CRDS-FY2013-WR-03

科学技術未来戦略ワークショップ

データを活用した設計型物質・材料研究（マテリアルズ・インフォマティクス）ワークショップ報告書

2013年2月11日（第1回）、6月1日（第2回）

独立行政法人科学技術振興機構 研究開発戦略センター
Center for Research and Development Strategy Japan Science and Technology Agency

目　次

1. ワークショップ開催の背景と趣旨

平成 23 年度にスタートした第 4 期科学技術基本計画は、我が国が取り組むべき社会的課題を設定し、それを解決するための戦略を策定する一連の流れの中で、実効性のある研究開発課題を設定する課題解決型の政策を求めている。

現在、我が国の素材・材料分野は、産業の発展に大きく貢献しており、国際的にも優位な位置を堅持している。材料開発にはこれまで経験と勘に裏打ちされた実験的手法が大きく貢献してきたが、新物質の発見から材料としての実用化まで非常に長い時間と費用を要しているのも事実である。今後もグローバルレベルで産業競争力を発揮し続けるためには科学技術を総動員して材料開発に要する時間と費用を短縮すると同時に社会的課題に応える材料を開発することが望まれる。

このような観点から、世界の研究者の関心は、所与の機能を持つ材料を理論的に探索・設計した上で合成・評価するという方向に注目が集まりつつあるが、その方法論は確立していない。一方で、ある物質が優れた特性を持つことがわかっても、その構造と物性の相関、物性を支配する原理（パラメータや関数）が不明なものも多く、これを解明する科学にチャレンジすることには大変意義がある。

このような問題を解決するためには、従来の理論・計算、物質創成、計測・解析という 3 本柱の取組みに加えてインフォマティクス（データのマネジメント）の活用に真剣に取り組むことが重要ではないかと考えられる。すなわち、計測機器や計算機の進展にともない、短時間に詳細なデータが大量に得られるようになったものの、そこから意味のある情報を抽出する方法論は未だ確立されておらず、大量のデータを統合的に活用する枠組み、使用目的に合わせた材料開発に活かす枠組みを構築することが望まれる。大量のデータをマネジメント（蓄積・共有・循環）し、インフォマティクスを活用することで、高効率な材料探索等が可能となる。さらに、大量データから導き出される物理的・化学的法則の発見という基礎科学的な可能性も期待され、データ作成・蓄積・活用における物質科学者による挑戦に期待が寄せられている。

これらの課題に挑戦する際には、専門性の深化に伴う異分野研究者間のコミュニケーション不足をどう解消し、基礎サイドと応用サイドの認識のギャップをどう補っていくかが肝要となる。

従って、JST 研究開発戦略センター（CRDS）のワークショップにおいては、(1) データを活用した設計型物質・材料研究（マテリアルインフォマティクス）に関連する各分野の第一線で活躍されている皆様の意見交換の場、かつ異分野研究者間の相互理解を深める場、(2) 今後のマテリアルインフォマティクスにおける研究開発の推進方策を探る場、そして、(3) そのために実験・計測科学者、計算科学者が今後なすべきこと、あるいは成し得ること等について産業界の有識者も参加して考える場、を提供するものとした。

具体的には、2 月 11 日と 6 月 1 日の 2 回にわたって開催し、1 回目は、物質と材料の両者について実験と計算科学者に登壇いただくとともに、データ科学の現状、企業の現状などについて話題提供頂いた。2 回目は、主にデータ基盤（データベース）の話を話題提供いただいた。

本報告書はこれら 2 回のワークショップの内容を取りまとめたものである。

2. 第１回ワークショップ議事

２－１　開会挨拶

田中一宜（科学技術振興機構 研究開発戦略センター）

私はこのセンターに勤め始めてから、今年３月で８年になります。研究を離れて行政に近いところ、科学技術の戦略に関連するところで働いてきたわけですが、最近の傾向を申し上げると、もはや科学技術が一部の先進国のものではなくなったということです。たとえばナノサイエンスやナノテクノロジー、あるいは物質科学というところでいうと、政府のそうした分野への投資総額を米国、欧州、アジアで比較すると、アジアに重心が移っています。また研究人口も最近の統計では重心がアジアに移っています。本当の意味で、大競争時代に入ったという気がします。

20 ー 30 年前に、とても面白い本がありました。『科学者たちの自由の楽園』というタイトルで、昔の自由闊達な研究者を描いたようなものだったと思います。いまや研究者は、そういう状況にはよくも悪くもいられず、科学技術もスピードを競う時代になっているわけです。そのような時代に、おそらく一番大きな役割を果たすであろうと思われるものの１つが、インフォマティクスであり、本日のワークショップのテーマだと思います。今日のこういうインフォマティクスというのは、データをいかにうまく利用して、物事を早く進めるかということになると思うのですが、産業のスピードアップにほとんど直結する大きな影響力をもっている分野ではないかと感じています。

実は日本人はそれが一番苦手なのですね。要素技術その他、あるいは知識にしてもいろいろないいものがあるのですが、それを１つの目的に向かって統合して、１つ１つのプロセスを全体のプロセスとして効率よく流して、スピードアップしていく。そのようなシステム的なコーディネーションが日本は非常に苦手なのです。そのためにずいぶん損をしているわけです。

とにかくあふれ返るような基礎研究の情報を効率よく広く集めて、それを整理して全体を俯瞰していくといったシステムが必要ですし、それに基づいて社会が我々に対して求めているいろいろな機能を最適に組み合わせて、設計して実現していくといったようなことが必要になってきているわけです。科学をアナリシスのタイプとすると、設計、デザインの時代に入っているのではないかという気がします。こういうテーマを扱ったワークショップをオペレートしていくためには、そういった俯瞰的な見方ができることが必要です。

ちょうど 20 年前に当時の通産省の中で「アトムテクノロジープロジェクト」というナノテクノロジーの国家プロジェクトの走りのようなものが始まりました。そのときに、全体として実験と理論をうまくつないでいくという意味で、理論・シミュレーションの支柱としてお願いした寺倉先生に、今回の問題を半年から１年ほどかけていろいろ考えていただきました。本日は公式に初めてその問題を扱うワークショップです。よろしくお願いいたします。

２－２　趣旨説明

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）

私は物性理論の立場でお話しします。計算科学、情報科学、数理科学の連携で、科学技術研究の新しい研究の流れをつくるのが今回の狙いです。そのためには、図 1 に示すように、中心に「データ科学」を置いて、「シミュレーション」、「実験」、「物質開発・材料開発・産業活動」をうまく連携し、全体的に相補的に協働させる体制をつくらなければなりません。これには、2000 年初めに産総研に計算科学研究部門を設立したときに私が作ったモットー「解析から予測へ、予測から設計へ」と、2012 年度から発足した北陸先端大のシミュレーション科学研究センターの狙い「データ処理、シミュレーション、数理科学、実験の連携」を具現化していきたいという思いがあります。

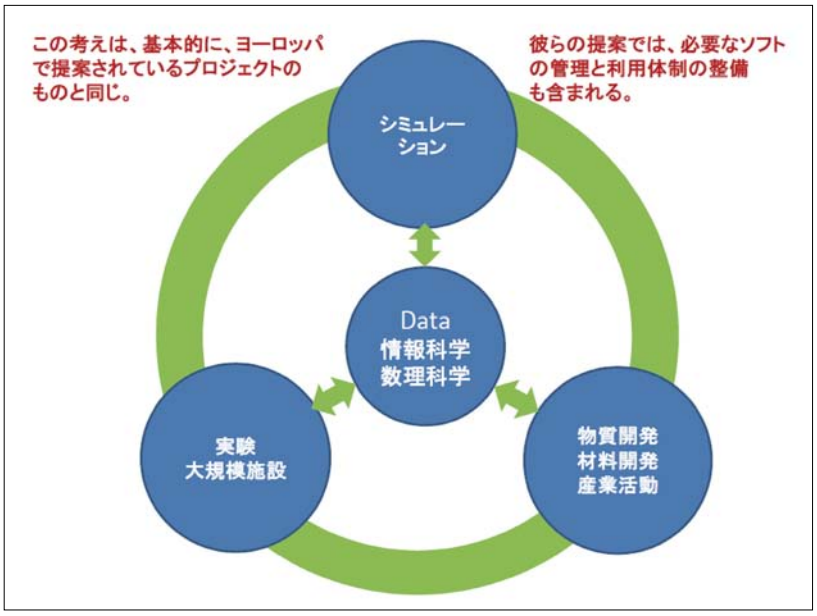


図 1

この考えは、実は、欧州で - 凝縮系物理学者が中心になって提案しているマテリアルインフォマティクスのプロジェクトの概念と同じです。このプロジェクトには、研究遂行に必要なソフトの管理と研究体制の整備も含まれています。

データ利用における効用、計算科学と実験科学との連携の重要性を述べ、何を問題にするのかも、少しだけ私の考える範囲内でお話しします。

図 2 の左上に示す結晶構造は細野先生が発見された鉄系超伝導体ですが、この関連物質にはいろいろな超伝導転移温度 T_c を持つものがあります。これらの T_c は、FeAs 層を構成する $FeAs_4$ 四面体における As-Fe-As のボンド角 α と相関関係があり、ほとんどの物質が一つの曲線に乗ることを、産総研の Lee らが見つけて、Lee プロットと呼ばれています。この系の T_c を記述する非常に良い記述子はこの四面体の頂角 α であるということ

を、Lee プロットは示唆しています。

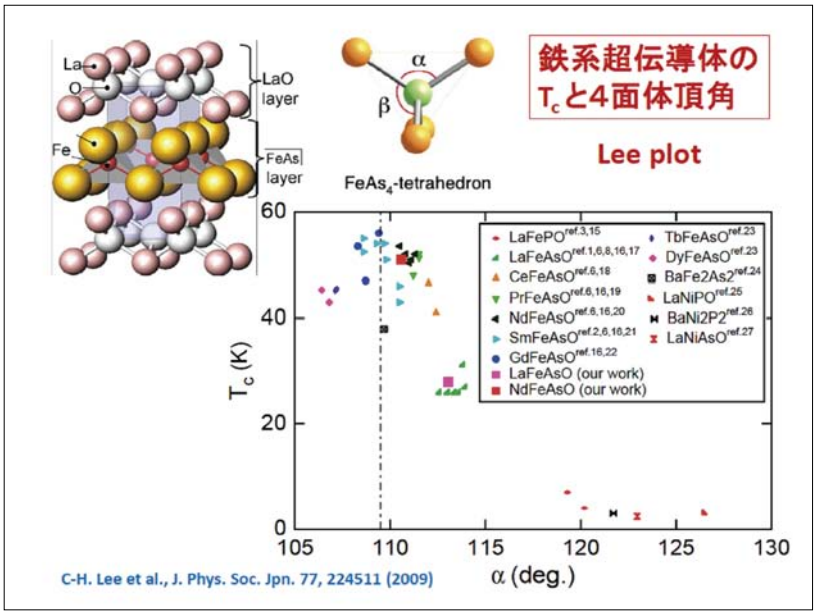


図 2

2 番目の例としては、2 元合金の生成エネルギーに関する Miedema 則が挙げられます。これは随分古い話ですが、2 つの元素の電気陰性度の差Δφと、それぞれの元素単体の基本単位胞（Winger Seitz cell）境界の電子密度の差Δn_{ws}、という 2 つのパラメータを使ってマッピングすると、2 原子が混じり合って合金を形成するものと形成しないものに明確に分かれることを Miedema が半経験的に見出しました。Miedema 則の物理的内容は、その後、理論的に非常に熱心な議論を経て、微視的に再解釈され、こうしたデータの整理の仕方が非常に面白い問題だとして浮かび上がってきていました。

ここから私自身の研究課題である固体高分子型燃料電池の話をします。今一番問題になっているのは、カソードにおける酸素の還元反応が非常に遅く、そのために触媒として大量のプラチナが必要だということです。我々が研究しているのは炭素系物質であり、それに窒素をドープすると触媒活性が上がるということが分かっているのです。ところが、炭素系触媒の合成には鉄を含んだ物質をまぜ込むので、実際は遷移金属元素がどう働いているのかも未解決問題になっています。酸素分子の非解離吸着の場合を考えることにして、標準水素電極電位を基準とした電位がゼロだとして、反応の中間状態が平衡状態にあると仮定しています。そうすると、$O_2+4H^++4e \rightarrow *OOH+3H^++3e \rightarrow *O+H_2O+2H^++2e \rightarrow *OH+H_2O+H^++e \rightarrow 2H_2O$ の各ステップの自由エネルギーの差ΔG を計算することができます。最大起電力は 4 つの電気化学過程の最小のΔG になるので、自由エネルギーの変化を等分にできれば、一番効率性の高い燃料電池が開発されることになります。ΔG を計算すると非常に簡単な式になって、各ステップのΔG が OH の吸着自由エネルギーΔG_{OH}だけで書くことができます。要するに各ステップのΔG をΔG_{OH}でプロットすることができ、最適な OH の吸着エネルギーが実現できる材料を探すことで、材料のスクリーニングができるわけです。この解析の基本的考え方は、Nørskov らによって提案されたものですが、

そのポイントは、多くの計算結果から、O, OH, OOH の吸着自由エネルギーの間の簡単な関係式を得たところにあります。

話題を変えて、計算と実験の連携は非常に重要だということをお話しします。最先端デバイスの革新的創製を目指すためには、多くの場合に革新的にきわどい性質を示す物質・材料が重要になります。たとえば、超伝導、巨大応答、マルチフェロイック、異なる磁性体間の界面､触媒､等はすべて非常に微妙なきわどい性質を持っています｡「シミュレーションでは、入力したものから得られる結果しか出てこない」というのは、実は 1980 年代に久保亮五先生がある報告書に書かれて､この当時私はこの言葉に非常に反発を感じました。しかしながら、今では、本当にそのとおりだと思っています。計算モデルの中に、考慮すべきものが始めからきちんと取り入れられているかどうかによってシミュレーション結果が変わってしまうのが事実です。それから、微妙なきわどい性質に関する問題に耐えるだけの計算の信頼度は、今はまだ得られていません。

たとえば、図 3 に示すように、非常に簡単な FeO という NaCl 型の遷移金属酸化物があり、Fe^{2+} イオンは d^6 というアップスピン 5 個、ダウンスピン 1 個の電子配置をとっています。多数スピン状態が 5 個全部つまっていて、少数スピン状態に電子が 1 個あり、電気的には絶縁体です。理想的な立方晶では、ここの三重縮退したスピン状態に電子が 1 個あり、通常のバンド計算では金属であるという間違った結果を導いてしまいます。この問題を解決するには、まずは結晶が立方晶でも、反強磁性秩序により電子状態的には菱面体の対称性になっており、三重縮退が一重状態と二重縮退状態に分かれることを考えなければなりません。ただし、通常のバンド計算の近似では、この分離がバンド幅より小さくて金属のままです。しかし、少し手の込んだ手法を使うと、一重項状態と二重項状態が大きく分離し、図 3 のエネルギーダイアグラムの左の rhombohedoral の解に落ち着きます。実験データとの対比を怠り、単純に考えてしまうと、下に来る一重のダウンスピン状態に電子が 1 個入るので絶縁体となるという一見問題を解決したかのような計算結果を得てしまうのです。しかしこれは誤りです。このような間違った論文が 2009 年に PRB と PRL に 1 報ずつ出ています。実際には、二重縮退状態が下に下がり、一重が上になり、さらにスピン軌道相互作用を考慮すれば、二重縮退状態がさらに 2 本に分かれて、この一番下のダウンスピン状態に電子が 1 個に収まるという状況が導かれます。これは 1957 年に金森先生の論文に書かれており、1999 年の我々の論文でも書いていますが、何十年も経ったときに実験データとの対比を怠り単純に考えると間違った結果になるわけです。このことは、実験をきちんと理解していないと誤った結果に導かれるという計算の危険性を示しています。

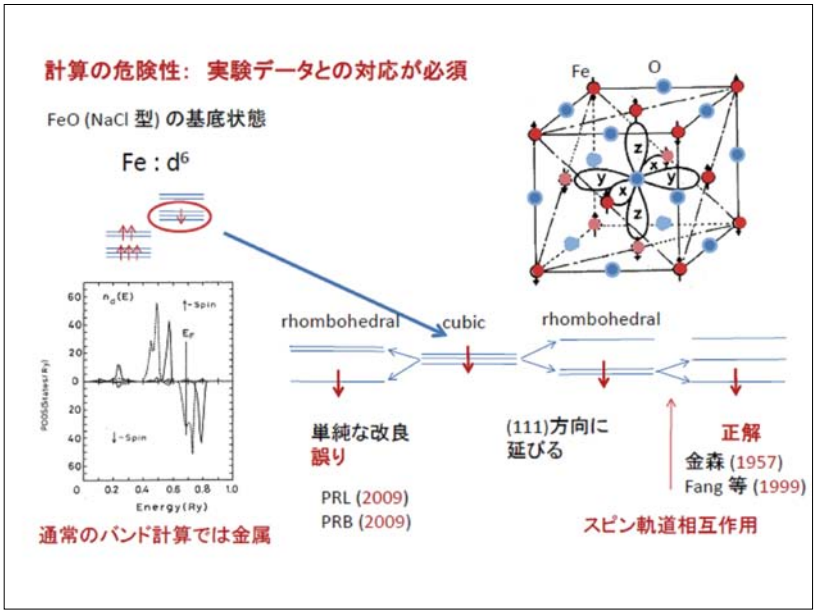


図 3

もう 1 つの例は、鉄鋼では非常に重要なγFe の問題です。鉄鋼では高温からクエンチするので、高温で安定な fcc（面心立方）相が重要なわけです。ところがこのγFe は非常に複雑であり完全な fcc 相だとすると、鉄の強磁性状態と反強磁性状態のエネルギー差は数 10meV あるのですが、fct（面心直方）に少し構造緩和すると、このエネルギー差は数 meV になります。さらに単斜晶にすると、強磁性が一番安定だという結果が出てきています。このように計算結果は非常に微妙なので、高圧にして fcc 構造が安定化されるような領域において何らかの実験と比較しないと、なかなかγFe の磁性の実態がわからないのではないかと思います。電子相関のことや、温度の影響を考えると、計算だけでは正しい答えを得るのは非常に困難です。

もう 1 つは、70 年間未解決であったマグネタイト（Fe_3O_4）の低温構造に関する問題ですが、これを解くきっかけになったのは、軌道放射を使った X 線回折と中性子線回折の併用による構造解析で、2001 年、2002 年になされた実験によります。それを受けて理論計算がされ、これまでの物理描像とはかなり違う電荷整列・軌道整列という電子状態結果を得ています。マグネタイトはフェリ磁性体であると同時に強誘電体にもなるのですが、この計算結果はそれを同時に説明するマルチフェロイクス（強磁性強誘電体）の理解の基盤にもなりました。

図 4 は、北陸先端大の Dam らが計算した Mn_4 単一分子磁石で、この単一分子磁石内の A サイトの Mn^{4+} と B サイトの Mn^{3+} との間に働く交換相互作用がどういうもので決まっているかを機械学習で解析しました。交換相互作用の大きさ、ボンド角やボンド長、エネルギー等いろいろなパラメータを機械学習すると、図 5 下に示すような、交換相互作用を決めるパラメータとして、どういうパラメータがどの程度依存しているかを表したグラフが得られます。

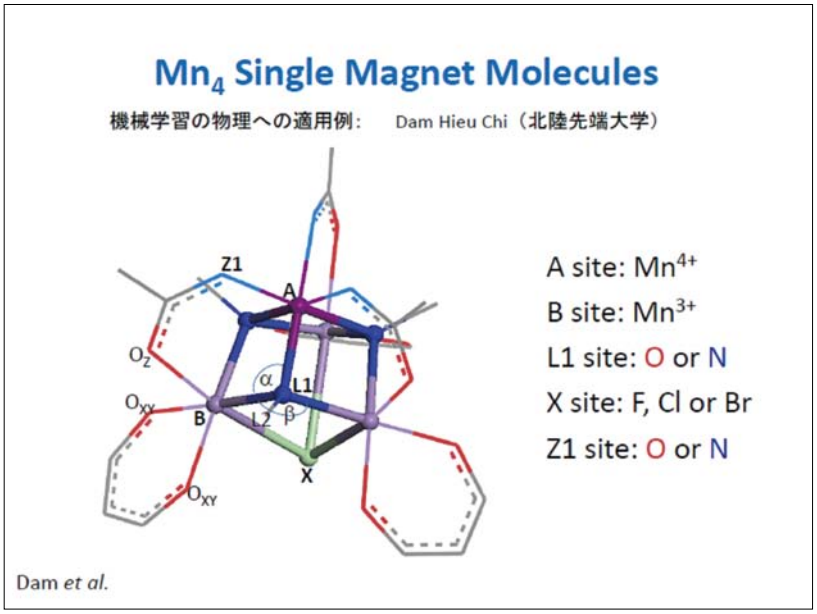


図 4

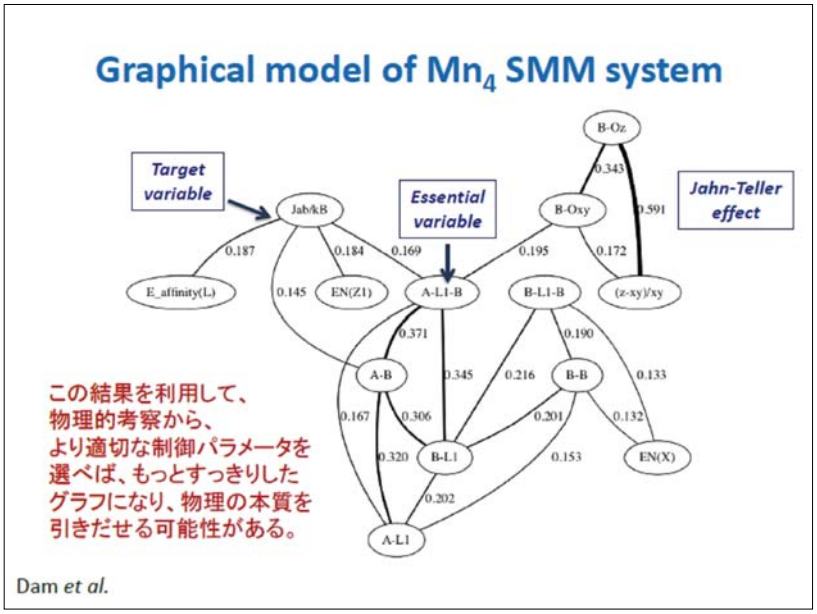


図 5

今の段階では、このダイアグラムは複雑で、適切な制御パラメータはなかなか見づらいのですが、この結果を利用して物理的考察を導入し、より適切な制御パラメータを得られれば、もっとすっきりしたグラフになって、物理の本質を引き出せる可能性が出てくると思います。こういうことが機械学習により自動的にできるというのは非常に素晴らしいことで、これから非常に有望なことだろうと思っています。

さて、物性を制御する物理パラメータにはいくつかあり、たとえば鉄系超伝導体の T_c であれば $FeAs_4$ 四面体における頂角（As-Fe-As のボンド角）α で決まりますが、2 元合金の生成エネルギーは 2 つの元素の電気陰性度の差と単位胞境界での電子密度差で決ま

り、酸素還元反応はOHの吸着エネルギーで決まります。これら、ある現象を制御するパラメータは記述子と呼ばれます。

最後に、マテリアルインフォマティクスで何をするか、について話します。

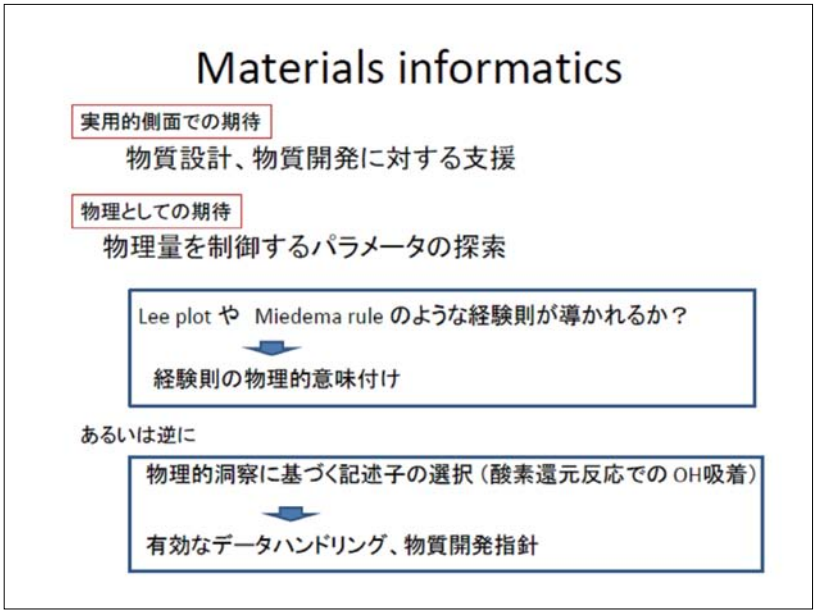


図6

まず1つは実用的側面での期待としては物質設計、物質開発に対する支援があり、物理としての期待としては、重要なターゲットである物理量を制御するパラメータを意味のある形で探索できる方向に進めていきたいと思っています。たとえばLeeプロットやMiedema則のような経験則が導かれるのか？あるいは逆に、物理的洞察に基づいて記述子を選択すれば、非常に有効なデータ処理が可能となり、物質開発の指針になるのではないかと思っています。

要するに、これから非常にたくさんのデータが出てきますから、強力な連携体制をつくり上げて、大量のデータを有効に活用し、全体としての物質開発や物質科学に貢献できるようなシステムを作っていければと思っています。

【質疑応答】

質問者：強い電子相関が効いている系で、今おっしゃったようなことがプラットフォームにそのまま乗るような状況になっているのかどうか？その背後にある問題として、デザインしている小さな系では、強い電子相関が無視できなくなり、その問題がやはりいつも顔を出すことになるのではないか？と想像します。今の段階では、このような手法では、正しい答えに行くとは限らないということですね。

寺倉：Mn_4単一分子磁石における制御パラメータの探索では、データマイニングの手法を使って、現状では粗い計算ですが、支配パラメータを自動的に見出すことが可能であるということをお見せしただけで、実際にはもっと先端的な計算をしないといけないと思っ

ています。要するにデータマイニングの技術を使うと、この範囲内でいろいろな相関が自動的に出てくることをお見せしただけで、まだ物理としては不十分です。さらに、実験との連携や高度の計算が必要なので、きちんと専門の物性物理や量子化学の方がこの問題に関わってくださらなければ､データの質の保証はされないという問題意識を持っています。最先端のデバイスや触媒など、いろいろなことをやろうとすると、やはり非常にきわどい話が多くなります。非常に高度な計算をしなければいけないけれども、高度の計算は複雑な系には適用できるほど力がないので、実験と連携して、どの位の系までならどの程度信頼できるかをしっかりチェックしながら研究を進めなければならないというのが、1つの大きな問題です。

たとえば、鉄系超伝導体においてスピンが重要な問題になっていますが、スピンのことを議論する理論計算の多くには、電子・格子相互作用が入っていません。何のパラメータを入れて、どこまで考慮しているかによって計算結果は大きく左右されてしまいます。それらの物理パラメータをきちっと全部考慮するのは人間には不可能なので、たとえばある物質の性質を計算すれば、すぐに実験データが出てきて、結果が妥当なのかどうかをすぐに判断できるような仕組みを入れる必要があります。そうしなければ、だんだん計算にのみ熱心になってしまい物理を忘れがちになってしまいます。計算すると、関係する実験結果がすぐ出てきて、すぐに比較し、物理的にも正しいから計算データをデータベースに登録しましょうという仕組みになっていないと、溜め込んだデータの信頼性がなくなってしまうことを非常に気にしています。

２－３　話題提供

「機能に基づく材料の設計　ー理論、実験、計算、データの統合への期待ー」
細野秀雄（東京工業大学）

はじめに、「機能に基づいた材料設計」において、私が考えてきたことや思いを述べた後で、ケーススタディとして私たちの研究の話をさせていただきます。私は画期的な物質の誕生は、従来もこれからも、「研究者個人＋用意された偶然」に依存する割合が高いと思います。ただ問題は、自分の時代はいいのですが、今後はこれではまずいだろうと思っています。まず個人の力量の部分、あとは偶然の部分を減らす必要があると考えます。それから合成された物質の数と物性が膨大になって、結晶構造と分子式などの従来の整理要素だけでは、もう見通しが悪くなって仕方がない状況になってきています。これでは、いくら何でももう革新的な次世代の材料設計・探索は無理だというのが 2 番目です。3 番目は、第一原理計算が気楽にできるインフラが揃ったので、実験屋の立場としても、物質選択をするときに、まずラフに第一原理バンド計算を行って、そのバンド構造を見てから実験をやったほうが効率的というのが、いまの研究室の状態です。それから 4 番目は社会的な問題ですが、必要な機能をいろいろな元素の組み合わせで実現しようという要求が高まっていると思います。ディスプロシウム等の元素が足りないことに対する対処療法だけでなく､必要な機能をどういう元素の組合せで実現できるかという課題を解決するために、いろいろな材料創製ルートを開拓する必要があるということです。これは、社会的な非常に強いニーズであろうと思います。

では、何を目指すかということですが、膨大な物質データをより包括的に整理できる高次のコンセプト（たとえば、物質のフェルミ面やそれを記述する数学的概念）を見つけないといけません。そして、そのコンセプトを使った新材料の創出を目指すべきであると考えます。具体的には、研究に関しては、工学的には特にアモルファスは、いくつかの性質を絶対に足し合わせなければならず、材料設計としては、物性値のマトリックスとしての最適化が必要になります。マトリックスとしての材料設計は、工学に関する重要な考え方であると思います。

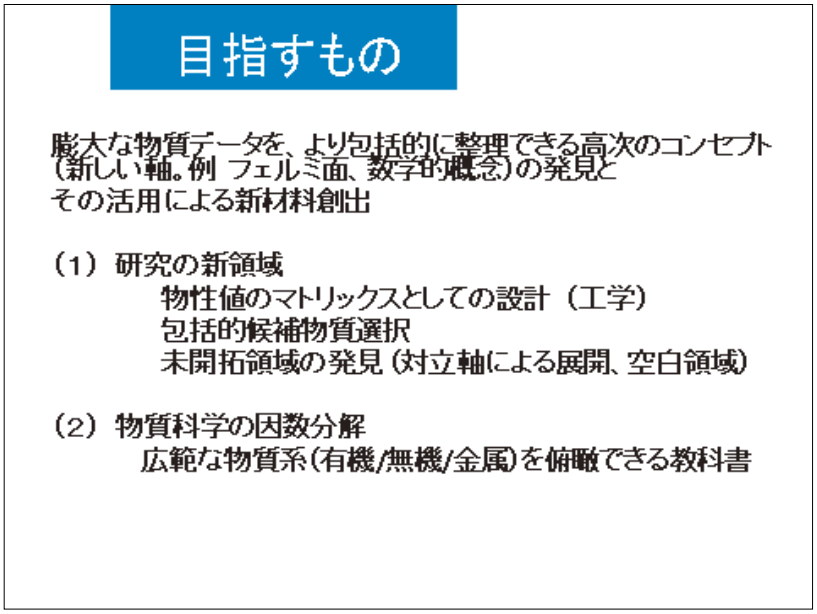


図 1

それから物質を探すときに包括的な候補物質選択。これは、ある程度のイメージでやっていると、データベースを多少は見るのですが、だいたい自分の頭の中にあるものからエイヤッと候補物質を決めてしまっている。後から考えてみると結構抜け落ちている場合もあるのですが、そのような抜けをなくすという意味では、マテリアルインフォマティクスは非常に重要なことだろうと思います。それから未開拓領域の発見。これは何か新しい軸が出ると、その直交軸が必ずできます。そこで新しい展開ができる。それからデータを整理してみると、空白領域があることに気づくと思います。こういうところは、やはり研究領域として、これから非常に可能性のある領域であるはずです。

それから 2 番目は、実は教育的要素が非常に強いのですが、あまりに膨大な物質が蓄積された物質科学を勉強しているうちに「因数分解」して見通し良く物質を捉える力を身につけられるように鍛錬する必要があります。有機、無機、金属という分類でものを見ていくのは、各論としては意味があると思いますが、もっと俯瞰できるような新しいコンセプトを作らなければ、新しい時代の物質科学に入ってくる人がいなくなってしまいます。もう少し、教育というものをサイエンスの意味で見直して、「物質科学の因数分解」を行ってもっと見通しのいい形にすることが、マテリアルインフォマティクスの目指すもう 1 つのものだろうと思います。

ここから少し各論に入ります。なぜ、私がマテリアルインフォマティクスに興味をもったかというと、米国の友人 Alex Zunger が米国エネルギー省（DOE）のファンドをもらったときに、Center for Inverse Design という研究センターを作りました。そのときの概要には、「歴史的に画期的な物質は、偶然に見つかったもの“accidentally-discovered materials”であり、その典型が超伝導体である」と書いてあります。そして、「今までは新しい物質を創り、それらの機能をいろいろと調べていたが、実際に、我々が欲しいのは『物質』ではなく『機能』である。欲しい『機能』を求めるためには、まず必要な電子状態を決め、それに対し実際の原子配置のものを作り、これで新材料あるいは画期的な機能

を持つ物質を設計しよう」と言い出しました。このような逆デザインの概念は、なかなか米国人らしいコンセプトで作られていて、言っていることは新しい装いをとっていますが、私にとっては常識的なことばかりです。

