTAACT
ATTAATC A AAGA C CT
GA C CTAAC T CTCAGACC
0011 1110 000
00 11 001010 1
0011 1110 000
0100 11100 11100 101010000111
001100 110010
0001 0011 11110 000101