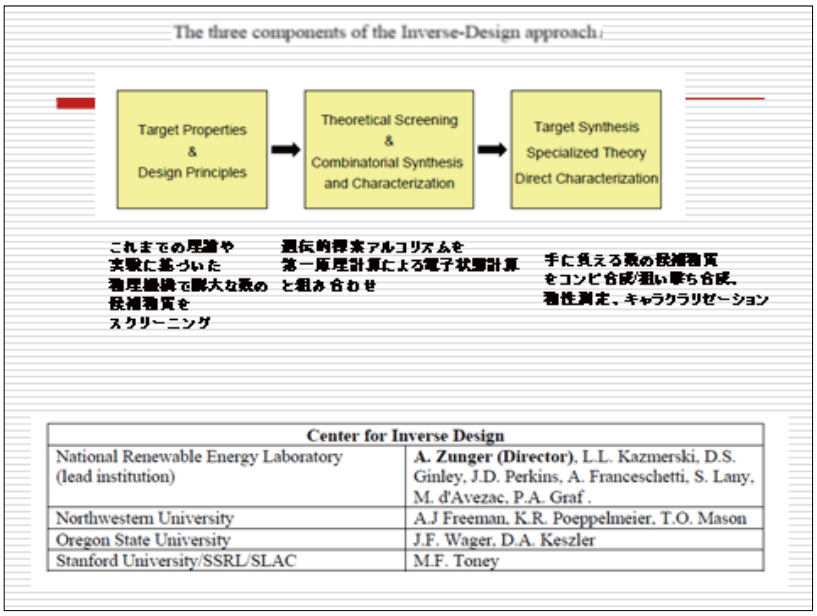


図 2

言い換えれば、物質を実格子空間（実際の原子配置）から決めるのではなく、機能発現に必要な逆格子空間（運動量空間における電子状態）から決めて材料設計をしているということです。実際には、欲しい性質を決めて、常識的な物理や化学でスクリーニングし、いくつかスクリーニングしたものから理論的なスクリーニングをして、選び出した物質群をコンビナトリアル合成法で一度に合成し、評価した結果残ったもの（候補物質）をきちんと研究する。図 2 に示すような、このスキームは非常に当たり前のアプローチですが、理論的な遺伝的探索アルゴリズムの部分に関しては、ある意味で最近の進歩だと思います。逆デザインで最初の実例が出てきたのが、若い Perkins らの“Inverse design approach to hole doping in ternary oxides: Enhancing p-type conductivity in cobalt oxide spinels”というタイトルの論文［Phys. Rev. B 84, 205207（2011）.］で、p 型の酸化物半導体を逆デザインして作ることを目指したものです。いくつかのバンド構造を見て、ドーパントを計算で決めていって、それでだいたいのところを決めて、先ほどのスキームに従って候補物質を探求するわけです。この逆デザインアプローチで出てきたものがスピネル型酸化物 Co_2ZnO_4 であり、一番の代表例になっているのですが情けないと思います。実は、いち早く我々が p 型の酸化物半導体 $ZnRh_2O_4$ を作っており［Appl. Phys. Lett. 80, 1207（2002）.］、彼らは周期律表で Rh（ロジウム）のすぐ上に位置する Co（コバルト）に換えた Zn（亜鉛）スピネル酸化物をやっただけだと言えます。もう少しこれからリファインされてくると思いますが、現状はプリミティブな段階ということだろうと思います。

材料設計という言い方をすると、一番データマイニングに適した物質系は、実はアモルファスなのです。物性値がなめらかに変化するので、多変量解析が非常に有効となります。

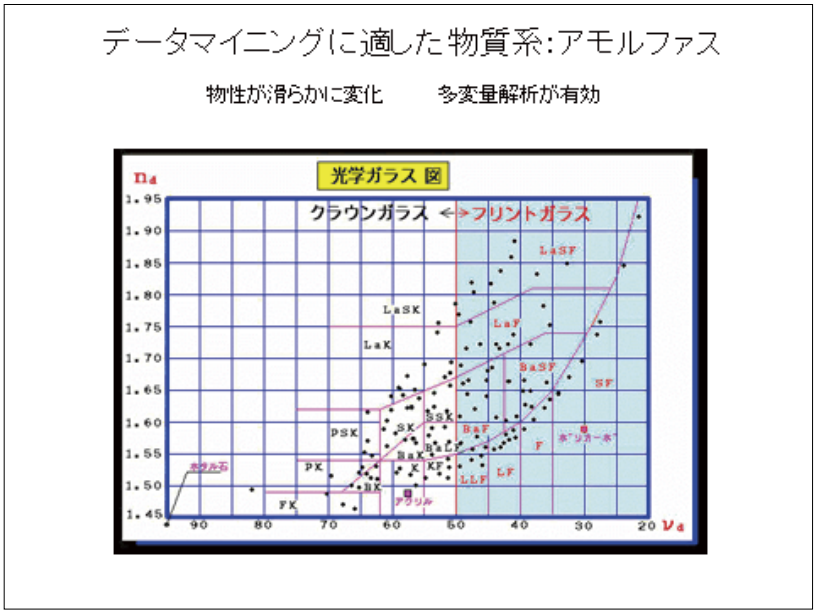


図 3

図 3 には、様々な光学ガラスの物性値として、縦軸に屈折率、横軸に波長分散の逆数をマッピングしています。（※屈折率が大きなレンズほど、光を強く曲げることができ、波長分散の逆数が大きなものほど、色の散らばり具合が小さくなる）マッピングしたプロットは、日本列島のように 45 度傾いた線上に並んでいて、高屈折・高分散ガラス群（クラウンガラス）と低屈折・低分散ガラス群（フリントガラス）に大別できます。したがって、レンズ設計に際しては、欲しいレンズの特性がどれか？（この日本列島的マッピングのどの方向にあるか？）を考え、探索するのが大事だということです。このように、物性値をプロットしてみると、いろいろな相関関係が見えてきます。実はこの背後には、屈折率と密度には単純な物理法則があるのですが、波長分散に関しては、その起源を考えないと、このダイアグラムの意味は理解できません。

これから私たちがやっていることについてお話します。今まで、私たちは計算も活用したり、あるいは新しいマテリアルデザインコンセプトも出しながら、材料を世に送り出してきたつもりです。まずは、IGZO（イグゾー：In-Ga-ZnO 系アモルファス酸化物半導体の略称）。これは、透明酸化物半導体ガラスを使ったトランジスタの研究から出てきて、最近では頻繁にイグゾーとしてテレビコマーシャルで流されています。それから 2 番目は、セメントの構成成分の一つである $12CaO \cdot 7Al_2O_3$（C12A7）に電子を取り込んだエレクトライド。これが、透明金属になって、超伝導体にも変化します。さらに高活性なアンモニア合成触媒として、100 年ぶりにハーバー・ボッシュ法を超えられる可能性が出てきて、今かなり一生懸命やっているとところです。それから鉄系超伝導。この 3 つについて、どんなことを考えてきたかという話をさせていただこうと思います。

私の研究の好きなやり方を、将棋の駒に例えるならば、「桂馬」（全ての駒の中で唯一他の駒を飛び越えて敵陣へ進撃できる駒）と、敵陣でひっくり返った「成桂」（桂馬の成駒）です。私は、「歩」のように、一歩、一歩進むのが嫌いで、そのような研究アプローチは

取りません。「桂馬」のように少し斜めに他の駒を飛び越え、敵陣で「成桂」になり「金将」として戦いを制する。これが私の一番好きな研究のやりかたです。これでうまくいく場合はいいのですが、うまくいかないことが実は出てきます。

図 4 には、2006 年の LaFePO（T_c=4K）と 2008 年の LaFeAs（O,F）（T_c=26K）を初めとする鉄系超伝導体の進化を示しています。LaFeAs（O,F）は、リン（P）をヒ素（As）に変えてフッ素（F）ドープしたら T_c が高くなったということなのですが、この発見後、実は大変痛い目にあいました。ランタン（La）系（LaFePO、LaFeAsO）で超伝導を見つけ、T_c が高いものが出てきて、それをいろいろな方法で T_c を高くしなければいけないわけですが、周期律表で La の隣のセリウム（Ce）ーサマリウム（Sm）等の同じ希土類元素で全置換した SmFeAs（O,F）等の物質を合成する戦術は、将棋の「歩」のようなので、取りませんでした。もっと「桂馬」のように飛んで、ルテチウム（Lu）でやるか、下のアンチモン（Sb）でやるかということで、研究グループのメンバーにお願いしたのですが、ことごとく失敗し、物質すら未だにできない状況です。ところが、周期律表で La の隣でやったら T_c がいきなり 43K、その隣でやったら、55K になったという「歩」のような研究報告が…。頭の中で全部ストーリーを書いて「これでやれ」と指示をしたものなのですが、ホールドープ、電子ドープの方向が逆であるなどの理由で、T_c を飛躍的に上げる戦にはすべて負けたわけです。その敗因は、私の考え違いでした。どういうことかというと、半導体のセンスで超伝導体の研究をやってしまったことです。隣の元素にちょっと変えたぐらいで、とくに La を他の希土類でちょっと変えたぐらいで T_c など変わらないだろうと思ってしまったわけです。実のところ、超伝導というのは、もっとエネルギースケールの極めて小さいところだという認識がなく、そのために指示を間違ってしまった。結果的に、隣をきちっとやったグループが合成した物質の T_c が高くなったという非常に痛い思いをしています。

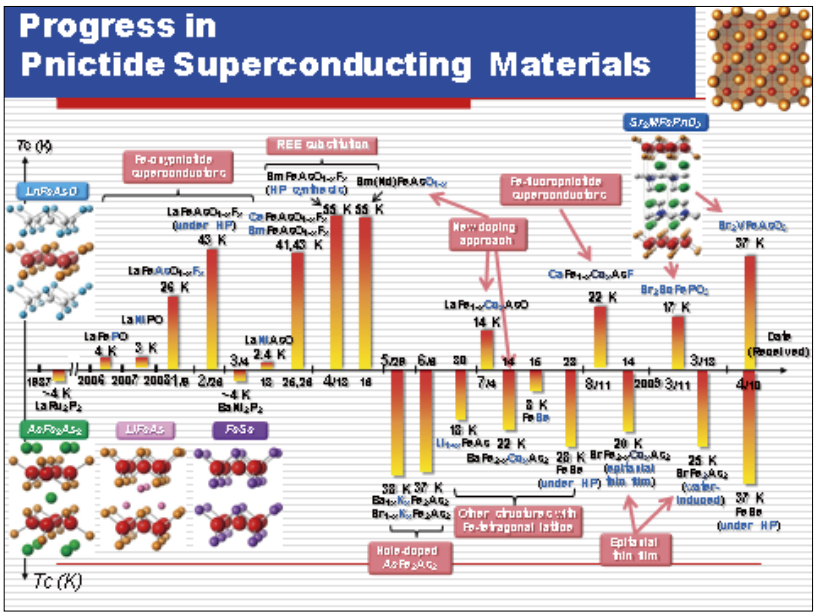


図 4

今日これから、計算でもこの系では T_c の予測はできないということをお話しします。

実のところ、超伝導になるかどうかすら予測できていないのです。図 5 に示すように、鉄系超伝導体は共通の構造である鉄の平面格子を持っていて、フェルミ面に絡んでいるのは、ほとんど鉄の 3d 軌道であることが分かっています。鉄系超伝導体にもいろいろな母相があり、いま 50 種類ぐらいの母物質が存在しています。

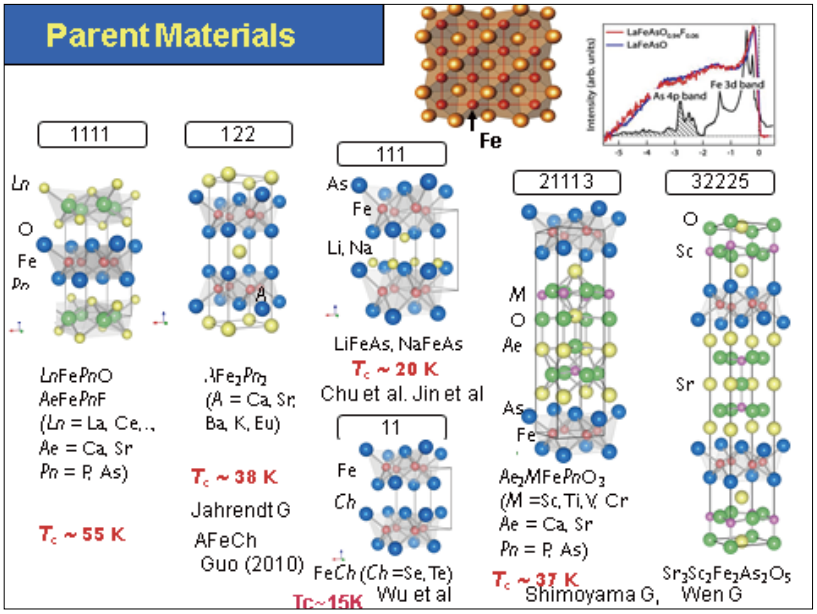


図 5

先ほど寺倉先生も示された Lee プロット（T_c と $FeAs_4$ 四面体の頂角 α の相関図）によれば、正四面体（頂角 α =109.47 度）のとき T_c が一番高くなっています。これに対する理論的解釈はどうなっているか、スピン揺らぎに基づく黒木（電通大）らの仕事についてお話します。Fe の 3d 軌道のスピン帯磁率が大きくなり、そこでスピン揺らぎが強くなってスピン揺らぎを媒介とした超伝導が発現します。フェルミ・ポケットの構造は、FeAs 結合角が小さくなるにつれて変化し、3 枚（最大枚数）のホール面を持つ正四面体の角度近傍において T_c が極大となるということです。第一原理バンド計算により、ホール面と電子面から構成されるフェルミ面を求め、T_c に対して実際に意味があるのが、$FeAs_4$ 四面体の頂角ではなく、ヒ素の鉄平面に対する高さであるということを彼らは計算から導き出しています。そうすると、ヒ素の鉄平面に対する高さが低い場合には、電子面が消えてしまいますので、ホール面と電子面のポケット間に働く強いネスティングが起こらなくなってしまうわけです。ですからネスティングベクトルが具合よく書けるか、書けないかということで、T_c の物理的解釈がなされています。これだけ見ると非常に完成されたように見え、新しい超伝導体はどんどん見つかるはずだと思って、実験家は頑張るわけです。

図 5 に示す超伝導母物質の中に、111 系と呼ばれる LiFeAs 等の物質があります。これは、形式価数が $Li^{+}Fe^{2+}As^{3-}$ で、ドープしなくても T_c が 20K ぐらい出ます。これは、我々のところで昨年 PRB に書いた論文ですが、先ほどの LiFeAs とほとんど同じフェルミ面を持っている物質を、第一原理バンド計算によりフェルミ面を出しながら探索したわけです。MgFeGe は、超伝導体 LiFeAs の等電子物質であり、構造的なものも四面体の要素に比べると、四面体の角度がそれほど悪くない物質です。実際にこの計算をしてみると、非磁性

状態に比べて、フェルミ準位における状態密度（DOS）へのかかり方が、反強磁性状態の方が安定化します。超伝導体 LiFeAs と等電子物質 MgFeGe の電子状態を比較すると、バンド分散、状態密度、フェルミ面でかなりの類似が確認できます。パッと見ると、そっくりではないかと思うわけです。ところが実際に試料を合成して測定してみると、MgFeGe は超伝導にはならないのです。フェルミ面を見ても、円筒状のフェルミ面が立っており、真ん中にホール面があって、電子面がエッジにあるということなので、非常に似ています。しかしながら、片方は超伝導になり、もう片方は超伝導にならない。違いがどこにあるかというと、ホール面の数ぐらいしか違わないわけです。これはほんの一例なのですが、我々は実験をする前にこれぞと思われる候補物質については、第一原理バンド計算によりフェルミ面ぐらいもちろん書いているわけです。それで一生懸命に候補物質を探すわけですが、当たる確率は極めて少ない。いわんや T_c の高い物質探索に至っては、ほとんど現状の計算ではホープレスに近いですね。計算から出てくるものは、現状ではとても T_c の高いものは予想できないだろうと考えています。実験から出てきた物質の解釈で精一杯だろうと私は思います。

最近我々のところが発表した 2 次元エレクトライドについて話します。エレクトライドというのは、電子が陰イオンの役割を電子が担う物質ですが、これまで電子が存在する場所はケージ構造（0 次元）に限られていましたので、コンセプトを拡張し、電子が 2 次元の隙間に存在する物質を探したわけです。エレクトライドは 1983 年に有機物で最初に見つかって、安定なものが 1 つもなかったため、2003 年ぐらいに我々のところでセメントの化合物を使ってケージの中に電子を入れることによって C12A7 エレクトライド $[Ca_{24}Al_{28}O_{64}]^{4+}\cdot(e^-)_4$ を作ったわけです。2 次元で何を考えているかというと、2 次元電子ガス（2DEG: 2-Dimensional Electron Gas）の結晶を作りたいわけです。GaAs と $Al_xGa_{1-x}As$ の 2DEG 界面、これはケミカルポテンシャルが違いますから、ここにたとえば Si をドープすれば、ドープされた Si から出てきたドナーが界面に溜まるわけです。ドーパントと出てきた電子の空間的位置が違うので、イオン化不純物散乱が起こりません。これが実は高電子移動度トランジスタ（HEMT: High Electron Mobility Transistor）の原理であるわけです。たとえば酸化物では、Harold Hwang がやられたペロブスカイト系 $LaAlO_3$（LAO）-$SrTiO_3$（STO）界面でも 2DEG が出ると、自然にあるもので 2DEG の結晶があるのではないか？それがおそらく 2 次元エレクトライドであろうということで、ものを探すわけです。

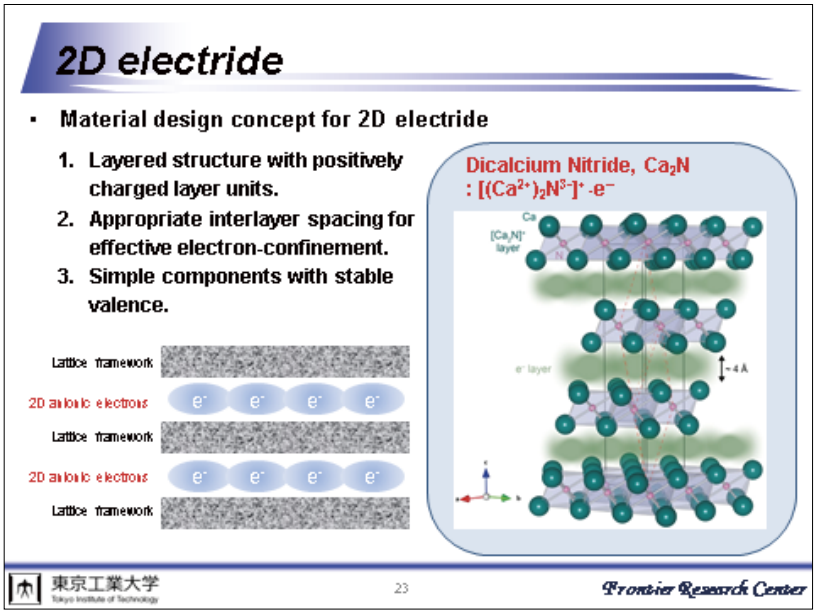


図 6

どういうデザインコンセプトでいったかを図 6 に示します。1 番目は、層状構造で、層が正に帯電していなければいけません。2 番目は、層と層との間隔が、今までエレクトライドになったものと同程度（約 4 ー 5Å）の大きさのものを探します。それから 3 番目は、層の原子価が変わっては駄目なので、層を構成する元素はご法度の（原子価が変わりやすい）遷移金属ではなく、ある意味でありきたりの典型元素が一番良く、そういうものからできている結晶を探します。その結果、いくつか候補を見たのですが、一番単純なものでいこうということで、Ca^{2+} と N^{3-}から構成されている層状ダイカルシウムナイトライド（Ca_2N）を選びました。この形式価数は（Ca^{2+}）$_2$ N^{3-}ですから、層は［Ca_2N］$^+$ となって電気的に電子が 1 個合わず、これから見る限り隙間に電子が 1 個存在する可能性があると考えました。この物質の結晶構造は既知でしたが、試料に問題が多く、物性を正しく評価ができていなかったので、単結晶を合成して、物性を正確に測っていくわけです。もちろん、その前に、同時に第一原理バンド計算もしていきます。計算をすると、フェルミ面はきちんと 2 次元の円筒状になっており、電子密度を計算すると、きちんと層間に電子があるような形になります。そしてワニエ関数を使って計算してみると、きちんと層間に電子があることが分かるということで、Ca_2N でやればエレクトライドができるのではないかとなるわけです。

単結晶 Ca_2N の輸送特性を測ると、不純物ドープしたシリコン半導体とは違う挙動を示します。例えば、シリコンにリンをドープした場合などがそうですが、温度を下げていくと、まずフォノンの影響で移動度が上がってくるのですが、あるところからイオン化不純物散乱が効いてくるので、逆に移動度が下がり、その結果移動度は上に凸のカーブを描きます。ところが、この物質は、フォノン散乱が 100K くらいまでは効いているのですが、それより低温域になると、電気抵抗率や移動度の値がずっと一定になってしまいます。イオン化不純物散乱がまったく見えないのです。それから、これのホール効果で電子濃度を測ると、1.4×10^{22} /cm^3、これは化学量論比組成：［Ca_2N］$^+$・（e^-）で合わせた予想される電子の量とほとんど合います。

それから、低温で、電子濃度 n が $10^{22}/cm^3$ もあって、移動度 μ が 520 $cm^2V^{-1}s^{-1}$ もいくということは、普通の状態では考えられないわけです。今まで知られている 2DEG の結晶、GaAs-$Al_xGa_{1-x}As$、LAO-STO、グラフェン、と Ca_2N を比較してみますと、キャリア濃度が桁違いに大きいわけです。移動度 μ は、キャリア濃度 n が大きい割に、大きく保っていると。両方の値を掛算して電荷の値（e）をかけると、これが実は電気伝導度 σ （=ne μ）なのですが、一番この中では Ca_2N の電気伝導度が高くなります。こういうものが、グラフェンの場合は 1 枚の炭素シートで起こり、また他の場合は界面だけで起こりますが、Ca_2N の場合はバルク結晶で得られるということです。そういう物質がきちんとあるということです。磁気抵抗をとってみると、層に平行の方向に磁場をかけると抵抗率が増加するのに対して、層に垂直の方向に磁場をかけると、抵抗率が減少し（負の磁気抵抗）、2DEG として期待通りの著しい異方性が観測されました。

最後の例は、アモルファス半導体です。今日おられる田中一宜さんがちょうどアモルファス半導体の光構造変化をやっている頃で、私は若い頃にガラスの研究をしていたので、私も田中さんの仕事に憧れて、将来はどこかで絶対に酸化物のアモルファス半導体をやりたいという思いがあり、1993 年ぐらいから始めたのです。

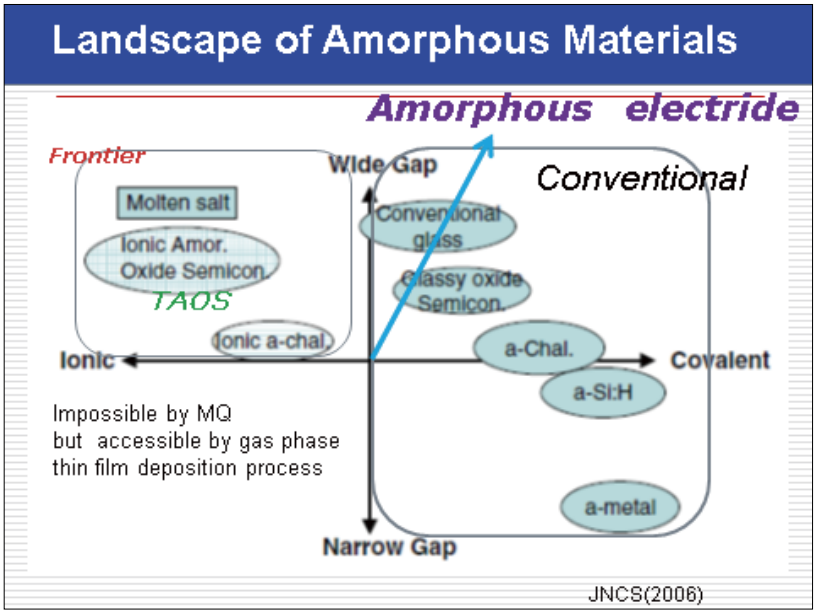


図 7

そのとき使ったコンセプトは非常に簡単で、イオン結晶性酸化物を使って、共有結合性のアモルファス半導体ではできないものをやってみようということで、実は空間的な拡がりの大きい球対称な s 軌道を最低非占有準位として持つ In や Cd のような重金属イオンを使います。これらの酸化物半導体では、伝導帯端が重金属の空間的な拡がりの大きい球対称な s 軌道で構成されているために、アモルファス構造の乱雑性によっても s 軌道の重なりは大きな影響を受けにくく、比較的高い移動度を実現できる系があるはずだということで、透明アモルファス酸化物半導体（TAOS: Transparent Amorphous Oxide Semiconductor）を作ったわけです。それが実用になったわけですが、それをやるときのコンセプトは、図 7 に示すように、横軸に結合性（共有結合、イオン結合）をとって、縦

軸として透明、不透明をとります。すると、今までのアモルファスは、皆共有結合性側なのですね。イオン結晶性の酸化物は、通常ではアモルファスにならないのですが、スパッタ法を使うと、普通の汎用の合成法ですが、冷却速度が非常に大きいので、いろいろなものがアモルファスになります。結晶をやっている人にとっては出来そこないのアモルファスだったわけですが、その中でアモルファスシリコンの 20 倍の移動度のものが出てきて、それが高精彩で省エネルギーのディスプレイに、要求と合って使われ出したというのが本当のところです。

次は、図 7 に示す平面に対して垂直の方向に展開したい訳です。今までのアモルファス半導体のように、全部電子軌道で電子の状態を書くのではなく、どこの軌道にも属していない電子が界面にあるような半導体ができないかということを考える訳です。それが、最近凝っているアモルファスエレクトライドです。なぜこんなことを考えたかというと、先ほどの C12A7 エレクトライドを高温、1600℃で溶かしているのですが、なんとフェニル基を制御して溶かすと、融点になっても、実はまだ若干電気伝導度は落ちますが、液体金属の状態ができていることを発見したからです。C12A7 は単なる酸化カルシウムと酸化アルミニウムからなる物質なのに、電子が $10^{21}/cm^3$ 入っている液体金属が 1600℃で出来たのです。きちんと色もついています。普通の電子が入っていないカルシウムアルミネートは、当然のことながらイオン伝導で溶けます。だったら、これをクエンチすれば、アモルファスエレクトライドができるのではないかということで、昨年の International Display Workshops（IDW）で発表しましたが、（有機物よりも電子輸送層として非常にいいだろうと思っている）非常に透明で $10^{21}/cm^3$ の電子が入ったアモルファスエレクトライドが実際に出来ました。

これは、まだ海のものとも山のものともわかりませんが、全部世の中で言われていることの反対、対立軸に従って、マテリアルデザインしたものです。共有結合からイオン結合に持っていき、次は原子に属している軌道から、今度は逆にどこの軌道にも属さないアモルファスを作ろうということで、2 回の対立軸に従ってアモルファスエレクトライドが作られたわけです。

これから物性開拓をしていかなければいけないのですが、やはり新しいコンセプトが出てくると、それに対する対立概念が出てきて、それで新しい道が開けると思います。そういう意味で、データマイニングあるいは物質設計というところから期待するのは、とくに理論家の方々、あるいはデータを解析する方々に期待するのは、多少粗くてもいいので、わかりやすいコンセプトを提出していただくことです。

【質疑応答】

質問者：今日の主題のデータマイニングという観点から質問します。いろいろあったお話を整理すると、細野先生にとっては、脇にある膨大なデータを一体どういうふうに使われているのですか。

細野：実際には、発表ではうまくいったところだけを示していますが、その陰には膨大な失敗があるわけです。先ほどの物質の、失敗例をきちんとした論文にしたのは MgFeGe ぐらいなのですが、結晶のデータベースを見て、有力候補についてはフェルミ面をいちおう計算しているのです。一番単純な方法ですが、いつも無機結晶構造データベース ICSD

（Inorganic Crystal Structure Database）を眺めているわけですよ。そうすると、ああ、世の中にはこういう物質があるのだということが、結構あるのです。例えば、我々はハイドライドで超伝導体のドーピングをやったわけですが、最初に気づいたのは、希土類は3価ですが、希土類ハイドライドは全部 LaH_2、SmH_2 の形で存在することでした。ということは、ああいう状態の中では、もしかすると面白い状態があるかもしれないと考えました。SmH_2 のサマリウム（Sm）の価数は3価ではなく2価であり、原子価が単純に変わりやすくないものが変わるような系があるのですね。

質問者：確かにそういうものがあるとわかってしまえばそうかもしれません。そこに行きつくところが、普通の人にとっては大変。そこには、今日の主題であるデータマイニングという膨大なデータから何か必要なものを引っ張り出すという難しい課題が…。

細野：でも、それは意図が無ければ引っ張り出せません。直観というのは極端にいえば偏見ということです。あっている、あっていないは関係なく、どこかに踏み出すための確信。それがないと踏み出せないのですよ。動いているうちに軌道修正ができるので、そういう意味ではコンビナトリアル合成化学が出たときに、1年で1世紀の仕事が可能という意見がありましたが、あれから10年経ったのですが、材料の分野で何が出て来たのでしょうか。

質問者：細野先生がこのインフォマティクスに関心をもたれて、注意を払われた理由は結局、何なのですか。

細野：結局、そういうところでデータマイニングという方法も道具の1つにしておかないと、ものを探すときに我々のように「黙って睨めばピタリとわかる」というのは、これはどう見たって源平盛衰記の頃の戦法です。やはり近代戦にするためには、もう少しシステマティックなことを考えなければいけません。もちろん最後にやるのは人間だと思いますが、源平盛衰記的戦法はさすがにもう、まずいでしょう。

質問者：細野先生自身は、今日は素晴らしいお話、結論だけを仰ったけれども、そこに至るときに、やはりいろいろとインフォマティクスを使ったということですね。

細野：インフォマティクスというかどうかは別として、データを眺めないで出てくることは、ほとんどありません。

質問者:1つを見つけると他の回答を探す意欲が激減する。これは、その通りだと思います。ですからコンピュータが必要ですが、コンピュータの特徴は網羅性ですよね。では、いかに網羅させるかというのに逆解析があります。では、逆解析をするためには何が必要かというと、モデルです。それはたとえば統計モデルかもしれません。では、データがたくさんあるようなことを皆さん思っているかもしれませんが、モデルをつくるために必要なデータは最初少ないかもしれないのですね。先端的な研究ほどそうですよね。少ないデータで、次にどういうふうに実験をして、目的に到達していくのか。これもインフォマティクスの課題だと思います。私自身は材料設計や分子設計、ケモインフォマティクスの研究を30年やってきましたが、痛切に感じているのです。

あと最後の先生のお話は、情報、要するに設計のための提案をしてくれるのはいいのだけれども、欲しいのは具体的にどういう構造で、どういう操作を行えばいいのかといったことにつながるような情報を、逆解析等を通して示してほしいという話だと思うのです。そのためにどういう必要があるかというと、データ、構造をそれにつながるように記述しなければならないわけです。パラメータ、要するに記述子の問題ですよね。おそらくその

辺りの考え方があやふやなので、こんな情報では、どんなふうに実験していいか、さっぱりわかりませんということになりがちなのですね。ですから、データをどういう記述子で表現するのか。それは理論、計算から出てくるパラメータでもいいですし、単純に温度、圧力という操作条件でもいいのですが､それらが一緒にセットになって提案されなければ､どっちに向いていいのかわからないのですね。そういったことをおそらく企業も含めて、インフォマティクスからの提案として求めているのだと思います。

細野：私たちは典型的に全然「網羅」できていないのです。1つ見つけると、もうだいたいこんなところでいいやと、すぐに他へ行ってしまうわけです。網羅するというのは、つまらない仕事なのですね。でも、実際に産業活動を考えると、それこそ重要ですよね。

質問者：人が得意なことと、計算機が得意なことをきっちり分けて、計算機が得意なことは何なのか。それを計算機にさせる。たとえば網羅させるなど、そういうことをしていく中で、網羅させるといってもパラメータにはそれぞれ範囲がありますから、馬鹿みたいな網羅はしないのですね。制約をかけると、当然ある程度の範囲に収まるかもしれないと思います。

「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）

今日は、データ科学によるマテリアルインフォマティクスへのアプローチに関して発表いたします。第 4 のパラダイムとしてデータ中心科学が最近注目を浴びており、これは経験科学・理論科学・計算科学に代わる 21 世紀の科学といわれています。従来は直感・経験に基づいて仮説生成を行い、それをデータに基づいて検証していたものを、これからはデータからの知識発見による仮説生成を行い、実験やより詳細なデータによる検証を行うという科学です。このような試みはいろいろな分野でなされていて、とくに生物学ではバイオインフォマティクスという形で長いこと行われてきました。最近、「マテリアルゲノムイニシアティブ」という構想が米国で始まり、材料分野でも非常に注目を浴びています。

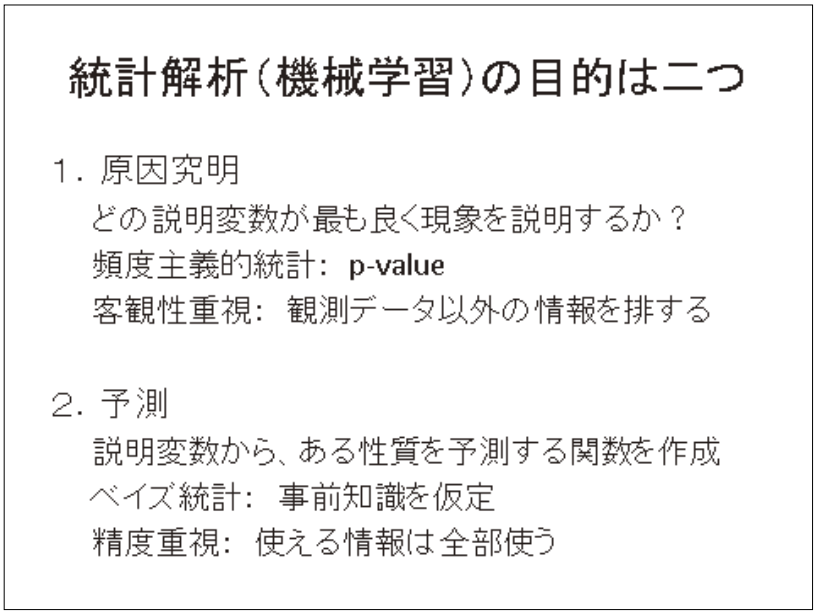


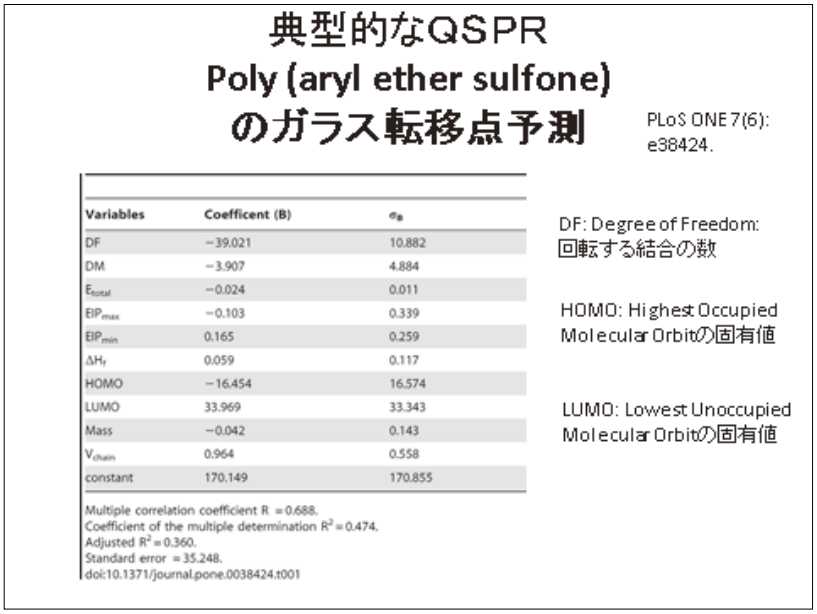
図 1

統計解析の目的は 2 つあります。1 つは原因究明、もう 1 つが予測です。原因究明は、いくつか説明変数のある現象があり、多数ある説明変数の中から、どれが効いているかを見つけ出す。原因究明を行うには頻度主義的統計、p-value などをよく使います。原因究明に関しては客観性重視ですから、観測データ以外の情報をできるだけ排します。一方、予測は説明変数からある性質を予測する関数を作成するのが目的です。たとえば、新しい物質の物性を予測するとき、その物性をできるだけ正しく予測できる関数を追求するのが予測です。そのためには、たとえばベイズ統計のような事前知識を仮定するものを使ってもよく、精度を重視するため、使える情報は全部使う。とにかく訳がわからなくなっても使うという立場があります。原因究明と予測という 2 つの目的を混同すると訳がわからなくなるので、区別して考える必要があります。

まず原因究明からお話しします。化学構造が与えられたときにその物性を説明する変数を探し出すというのは、構造・物性相関（QSPR：Quantitative Structure-Property

Relationships）と呼ばれるもので、この方法では物性をいくつかの説明変数、説明変数には記述子というものがあって、記述子に関することを説明するというものです。つまり、説明変数が n 個あり、それが何らかの関数になっていて、その関数のアウトプットが物性となります。QSPR では物性と説明変数の値のペアでわかっているものがたとえば 100 組あったとき、基本的には線形回帰分析を行います。ですから各説明変数に重み係数ベクトル W1 ー Wn をつけ、W の絶対値が大きいものがより物性に関与しているとするわけです。これが、典型的な特徴選択の方法です。基本的に、できる限り少数の変数を選びたいわけです。最終的に理解を目指しているので、全体に重みベクトルを変えてしまうと訳がわからなくなるので、それはあまり歓迎されない。これはなぜかというと、科学理論はシンプルでなければならないとする「オッカムの剃刀」という考え方があり、、1 個や 2 個のほんの少数の説明変数に依存するようなものを見つけ出せれば、これは非常に正しいということになります。アインシュタインが言ったように “When the solution is simple, God is answering.” ということで、できるだけ少数のものが見つかると嬉しい。

図 2 は典型的な QSPR の論文をコピーしたもので、ポリアリルエーテルスルフォンのガラス転移点予測の問題を解いています。



Variables	Coefficent (B)	σ_B
DF	−39.021	10.882
DM	−3.907	4.884
E_{total}	−0.024	0.011
EIP_{max}	−0.103	0.339
EIP_{min}	0.165	0.259
ΔH_f	0.059	0.117
HOMO	−16.454	16.574
LUMO	33.969	33.343
Mass	−0.042	0.143
V_{chain}	0.964	0.558
constant	170.149	170.855

Multiple correlation coefficient R = 0.688.
Coefficient of the multiple determination R^2 = 0.474.
Adjusted R^2 = 0.360.
Standard error = 35.248.
doi:10.1371/journal.pone.0038424.t001

図 2

ガラス転移点の物理的な基礎はまだわかっていないらしく、未解決問題の中に含まれているようですが、QSPR の中では非常に予測精度が高く出るということで、何本も論文が出ています。これを見ていくと、「Variables」の欄にあるのが説明変数で、10 個ほどある中で線型回帰分析をしてみると、DF、HOMO、LUMO にウエイトが集中しています。DF が回転する結合の数、HOMO は Highest Occupied Molecular Orbit の固有値で、量子力学的な量です。これをもとに、この現象に関しては、DF、HOMO、LUMO が効いていて、それをプロットしてどうなっているか見てみようというのが、典型的な QSPR です。

先ほど例では 10 個ぐらいしか説明変数、記述子がなかったのですが、いろいろなものが考えられます。記述子を計算するプログラムはいろいろ出回っていますが、図 3 は

CODESSA という商用のもので、米国企業の製品です。マテリアルサイエンスに使われているありとあらゆる記述子を計算できるプログラムです。

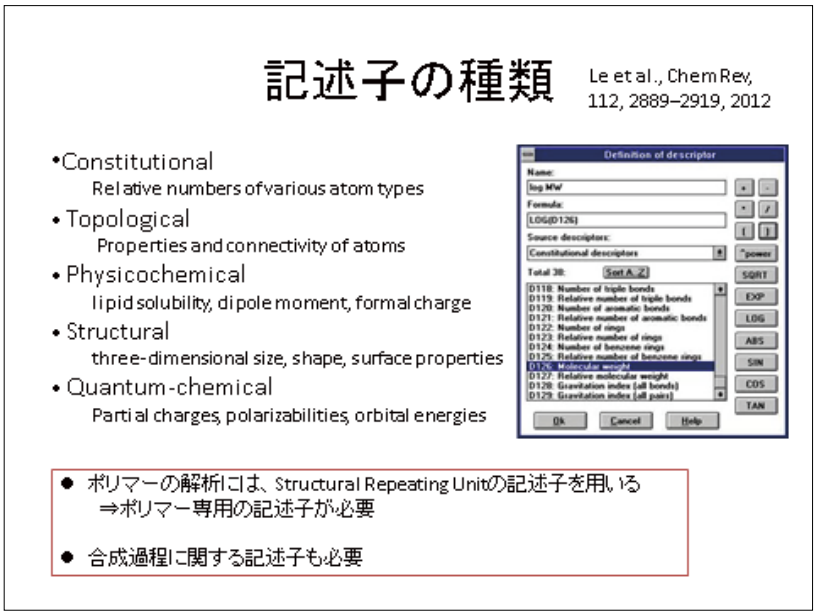


図 3

記述子の種類は大きく分けて 5 つあり、Constitutional というのが元素の相対的な割合です。Topological というのがコネクティビティに関するもので、化学構造式をグラフで表したときに何と何がつながっているか、その頻度など、といったものです。エネルギーに関連する Physicochemical なものもあります。Structural は、結晶構造などがわかっていると三次元座標がついていますから、その三次元座標に基づく特徴量です。最後に Quantum-chemical は量子的な特徴量、記述子となっています。CODESSA を使うだけでも何千という記述子が得られるわけです。ただ、これでもう研究が終わりではなく、ポリマーの解析には 1 単位である Structural Repeating Unit の記述子を用いることが多いようですが、当然ポリマーは全体につながっていますから、それを 1 つのユニットの記述子を用いてやるというのはかなり乱暴であったり、たとえば合成過程に関する記述子も必要であったりするので、まだ研究する余地があるし、これが進んで行くと、さらに知識発見の可能性も増えてくるということです。

一方、今までの QSAR の論文などを見ていると、それほど大したものはなく、まだすごく小規模です。少数のデータ、少数の説明変数しかない。私はあまり材料はやったことはないですが、いろいろなデータ解析のタスクをやっていると、データ解析はすぐできるのですが、データを取得するほうが問題になります。どうしてもデータ解析はトライアルアンドエラーなので、いろいろな記述子を集めてやってみて、いいのが出なかったのでまだやってみる、一発で終わることはまずありません。ですから生産性を上げるためには多数の説明変数を、インターネットを通して瞬時に用意できるというのが夢であるわけです。今それが実現できているわけではありません。とくに自分が実験して得たデータと、パブリックなデータを混ぜて統合して解析したいわけです。

そのために、最近 Open linked data という流れがあり、データベースを用意するとき

に統合可能なように整備しようという動きがあります。これは RDF によるデータの表現ということになります。RDF（Resource Description Framework）は、2007 年に W3C というウェブのテクノロジーを決める標準化団体によって標準化されています。発想は簡単で、通常データベースは表であり、データベースと銘打たれていても単にエクセルシートであったりします。そういうものではなく、たとえば「エタノールの沸点は 78.37℃」ということを図 4 のように主語、述語、目的語の組み合わせで表現するのです。だから 1 つの表に 10 × 10 のデータがあればもう 100 個、右のトリプルと呼ばれるものができるわけです。

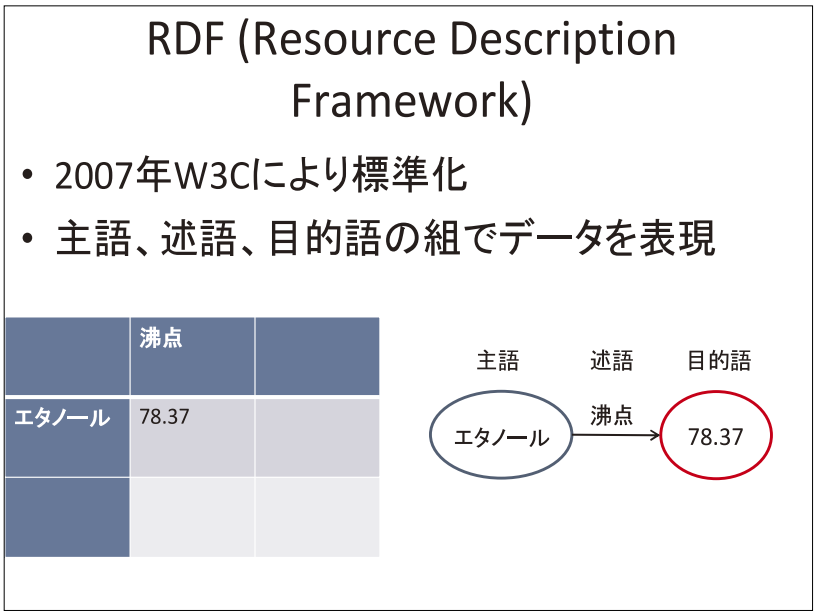


図 4

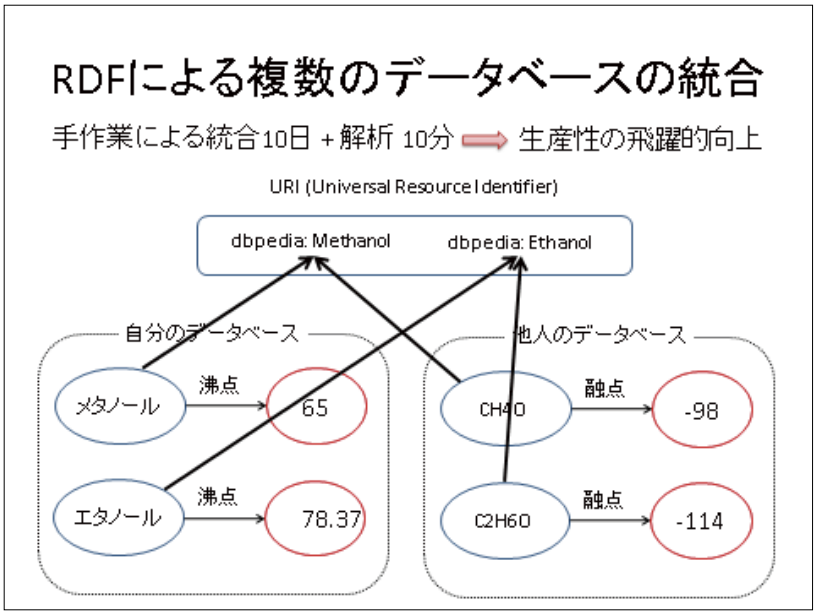


図 5

RDF では、各主語、述語、目的語に URI（Universal Resource Identifier）をつけることが義務づけられています。たとえばメタノールであれば、メタノールとカタカナで書いていても、CH4O と書いていても、同じところにリンクが貼ってあるということで、同じものだとわかる。そういうフレームワークになっています。もちろんオントロジーなどいろいろなものがあるので、本当はもっと複雑ですが、基本はこういうことです。最近いろいろなデータベースが RDF 化されていて、それを一気に統合して解析できるようになっています。

データ科学をいろいろな科学分野に導入するときに、やはりいろいろな反発があります。なぜかというと、機械学習でたとえば現象をうまく分類・説明・予測できるようになったときに、我々は何を理解したといえるのか、結局何もわかっていないではないか、そういう疑問があるわけです。たとえば昨年、CMU の哲学専攻で「オッカムの剃刀」ワークショップがあり、データ・複雑性・真実、つまりデータがあり、どんなデータであれば真実が出せるのか、複雑性が少なければ出せるけれども、複雑性とは何ぞや、といった議論がなされています。ただ前提条件としてアインシュタインの時代とは違い、我々に残されているのは複雑な現象のみなのです。還元論でどんどん分割していって、1 つずつ微分方程式で表してもいいし、それをきちんとやるべきなのですが、本当にそれですべての現象が理解できるのか。我々の立場は、Vapnik という機械学習の創始者のひとりが言っていることに近く、それは、「アインシュタインは法則がシンプルであれば発見できると言ったが、実世界とは何だ？それは本来、単純なのか、複雑なのか？機械学習を用いれば、単純世界とは異なる立場から複雑世界にアプローチできるのだ」ということです。つまり複雑世界においては、予測可能性のために説明可能性を諦めなければならない。こういう考え方を道具主義といって、完全に理解できずとも精度の高いモデルには実用上の価値があるということです。

ここからが予測の話ですが、機械学習ではデータから対象の性質を予測できる。マテリアルゲノム計画では新規材料設計のコストを削減するといっています。予測をどのようにしてコスト削減につなげるか。ここで、クリギングというテクニックをご紹介します。これは Krige によって 1951 年に考案されたものです。ベイズ推定を用いて未知の関数の最大値をできるだけ少ない観測で発見するということで、探鉱などに利用例があります。ある土地があって、どこに金鉱があるかをできるだけ少ないボーリング調査によって発見するというやり方です。

マテリアルのデータを用いた計算実験の結果を紹介します。説明変数を変えながらいくつか観測を行ない、説明変数がある値をとったとき、観測値が最高値をとることが分かっているときに、観測値が最大になるポイントを見つけ出したいとしたら、次にどこを観測するかという問いに答えるのが実験計画です。それに答えるために、まずベイズ推定を行うと、図 6 のように予測関数が得られますが、予測関数の確からしさを表す予測分布も得られます。そしてクリギングで次の実験点をどうやって選ぶかというと、これまでの最高値があるときに、それを超える確率、つまり今図 6 のように予測したわけですが、この予測した分布に基づいて、これまでの最高値を超える確率が一番高いところを探します。

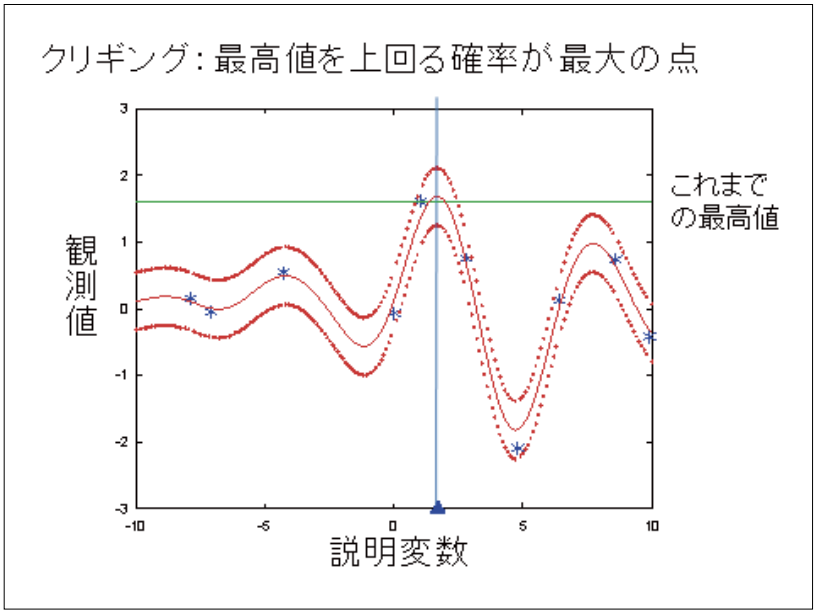


図 6

そうすると、図 6 で▲で示したところを選ぶということになります。これをどんどん繰り返していくわけですが、なぜこれが働くかというと、観測を繰り返してデータリッチになってくると、予測分布がシュリンクして狭くなるので、結局説明変数の他のところのほうが最高値を超える確率が高くなるのです。そういう形でうまく同じ説明変数領域に集中することなく、観測ができます。

計算実験ですが、ここでは 226 ある材料の中から融点が最高のものを発見するという実験を行いました。基本的には 5%をランダムに選んで融点を観測します。その後、クリギングを用いて観測順を自動的に決定していきます。ここで使った説明変数は 17 個あります。

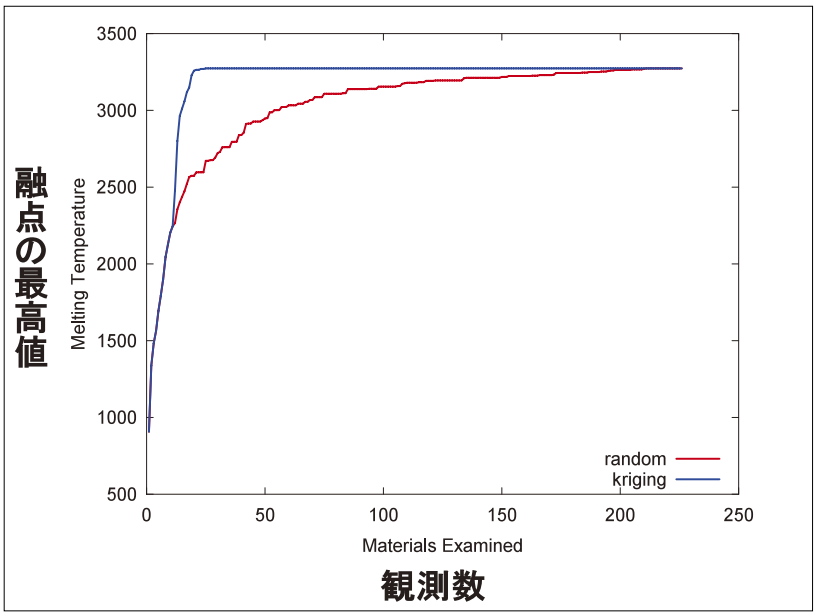


図 7

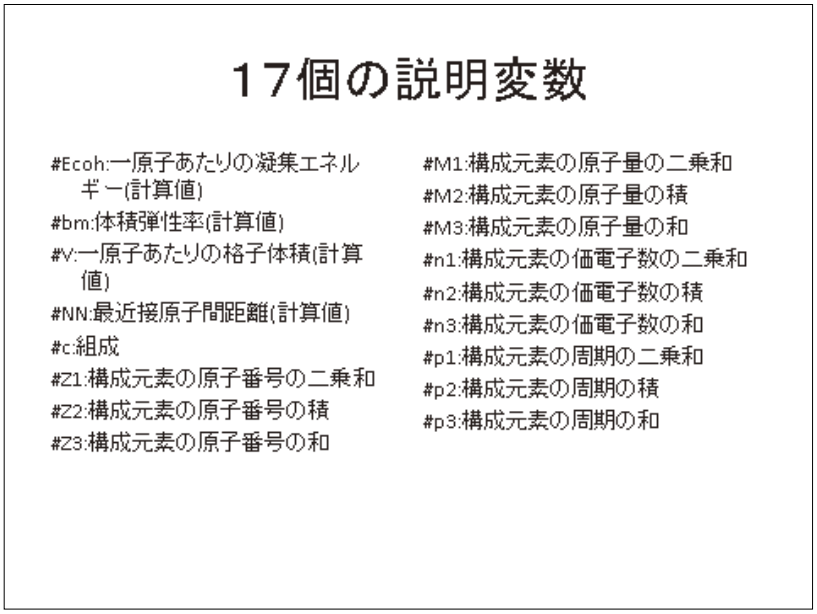


図 8

観測数とその中で得られた融点の最高値の関係を示したのが図 7 です。5%まではランダムに観測しています。そこからクリギングで実験を計画すると、一気に上がります。つまりランダムに比べると一気に上がり、融点が一番高いものを探すにあたって、より少ない観測で探し出せています。最高融点の材料を見つけ出すまでの平均観測数は、クリギングが 16.1 回に対しランダムが 133.4 回ということで、ランダムに比べれば非常にうまくいくということがわかります。こういうことができると、第一原理計算とクリギングの組み合わせによる自動設計のようなものも視野に入ってくるわけです。

物性科学を進めていく上では、原因究明がいろいろなインスピレーションを得るという意味で大事ですが、原因究明のやり方に何か 1 つ決まったやり方があって、この記述子を使えば全部できるということは一切ないので、できるだけ大規模な実験をできるだけ高速に繰り返しながら何らかのインスピレーションを得るという、たとえばグーグルサーチのようなイメージでいろいろサーチしながら探していく。だから、一発ですべてわかるようなことは、まずないので、そういうことができるインフラ、Open linked data やデータベースの RDF などが必要だということです。

予測に関しては、予測しただけでは意味がありません。予測に基づいたアクションを起こさなければならない。たとえば実験計画のように、次にどこを実験するか決めるといったアクションを起こす必要がある。予測精度自体が目的ではないため、予測精度がもし悪くても、予測に基づいて起こしたアクションの結果には価値があるかもしれないということです。

最後のポイントですが、第一原理計算などのシミュレーションと今回の予測は方向がまったく逆であって、つまり、コンピュータを使っているというところは一緒なのですが、ルールからデータをつくり出すのがシミュレーション、逆にデータからルールを作り出すのが我々のデータマイニング、機械学習なので、協力できるのですね。そのループをいか

にうまく回すか。第一原理計算とクリギングのような機械学習の方法との組合せが、本当は非常にパワフルなはずです。

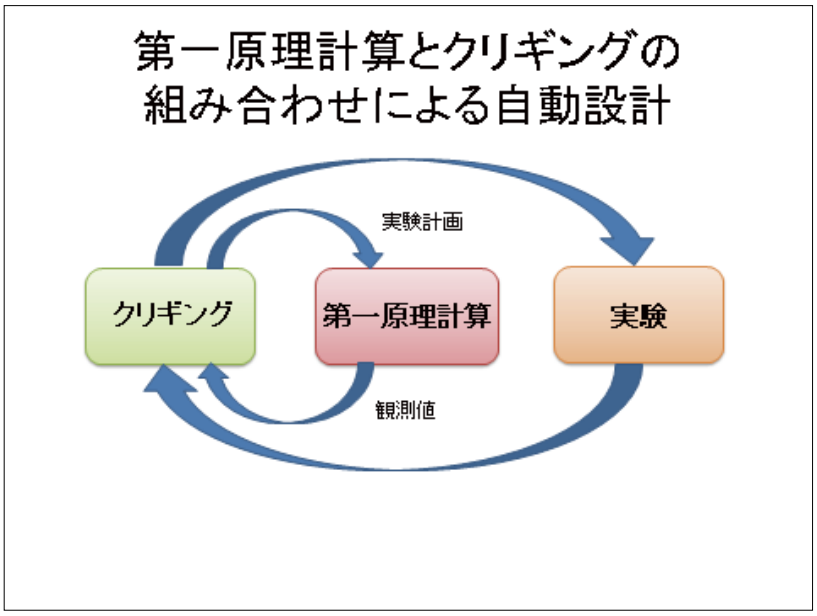


図 9

【質疑応答】

質問者：機械学習については、材料の分野では導入がまだされておらず、これからだと聞くのですが、それはどうしてなのでしょうか。本質的にカルチャーが違う、それとも整ったデータベースが無いからなのでしょうか。

津田：モレキュラーインフォマティクスを名乗っている研究者は、ほぼいませんね。ただドラッグディスカバリーなど、有機のほうはすごくたくさんいらっしゃいます。私が思う理由は、たとえば第一原理計算は非常に正確ですね。バイオではシミュレーションをやっても本当に当たらないです。だから、仕方がないのです。つまり機械学習というのは、ある意味で本当に物理法則をきちんとわかって、そこから現象を再現できるならば必要ないのです。材料科学の分野は、それがある程度、かなりのところまでできるのです。ただ今後を見ていったときに、やはりできないこともある。先ほどのガラス転移点のような、そもそもプロセス・原理がわかっていないようなことも浮上してきているので、いま注目を浴びているのかなと思います。

質問者：先ほど、細野先生からのお話で、最後に強調されていたのは、何となくはっきりしないような結果であればいらないというか。はっきり研究者の背景知識に照らして、説明がつくような知見が欲しいというのが、物性研究者の切実な要求だと思うのです。それに対して、従来のこういった機械学習やデータマイニングがどこまで答えられると思いますか。

津田：先ほどの回帰分析などをしてみて、本当に 1 つの変数しか効いていない、2 つしか効いていないという場合もあります。それでハッピーに話が進む場合もありますが、そういうものを探していくのであれば、スループットがすごく大事ですよね。つまり、いろい

ろな変数を入れて解析するというサイクルを、たとえば1時間に1回ぐらいできるようになれば。そして物性はものすごくたくさんあるわけですから、いろいろな物性やデータに対して、すごく早く回していけば、そういういい例が見つかると思います。ただ、私の経験からいくと、なかなかそこまできれいに2つ、3つとピシッと説明できるのは稀ではあるのです。ですから、いつも成功するわけではないということですね。

質問者：先ほど船津先生が仰ったように、本当に先端の研究分野にいくとデータがないのです。ギリギリのデータ数でNatureやScienceに投稿して載せている研究者はたくさんいるわけです。それに対しては、どう思いますか。
津田：もちろんデータがあってのデータマイニングなので、それは難しいのですが。バイオのデータとマテリアルのデータの違いは、マテリアルのデータのほうが扱いやすいです。つまり、バイオのデータではノイジーで訳がわからないものが多いですが、こちらはわりと法則性があって当たります。ですから、そういうことを使って本当に、先ほどの実験計画のようなこともできるだろうし、データを実験の人と協力して作り上げていくということです。

質問者：企業の中の計算機工学屋にとってはまさにQSPRです。計算科学でいろいろな記述子をやっていきます。HOMO、LUMO、その各々の精度が違うのですね。それは原理的な物理量を計算するものが違うというのと、いま自分たちがやっていることのために違うという、いま自分たちが限られたメソッドでやっているために違うというものと、原理的に第一原理計算といえども神ではないので、ある精度があるわけですね。結局そこでも経験的にやっているのですが、いま言った計算科学が本質的に持っているエラーバーを、このデータマイニングに取り込む仕組みというのは、どういうロジックがあるのでしょうか。
津田：それは、すべてをベイズ統計にしてしまえばいいのです。今のお話は、観測値自体が確率分布だという意味なのかと。
質問者：そうです。計算した値そのものが、自分たちが実験のようにエラーバーを定義、よくわからないまま経験的なエラーバーを持っている。
津田：それはたとえば適切な値を設定してもらって、たとえばベイジアンネットなど、そういう機械学習の手法を使えば、それも考慮することができます。今回の話では本当に簡単なことしか言っておらず、結局、機械学習やデータマイニングの本質は、いかに不確実性のあるデータから何とかして確実性ある結果を出すかという話なので、そういう研究もあります。

「スパコン京が開くインシリコ創薬の未来」
奥野恭史（京都大学）

私は、製薬会社に 10 社ほど集まっていただき、スーパーコンピューター「京」による創薬応用のプロジェクトをやっています。製薬会社の医薬品開発ではお金も年月も非常にかかります。実際、出発点から最後にものになる確率はだいたい 2 万分の 1 といわれています。こういったコストパフォーマンスが非常に低い状態がありますので、その向上は切実な死活問題になっていて、少しでも可能性があれば、インフォマティクスであろうが、いわゆる計算科学であろうが、何でも適用してものを作りたい、どんどん計算を使おうというのが創薬分野の業界です。

今日の話題は、創薬、創薬における計算機の役割、そして、インシリコ創薬、計算創薬の現状と限界です。インシリコ創薬は計算科学的アプローチと情報科学的アプローチの 2 つがありますが、その 2 種類の関係もご説明します。さらには、将来を見据えて、「京」による次世代型のインシリコ創薬とインフォマティクスとシミュレーションの融合についてお話します。

創薬の考え方は非常にシンプルです（図 1）。たとえばがんというのは細胞が異常に増殖する病気ですが、細胞の増殖に関するタンパクのポケットに ATP という物質がはまり込むと細胞が増殖します。これがはまり込まないように別の物質（化合物）を作ってブロックしてやることができれば増殖を防げます。この物質（化合物）が抗がん剤ということになるわけです。

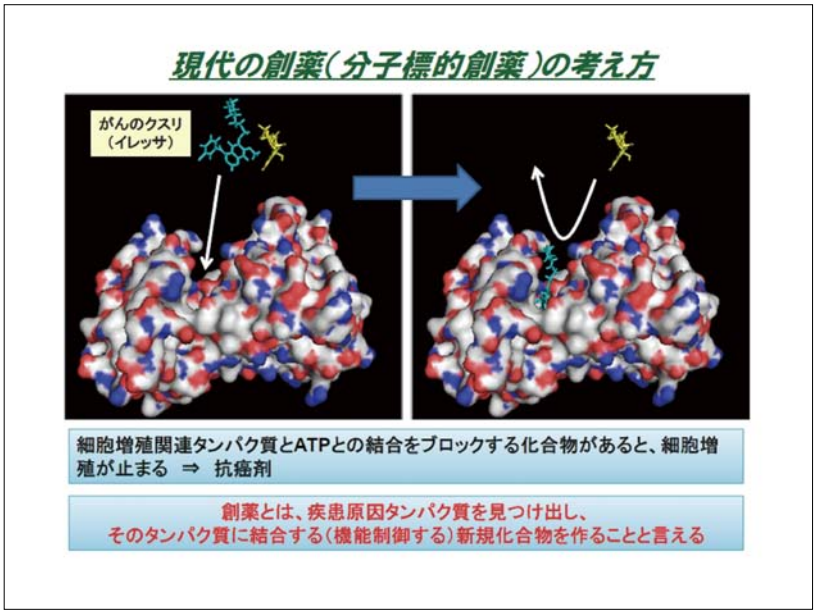


図 1

ですから非常に簡単に言うと、創薬とは、まず病気の原因となるタンパクを見つけ、次にそのタンパク質に結合して機能を制御する新しい化合物をデザインしてやるという、これだけの話です。

したがって、創薬の流れ（図 2）の始めの段階は、たくさんあるタンパク質の中から、病気の原因になっている創薬の標的タンパク質を見つけてやることです。この段階は通常、実験あるいは動物実験等でなされるのですが、最近はヒトゲノム計画以来明らかになったゲノムの情報をベースにした網羅的な実験技術が普及し、この辺のタンパク質あるいは遺伝子の非常に膨大な情報が高速で出てくる時代になっています。

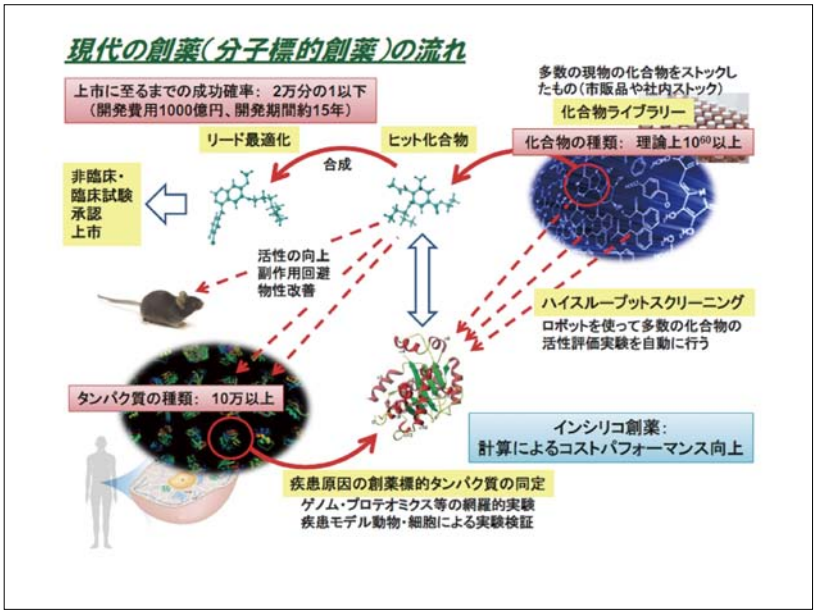


図 2

化合物の探索については、かなり原始的な方法ですが、病気を模した細胞等の評価系に多数の化合物をひたすらぶっかけて活性のあるものを探します。大手製薬会社ではロボットを使ってたくさんの化合物を一遍に活性評価するといういわゆるハイスループットスクリーニング実験をおこないます。化合物の候補としてよく使われるのは、世界中で売られているものをベンダーが集めて売っていたり、あるいは各社内で過去に合成してきたものをストックしていたりする、化合物ライブラリです。それでまず、たとえ弱いものであっても活性のある、いわゆるヒット化合物が見つかってきますが、その後、この活性が弱いヒット化合物の最適化をしていきます。元の構造をベースに、いろいろと構造変換をかけていきますが、ここは 1 つ 1 ついろいろ変えて合成していくため、ロースループットにならざるをえません。この段階では、「活性の向上」、「副作用回避」、「物性改善」の評価について実験しながら合成していきます。たとえば、「副作用回避」では、実際の標的の疾患関連タンパクとは違うタンパク質との相互作用も見ていき、別のタンパク質への作用による副作用についても評価しなければなりません。こうして最後には、実際に臨床実験をして市場に出していくというのが、創薬のプロセスです。

現在、遺伝子の種類は 2 万 2000、タンパク質にすると 10 万以上あるといわれています。化合物の種類も、売っているものは数百万、600 万個ぐらいあると言われていますし、低分子化合物、いわゆる有機分子も理論上 10 の 60 乗以上の組合せがあるといわれています。この膨大な数の中から医薬品の候補化合物をどう選ぶのかが問題で、実験をずっとやり続けていたら大変なことになります。

次に、創薬の現場で使われている計算技術として、まず、単純に結合を見ていくドッキングシミュレーションがあります。ドッキングシミュレーションは、これまでインシリコ創薬を牽引してきた技術です。また、疾患原因のタンパク質を見つけるバイオインフォマティクスがあり、さらに化合物を見つけていって最適化していく部分にはケモインフォマティクスがあります。

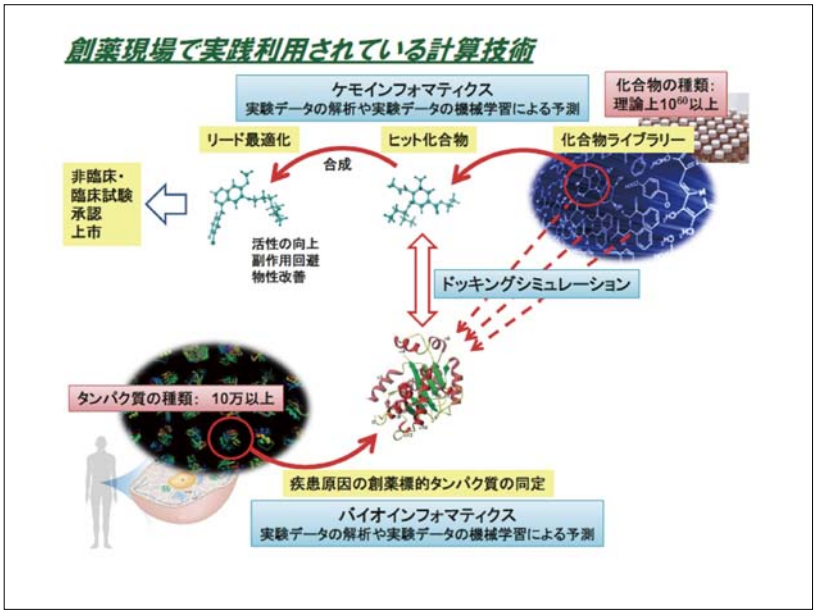


図 3

ドッキングシミュレーションとは、たとえばある構造のタンパク質のポケットにいろいろな化合物をはめ込んでいくのを自動化することです。ただ、これまでの予測正答率は非常に低く、5%といったヒット率です。それでも 5%でも当たるならば実験をするよりも効率的なので、やろうということです。なぜ正答率が悪いかというと、タンパク質のポケットをほとんど動かない状態にして化合物を当てはめているにすぎないからです。また実際に結合の強さ、結合自由エネルギーを正確に見積もれないからでもあります。結合自由エネルギーがうまく見積もれないのは、今の計算機では計算時間が追いつかないために、真空状態での水が存在しない単純な当てはめ計算しかできていないからです。あるいはスコア関数が非常に不十分であるという問題もあります。また、ドッキングシミュレーションするためにはタンパク質の立体構造が必要ですが、そういった立体構造が明らかになっていないタンパク質を対象とした計算が出来ないという問題点や、計算できたとしても、今の計算機で処理できる化合物は大体数千万個ぐらいで、10 の 60 乗からは程遠いという現状です。

こういった多くの課題のため、ドッキングシミュレーションはこの状態で頭打ちしているという現状があります。これらのうち計算時間の問題については、「京」を使ってこれをクリアしようといま行っているところです。

実際に「京」で計算した結果の例では、いわゆる水の存在下で長時間の MD 計算を行うことが出来るようになり、先ほどのドッキングシミュレーションと比べると、かなり先進的な計算を行っていると言えます。これまでの通常の汎用機ではマシンパワーがなかっ

たので 50 個ぐらいの化合物を計算するのに 20 年かかるものを、「京」は 1 週間程度で答えを出してしまうのです。このように「京」を使うことで、精密な結合自由エネルギーの計算が現実的な時間で出来るようになります。例えば、東大の藤谷らによる結果では、結合自由エネルギーの実験値と計算値がほぼリニアに乗っています。このように、ドッキングシミュレーションの限界の中でも、正確に結合の自由エネルギーを見積もるという問題に対しては「京」の計算力で何とかクリアしていけるのではないかと思っています。

この他にも、化合物が膨大であるという問題や、タンパク質の結晶構造、立体構造がないために先ほどのシミュレーションができないという問題もありますが、それらに対しては、いわゆる機械学習のアプローチがあります（図 4、図 5）。

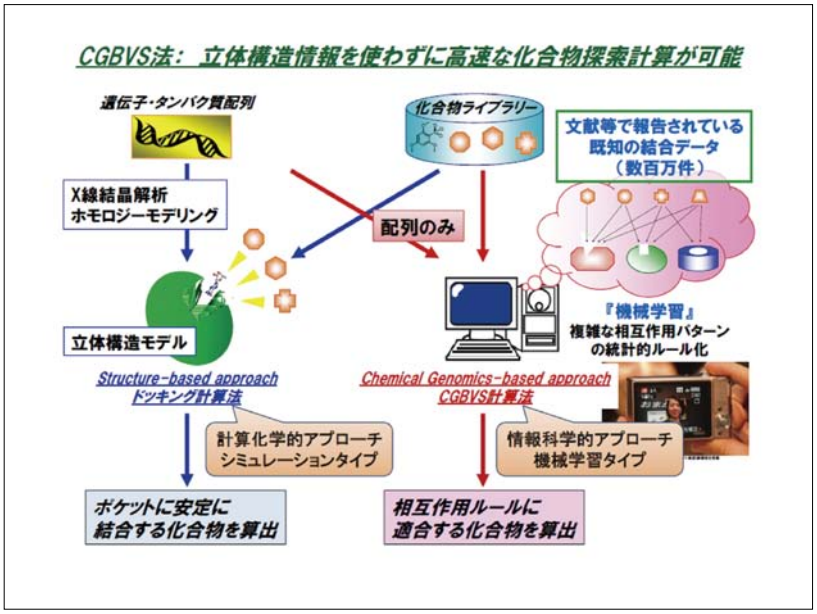


図 4

この機械学習のアプローチでは、過去に文献で報告されているような化合物のデータをまず計算機に学習させ、それに基づいて新規の化合物を予測します。化合物を記述子で表現し、molecular weight や log P、炭素の数、水酸基の数など、構造や物性を表すような数値列にします。タンパク質の方も、バイオインフォマティクスの分野では、アミノ酸配列に基づいたタンパク質の性質を表す記述子があります。重要なのは、化合物の構造が似ていれば、この数値列が似てくるということです。タンパク質の記述子表現についても、配列や構造が似ていれば数値列が似てきます。

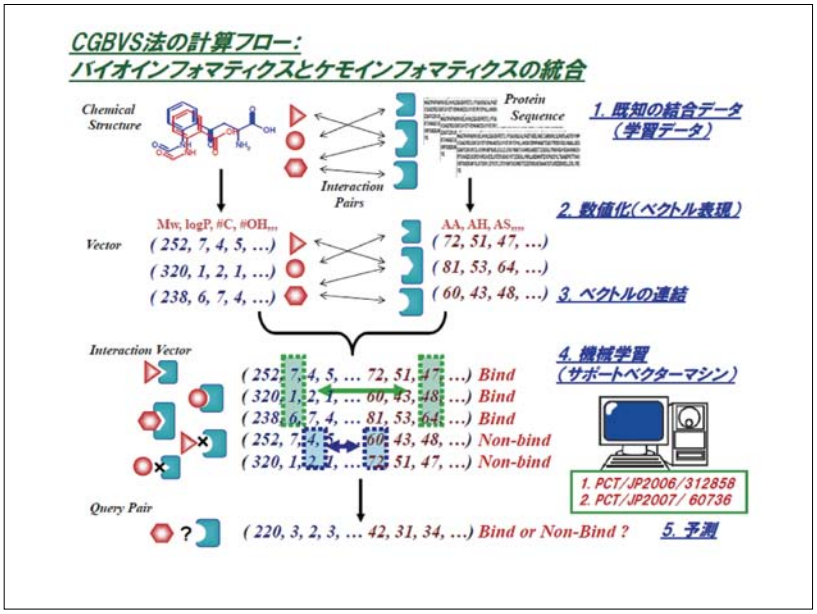


図 5

こういった状態からどうするかというと、結合するものは結合するもの同士で、ベクトルをくっつけて「結合します」というラベルをふり、逆に結合しないものに関しては、結合しないものでベクトルをくっつけて「結合しない」というラベルをふります。そこで計算機に、この数値列に内在している数値パターンを学習させます。たとえば結合するというベクトル群と、結合しないというベクトル群があったときに、結合するというベクトル群に出ている特徴と、結合しないというベクトル群に出ている特徴が違う場合、これは結合しているというベクトル群に出てくる特徴だとみなして、そういった重みづけをするのが機械学習です。こうして一旦学習をさせると、予測の段階では、未知の化合物とタンパク質の組み合せも同じように数値表現ができますから、この未知の組み合せの数値ベクトルを作ってやって、このベクトルが「結合する」というベクトル群に近いのか、「結合しない」ベクトル群に近いのかを判断します。「結合する」というベクトル群に近ければ結合すると予測するわけです。

立体構造モデルを作って予測するという従来の立場では、こういった数値化予測の方法で本当に予測できるはずがないだろうと思われるのですが、しかし実際この方法はかなりパフォーマンスがあり、立体構造を使ったドッキングの正答率がたかだか 5%いけばいいと言っていたのに対して、こういった機械学習のアプローチでは 10 ー 20%ぐらいのヒット率になります。さらに、X 線結晶構造解析で立体構造がなかなか解析できないタンパク質群に対しても、こういう機械学習のアプローチは実際に適用でき、ヒットがきちんと出てくるのです。何が起こっているかわからないけれども、とにかく現場では役に立つものとして使われているのです。

このアプローチを、我々はいま「京」を使って、先ほど言った膨大な化合物に適用することを実際に行っています。たとえば 105 億のタンパクと化合物の組み合せを計算するのに汎用機では 20 年かかるところ、「京」をフルに利用できた場合は 20 分で終わってしまうというパフォーマンスを示しています。

現状の我々が使っているインシリコ創薬の計算技術は、タンパク質の情報、構造、配列、そして化合物の化学構造を入れてその答えを返してもらうといった使われ方をしています。しかしそこには、10 の 60 乗個という膨大な化合物から始めに何を選べばいいのかという問題があります。あるいは何かヒット化合物があるけれども、次はどんな化学構造にすればいいのか、そういったことがわからないという問題があります。それはひとえに、予測するのに化学構造をまず人間が考えて入力しなければいけないといった問題があるからです。

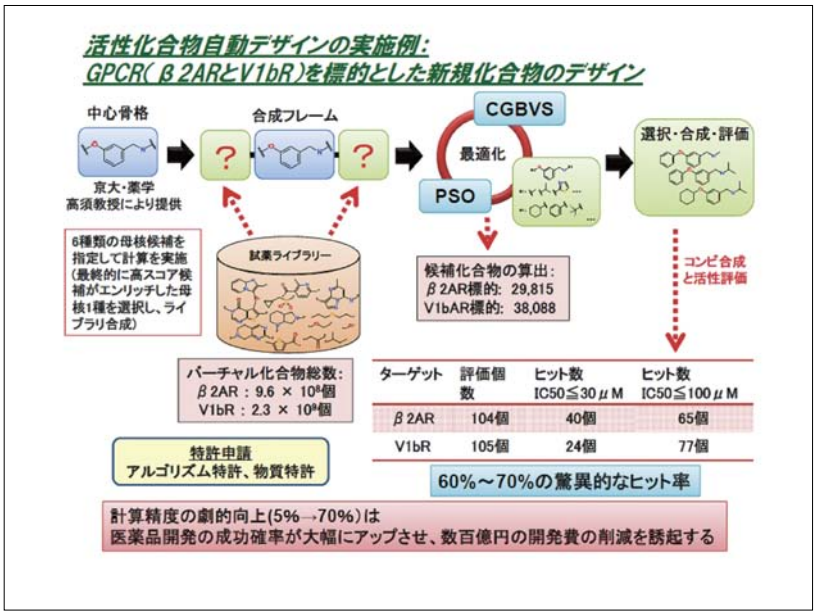


図 6

そこで我々は、化学構造自体を計算機が自動的にデザインしてアウトプットしてくれるというアプローチをいま開発しています（図 6）。これは、標的のタンパクの名前だけを入れれば、こんな化合物に活性があるかもしれないということを自由にデザインしてくれる、つまり計算機が能動的に考えて化合物を提示してくれるといったものです。たとえば、真ん中の骨格だけが決まっている素材があって、そのまわりにどういった構造を結合させるか、まわりのフラグメントをどうくっつけてやるか、どれを選んでやるかということを考えるときに、（これを実際に売られているものから計算すると、2 × 10 の 15 乗ぐらいの組み合わせがあり、その中からどのように活性化合物を手にするかが課題になるのですが、）まずランダムに 100 個の化合物を発生させ、これらの化合物を一旦、計算機で予測させると、活性のスコアが出てきます。その中でスコアの高いほど活性化合物である確率が高いので、スコアの高い化合物に着目して構造の共通性を見つけ、その部分構造を残したまま別のところの構造を変えてまた次の化合物を自動的に発生させてやる。こういったことを繰り返して化合物をどんどん計算機上で作りかえていくうちに、スコアの高いものに共通する特徴、構造がよく保持されるようになります。いわゆる最適化アルゴリズムを使うのですが、計算機の中で化合物が進化していってやがて活性の高いものばかりでマチュアされていくというものです。実際に我々はこういった計算をして、それを実際に合成し、生物活性の評価をしたのですが、2 種類のタンパクに対して 100 個ずつ化合物を作っ

てやって、アッセイをしたら、60 ー 70%ぐらいの驚異的なヒット率に跳ね上がるという結果を得ています。

最後に、問題点を、まとめとともに示します。この分野では、扱う化合物あるいはタンパク質の数がとにかく膨大ですので、必要に迫られてインフォマティクスを使わざるを得なくなったという背景があり、いわゆる実験結果の予測や解釈にバイオインフォマティクス、ケモインフォマティクスが使われてきています。

こういった状態を「京」がどう変えていくかというと、「京」でシミュレーションすると、驚いたことに、タンパク質のスタート構造は、数百 Kb 程度のデータ量しかないのですが、長時間 MD をすると数百 Gb と膨大にデータ量が増えるわけです。それがなぜかというと、時間軸が出てきたからです。今までは物質の数だけの勝負だったのですが、今後は時間も見ていかなければいけないのです。それで、いわゆるシミュレーションした結果を解析するインフォマティクスが必要になってくるということを切実に感じています。最終的にはこれまでのインフォマティクスと次世代のシミュレーションとが合わさって、実験ではとらえられない事象が見えてくる、あるいは実験に近づく予測が出てくるのではないかと感じています。

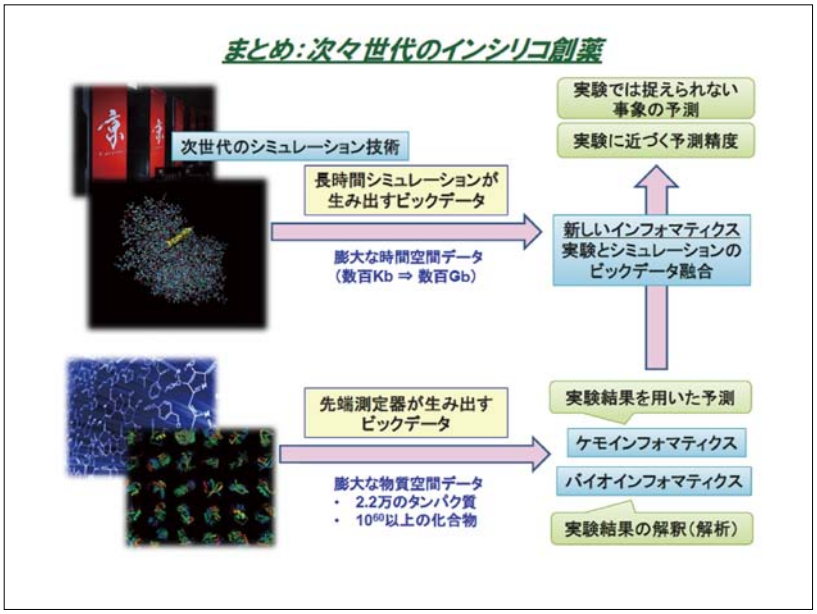


図 7

先ほど、先端の分野ではそもそも学習するデータがないとの話がありましたが、同感です。いわゆる第一原理計算に基づくようなシミュレーションの結果を「学習」をおこなって予測していくという時代が、そう遠からず来るのではないかと感じています。

【質疑応答】

質問者：有機の場合、やはり分子概念があるということですね。無機の場合、分子概念でなくユニットセルですから、そこでやはり本質的に随分違いがあると感じました。それから、いわゆるコンデンスマターの材料と比べると、薬の場合はあまりにも「風が吹けば桶

屋が儲かる」で、結果の論理がまったくわかりません。物性物理では、それがわからない者は馬鹿だというぐらい、ある意味ではわかるわけです。そこが、非常に本質的な違いだろうと思います。ただ、それだけに頼っているときりがないので、そのための情報が無機、特に金属化合物などでは膨大に溜まっていますので、それを利用しない手はないだろうなということを感じました。
奥野：たとえばタンパク質は一種のポリマーですので、アミノ酸をベースにしたユニットとして考えるわけです。セル、ユニットとして考えると、無機用の記述子を作らなければ駄目だろうなと思いました。
質問者：確かに記述子を区別するのが最初ですね。
寺倉：これはおそらく非常に大きな問題で、無機材料のインフォマティクスのときの 1 つの大きなバリアだと思います。しかし何でもやり始めると、10 年経つとまるっきり変わりますので、あまり尻込みしないでやっていかなければならないと思います。

質問者：要するに、注目しているものだけにこだわってしまうと全貌は見えない、予測はおそらくできないと思うのです。先ほど、データはあるようでないという話がありましたが、実はあるわけですね。そのごく一部のデータで相関モデルを作ろうとしても、誤差の大きいモデルしかできない。ところがその前後にデータがあり、それをつないで、ひとまとめにしてモデルにすることができれば、わりと精度のいいモデルができることがあるのですね。要するに全体を俯瞰することが大事で、そのためにどういうデータが結合できるのかといった考察を経なければいけないのですが、その意味で、先生のお話は創薬に限らず、材料ポリマーのデザインといったものにも応用できると感じました。

寺倉：今のお話で非常に印象深かったのは、活性化合物の自動デザインのところです。元素の組成を与えたときに構造を探す方法はいろいろ発展していて、最近読んだ論文の中で一番印象的だったのは、「あるコンポーネントを決めたら、その配置を与えて、それに空間群の 230 個のオペレーションを全部操作して、可能な結晶構造を全部サーチする。そして、それぞれの空間群の枠の中でまず最適化し、それぞれの与えられた空間群の中で最適化されたものの間で比較しながら、またそれにランダムなディスターバンスを加えて、最適化していく」というものでした。我々の分野でも、そういう計算機の力に任せてやっていくことはそれほど遠くない時期に可能になるのではないかと思います。

「第一原理計算に基づいたマテリアルズ・インフォマティクス」
田中功（京都大学）

インフォマティクスという話をすると聞き手のリアクションはさまざまですが、「データベースはあまり役に立たないのですよね」というリアクションがマジョリティです。たしかに実験のデータベースを見たときに、情報が非常に限定的であって、物理モデルも過度に単純化され、あまり適切なものでないので、それを元に材料探索をすることには、多くの人がネガティブなイメージを持っていることは間違いないと思います。それが、たとえ第一原理計算に代わったとしても、状況はあまり変わらないわけです。物理モデルを使って材料探索をするという意味では、それほど明るい未来があるような気がしません。ただ我々が必要なことと考えているのは、このようなことです。

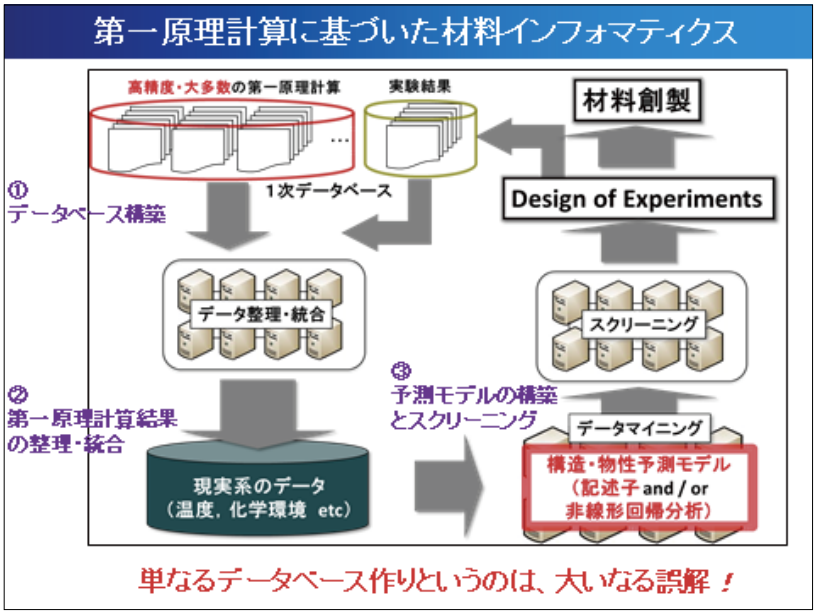


図 1

もちろん第一原理計算はたくさんやる必要があります。それで、データをうまく活用していかなくてはなりません。有限温度あるいは様々なケミカルポテンシャルのもとでのデータにしなければ、まず役に立ちません。でも、それでも駄目なのですね。何か分かった物質についてインプットがあって、それを計算しているだけで、結局計算をやっている人間は、何か後付けで説明するだけだという批判を繰り返し受けているわけです。計算をうまく使って材料創成につなげていこうというところは、記述子を使ってスクリーニングする、また実験をデザインするというプロセスがあって、必要に応じてまた実験にフィードバックする。こういうループをうまく回していくことが必要になってきているのではないかと思います。

言いたいことはこの３つです。単なるデータベース作りではないということ。データベースを作るのは本当の最初のステップであって、それで終わってはまったく意味がありません。我々は、いろいろな産業界の方とお話しをすることがあります。その要求は、何か良

い特性を持つ組成や構造を見つけ出して、何を混ぜればいいですか、というものです。そういうことに対して、我々はなかなか答えを持っていなかったのですが、そういうニーズはものすごくあるわけです。それは産業波及効果があるからです。ですから、そういうところにうまくいけば大きなインパクトがあると考えています。それからもう 1 つ、これも強調したいと思っているのですが、我が国はほとんどこういうアクティビティで非常に遅れている気がしますので、そういう問題意識を持っています。では、それぞれの個別について話をしていきたいと思います。

最初のデータベースの構築のところですが、米国の取り組みが随分進んでいます。「マテリアルゲノムイニシアティブ」の中心になっている 1 つが、MIT でやっている「マテリアルプロジェクト」です。その中心にいるのがこの MIT の Ceder と Pearson の 2 人ですが、このデータベースがどういう状況か、簡単に見たいと思います。これはオープンソース、オープンライブラリになっていますので、誰でも見られる状況になっています。いま 3 万件のデータが登録されているということです。この Materials Explorer に、たとえば鉄の酸化物が欲しい場合、データを入力すると、こういう結晶多形が出てくる。それぞれクリックをすると結晶構造が出てきて、普通の GGA + U レベルですが第一原理計算のエネルギー、構造、X 線回析のプロファイル、あるいはそのバンド図が出ている。そういう計算した結果が 3 万件ストアしてあるわけです。

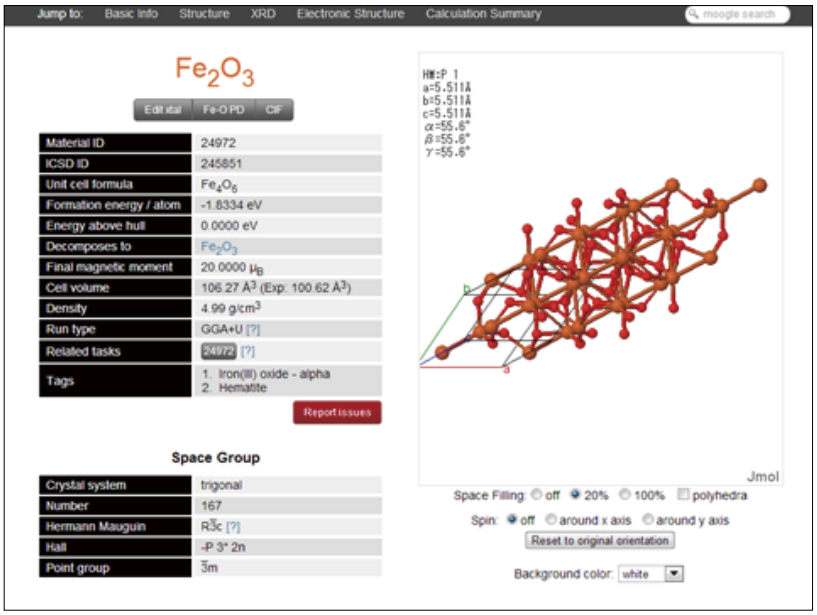


図 2

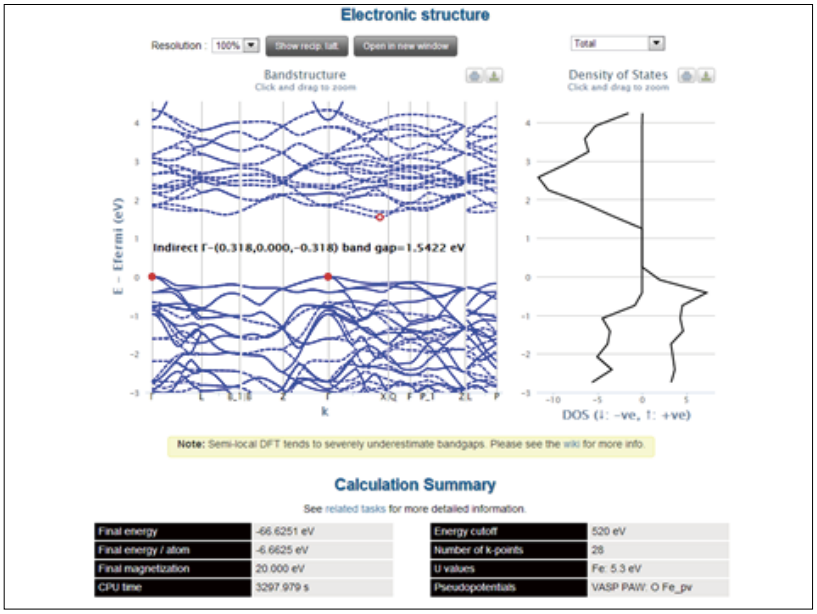


図 3

あるいは、Phase Diagram Application というものがあります。材料科学では状態図が非常に大事です。ここをクリックすると、たとえば擬三元系の状態図にあるような様々な化合物について、ICSD のデータベースに載っている構造について全部、彼らが計算したデータをストアしてあるわけです。その中で安定なものが、こういう赤で示したような化合物であって、そのエネルギーなどが書いてあります。

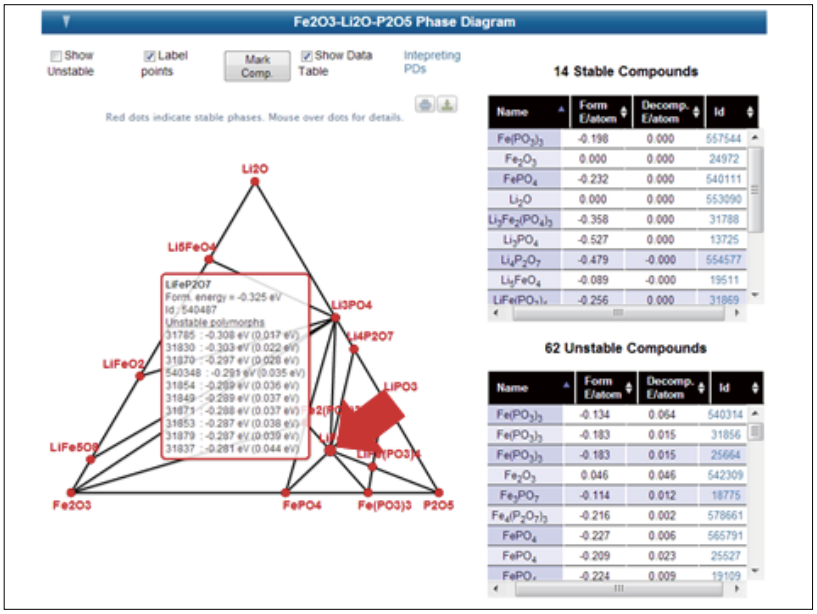


図 4

さらにこのブルーのものは、実験としては報告されているけれども、計算をすると生成エネルギーがプラスになるようなもの、つまり準安定な構造であるといったデータについても、提供されているわけです。数値データとしては少し問題のあるところもありますが、

ざっとこういう全体像を見るという意味では、非常によくできているものだと考えています。

同じように、これも MIT にいた人ですが、デューク大学で Curtarolo という人がデータベースを作っています。たとえば Fe_2O_3 を入れると Fe_2O_3 の構造が出てくるというのは同じです。もう少し詳しく、たとえば有効質量はどうなっているか、といったことをデータとしてストアしてあるわけです。

彼らは、そういうものを使っていろいろなことをやっているわけですが、たとえばこれはγ線検出シンチレーター材料の光のシールドを最適化したいという目的で、スクリーニングをやった結果ですが、バンドギャップや電子、正孔の有効質量を網羅的に計算したデータベースを持っていて、それを割と粗っぽい記述子に放り込んでやって、そういうデータの中から、たとえばこういう化合物がいいのではないかといったことを提案しています。

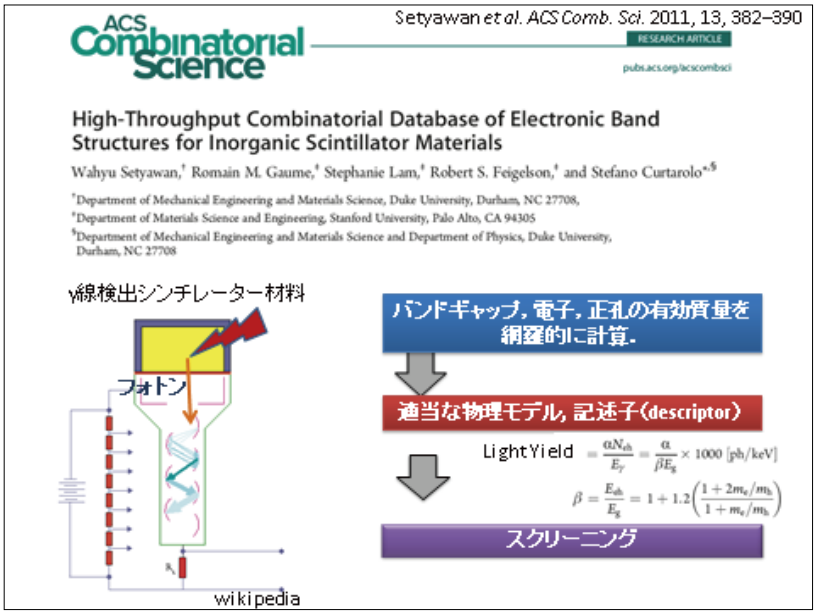


図 5

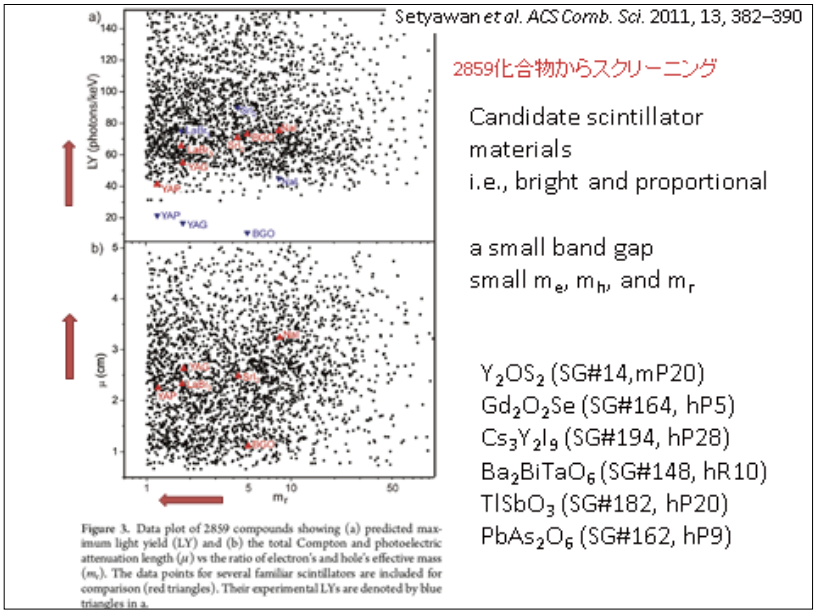


図 6

同じようなアクティビティが、先ほど細野先生が紹介された Alex Zunger の Inverse Design センター、あるいはアイオワ州立大で Krishna Rajan たちがやっているマテリアルインフォマティクスなど、少し前からあります。たとえば、これは低摩擦係数物質を探索するということで、トライボロジーなどの分野では、あまり第一原理計算を使う、あるいは固体科学的なアプローチはされておらず、ほとんど経験に頼られていたような分野のようです。主因子解析などをして、いい相関のあるパラメータを見つけ出し、それをもとに低摩擦係数の物質を予測するといったことをやっています。

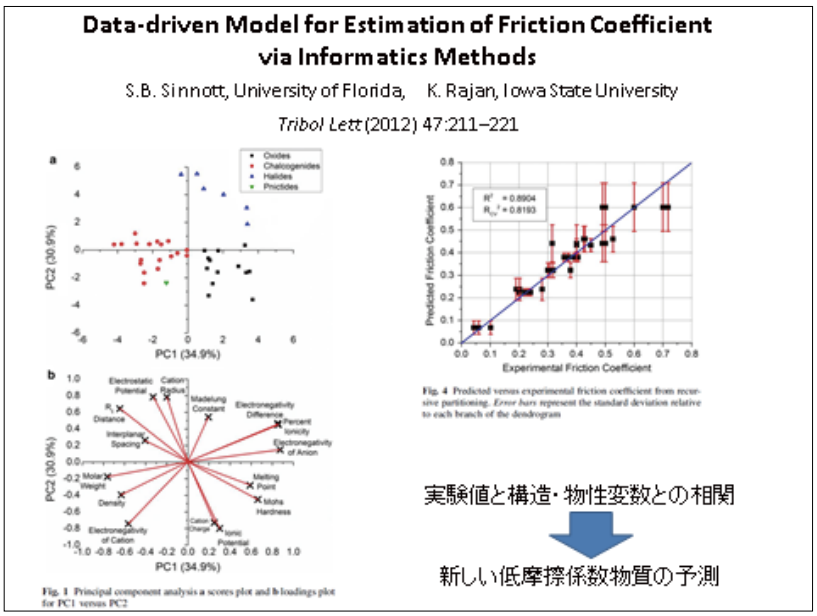


図 7

米国の企業、とくに BOSCH などは非常にそういうことを精力的にやっており、彼ら

はAIDA（Automated Infrastructure and Database for Ab-initio design）というインフラストラクチャーを、第一原理計算をもとにしたデザインのシステムを作って、リチウムバッテリーや熱電や圧電といったものの材料設計をかなり精力的にやっています。

欧州でも、たとえばこれはCECAM（Centre Européen de Calcul Atomique et Moléculaire）という国際シンポジウムですが、こういうものが毎年同じようなマテリアルインフォマティクスのようなタイトルで開催されています。豪州でも最近こういうシンポジウムがありました。韓国でも、KISTでMIDASというデータベースを作って、それをもとにいろいろな材料設計をしていこうということを聞きました。これらが世界の状況です。

我々の取り組みとしては、こういうデータベース作成を、自動化することが大事だと思い、多重処理の一般的なプログラムを作りました。これはオープンソースとして公開していますので、どなたでも使うことができます。たとえば、これはペロブスカイトの化合物について、合計530個の第一原理計算を一遍にやったということです。530種類の計算でも8ノードぐらいのPCクラスターによって6時間で終わってしまいます。たくさんの計算をするのは、もうそれほど大変なことではないのですね。逆にいうと、データベースを作ることは、それ自体ではそれほど時間がかかるわけではない。結晶のデータベースがありますので、そのデータのフォーマットを変換して、計算機に流すだけなのです。たとえば、これは単にVolume（体積）とBulk Modulus（体積弾性係数）だけですが、たとえばこの中からユビキタス元素からできているものはどれだけあるかというと、64種類で遷移金属のシリサイドが割と硬くて体積が小さく、フローライトなどは体積が大きく体積弾性係数が小さいということがわかります。それだけでは何の意味もありませんが、こういうことができるという一例です。

その後、どうやって使っていくかということです。1つは、このデータの整理と統合というところで、まず現実系のデータ、温度や化学環境に直すところです。これについては、1つは自由エネルギーの温度依存性を出すときに、フォノンの寄与というのが本質的ですので、フォノンの計算をすればいい。それから合金の構造についても、固溶体の自由エネルギーを出すなど、あるいは構造探索をしていくといったこと、これもすべて物理の法則に従ってプログラムを作っていけば、それに従って計算できるわけです。たとえば化合物安定性ですが、AとBという組成があり、ABという化合物がある。いろいろな構造について計算すると、エネルギーがこれだけ出てきます。これが安定か不安定かというと、このAとBというエンドメンバーについてエネルギーを出し、そこに線を引いて、それを安定だということなのですが、実はそれは正しい答えではなく、いろいろな組成について計算し、このconvex hullを作ってやると、この一番エネルギーの低かったものでも、実はこれとこれに層分離した方が安定である。だから、これは準安定にしかすぎません。ですから、この系について理解しようとすると、全範囲についていろいろな構造を計算しなければいけないということです。こういうことを有限温度でやらなければいけないわけです。ですから1つはフォノンで、このフォノンは第一原理計算についてたくさんの計算をして、それを情報統合することで、第一原理計算の精度で自由エネルギーを計算することができます。これもオープンソースになっており、すでに世界的に使われています。

たとえば典型的な例として、酸化ジルコニウムの相転移について自由エネルギーを計算してみると、低温側では monoclinic（単斜晶）の相のほうが自由エネルギーは低いけれども、こういうところで自由エネルギーがクロスして、ここから高温側ではこちらの tetragonal（正方晶）の構造のほうが安定になる。相転移温度はこれぐらいということが、いまルーチンで計算できるわけです。そうすると、こういうものを熱力学、統計力学のプログラムとうまくコンバインするということで、我々はまた別にこういうクラスター展開のプログラムを作って公開しています。たとえば 1 つの例ですが、酸化マグネシウム、酸化亜鉛といった擬二元系について状態図を計算した結果です。第一原理計算は Rocksalt、Wurtzite といったものでやるわけですが、ここにマークしてあるようなものは全部の配列です。合金配列 15 万通りといったものについてクラスター展開し、全部エネルギーを出して、先ほどの convex hull というものを得て、あるいはこれについてモンテカルロ計算をして、こういう状態図の固溶限を出していくわけです。

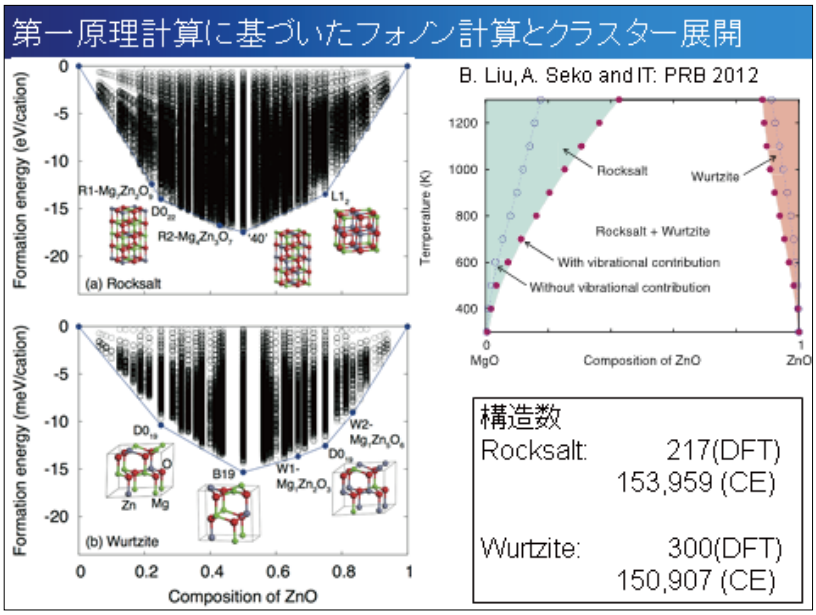


図 8

この系については、すでに実験的に知られていますので、実験と比べることができるわけですが、多くの酸化物では実験情報がない場合が多いです。金属は割と状態図がわかっているのですが、それでも多元系になると、なかなかわかっていない。それについて、こういう計算をすると第一原理計算の精度で、実験の人が欲しいデータを出すことができ、現実系のデータができてくるわけですが、これでもまだ単に何か与えたものについて計算しただけになってしまうわけです。

やはりそこから材料創成、あるいは何か発見に持っていくためには次のステップが必要になってくるわけです。次のステップのところを、津田さんともいろいろディスカッションしながら、少しずつやり始めているところです。たとえば 1 つは、融点の予測です。

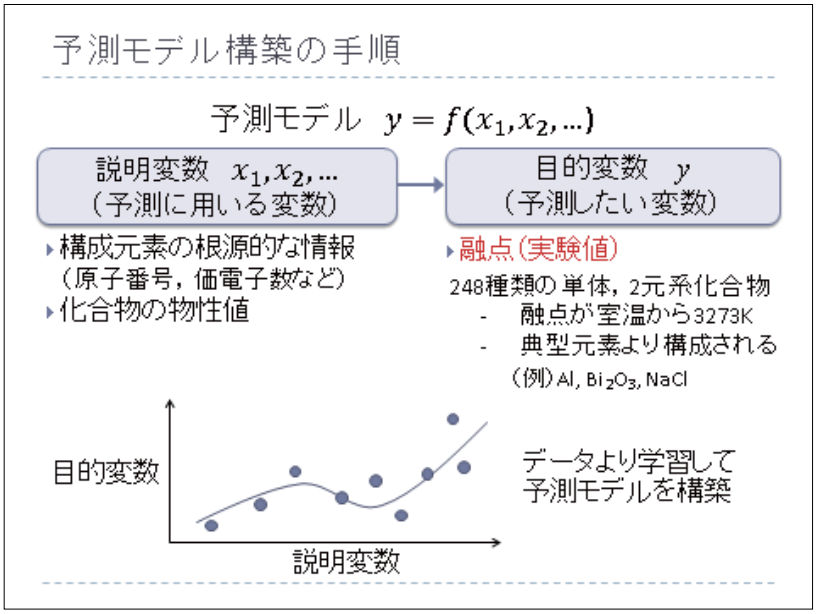


図 9

融点とは、ご承知のように 1910 年に Lindemann 則が提唱されています。固体内の原子の熱振動振幅がだんだん大きくなって、ある程度以上になると融解するということで、実験データをプロットするとだいたい説明できるということで、多くの固体物理の教科書には、この 100 年前のルールが載っているわけです。こういうものをどの程度計算・予測できるであろうかということです。対象としたのは 248 種類の単体あるいは 2 元系化合物についての融点です。なぜ融点かというと、実験をやると比較的誰が測っても同じように測定できるものです。それから比較的データが揃っている。また、融点を第一原理計算するということはそう簡単ではないので、こういうものを予測するのは意味があるだろうということで計算をしました。しかし、これは第一原理計算をしているのではなく、説明変数を持って来てそれを予測するということです。説明変数としては、構成元素の情報として、周期律表の場所ですね。ですから原子量や原子番号、周期律表の属、周期、荷電指数、あるいは周期律表に載っているようなファンデルワールス半径や共有結合半径などだけを使っています。もちろん化合物の場合には組成というのもあります。それから、あとは第一原理計算でもかなりプリミティブに得られるようなもの、たとえば格子体積、体積弾性率、最近接原子間距離、凝集エネルギーといったものを使ってどれぐらい予測できるかということをやってみました。32 個の変数を全部使うわけではないのですね。よい変数をうまく選び出していくということで、変数を 1 つずつ加えていくことや、全部の変数を入れて 1 つずつ減らしていくという方法もありますし、変数の選択はいろいろあるのですが、我々は 1 つずつ増やしていくということで、サポートベクトル回帰を使って分析をしました。

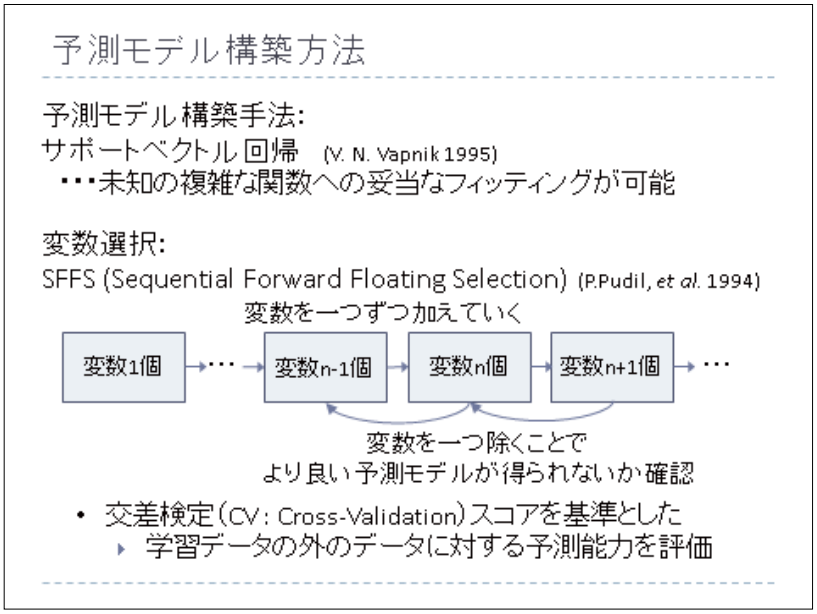


図 10

まず、先ほど第一原理計算の情報を使わずに、単に周期律表の何丁目何番地という情報だけを使って、これを線形回帰したとき、サポートベクトル回帰したとき、それがこの 2 つのプロットです。横軸が実験のデータ、縦軸が予測値です。学習が終わった後、テストデータについてテストする。サポートベクトル回帰をするとテストの点数の結果が 299 K ということで、線形回帰よりも随分いいのですが、たとえば融点がマイナスを予測するなど、少しばらつきはあるわけです。ここに第一原理計算の結果を加えてサポートベクトル回帰をすると 249 K の誤差になる。249 K の誤差をどう考えるかですが、もちろん融点がきちんと定量的に出ているわけではないのですね。ただ、入れている情報も周期律表の何丁目何番地と、それから非常にプリミティブな化合物のデータですので、それぐらいのものを使って未知の化合物の融点がおよそ 250 K 程度の誤差で予測できると、それは場合によって非常に満足できる状況になると考えています。沸点も同じようにやると、沸点についてもこのように 249 K、これは偶然ですが、これぐらいの誤差で出てくる。あるいは室温の低圧比熱を予測したところ、これもこういう線上に乗っていくような結果になりました。ですから、比熱や融点・沸点はそれほどエキサイティングな物性ではありませんが、こういうものをうまく使っていって、何かの予測に使える可能性はあると考えています。

もう 1 つの例は、リチウムイオンの固体電解質について、こういう手法を適用した話です。これは LIthium SuperIonic CONductor（LISICON）オキサイドで、もう 30 年ほど前から研究がされています。少なくとも Li、Zn、Ge、O という 4 つの元素を含んでいるような複雑な系でうまく組成をコントロールするとイオン伝導度が大きく変化するというもので、こういうものを使って全固体電池に使いたいという話があります。元素設定や化学組成に大きく依存しますし、報告者によっても随分バラついている。ですから明確な材料設計指針がなければ、単にこのバラつきを増やしていくだけのようなところもあります。我々は、これを第一原理分子動力学の計算をしました。高温で分子動力学（MD）

の計算をして、この傾きから拡散係数を出すということをやります。そうすると、こういう MD 計算は高温でしかできないのですが、実験は低温でしかないので、実験の結果と高温から低温を結ぶと、まあ一致しているようである。ところが実験結果をよくみると、この辺りで折れ曲がるのですね。これは合金元素、溶質が秩序・無秩序相転移を起こすわけですが、こういうことまでは高温の MD 計算からは予測できないわけです。ですからこれはこれで別の計算をして、こういう相転移がどういう温度で起こるかを予測する。そういうものを 92 通りの組成について、たとえば体積やエネルギー、秩序・無秩序の相転移温度､それから拡散係数､こういうものを第一原理計算で求めました。こういう表を作って「さあ、どこが一番いいか」というと、これだけでは出てこないのですね。それは、先ほどの拡散係数を表す明確な物理モデルがないからです。それで、実験結果を学習して、実験結果と第一原理計算を合わせてサポートベクトル回帰をします。すると、これは縦軸が 373 K でのイオン伝導度の予測値です。従来の最高値に比べて、こういったところでおよそ 5 倍程度の伝導度が上がるものを予測できました。第一原理計算では直接求めにくい物性値を機械学習でうまく予測できる 1 つのモデルになっているものと考えています。

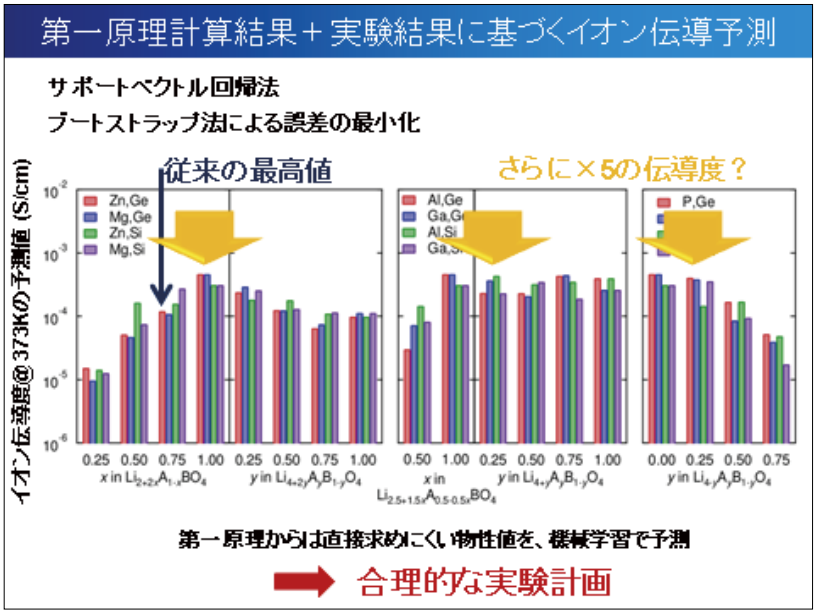


図 11

こういうものを使って合理的な実験計画を立てることができれば、こういうところで実験をし、それをうまく回していくと最終的に材料設計つながっていく。計算が終わったのが終着点ではなく、実験をうまくデザインしたということだと考えています。バイオインフォマティクスとの大きな違いですが、材料の場合は、そもそも ICSD のデータが 10 万点ぐらいですので、データ数として 10 万ー 100 万程度あれば、まずいいのかなと思っています。それは、追いつくことは十分可能で、それほど時間がかかるものではないと思います。しかし開拓すべきは、そのデータをどう使っていくかということ、それから材料の場合は第一原理計算がかなり使えて、物性値を計算できるということ、また物理モデルもわりとわかっているものがありますので、記述子として使える可能性があるということで

す。それから Design of Experiments に基づく実験が、バイオよりも材料の実験のほうが、ある程度数をこなすのは簡単であるという意見もあります。こういう方向にうまく政策誘導することが大事と考えています。

【質疑応答】
質問者：計算結果と実験屋さんとのシステマティックなコラボレーションは、具体的にはどのようにやられているのでしょうか。
田中功：一般論として言うと、こういう形でいいものが見つかりそうであると、その辺りを企業の人が合成の実験をする。合成の実験をするときも企業でやるとコストが高いので、さらに下請けに出してたくさん合成をするといったことをしています。その中でいいものが見つかるということは、実際に例があります。

質問者：先ほど線形回帰のところで SVM（Support Vector Machine）を使われるという話でしたが、ああいった精度が出てくるのは非常にいいことだと思います。第一原理計算するというのは、予測するという目的もあるでしょうが、やはり理解することが重要な目的だと思います。そのときに、線形回帰に比べると SVM はもっと複雑なので、予測はいいのですが、それをどう解釈して企業とやるときに実験、材料開発に生かしていくか。その辺のスキーム、考え方はどのようにとらえればいいのでしょうか。
田中功：我々としても、どうしてそうなるかが知りたいので、必ずしもいいものを作りたいわけではないのですが、ただああいう SVM などでうまく予測ができて、結果を整理していくと、その中から物理法則が見つかってくるということは十分あると思っています。まず整理をして、いいものを見つけるということも大事なのだと思います。
質問者：サポートベクトルを使った場合、たとえば融点では、どういう記述子が出てくるのでしょうか。
田中功：必ずしも説明できるような、なるほどという記述子ではないですね。ですからこの中の、たとえば属の二乗を足し合わせたようなもの、あるいは電気陰性度を単に足したもの、あるいはそれを掛けたものなどが出てきます。その物理的意味は、今の段階では諦めています。主因子解析のような手法を使えば、どのファクターの重みが大きいかといったことはわかると思います。ただ、そういう方法では、精度はそこまでは上がらないかもしれません。

“Data Driven Combinatorial Experimentation and Trends in the Materials Genome Initiative”

竹内一郎（メリーランド大学）

我々が研究を進めているコンビナトリアル実験について、新材料の探索ということで、ここ 10 年以上手をつけているいろいろなテーマについてお話し、それとつなげて最近、米国でよく話に出てくるマテリアルゲノムとは一体何かなどを話したいと思います。

最初に、簡単にコンビナトリアルの説明をします。1 つの実験に関して、まとめてたくさんの数の組成の違ったものをライブラリとして作ります。こういう実験をやることによって、今までわからなかったような情報を得ることで新材料の発見に結び付けていく、あるいは、組成 - 物性相関がわかるだけでも随分得られることがあるわけです。一番ボトルネックとなるのが、いかに並列させて評価するかです。最初に無機の材料でコンビナトリアルが始まったときに人気があったのは、蛍光材料の探索ですが、それはなぜかというと、やはりスクリーニングがすごく大変なわけですが、蛍光材料に関しては励起させてどこが光るか見ればいいだけということで、定量的な情報は得られませんが、新しいものもいろいろ見つかりました。

最近、我々が力を入れているのは、いわゆるスマート材料というもので、たとえばピエゾ材料や磁歪材料など磁性材料にいえることですが、その構造の相転移や構造変換、あるいはマルテンサイト変態が起こるようなところで機能が増幅されているので、実は X 線の解析をするだけでスクリーニングができて、どんどんライブラリを作って測れるという点で適しています。主に薄膜のコンポジションスプレッド法を使っているのですが、基本的にはたとえば図 1 に示したように Ni（青）、Mn（赤）、Al（緑）の違った 3 つの元素をまとめてスパッタします。普通の膜は均一に作りたいわけですが、逆にここはわざと組成に勾配をつけるように連続的に膜をつけるわけです。

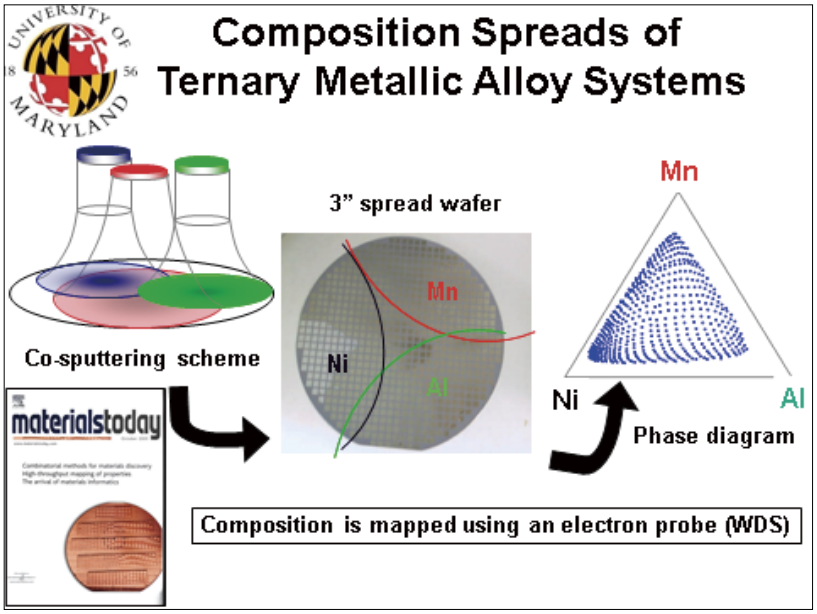


図 1

うまくいけば、3 つの元素に関して組成比を 0 ~ 100% の範囲で変えられるわけで、1 点ずつ測ってやると組成として 3 元相図にマッピングできる。

構造の解析については、シンクロトロンに行けばフラックスが数桁高く、スループットも上がるのでコンビナトリアルには非常に向いているということで昔から進めてきたわけですが、最近になってうまくシンクトロンのビームラインの研究者と協力できるようになりつつあります。というのは、ビームラインに普通に実験に行くと、だいたい 1 週間時間をもらえれば、3 ー 4 日はセットアップするだけで、下手をするとほとんどデータを取れずに帰ってくるというのが常識だったのですが、最近になって、我々は SLAC でやっているのですが、ソフトウェア、ハードウェアも全部サポートしてくれるようになりました。コンビナトリアルの実験に対する風向きが米国では少し変わってきたといえると思います。

データの数も 10 年前頃にやっていたときは、1 週間でうまくいって 1 つライブラリが測れる程度だったのですが、今では 1 つのライブラリを 2 時間で測れるということで、次から次へとデータが得られます。ここで問題になるのは、すごいスピードで取っていったデータをいかにうまく解釈し、解析するかということで、下手をするとデータに埋もれて何をやっていたのかわからなくなることもあるわけです。そこで、我々はデータマネジメント、データアナリシスについて、自分たちでソフトウェアのコードを書いているわけです。

では一体どんな実験をやっているかということですが、普通、磁歪材料はターフェノール D など希土類が入っているので、まず 1 つにその希土類をなくしていかなければいけない。もう 1 つは、最近 heterogeneous magnetostriction というまったく違った効果で磁歪が出る材料があるのではないかという理論が提唱されていたため、ではその実証も加えて一緒にやることにして、鉄コバルトの系を調べてみることにしました。うまく熱処理することによって組成を変えると磁歪が普通に測れるものの 3 倍ぐらいの値が現れることがわかりました。これは fcc と bcc の違った構造を持った層が封じ込められるようになったことによって現れた効果ですが、もちろんコンビナトリアルをやっているということで、組成が変われるということが強みで、これができなければ見つけられなかったといえます。

もう 1 つは昔からやっている実験で、いわゆる非鉛ピエゾ材料を見つけようという話で、よく調べられているビスマスフェライトの A サイトにシステマティックに置換していくことによって、結晶相境界の誘電率も上がり、もちろんピエゾ係数も 2 倍になるということがわかったわけです。最近、もう少し構造の細かいことを調べており、図 2 は電顕の絵ですが、元素の位置に変調がかかっており、ひずみが入っている。つまり最近、強誘電の分野でよく出てくる話ですが、フレクソエレクトリック効果がかかっていることも関係して、結晶相境界で物性の増幅が見られるということがわかりました。これもやはりシステマティックに物性が測れるというコンビナトリアル実験の強みが見られる世入れだと思います。

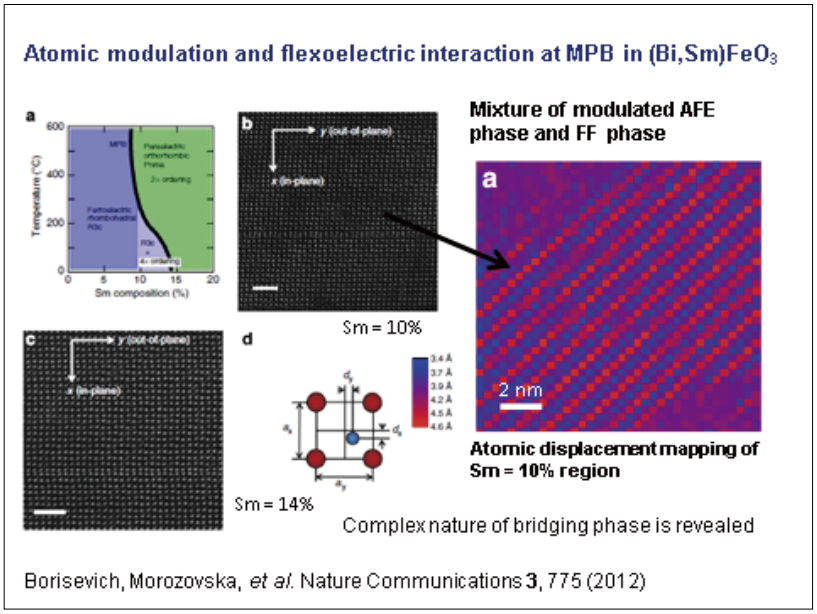


図 2

こういったわけで X 線のデータを次から次にとるわけですが、1 つのウエハにつき 300 ー 400 とると、とても 1 つずつ調べるわけにはいかず、まとめたらどうだろうということで、データの可視化について力を入れてやっています（図 3）。

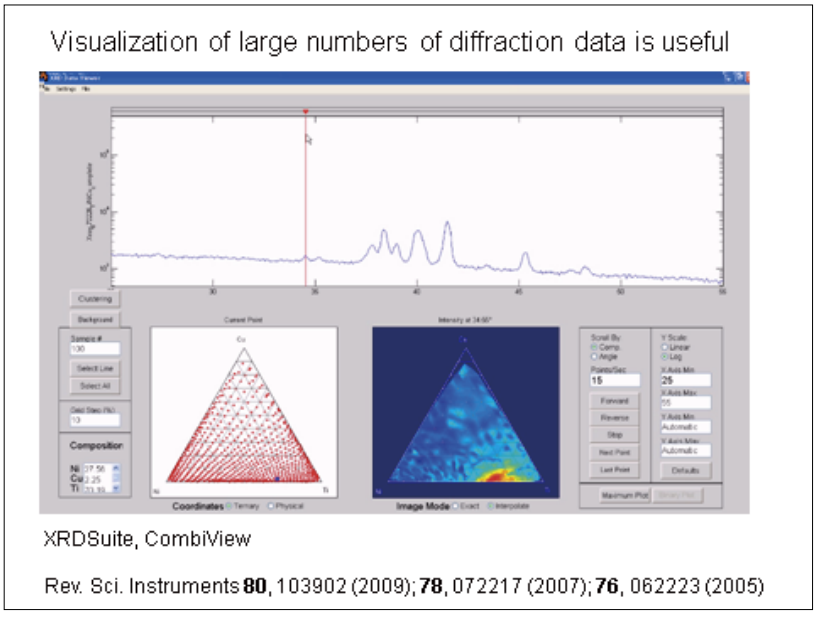


図 3

他にも似たような実験をやっている方がいますので、ソフトウェアを提供して、またフィードバックをかけてソースもオープンにして、同じ分野の他の方とも一緒に研究を進めています。

データがものすごい量で出るということで、たとえば X 線のデータでマシンラーニングをやると、結晶構造の分布が二元相、三元相などのデータを見ると、すぐに分かれてく

るのが見えてきます。それはクラスタリングをやることによって、この組成のこの地域ではある 1 つのパターンの X 線データが見えて、別のところではまた別のパターンが見えるということです。いろいろなマシンラーニングの方法が応用できます。ここで一つ面白いのは、X 線のデータをとっているわけですが、実はこれは完全に統計的な話で、格子情報が全くなくてもクラスタリングができ、全く構造の情報がない場合でも相の分離と分布が調べられるわけです。この例は鉄、パラジウム、ガリウムの三元系を見たときです。

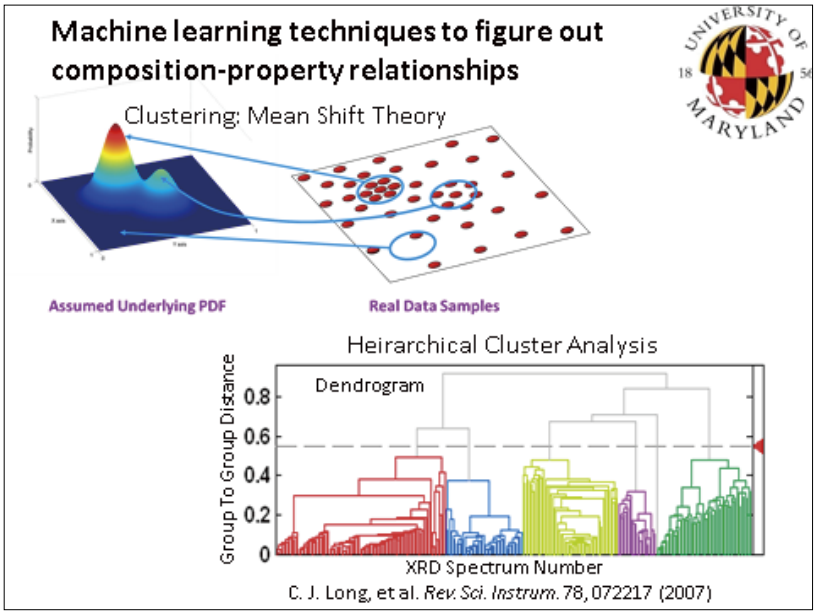


図 4

このようにコンビナトリアル実験は膨大な量のデータが出るということで、最終的にはコンビナトリアルデータを使ってデータベースを作るというふうにもっていかなければいけないと思いますが、とりあえずデータと存在する ICSD などのデータを一緒にして解析できないかといったことを考えてやっています。図 5 左上の挿入図にある 1 点 1 点が X 線回折のデータを表しており、これはそのミーンシフトというデータ解析のテクニックを使って分離した結果ですが、まずコンビナトリアルライブラリを作って、シンクロトロンで X 線回折のデータをとります。ICSD からもエントリーを引出し、シミュレートした回折の結果を出す。それらを一緒にして、マシンラーニングをかけることによって、もっと信頼性の高い結果を得られることがわかりました。

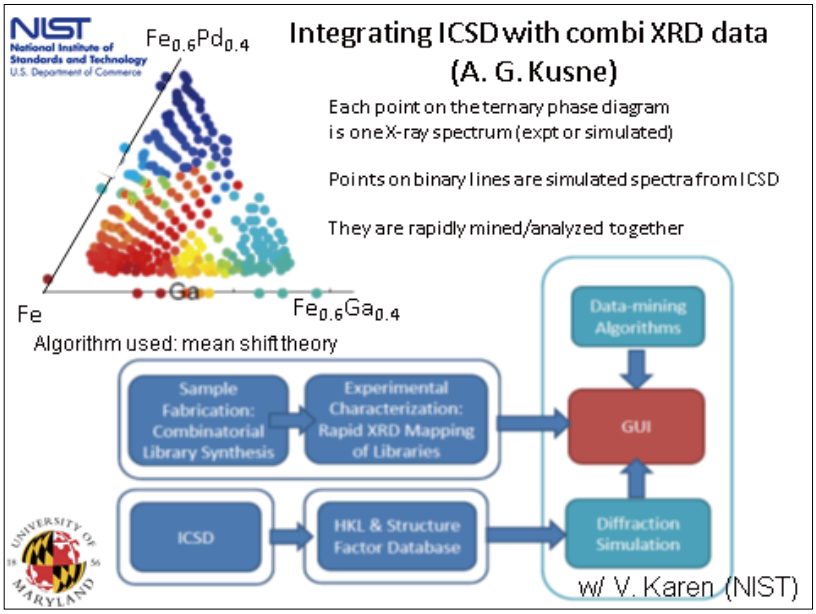


図 5

今までのコンビナトリアル実験の一番のチャレンジは、いかに迅速に評価できるかであったわけですが、これがある程度、たとえばX線の場合、シンクロトロンがうまく使えるようになって、次のチャレンジは、やはりアナリシスに時間がかかるということです。ここでは結晶構造の分布をダイナミックにその場で測りながら計算できないか、解析できないかということを考えた例です。このようなことが可能になり、解析にも迅速化できることがわかりました（図 6）。コンビナトリアルの研究方法のエボレーションについて簡単にまとめたものが図 6 です。

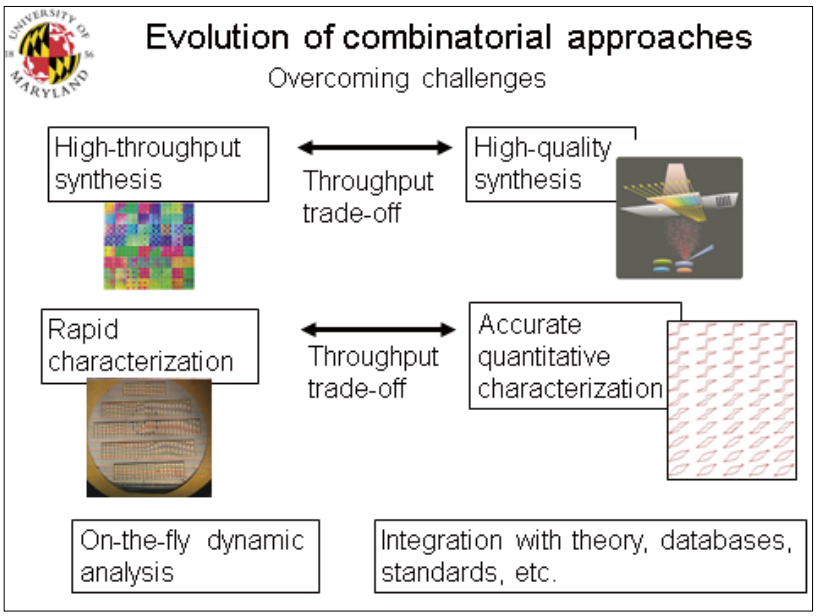


図 6

最初のチャレンジは、一般的によく言われる試料をたくさん作っているということで、

材料の質の面で妥協しているのではないかと言われていたことです。ここで高品質でコンビナトリアルライブラリができないかと、知京さんや鯉沼先生などが開発された手法が出てきました。この手法では１つ１つエピタキシャル成長をしながら作っており、一度に調べられる相図の領域が限られているわけですが、それでも高品質の材料で実験ができるというのがアドバンテージです。そして次のチャレンジは、やはり迅速な評価方法です。しかも定量的な情報が得られなければ、あまり意味がないということで、それが今でも新しい実験を始めるたびに苦心するところですが、最近では解析もその場で迅速にできないかというところまできたわけです。そして今までももちろん理論の予測を元に、それをガイドにして実験を進め、たとえばバリデートできるとか、本当に新しい材料なのかどうか、そういう話ができるように持っていったわけですが、これからもっとアクティブに理論の計算と一緒にコンビナトリアルの実験を進めていくべきだと我々は考えています。

続けてマテリアルゲノムの話ですが、米国は最近製造業の力がなくなってくるのではないかとよく言われており、これはオバマ大統領が打ち出したのですが、Accelerated Innovation あるいは発見からエンドユーザーまで持っていく時間をいかに短縮できるか、その辺を考えなさいということで、複数のエージェンシーが一緒になってファンディングを出すことになっていて、それが Materials Genome Initiative（MGI）というものです。MGI の三本柱をみると、実験方法のアドバンスも重要であるし、データベースももちろん取り入れなければいけない。それで、材料特性を予測する計算技術を進めるのであれば、それに対する実験技術の対応はどうあるべきかと考えると、その答えはやはりコンビナトリアル実験であると我々は議論します。

MGI に関しては、次から次へとワークショップが開かれつつあり、私が関連しているものでもすでに 3 つか 4 つありました。そこでマテリアルゲノムとは一体何かという話になると、「私がやっているのはマテリアルゲノムだけれども、あなたのものは違います」という方がいらっしゃいます。いろいろな定義の仕方があると思いますが、2 つに大きく分かれていて、その 1 つはいわゆるマテリアルデザイン、つまり計算をもっと進めてそれを実験で実証するというのが 1 つです。そしてもう 1 つはここ 10 年ぐらいやられている Integrated Computational Materials Engineering（ICME）という概念です。これは冶金の分野でおもににやられているのですが、そこでの計算の話は、たとえば有限要素法や固体力学などが含まれます。これは First Principles を使ってマテリアルデザインを進めるのとは少し異なり、材料の微細構造を計算科学を使って理解し、新しい高性能合金などを作るようにもっていくという話のようです。

いま米国は財政がすごく問題になっており、たとえば今年の会計年度は 10 月から始まってもう半分近く過ぎそうですが、まだ予算がついていない。NSF は実は財政に問題があると、一番しわ寄せが行きやすいところなので、予算も大きく削減されて、MGI 関係のグラントについてもおそらく応募が 300 件ほどあったと言われている中で 20 チームしか選ばれず、これではほとんど意味がないではないかと言われる方もいらっしゃいます。いつもの usual suspects（札付き）が出てきて、うまくチーミングができて、いいプロポーザルが書けていれば、ゲノムであろうがなかろうが採択されたのではないか、とも言われています。他のエージェンシーもグラントを出しており、たとえば DOE は、データとソ

フトウェアの開発に力を入れているグループにグラントをつけているようです。

マテリアルデザインの世界では、今までなかったような新しい方法、たとえば材料の特性の計算予測に機械学習などを組み入れて、調べていこうという試みがあります。私が共同研究を進めている、デューク大の Curtarolo のグループでは、太陽電池材料、トポロジカル絶縁体、熱電材料、圧電材料などいろいろな材料に対して電子構造などを計算してデータベースを作っています。図 7 は遷移金属の二元合金を全部並べて、計算と実験の結果がが合っているどうかを比べたものです。緑色は合っていたことを表しており、これによっていかに計算技術が進んできたかを知らしめる図です。

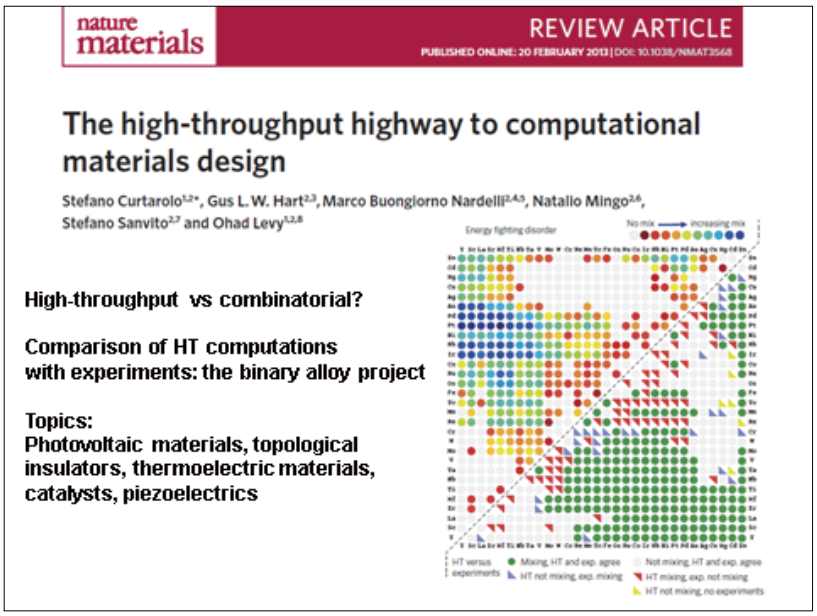


図 7

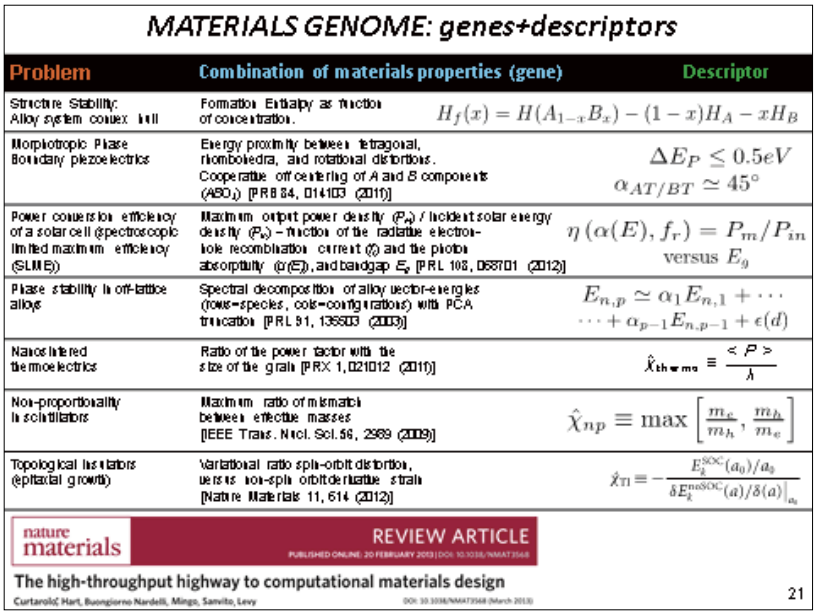


図 8

図 8 はそこで用いられた記述子ですが、たとえば熱電材料で一番重要な情報はパワーファクターと粒の大きさであるといったことを決めて、どんどん計算していくわけです。その結果を保存し、パブリックにアクセスできるような情報をつくっているわけです（図 9）。

そしてデータベースの話ですが、計算してつくるデータベース、コンピュテーショナルデータベースに関しては、たとえばデューク大では、ICSD をもとに計算を行っており、ある意味実験結果をもとにした計算を使っています。

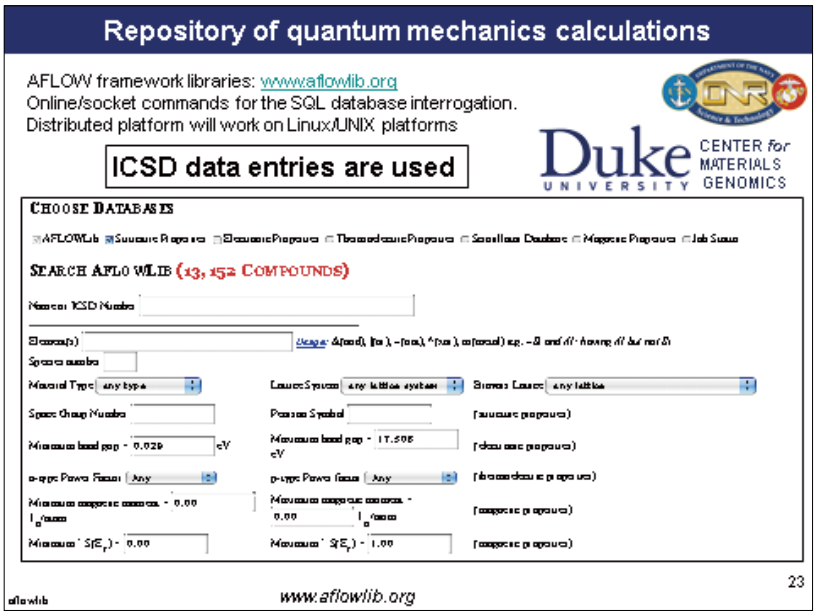


図 9

ここ 10 ー 20 年の間、計算科学で一番よく調べられているのは電子構造であって、これが面白ければ熱力学的安定性はどうでもいいような話もときどきある。デューク大の Curtarolo たちは、熱力学的安定性が一番重要なところだと言っており、きちんとすべてのことを考えてデータベースを作ることを進めなければいけないと主張しています。

その次のステップが非常に重要で面白くなってくるわけですが、計算されたものからデータマイニングをして、今までなかった物性の相関を取り出していかなければいけないわけです。たとえば二元合金のプロジェクトに関しては計算で予測されたもののうち数十個は実験でベリファイされたそうです。最近、トポロジカル絶縁体が流行っているのですが、それについても何十から何百もの合金について計算が行われ、最近になって 1 つ、2 つ、作ってみたら確かにトポロジカル絶縁体になっているようだという報告があるようです。

ここで、実験のデータベースの話をしたいと思います。実は、実験のデータベースはほとんどない場合が多いわけです。本当に知りたいもののデータベースはバラバラに存在し、しかもパブリックにアクセスできるものはほとんどない。しかし、実験家の我々の立場からみても、実験データのデータベースをデータマイニングすると重要な相関のデータが出るはずで、それが一番やりたいわけです。ICSD は 15 万エントリーあって、結晶構造の

話で非常に重要なわけですが、その他の例としてすごく重要なデータベースは、やはりNIMSのMatNaviだと思います。これは文献を1つ1つ拾ってきて、データベースが作られているわけで、世界中他にパブリックにアクセスできるようなデータベースは存在しないと思っています。最近、米国でも、ASM（米国材料学会）を中心にコンピュテーショナルマテリアルデータネットワークの名のもとで、計算予測の結果がが半分ぐらいになってしまうのかもしれませんが、データベースを作ろうという試みも出てきています。

たとえば5、6年前に我々がやった話ですが、酸化物の磁性材料について実験のデータベース作り、そのマイニングをやりたいが、そんなデータベースがないから、自分たちでデータベースを作ろうとことで、学生を毎日図書館に送ってランドルト・ベルンシュタインの本を3冊ぐらい引出してきて、データベースを作らせたわけです。時間はかかったのですが、1年ぐらいかけてペロブスカイトとスピネルに関しては、すべて引き出しができ便利なレファレンステーブルができました。いつも実験するときに、この材料は今まで使われたことがあるのかをこのデータベースで調べるわけです。

やはり素晴らしいと思うのはMatNaviなのですが、たとえば超伝導のデータベースでエントリーが2万9000個ぐらい入っています。すごくいいのはダウンロードできることで、うまくいけばそのままデータマイニングできるレベルまで持っていけます。これを何とか使えないか、もっとシステマティックにできないかと考えて、いま進めているところです。

もう1つデータベースを作る話ですが、希土類なしの永久磁石を見つけなさいという大きいプロジェクトをエネルギー省からいただいており、20人ぐらいでやっているのですが、最初に集まって、ではどの系を調べるかという話に最初になったときに「では、この系はどうか」と言うと、必ず誰かが「それは私がやったよ」、「それは30年前にロシアでやったよ」、「それは日本で調べられたよ」と、そういううわさ話ばかりで、実はどこかに包括的なデータベースがないのか、ということになります。それがなければ前と同じことを調べるだけだから、では、実存する文献すべてをもとにデータベースを作らなければという話になりました。

それを考えたときに、MatNaviのように人手で1つずつ調べるのではなく、マシンリーディング的なものがあり、文献のスキャンも自動化できるし、それをうまく使ってやったらどうだという話が出てきました。他の分野ではもう進んでいるらしいのですが、材料の分野ではまだやられていないわけです。文献を次々にスキャンしていくという話になると、コピーライトの問題も出てくるわけですが、セマンティックオントロジーについて調べるわけで、たとえば英語の文章でなくとも中国語でも日本語でも、1つのアルゴリズムで解析が可能です。たとえば磁性の話であればキュリー温度、磁化、保磁力について調べさせるわけですが、パイロットプロジェクトで500本の文献に対してこの3つの物性を見つけさせると、うまい具合に表ができることがわかりました。そこで問題になったのは、実は論文を見ると一番重要な情報は文になっておらず、グラフの中に入っているわけで、グラフをコンピュータに解釈させてテーブルに変える必要がありますが、そこまではまだやられていない。そういう試みがあるという話も聞きいてはいるのですが、ものになるところまではいっていないようです。

すでに100年以上ものマテリアルサイエンスの文献が存在するわけですが、それらすべてに関してマシンリーディングを用いデータベースをつくることができれば、それこそ

マテリアルゲノムではないかと我々は考えています。最近の文献は PDF になっていて、すべてオプティカルキャラクターリコグニション（OCR）でうまいこと文字データが読めるようになっているらしいのですが、10 年ぐらい前になると、変換してクリーニングのプロセスをやらないと読めないという話で、30 年、40 年前の論文などは全部スキャンされて人間が見ることはできるのですが、そこからリーディングするのはちょっと不可能だということです。だからアイデアとしてはいいものかもしれませんが、課題はいろいろあります。

そういうわけで、ハイスループットの計算科学が進みつつあり、、また我々はハイスループット実験を進めてきたというわけですので、ここで一番いいのはこれを結合させれば、スループットも上がるし効果的であろうということで、図 10 のようにたくさんの方法を開発しています

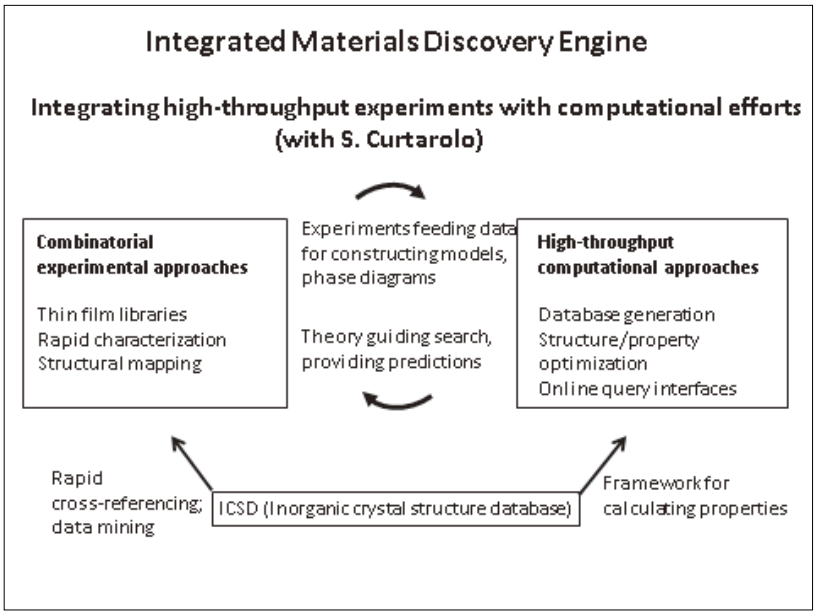


図 10

たとえば理論で予測された組成はピンポイントで出ることもあるけれども、そうでない場合も多いわけで、コンビナトリアルを使ってその近辺の組成を振ることが重要なわけです。また、計算する上でモデルを作るのにも、実験のデータがあればあるほどうまくいくわけです。パスが 2 つあり、1 つは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ってもらい、それから新しく何か予測してもらい、そこでまたコンビナトリアル実験を行う。また、それと逆に、最初から予測あるいはデータベースを計算していただき、その材料に関してコンビナトリアル実験をやって、その結果を元に、また計算を進めるというのがもう 1 つのパスです。こういったシステマティックなアプローチをを我々は、マテリアルディスカバリーエンジンと呼んでいます。

最後に 1 つ、材料の未来は何かということで、よく話が出るわけですが、ポストナノというわけではないかもしれませんが、次は何かというと meso らしいのです。Meso と

いうのは、もちろん 80 年代、90 年代に物性物理でたくさんやられたわけですが、そういうものとは話が違って、meso のスケールで今までできなかったことを、いかにコントロールできるようにするか、たとえば欠陥をいかにコントロールするか、それによって、マニュファクチャリングにつなげていくというのが次世代のマテリアルサイエンスのメインのテーマになるはずだということを、DOE のレポート“From Quanta to the Continuum: Opportunities for Mesoscale Science”では唱えています。

【質疑応答】
質問者：今日の最初のほうで津田先生も言われていた RDF といったデータベースは、基本的にナショナリティを超えて海外の人でも使えるものなのでしょうか。質問の意味するところは、今回のこの会議は日本国が非常にそういった分野で遅れているので、日本の国際競争力を高めるという意味でやろうとしていますね。一方、海外で行われているアクティビティがもしも国際的に開かれているときに、国際協力を行って全世界的な規模で物事を進めるということもあるかもしれないなと。
竹内：データベースというのは実はすごく泥臭いところがあり、昔からビジネスになっている面があるためオープンにできない。もちろんコンビナトリアル実験などは触媒でもちろんそうなのですが、産業界で米国でもあちこちやっているわけですが、その結果はどうしても出せない。結果自体が知財であるわけで、うまくいったものもうまくいかなかったものに関しても、絶対にオープンにできないとよく言われます。
津田：RDF というのはフォーマットで書き方の問題なのですが、今までそういうデータを書き表す標準がなかったのです。それが 2007 年にできたという話で、それに準じてデータを用意しておくと、データ同士の統合がやりやすくなるということです。文献に載っているようなものはもうオープンな情報ですよね。そういうものはおそらく RDF で共有されることになると思います。ポイントは、自分の作っているデータとパブリックなデータを統合して解析することが目的なので、たとえ自分のところで隠して持っている場合でも、RDF 化しておくと他とすぐ混ぜられるというところだと思います。

寺倉：データベースの国際化に関して、田中先生からご意見ありませんか。データそのものではなく、記述子が売りなのだと、以前おっしゃっていませんでしたか。
田中功：おそらく米国の計算した結果、MIT にしても、デューク大にしても、全部オープンになっていると思います。ただ記述子の部分だけを自分たちが持っていて、それはオープンにしないのだというのが彼らの言い方ですね。それは 1 つのやり方かもしれないと思います。データはある程度シェアしないと、いい質のものにもならないと思いますし、彼らがオープンにしているときに、こちらだけ隠すというのはあまりいい戦略ではないと思います。

「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）

私は独立行政法人　物質・材料研究機構でのユニット長としてコンビナトリアル手法を使った材料開発を進める傍ら、独立行政法人　物質・材料研究機構発ベンチャーである株式会社コメットの CTO もしています。この株式会社 COMET はコンビナトリアルツールを使って材料の委託開発をしている会社で、実際のビジネスに触れる機会があります。そこで今日は、コンビナトリアルツールを研究開発の現場で使用している立場と実際のビジネスの観点から見た 2 つの側面をご紹介したいと思います。

まず、コンビナトリアルツールが出てくる時代背景についてお話します。1 つ目は、企業では製品開発の前段階で､非常に多様な材料開発を短時間で行う必要があります。実際、我々の研究開発でも委託ビジネスでも、いろいろな材料開発のスクリーニングを行っています。これは基本材料に材料 A という材料を混ぜたらどうなるのか、B を入れたどうなるのか、そのときの結果が早く知りたいというニーズが背景にあります。エレクトロニクス分野はその典型だと思います。これまでのシリコンを中心とした LSI の中でも、最近は High-k、メタルゲートといった新しい材料が登場するようになりました。これからのトンネル FET や新しいタイプの三次元デバイスでも新材料が必要になってきます。

実は昨年 9 月まで NEDO プロジェクトで、Fin 型のフラッシュのメモリのための新材料と実デバイス化を企業、独法とやっていました。材料探索が最初にあるわけですが、そのときに最初に使ったツール MatNavi です。この中で基本になりそうな材料を選び、次に第一原理計算でバンド計算、そしてどのような機構でチャージトラップが行われるかを理解した後に、コンビツールで材料組成をスクリーニングを行い、最終的には産総研に持ち込んで Fin 型のフラッシュに仕上げました。その技術は最終的には企業に技術移転をしようということで現在も企業側で開発が進んでいます。コンビツールというのも 1 つのツールで、これですべてができるわけではありませんので、出発点が重要になると思います。また、トランジスタ単体をとってみても、かつてのシリコン、SiO_2、多結晶シリコンという時代はもう過ぎ、いまやグラフェンやゲルマのチャネル、High-k もいろいろ候補があがっていますし、これからはダイレクト High-k だという話もあります。メタルゲートにいたっては、非常に多くの材料がある。こうした新材料の探索、組成制御、成長条件の探索、界面の制御などいろいろなことを短時間で行っていく必要があります。

もう 1 つの時代背景は、R&D のモデルが変わってきたことがあります。1980 年代前半は、まさに中央研究所の時代でした。日本には日立中央研究所、東芝総合研究所、NEC 基礎研究所、海外では IBM のワトソン研究所、ベル研究所など名だたる中央研究所があり、そこでシーズが生まれ、それが開発され、実際に生産部門に移っていった過程はご存知だと思います。こういう中央研究所の時代は、非常に多様な人材が 1 カ所に集約されており、目的は共有されているため非常にうまいコミュニケーションをとっていたと思います。事実、東芝や NEC などでは活発な基礎研究が行われており、国際的な発表も多かったと思います。ただし短所として、シーズ研究が多くなりますので維持コストがかかる。それからシーズから生産までに時間がかかるという問題点があったと思います。実用化されなかったシーズも多くあったとおもいます。

さらに、収益に対する開発のコストが非常に上がってきたという事実があります。これ

は半導体の例ですが、たとえば 2030 年ぐらいには会社の開発費と収益が同じになるといわれており、いわゆる従来型の開発では産業自体が成り立たなくなるという危惧があります。したがって、R&D のコストをできるだけ下げたいという背景があります。そこで考えられたのが中央研究所をなくして、大学あるいは国研で生まれたシーズをうまく産業界に持ってくるということが期待されたわけですが、現実にはなかなかうまく行っていない。シーズを実際のものにしていくためには、いろいろなギャップを超えていく必要があります。いわゆる死の谷といわれているところですが、大学の研究者と企業の開発者の間のコミュニケーションがうまくいかない、意識が違う、いろいろな面で問題が発生している。たしかに「中央研究所の終焉の時代」という本が出ましたが、では、このモデルが必ずしもうまくいっているかというと、必ずしもそうではないと思います。

理由の一つは大学、独法などで生まれたシーズと企業が求める技術の間に、テクノロジーギャップが存在するためです。例えば、大学では 10mm 角の試料に材料を堆積させて評価をした結果、有望な材料であることがわかったが、組成や条件などが詰め切れていない。これを企業で追試しようとするとその研究開発ラインが企業にない。このギャップを埋めるためにコンビナトリアルという方法が非常にうまくマッチしています。

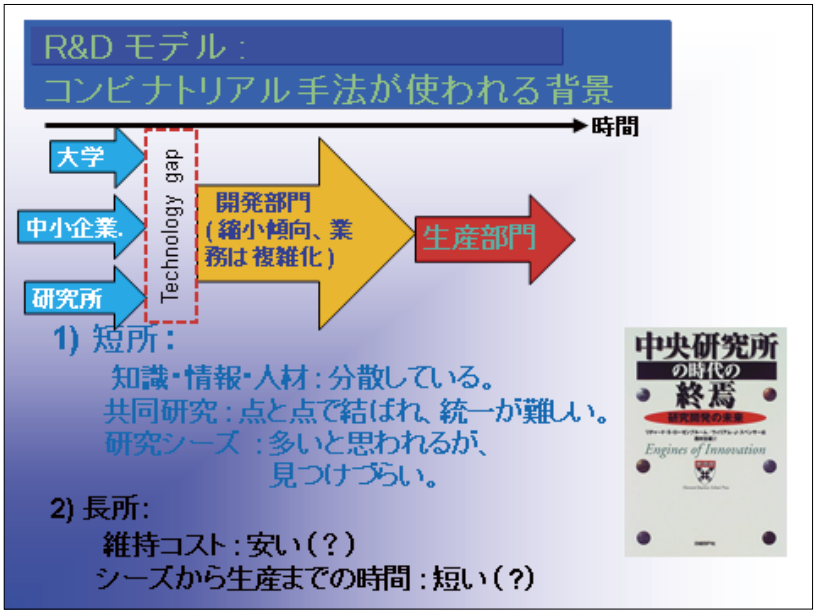


図 1

また、別の側面として、細分化してきた学問体系から分野融合が進んできたという背景もあります。たとえば、かつての大学の学部は、金属工学科、電気工学科、機械、化学など、それぞれに対応する代表的な産業があり、たとえば私は電子通信学科の出身ですが、そこでやっていたのは回路と電磁気が中心であり、電子通信学科という名前と同じ学会が存在するという形で非常に垂直型の体系であったと思います。ところが最近では分野融合が進み、実にいろいろな人たちがいろいろなことをやっている。たとえば金属工学科や材料工学科でも光材料をやっていますし、半導体といっても金属工学を含む材料工学のセンスは必要になっている。東工大で酸化物トランジスタを開発された細野先生はもともとガラスの研究からスタートされたわけですが、いまは半導体関係、IGZO などを研究されて

いるというのも分野融合の例です。つまり、こういうことを可能にするようなツールが必要になってきているという時代背景があると思います。

実際にコンビナトリアルのツールがどう変遷をしてきたかですが、出発点はやはりコンビナトリアル・ケミストリーです。最初は有機合成を模倣して無材料合成もスタートしたのですが、やはり薄膜特有のスパッタなどの合成技術が必要であろうということで新しいツールが開発されました。これが移動式のマスクと基盤回転を使って三元相図をオートマティックに合成する方法で、2000 年ぐらいに開発し、いま現在でもこのツールが生きています（図 2）。

薄膜系のコンビナトリアル材料合成の歴史をみていくと、いくつかの点が明らかになってきます。まずコンビナトリアルの合成自体は 1960 年代に、もうなくなってしまった会社ですが RCA で行われたのが最初です。ハナックという人が最初にスパッタ法で自然にできる組成傾斜膜を利用して新材料を見つけようとしたのが最初です。発表したのが 1970 年ぐらいです。ただ、それから竹内先生、Xiang、東工大の鯉沼先生、川崎先生が実際にコンビを始めるまでに、実に長い時間がかかっています。これは材料を合成ができても、それを分析する手段がなかったことが理由と言われています。その後、1998 ー 99 年にかけて、このコンビナトリアルという方法が日米でほぼ同時にスタートしていくわけですが、その時代背景の中にはコンピュータを使った自動制御が可能になってきたということと、ICT 時代、つまりインターネットが普及し始めたということ、それからシミュレーションも含めたコンピューテーションが可能になってきたということがあります。こういったコンビナトリアルツールとインターネット、コンピューテーションを融合すれば新しい材料分野が創成できるのではないかという大きな期待があったわけです。したがって、その後、米国では NIST が中心となったコンビナトリアルのためのコンソーシアムができ、日本では旧科学技術庁が中心になって東工大、旧金材研、旧無機材研が参加し、COMET というプロジェクトが 1999 年にスタートしています。その後、このプロジェクトで、膜厚を一定に保ちつつ、3 元連続組成変化膜を自動的に合成するという方法を開発し、これがいま現在の私どものビジネスのツールになっています。

同時に米国ではコンビを実際にビジネスに展開しようということで Symyx という会社が生まれ、その後 Intermolecular という半導体材料に特化したコンビナトリアルツールとインフォマティクスの会社が 2007 年ぐらいに生まれています。これはテクニカルアドバイザーの 1 人がスタンフォード大の西義雄先生、インターモレキュラージャパンの社長が元半導体メーカー関係者ということで、半導体関係者が非常に多く関係している会社です。同時に今日ご紹介した我々の会社コンビナトリアルのコメットという会社もスタートし、現在に至っています。

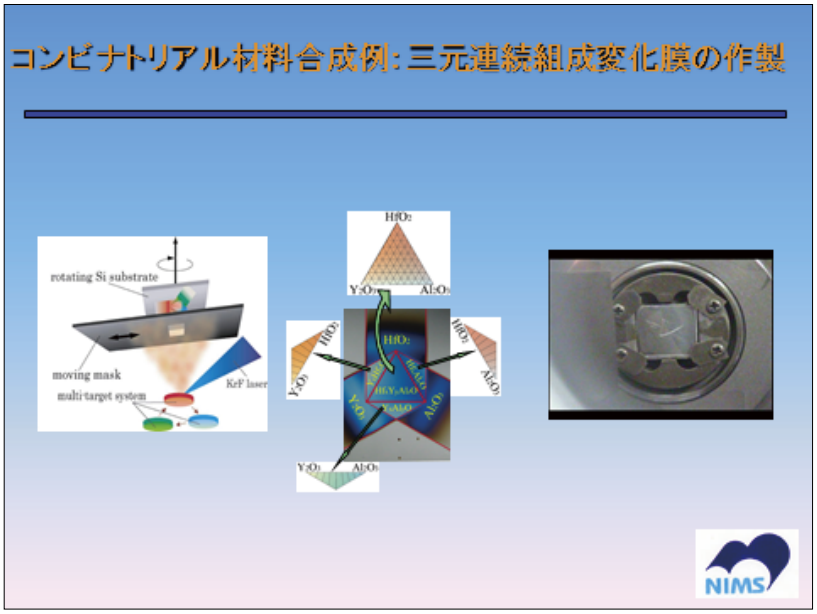


図 2

最初は、非常にプリミティブなスパッタ装置で非常に使いづらかったのですが、現在はグラフィカルユーザーインターフェースで誰でも使えるようなツールになっています。どの材料が堆積されているか、どのスパッタ銃が動いているかが実際にイラストレーションで表示されるので、今どのように状態になっているかが一目でわかります。成膜も基本的には材料を選択してスタートボタンを押せばコンビナトリアル合成が始まります。

制御系をコンピュータ化することによって、日時、条件材料がログとして自動記録されます。無線 LAN でネットワーク化し、画面を自分の机の PC で見ることも出来ますし、他のユーザとデータ共有もできます。評価装置については、現時点では X 線回折が共通のツールとして出ています。Bruker の D8 DISCOVER というシステムがあるのですが、多くのコンビナトリアルのユーザーが使っており、事実上のデファクトスタンダードになっています。こういったデータの共有化を図ることで将来の材料データ蓄積につながらないか期待しています。実際、我々のデータを竹内先生に送ってデータマイニングしてもらった事例もあります。

やはりコンビナトリアルのツールのよさは、最初は計算科学や MatNavi のようなデータベースを使って、この材料が基本になると理解して、実験計画をたて、それから合成するという過程で他のデータもシステマティックに得られる点にあります。これは別の側面もあります。コンビナトリアル材料合成では目的とする材料の発見だけでなく、系統的なデータが得られます。この中に本来の目的ではない新材料の発見や他の応用からみて有益な材料がみつかることがあります。コンビナトリアル材料開発のループを回すことによって、自分たちのゴールを達成すると同時に関連するデータも一緒に集められることになります（図 3)。これを私たちは「ディスカバリーループ」と呼んでいます。

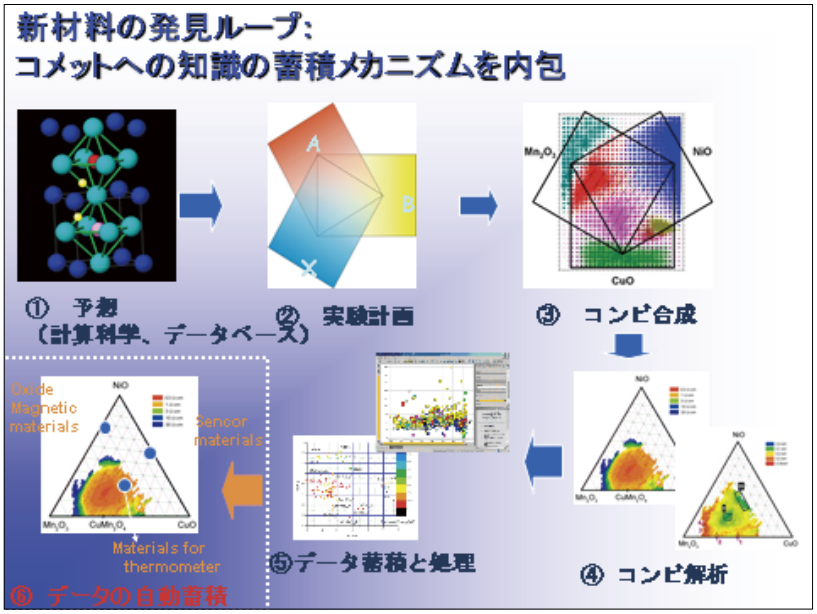


図 3

コンビナトリアルのビジネスを見ていくと、米国では Intermolecular 社が中心となって、実際の半導体生産現場に近い 300 ミリラインを使った材料のスクリーニングとインフォマティクスをやっており、かなりの成功を収めているようです。私どもコメットはどうなっているかというと、発足した時点ではリーマンショックなどもあり、大変だったのですが、その後は受託が増えて収益がどんどん伸びています。コメットはどういうところをやっているかというと、「大学の先生がこういう材料を見つけたけれども、その材料の中のこの材料を他の材料に変えたいが、どうすればいいか。そのためのシステマティックなデータを実際にとりたい」というニーズに対して、ビジネスを展開している。つまり、テクノロジーギャップを埋めるためのツールとしてコンビを使っているのが現状です。

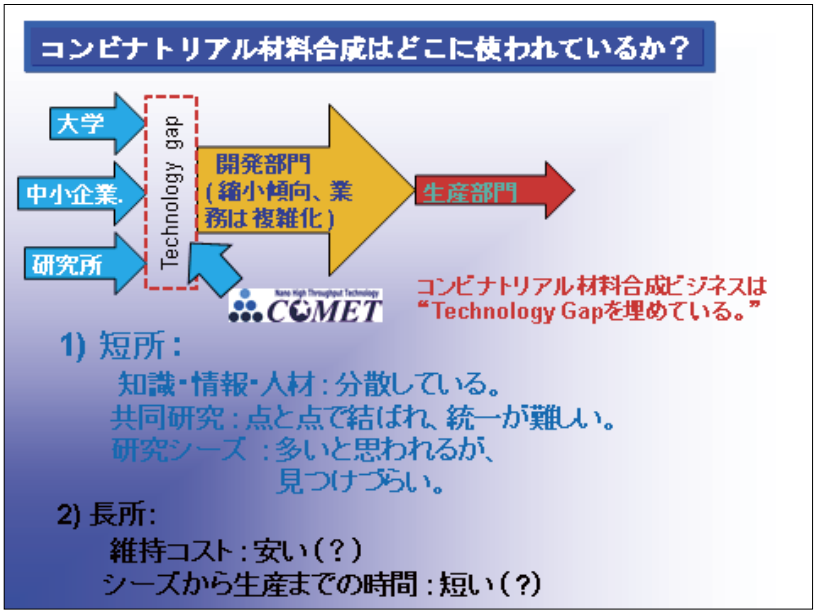


図 4

まとめとして、2003 年当時、私たちはコンビナトリアルに対して非常に大きな期待を持っていました。すべてをネットワークし、データマイニングし、いろいろなことができるのだと思ったのですが、現在できているのはツール、とくに薄膜成長装置の関連とそのネットワーク化はできていますが、データマイニングはまだ不十分な段階です。しかし現在では､テクノロジーギャップを埋める手段としては非常に有効であり､プロセスパラメータの決定にも有効であることがわかってきました。

今後何が必要かというと、やはり材料データの蓄積です。材料はいろいろな側面を持っていますので、材料の多面性を扱っている材料データベース、使いやすい MatNavi のようなインターフェースをもち、既存のデータとコンビナトリアルデータとのリンクが必要と考えています。

【質疑応答】

質問者:今のテクノロジーギャップというのは、非常にぴったりした言い方だと思います。初めの誇張した表現がコンビナトリアルの普及をさまたげたのは、仕方がありませんね。もう 1 つこれを進めると、特許を出すときの実施例のサンプルとして使えるのではないでしょうか。ある意味では、一番これが面倒なのです。横の展開はやりたくない。1 個見つければ十分だと。大学でやるときは非常に慎重に、どこから見られても大丈夫なように特許を出すと、やっていられなくなってしまうのです。現場にいると諦めざるを得ないのかなと思っているのですが、こういう会社がどんどん出てきてくれると助かりますね。

質問者：いま 3 元でやっていますが、ニーズとしてたとえば 5 とか、6 とか、もっと高い次元の組み合わせも本当はやりたいのだけれども、現状はいろいろな制約があってできないという形なのでしょうか。

知京：実際に、特許の範囲を決める場合など、コンビナトリアル合成は非常に有効です。また、質問にありました擬似 4 元は可能ですが、たとえば 5 次元になったときの分布と

なると、当然ながら立体的になります。それを人間が直感的に理解できるかという話になるのです。つまりデータとしては取れます。たとえば5軸を書いて、それにマッピングするというのはコンピュータに入れればできるのですが、やはりコンビナトリアルのもう1つの点は、パッと見たときに、どこがいいかがわかるというところが重要であり、その点からすると3元が一番いいかなとおもっています。3元系を何層にもやって、多層な三元相図があると材料の範囲をカバーしつつ、直感的にも理解しやすい。たとえば温度を変えるか、Aという組成を変えるか。A、B、C、Dだとすれば、Dの温度だけは固定してABCは組成を変える。Dの温度を少し変えてABC組成を再度変える。これを組み合わせて見せたほうが会社の方も喜びます。

質問者：いまの三元連続組成の合成ですが、合成をさせていくとプロセスパラメータ、組成の情報を入れて合成していきますね。それに関するデータはどんどんログをとられていく。言ってみれば、データベースが自動的にできるわけです。今度は、そのデータを使ってy＝f（x）のようなモデルを作りますね。xがそれで適当かわかりませんが、ものを作っていくという意味でのプロセスパラメータという意味では使えると思うのです。どの程度の精度がモデルとしてあるかわかりませんが、もしある程度いいモデルができれば、そのモデルを使ってyを指定し、ある範囲を指定し、xをいろいろ変えて、変化させる範囲は自ずと制限があると思いますが、その装置を使って逆に合成させるということは当然できますよね。そういう場合に、データベースを作る道具としても使えるけれども、それでとったデータを使って、逆に目的の物性を満足するかもしれないところだけを狙って合成させることは、この装置でできると考えていいのでしょうか。
知京：たとえば三元相図の場合、最初はすべて0-100％でやりますが、この辺がよさそうだと思うと、ズーミングという機能があり、たとえば20ー40％だけのところをたとえばズームインという形で三角形に展開することができますので、そういう意味では目標とする周辺を見ていくことは十分可能です。
質問者：それはモデルを作らずに、特性がよさそうなところが見つかったので、そこを細かく見ていくという話ですよね。
知京:そうです。モデリングは今のところはかなり難しく今後の課題です。有機材料と違って無機材料は同じ材料でも使われる分野が多様だからです。現在はスタッフがいませんので、対応出来ていませんが、そういったモデルができれば面白いと思っています。
質問者：有機合成もロボットでできるのですが、データが溜まってきて構造と反応性のモデルを作ることで、このリアクタントとこの条件でこれぐらいの収率のものが、これぐらいの時間できるのではないかと。そういう合成ロボットをコントロールするモデルを作るというのがあったのですね。ですから、今は最適のところを見るだけかもしれませんが、データを取るだけでなく、発見的なことに使えるようになるのではないかと思ったのです。
知京：おっしゃるとおりです。もう1つは、実はこのコンビナトリアルを作ったシステマティックなデータは他の部分でも展開できることを、我々は経験したことがあります。high-k材料開発の観点から、高融点酸化物の3元を混ぜて完全にアモルファスの状態を作るという実験をやっていました。そのときの我々の目的は、アモルファスの誘電率の高い、いわゆるゲート酸化膜を探していたわけですが、同じ分野で実は耐熱合金、NIMSはニッケルアルミの超耐熱合金を作っており、そのコーティング材料として彼らが考えてい

たのがジルコニア系の非晶質酸化物だったのです。ハフニウムをジルコニウムに変えただけでアモルファスにするという目的は同じだったのです。ですから、違う分野のデータの参考になった場合もあります。コンビの目標は１つのゴールを達成するだけではなく、その周辺のシステマティックなデータも当然取れていて、それが他の分野にも展開できたということが大きかったとおもいます。

「『鉄鋼ゲノムの解明』について」
足立吉隆（鹿児島大学）

「鉄鋼ゲノムの解明」ということで、金属材料の中でも特に鉄鋼についてのデータを活用した材料開発のニーズと現状について、金属材料を開発していく上でどういったデータベースを作るべきか、データをどうやって実験的に求めていくかという観点から、モデリング、あるいは材料計算学の方々とどう協力していくかといった話をします。

最初は手法の開発についてです（図 1）。構造材料工学・科学における理想として、モデリング支援ハイスループット材料開発が必要ですが、研究をより早く商品化につなげなければいけない鉄鋼材料の分野においても、ますます認識されており、いかにモデリングをうまく使っていくかということが、ほかの材料と共通の課題となっています。さらに材料科学に基づいた材料信頼性の定量評価と向上、あるいは既存材料科学を包含するような一階層上の材料科学の構築を進めていきたいということです。

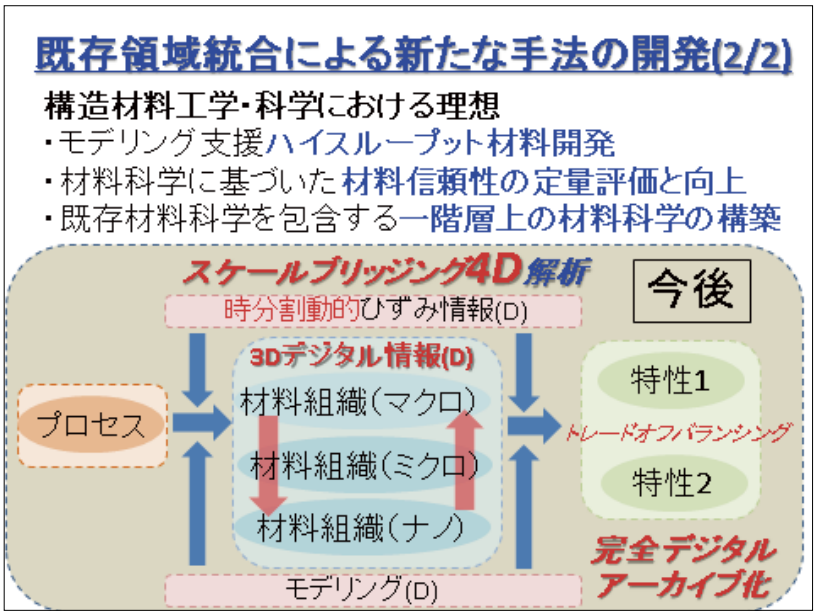


図 1

材料を作る際には、プロセスと特性があり、プロセスがいろいろな組織を通じて特性が発揮されます。この組織にはマクロ的、ミクロ的、ナノ的な特性をそれぞれ有する階層があります。従来、このプロセスと特性とを結びつけるために、組織を顕微鏡で観察したり、あるいは回折の方法で評価したりしていましたが、そのほとんどのデータがいまだにアナログでした。プロセスや特性などのデータがデジタルであるのに対してこの組織の部分がアナログであるために、なかなかモデリングにつなげられないという課題があるので、この組織のデータをいかにしてデジタル化していくかということが 1 つのキーです。

プロセスと特性を結びつけるときには、フェーズフィールド法などのモデリングがありますし、モデリングと特性を結びつけるためには、たとえばマイクロメカニクス、あるいは有限要素法等のモデリング手法があります。これらに加えまして、実験手法としてのプロセスと組織を評価するときに、組織を見ると同時に、ひずみをどのようにして評価する

かということが重要になってきます。従来はこれを静的に動かない条件下で捉えてきたところがあります。プロセスと特性を結びつける辺りのデータがアナログであったために、部分的にしかデジタルアーカイブ化されていないというのが金属材料分野の現状であると思われます。二次元で組織写真を撮って、その写真でデータベース化しているものは確かにあるのですが、そういうデータベースは、せいぜい結晶の粒径とか、あるいは体積率といったものの平均値でしか作られていないように思います。これをいかにしてデジタル化するかが重要です。

また､各スケールでの解析は行っていますが､そのスケール間の関係づけが十分でなかったように思います。そこで我々はいろいろな方々と協力しながら次のような取組を進めています｡プロセスと特性を結びつけるときに､まずこの組織を三次元で評価することによって、組織をデジタル化します。同時に、今までマルチスケール評価の関連づけが不十分であったことを反省し、スケールをブリッジングしながら評価することに特に注力しています。

従来は静的にとってきたひずみ情報につきましても、たとえばJ-PARCを使ってその場で実験ができるので、そういう時分割な動的ひずみ測定という最近の手法を種々取り入れたり、独自の手法を開発したりしながら進めてきています。これらを通じてスケールブリッジングに三次元、あるいはそれに時間軸を入れて四次元で評価するということで、完全デジタルアーカイブ化を目指して検討を進めています。これらを通じて、相反する特性をともに向上させるというトレードオフのバランシングにも注意しながら、このデジタル情報をどうやってここに反映させていくかということを現在検討しています。

これらを進めるにあたっては、次のようなグループで検討を進めています（図2)。

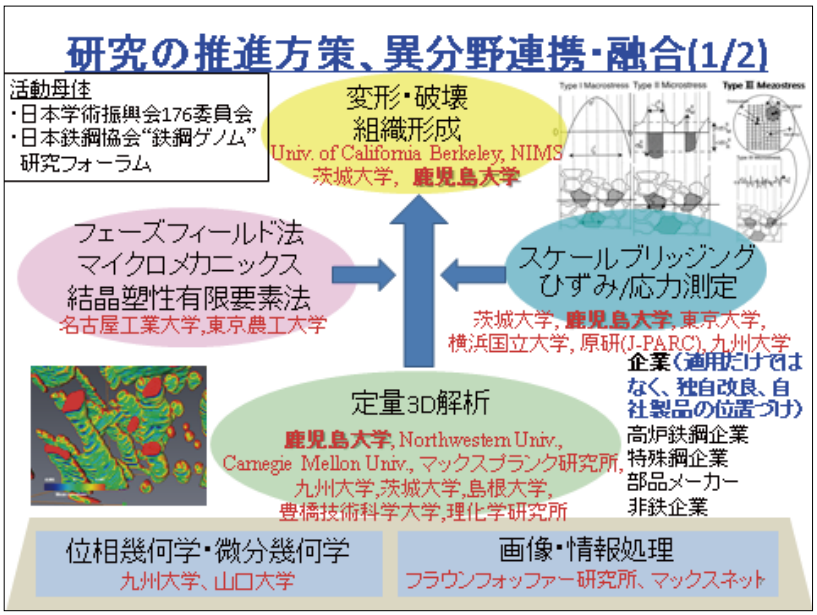


図2

定量の三次元解析については、私、鹿児島大学のほか、アメリカではマテリアルゲノムが進んでいる関係で､たとえばノースウエスタン大学やカーネギーメロン大学が強いです。

マックスプランクも大変強いです。様々な方法で三次元の解析をやっている九州大学、茨城大学、島根大学、豊橋技科大、理化学研究所とも連携しながら我々はこの三次元の解析を進めてきています。見るだけではなくていかに定量化するかということが重要ですので、たとえば九州大学、あるいは山口大学の数学の先生にも教えていただきながら形態の定量化を進めてきており、これを通じて組織をいかにデジタル化するか検討しています。これを変形・破壊や、あるいは組織の形成機構に結びつけるということで、あるいはNIMSやバークレーや茨城大学の先生方とも組んでやっています。ひずみ測定におきましては、J-PARCを活用し、たとえば茨城大学の友田先生たちと組んでいます。ここがデジタル情報になったので、モデリングの人たちと最初から連携できるようになってきています。そういう意味で、たとえばフェーズフィールド法、マイクロメカニクス、結晶塑性有限要素法を名古屋工業大学、具体的には小山先生、東京農工大学の山中先生と組んで、三次元イメージベースのモデリングで特性を予測したり、理解したりしようという取組みを進めています。

海外ではMRSで3Dの特別な取組みがありますし、TMSでも昨年始まったばかりです。また、先日、3Dのマテリアルサイエンスで初めての国際会議が開催されて、そこでも連携しながら話を進めていますし、日本ではたとえばJSPSが先日豊橋でISAEMという国際会議をやったときに、1つのセッションとして三次元ベースのモデリング等に焦点を当てた取組みを行っています。我々が研究を進めてきた母体は、たとえば日本学術振興会の176委員会で、ここでたとえば新日鉄住金の研究者の方、あるいは名古屋工業大学の先生方と一緒になって、デジタルアーカイブ化、標準化の検討を進めています。また、鉄鋼業界のほうでは「鉄鋼ゲノム」という研究フォーラムを進めておりまして、三次元、あるいは時分割測定とモデリングをどうやって組み合わせていくかという検討を4ー5年続けてきました。ただし、個々の材料や実験手法別の検討はなされていますが、実験手法、材料ごとの研究を横断的に取りまとめるような母体が国内にはないのです。3D/4D研究やデータのアーカイブ化は共通基盤技術であって、連携が研究の効率化に有効と思われるので、ぜひとも横断的に取りまとめるような研究母体を作っていければと願っています。

こういうことをやるときに、モデリングの方々、計算されている方々に、効率よく、しかも信頼性のあるデータを使っていただきたいということで、たとえば金属材料で組織をデジタル化するときに、三次元像が必要になってくることがあります。アメリカの空軍はロボットのアームを使って、研磨装置と光学顕微鏡との間を、試料を行ったり来たりさせながら磨いては見、磨いては見、ということを繰り返す装置を作っています（図3）。ただし、世界中でこれ1台しかありません。

図 3

そこで我々は私と社員 1 人の会社の社長と 2 人で全自動のシリアルセクショニング顕微鏡を開発しました。これが現在国内の大学や鉄鋼会社に複数台納入されています。実験室でいつでも誰でも三次元の組織図を得て、それをデジタル化するような土壌作りを進めてきました。これがいろいろなところに納品されていきますと、同じようなデータをいろいろなところでとることができるようになりますので、デジタル化したときの比較が比較的容易になるものと思われます。実際にこのようにして 500 μ m × 500 μ m × 200 μ m という組織をたった 1 日で評価することができます（図 4）。

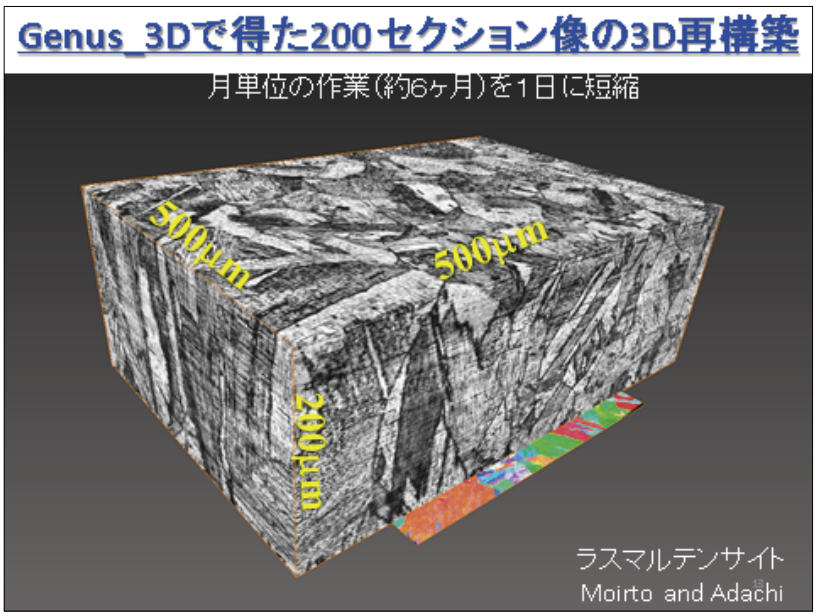


図 4

手でやると約 6 カ月かかっていたのが、たった 1 日でできますから、いろいろな材料を、

本当に手軽に三次元で評価できる時代になったと言えます。

三次元で評価する方法はいろいろあり（図 5）、まず中性子 / 放射光を使ったものはたとえば豊橋技科大の戸田先生たちが取り組まれ、たとえば耐熱鋼クリープのボイドを放射光で見ています。

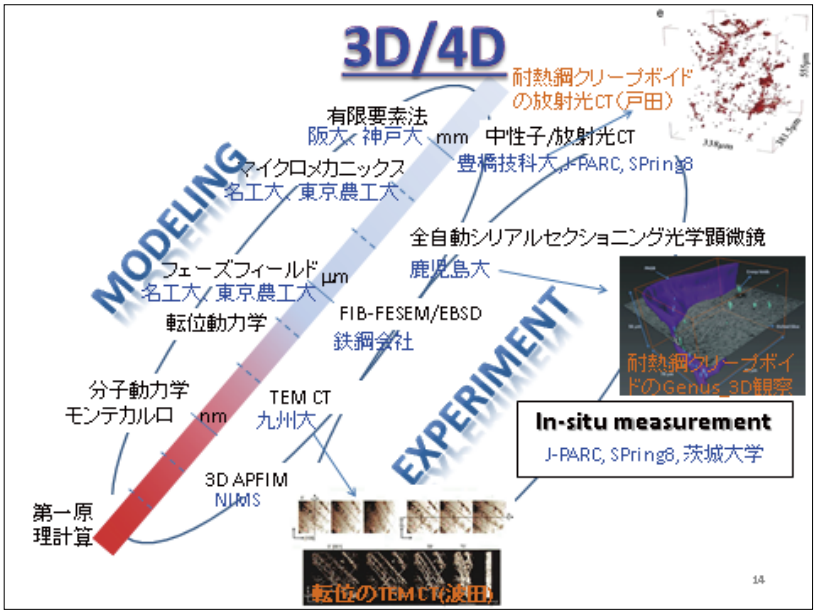


図 5

次は全自動セクショニング光学顕微鏡を使って組織を見ながらボイドを見るという方法で、放射光とセクショニングの方法をうまく組み合わせ、放射光で見たボイドと同じところをセクショニング方法でも見るという例です。もっと細かいところですと、たとえばイオンで研磨しながら電子顕微鏡で見る方法があります。透過型電子顕微鏡の中のトモグラフィーで見る方法は、九州大学に得意な人たちが多いです。このような格子欠陥を三次元で見る方法等が開発され、原子オーダーでは NIMS の宝野先生のグループが三次元像を評価することができます。こういういろいろな方法がやっと揃い、組織のデジタル化が進むようになってきたところです。

並行して、モデリングの方法でも、有限要素法から第一原理計算までもいろいろな方法があります。これまで別々にやってきて最後に組み合わせてきたのですが、これからは、同じ基盤に乗って研究開発の最初から議論しながら行ったり来たりし、計算とモデリングと実験を組み合わせながら材料開発を進めていく時代かと思われます。さらに J-PARC や SPring-8 を使った動的な評価ができるようになったので、ひずみ測定や組織発達の方法を動的に、また三次元で評価しながら、3D/4D で材料組織を評価していくことがやっとできるようになっています。

金属材料で重要な組織情報の例としては、大方の場合、これまでは体積率とか面積とか長さとかの平均値を評価してきました。二次元でもある程度はできるわけですが、信頼性が少し足りないので、それを三次元ですると、より精度の高いデータを得ることができま

す。また、位相幾何学的な特徴値（連結性、粒子数）につきましては二次元ではほとんど評価できていませんでした。これが三次元では曲率も使って評価できたり、今まで評価できなかったトポロジカル、あるいは Differential Geometry な特徴がやっと定量化できるようになりました。こういうものを使って、どのような形態のどこにどれだけのひずみや応力が集中するのかということを通じて、組織の発達機構や変形・破壊挙動をより理解できるようになりつつあります。たとえば三次元像があったときに、二次元で体積率を評価する際には、ある断面で切ったときの断面積から体積率を求めます。すると、観察した面によってデータがずいぶんばらつきますが、少しずつ対象領域を広げていくようにして体積率を測っていきますと、そのピーク値が信頼性のある体積率ということになります。また、どれくらいの領域を測定すれば安定した体積率が得られるのかというデータも得られるわけです。ただし、これを手動でやっていますと 1 年くらいかかりますので、全自動の解析ソフトを我々のところで開発しています。

また、材料特性を考える上では、上記のメトリック特徴値だけではなくて、トポロジカルな特徴が非常に重要で、オイラー＝ポアンカレの式という位相幾何学でもっとも有名な式が重要になります。オイラー＝ポアンカレで求められるオイラー標数は、ボディの数とボイド、内在している空洞の数と貫通している穴の数と、図 6 に示す関係式があります。

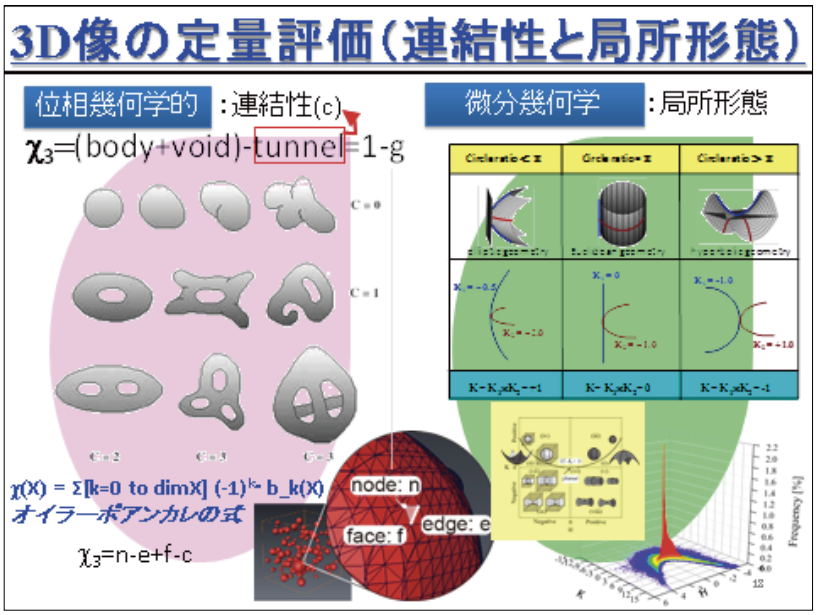


図 6

たとえばドーナツ形状のものは、1 回だけ切っても完全に 2 つに分けることはできませんが、2 回切ると 2 つに分かれます。ある物体を完全に分けることなく何回まで切ることができるか、というのが連結性 c の定義ですので、この場合の連結性は 1 ということになります。連結性を求めるためには貫通している穴の数を数える必要がありますが、そのためには三次元情報を構成しているメッシュの頂点と辺と面とその中に存在するセルの数を数えることによって、オイラー＝ポアンカレの式からオイラー標数を求めればよいのです。

すると自動的にトンネルの数が求まって、今見ている材料組織の連結性がいくらかということがデジタル的に評価できるわけです。

また、局所的な形態を定量的に求めたいときには、平均曲率とガウス曲率を用います。見ている物体の中で平均曲率とガウス曲率がどのような分布をとっているかの表示は、図7の確率密度プロットで得られます。

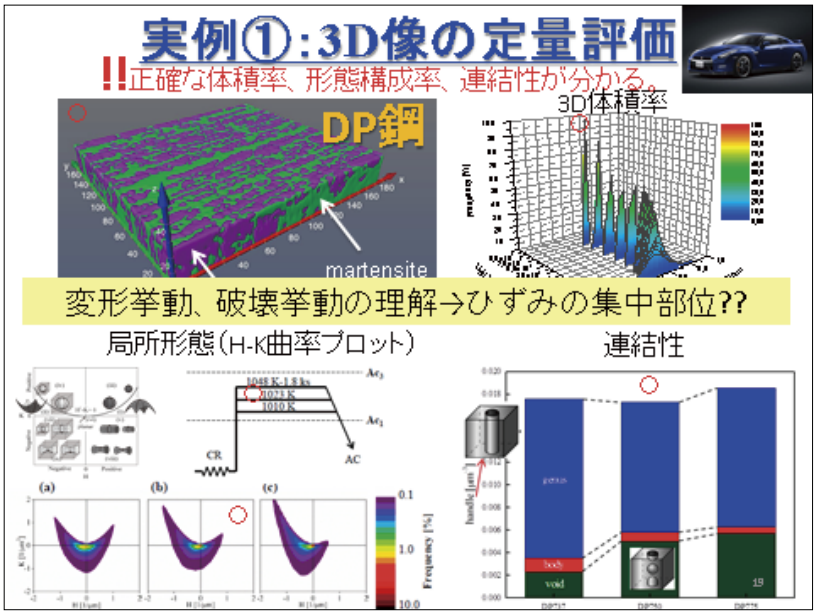


図7

図中のH0（ゼロ）、K0、つまり平均曲率、ガウス曲率が0というところは、平坦な形態を表しており、どういった形態が何パーセント存在するのかということが定量的に評価できます。いろいろな作り方ごとに、この形態がどのように変わっているかということを評価すれば、たとえば我が社の製品の組織の形態分布はどういったところに位置づけられるのかという評価ができます。この評価法を使いまして、自動車に使われているデュアルフェーズ鋼の体積率を実際に求めた例、先ほどのガウス曲率、平均曲率のプロット、確率密度表示した例、あるいは連結性の評価をした例もあり、体積率がいくらで、形態の分布がどういう分布をしていて、連結性がいくらで、ということがすべてデジタル情報としてとれるようになっています。

ひずみがどういうところに集中しているかということが、材料の変形･破壊を使う上で、あるいは信頼性を考える上で重要になってきますが、ここではデジタル情報を扱いますので、簡単に有限要素法に乗ります。つまり、結晶塑性の有限要素法というより高度な変形挙動のモデリングができるわけです。

一方で、ひずみ測定のほうもずいぶん進んでいて、たとえば中性子線の中で加熱しながら、あるいは引っ張りながら、変形させながら中性子線回折をすることができます。すると、材料の組織の中の複数の相のそれぞれがどういった応力を、あるいはひずみを分担しているのか、応力分配が評価できます。中性子のいいところは、バルクの材料からデータ

をとっていますので、信頼性のある平均値を測定できます。平均値しか想定できないのかというご指摘もいただくのですが、信頼性ある平均値をいかにとるかということは非常に難しくて、それが今中性子線でまさにできるようになってきています。なお、中性子線の場合は組織単位では評価できなかったのですが、最近は、たとえば走査型電子顕微鏡等の中に小さな引張試験器を、あるいは加熱装置を入れながら、応力を測定する特殊な方法があります。この結晶粒のどういったところに応力が集中しているのか、あるいは、塑性ひずみがどういうところに集中しているのかという評価もできます。ある意味では弾性応力も、弾性ひずみも塑性ひずみも組織単位で定量的に評価することができ、これと三次元の情報を組み合わせますと、どういったところにどういった種類のどれだけのひずみが集中しているのかということが定量的に評価できるようになってきています。

こういうことを本当はもっと細かく転位オーダー、格子欠陥オーダーでも三次元で評価していきたいということで、たとえば九州大学の東田先生たちのグループは、転位構造を三次元で見ています。それぞれの転位を特徴づけしまして、転位の周りでどういった応力分布が起こっているかがわかり、三次元像でこの転位の特徴を明らかにすることによって、非常に微視的なところでも応力の分布がどうなっているかという評価ができるわけです。これと、もう少しマクロ的な先ほどの中性子や、あるいはほかの電子線を使った方法でマクロ的な応力の分配も評価することによって、階層的に応力、あるいは塑性ひずみを定量評価できるようになりつつあり、これと三次元の情報を組み合わせれば、かなりのことが言えるようになっています。

たとえばデュアルフェーズ鋼の変形中にどういうところにボイドができるかという問題では、三次元像を基にどういったところに破壊の初期過程であるボイドの形成ができるかを見ることも可能です。この結果を基にして数えますと、たとえば固いマルテンサイトという中で割れている場合がほとんどである、というようなこともわかってきます。このような定量的な評価が、三次元、あるいはひずみ測定をやることによって可能になりつつあります。

ラメラー組織の球状化という過程も鉄鋼材料には非常に重要になってきます（図 8）。このパーライト組織というのは、瀬戸大橋のケーブルに使われています。二次元で見ますときれいなラメラー、色相構造をとっていますが、これを三次元で評価しますと、たくさんの穴が開いていたり、裂けていたりします。我々は実際にはこのような組織欠陥がたくさんある材料を作っているにも関わらず、きれいな組織だとしていろいろなモデリングに使ってきたわけです。

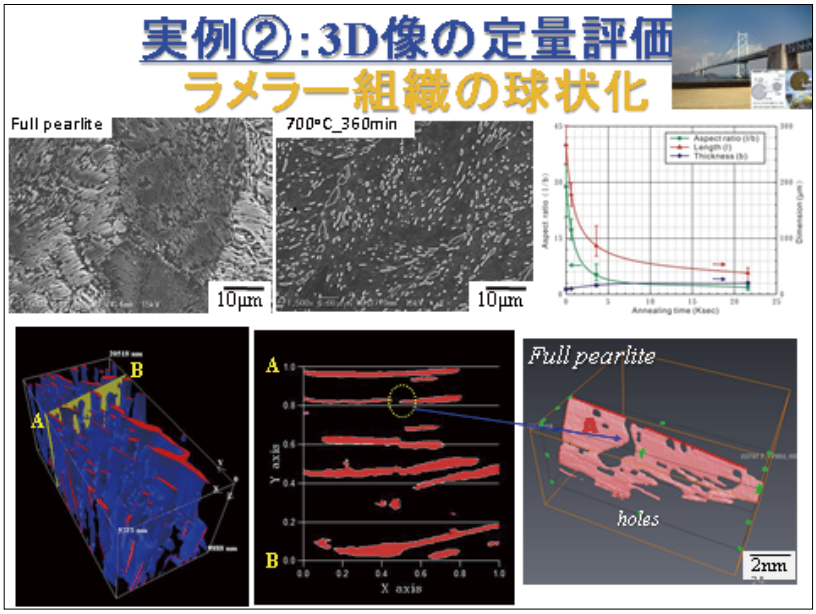


図 8

今後は、たくさんの組織欠陥があることを前提にしながら検討していく必要があると思われます。これをずっと熱処理していくとどうなるかといいますと、最終的には球になっていきますが、板のものがリボン状に変わって、ロッド状に変わって、それから球に変わっていくという形態変化があることがわかってきました。この過程で、実は初期の組織欠陥の縁の曲率が高いので、平衡組成濃度と大きく異なっています。Gibbs-Thomson 効果を通じて炭素が拡散し、それが組織変化を加速していることの本質であると考えています。

それを実際に見てみるために、先ほどの平均曲率、ガウス曲率プロットで材料を熱処理したときに曲率の分布がどのように変わっているかということを評価していきます。すると、ここの次元のところがまず減っていっているということは何をいっているかといいますと、裂け目が減っていっていますよ、ということをいっています。また、少し見にくいですが、真中のところも減っていっていまして、平坦なものも減っていっていることは定量的に評価できるわけです。それは位相幾何学的に評価しましても、穴の数が減っていっていますよ、ということを言っていますので、材料の連結性が減っていっているということは、ここからも定量的に評価できるわけです。どうしてこのような変化が起こるかということを考える際に、板状のものの中に穴が開いておりまして、これが成長して合体していくと、表面積がどのように変化していくかを数学的に求めていきます。あるところでは穴が成長して合体することによって、表面積が減りますよ、ということが求められます。実際に我々が三次元で見たときの穴の形状はこういうところにプロットされまして、ちょうどそれが合体することによって表面積が下がりますよ、というところに相当します。要するにラメラー組織の球状化過程のドライビングフォースは、この穴の成長・合体によって表面積を減らす成長である、とミクロ的には考えられるわけです。

マクロ的にもそういうことが言えるのかという検討は、たとえば三次元像でありますと、直接単位体積あたりの表面積を容易に求めることができ、（図 9、図 10）時間の 3 分の 1 乗に比例して表面積が減っていくことがわかります。これはオストワルド成長が、パーライト組織というものの、ラメラー組織から球状化に変わるときの組織の変化機構だという

ことを物語っています。

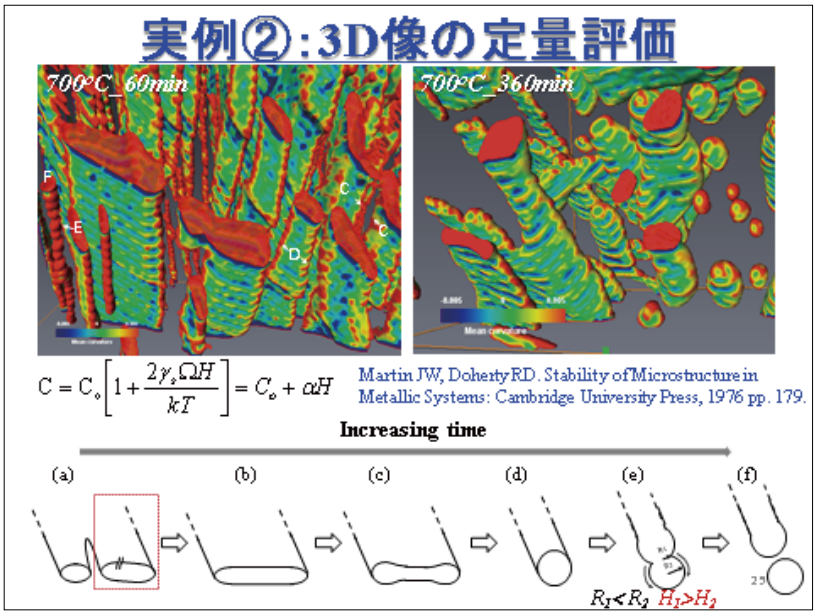


図 9

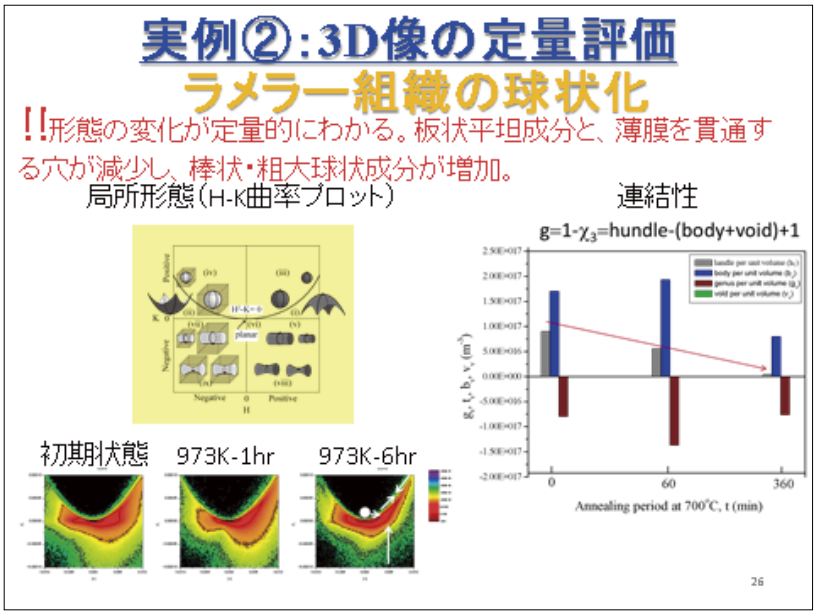


図 10

この変化が本当かということで、J-PARC の中性子線を使って測ったところ、確かにその傾向が同じであることがわかってきました。表面積が熱処理時間に伴って減少することを示していまして、間違いなく球状化処理によって表面積が減少することがマクロ的にも言えるわけです。マクロ的にも表面積は減っており、ミクロ的にも表面積が減ることがこのパーライトというラメラー組織の球状化過程の素過程であることがわかってきました。

こういうことが本当にモデリングでも再現できるのか、小山先生と協力してフェーズ

フィールド法によってラメラー組織の球状化過程を検討しました（図 11）。すると、やはり、たくさん開いている穴が合体することによって、板から、たとえば棒状に変化していくことが計算できました。

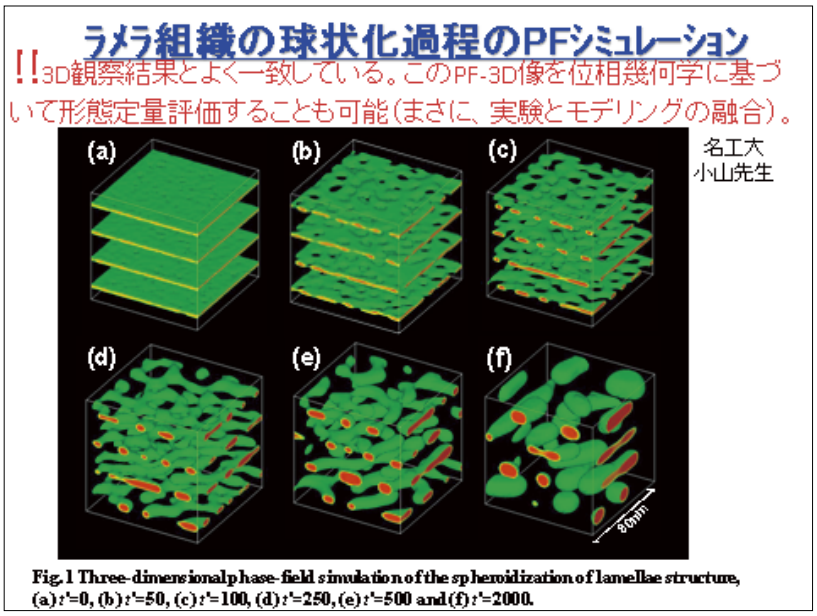


図 11

これは先ほどの三次元で実際に見たものとよく一致しています。さらにフェーズフィールド法をマイクロメカニクスと融合して、任意の形状のものの機械的特性の予測を検討しています。この任意形状、たとえば形状の変化によって、言い換えますと、連結性の変化によって、機械的特性が大きく変わりますよ、ということが計算できます。実際このような応力分布はひずみ測定のほうからもできますので、三次元探査とひずみ測定の実験と比較が容易になってきています。

現在、これまでのいろいろなデータをアーカイブ化しています。たとえば鉄鋼材料の中でどういった組織があって、その中のどういった組成のものに対して先ほど求めた位相幾何学的な組織の形態や体積率、あるいは機械的特性がどういうものか、ということが一覧表示できるようなフォーマットを作っていまして、日本学術振興会の 176 委員会のほうで進めています。

今後の展望ですが、材料組織の完全デジタル情報化の継続的取組体制の構築が必要です。皆が使える勝手の良いデータアーカイブ化の様式を、やはり情報処理研究者、あるいは NIMS のデータベースの方々と協力しながら進めていきたいと思っています。このときに、単に組織形態だけをデータアーカイブ化するのではなくて、塑性プロセス、あるいはひずみ測定値、あるいは応力ひずみ曲線などの特性についてもすべてアーカイブが必要ではないかと思う次第です。

従来、実験研究者とモデリング研究者は別々であって最後だけ合わせるということをやってきたかもしれませんが、これからお互いが理解し合って最初から組んで物事を開発

していくと、より一層短期で製品化までつながるような研究開発が進むのではないかと思われます。

【質疑応答】
質問者：ボイドの数云々、それから強度の関係しているとき、穴の数、コネクティビティ云々が議論され、ずいぶんいろいろなことがわかってきているということ、特にJ-PARC中性子がいろいろ重要なデータを出しているということで、喜ばしいかぎりです。ただ、物質を見ていたときに、どのサイズで穴と定義できるか、長さのスケールが500μm、サンプルのサイズ数百μm云々で、解像度があまりよくわからなかったです。ラメラーの云々のときにはナノメーターのスケールが少し図の中に入っていたりして、その長さのスケールがよくわからなかったのですが、特徴的に穴と定義するときの空間サイズはどのくらいなのでしょうか。ミクロに見ていったら穴か何かわからないですよね。
足立：私が申しあげている穴というのは、せいぜい0.1μmくらいの穴です。やはり穴のサイズによって測定方法も変える必要があります。比較的大きなところはやはり放射光が強いです。組織オーダーでたとえば0.1μmですとセクショニング方法が得意です。もっと細かくなってきますと、やはり電子顕微鏡を使った評価が必要になってきます。
質問者：強度、ひずみとその強度の問題をつなげるときにボイドのサイズが最終的に重要なファクターになると思うのですが、今日のお話のときにどのくらいのサイズが物性値に影響を与えているということだったのでしょうか。
足立：それもたとえば穴が小さくなってきますと、硬質層によって囲まれた軟らかい層の内部には、内部応力が発生しやすくなります。というのは、やわらかい層が内部にありましても、周りに拘束されていないと自由に変形できますが、周りに拘束されているとそれが変形できなくなりますので、軟質層であっても硬質化するわけです。その辺の評価が中性子線ではできてきまして、実際にこれを見ていきますと、穴の形態がどのくらいの大きさなのかというのは、先ほどの平均曲率とガウス曲率のプロットから、実験的に捉えられていれば捉えることはできますので、その穴が実際にどれくらいの内部応力の発生に寄与しているかという評価も実験できます。具体的に何μmだということは今ちょっと申しあげられないのですが、どういうオーダーであっても形態も評価できますし、どんなスケールであっても応力も評価できるようになっています。
質問者：強調されたいのは、実験的に本当にいろいろな長さのスケールでそれぞれ実験手段がほぼ整っていると。今日の機械的な特性に関しては。それに対応してモデリングのところでフェーズフィールドを使ったときにやはり長さのスケール、どこら辺に注目するかで取り扱いがずいぶんいろいろ変わってくると思うのですが、そこのボールのやりとりがこれから始まるという状況ですか。
寺倉：今の質問は非常に面白い問題で、シミュレーションと実験の非常にいい共同作業のテーマですね。
足立：フェーズフィールドで作り込みました穴があいている組織を先ほどの位相幾何学・微分幾何学を使ってコンピュータ上で作った組織はどれくらいの連結性かということを改めて実験することもできまして、それはまさに実験屋さんと計算屋さんの連携で進めています。
寺倉：フェーズフィールド法は連続体モデルだから、元々から長さのスケールに関しては

原理的に限界があると思います。それが、離散性が絡むようなレベルまでいくかどうかというのは、非常に面白い問題ですね。

質問者：特に 3D 像の、これは定量評価というところがあったのですが、たとえばボイドがどんどん集中していってそこにひずみが集まってきて、最終的に破壊までいく様子というのはシミュレートできるのでしょうか。
足立：たとえば大阪大学に大畑先生という方がいらっしゃいまして、損傷度評価ということで、モデリングのほうである程度できるというところまで進んできているようです。実験のほうも、これは放射光を使っていきます。放射光のいいところは三次元だけではなくて四次元ができます、時間軸の変化もできますので、ボイドの周りにどれくらいのひずみが集中しているかを捉えることができるのです。そうすると最終的な破壊との差が少なくなるものですから、理解しやすくなるという、実験的にも、モデリングもその分野はだいぶ進んできています。

質問者：曲率等の情報というのは、もともとは幾何ですので、滑らかなマクロなものを記述する道具だったと思うのですが、最近の数学はもう少し進んでいまして、特異点があったり、不連続なところがあっても、曲率を定義したりすることはできるようになっています。そういうことも使っていければもっと、特に破壊のメカニズムというのは使えると思います。あと今ミクロとメソとマクロという話もありましたが、その辺がおそらく最近の数学でいちばん進んでいるところですので、さらにこういうところで数学とモデリングと実験と計算といろいろ絡まるととても発展するのではないかと思いました。

「データ同化によるモデルの高度化　ー物質材料研究への応用ー」
樋口知之（統計数理研究所）

私は統計科学、数理科学等々が専門分野でして、幅広い領域の方々といろいろな共同研究をしています。いろいろなデータ解析、モデリングをやってきたのですが、物質に関しては、共同研究をやったことがありません。たぶんそれが今日私がワークショップに呼ばれた１つの理由ではないかと思います。

物性科学のパラダイムシフトにおいて、やはりセレンディピティが重要であるとよく聞きます。私もこれを否定するものではありませんが、機械学習等々は、このセレンディピティの獲得をなるべくコンピュータに置き換えていきたいということを目指しています。この物性科学コミュニティにおいてはデータマイニング、機械学習等がどのように活用できるのかというところが十分に認識されていないのではないかと、私は感じています。ただ、その辺りの認識がこの分野でどの程度であるのかが、私は正確にはわからないので、その議論、面白い点ですね。あと、新物質開発における地道な性能向上努力が必要なステージに滞留する時間をシステマティックに短縮するような努力が重要ではないかと思います。

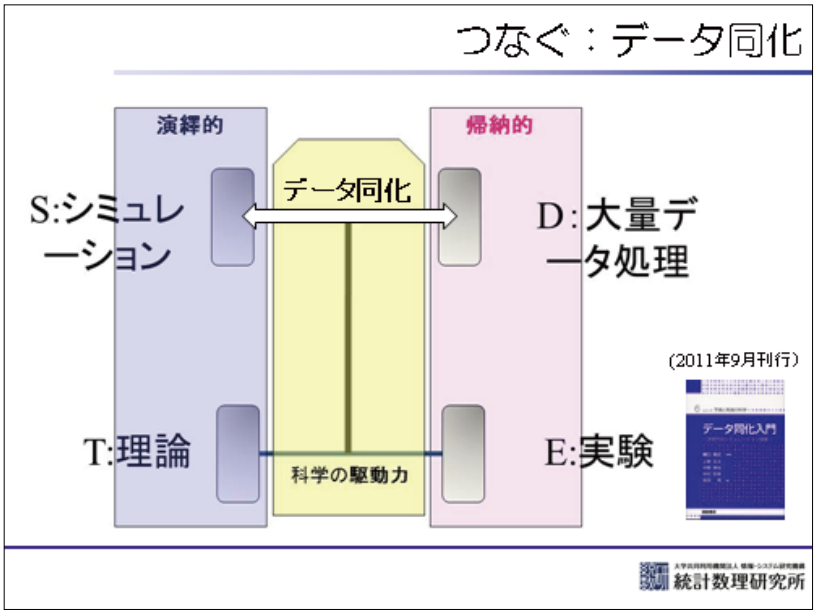


図１

本日はデータ同化についてお話します。データ同化は、シミュレーションと大量データ解析をつなぐ手法です（図 1）。シミュレーションや理論解析というのは演繹的な考え方ですが、経験や大量データから知識を産み出すプロセスは帰納的なもので、まったく違う考え方です。このまったく違う考え方をつなぐのは結構難しく、サイエンティスト自身が非常に戸惑う部分もありますし、日本の教育自体が非常に演繹に重点を置かれているところがあるという問題点もあります。帰納と演繹を考え方としてつなぐには結構大きなハードルがありますので、その話を少ししたいと思います。

ビッグデータは、一般論としては、いわゆる総量としての価値は大きいのですが、価値

密度はかなり低いので、それらをどう扱うかがビッグデータで重要な研究テーマになります。

昔からそうですが、日本では広義のデータマイニングは錬金術的なものとみられ、データ解析への懐疑的態度がとくに日本では強いのではないかと思います。あとはエキスパートへの過度な依存、これは日本独自のものがあるのではないかと思います。しかし、いつまでもプロジェクトXのような経験に裏打ちされた技術力の成功体験の感傷に浸っていては世界から取り残されます。エキスパートはすばらしい知識を持っているが、そういう方々はどんどん退職され、あるいは海外に行かれているということで、エキスパートの知識や経験をコンピュータに置き換えていきたいというのがマシンラーニング、統計数理の狙うところの1つです。もう1つ言わせていただくと、ここにいらっしゃる方々はたぶんほとんど『物理帝国主義』の教育を受けられた方だと思うのですが、ライフサイエンスはいわゆる第一原理というのが非常に弱いので、データからの知識獲得に軸足を置いた研究開発がされています。物質科学の分野は、第一原理にもとづく研究推進が相当な成功を挙げているということで、これまでデータからの知識獲得にあまり重点を置かれていなかったのではないかと思います。

データからの知識獲得も有効であるということを示すために、私と情報システム研究機構の北川機構長が、10年も前になりますが、JR東日本と一緒にやった仕事を紹介します。新幹線では強風になると安全運行のために止まりますが、その運行システムのアルゴリズムを私どもがJR東日本と共に開発しました。JR東日本は、最初物理モデルに立脚したコンピュータシミュレーションで風向予測をしたいということでしたが、ふつうに考えれば非常に難しいことがわかります。局所的な予測をするというのは非常に難しいわけです。ただ、風についていうと、その場で得られた計測値は、不完全だけれども有益な情報なので、我々は統計モデリングに基づくシステムを開発し、それが新幹線の運行システムに利用されています。

ではデータからの知識獲得はもうそれで十分かというと、やはり違うのです。第一原理といわゆる機械学習等のデータ解析法をどうミックスするかが実は本質的なのです。どの程度ミックスするか、それは分野や問題ごとに違います。ざっくりと言えば、機械学習、統計数理というのは基本的に内挿問題です。データがたくさんあったときに、近傍のデータから内挿する。これは、場合によっては第一原理等々使うよりもパワフルです。

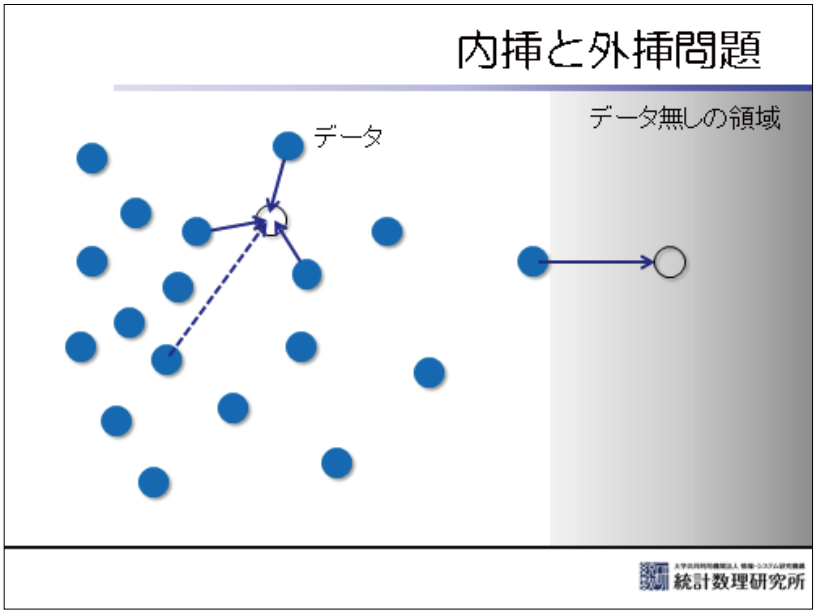


図 2

一方､第一原理に基づくところの大きな力は外挿問題で発揮されます。なぜかというと、データのないところを予測する能力は帰納法にはないことは論理的にあきらかです。こちらは第一原理の力が最大限に活用できるということですので、分野、データ、問題に応じて演繹と帰納をうまくミックスすることが成功の鍵だと思います。

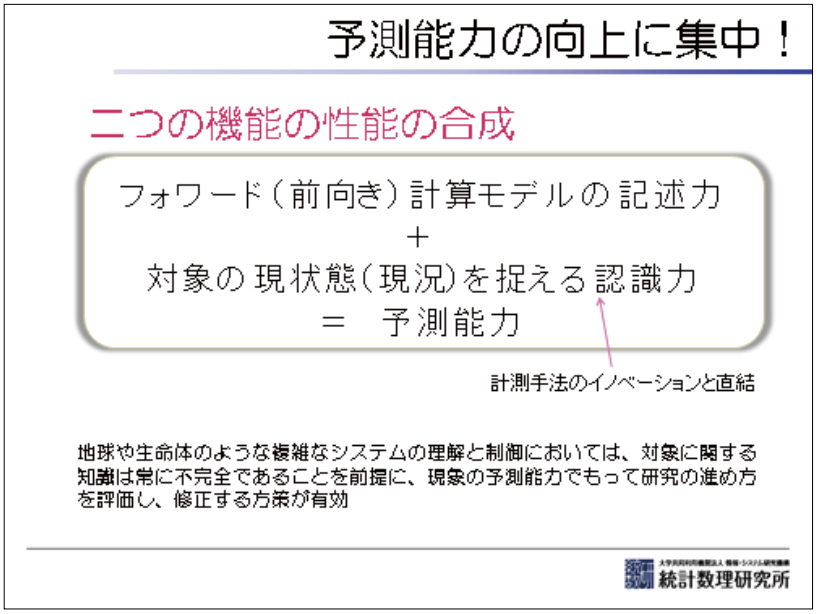


図 3

それはなぜかというと、津田先生の話に機械学習の利用法の 1 つとして予測性能というのがありましたが、予測性能は大きくわければ 2 つの機能の合成で構成されています。1 つは今どうなっているのかということを捉える認識力、もう 1 つは先ほどの外挿モデルのように前向き計算能力の記述力です。したがって、どっちだけやればいいというもので

はなくて、問題、状況等々に応じて適応的にこの 2 つを結合するのが成功の鍵であると言えます。それを行っているのがデータ同化なのです。

データ同化というのは、20 年前くらいに気象海洋学の分野で生まれた技術です。現況の気象予報ではこのデータ同化が非常に強力な道具となっています。気象海洋でなぜデータ同化が生まれたかというと、時間スケール、空間スケールの粒度にもよりますが、ざっくりと言えばこの分野のシミュレータは能力が“そこそこ”高いと言えます。また、データからの情報量を考えると、人工衛星等々、いろいろなデバイスが開発されましたので、“そこそこ”データが入ってくる。ということは、1 ＋ 1 が 3 になりやすいドメインであったためだと思います。気象・海洋学の分野でデータ同化の目的としてあげられるのは、初期条件や境界条件をどのように探したらいいか、再解析データをどう生成するか、などです（図 4）。気象データ、あるいは海洋データの再解析データは、実際の観測だけではなくて、極端に言えばバーチャルなものです。いくつかのデータにシミュレーション計算を同化させることによって、計算グリッド上でのアウトプットがあたかもデータのように使われている。これが再解析データです。津田先生がクリッギングの話をされていましたが、データ同化の目的の 1 つとしては、効率的な観測システムを構築するため、シミュレータを使って、ではどこを観測すればその全体像を得る観点から知識獲得を効果的にできるのか。それらがデータ同化の目的です。

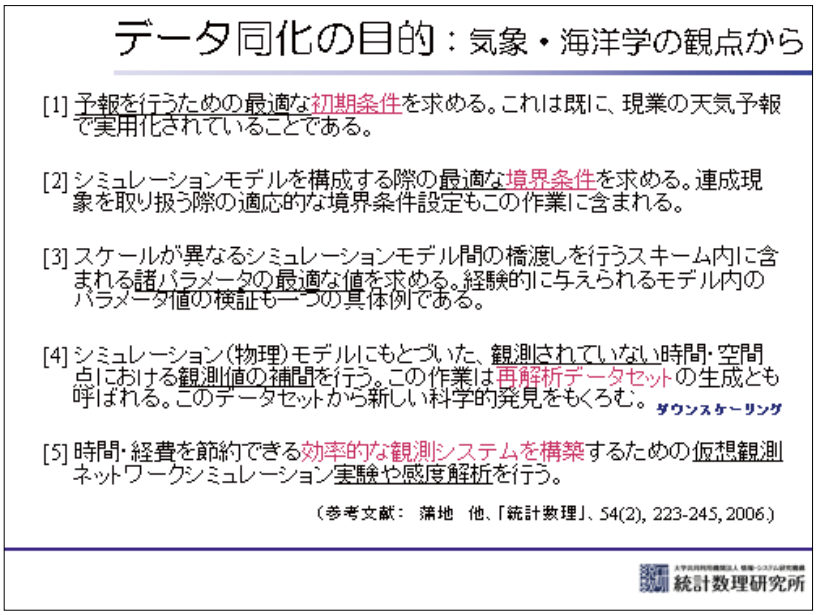


図 4

我々が南極でやっているダウンスケールのお話をします。南極には数十の観測ポイントしかありませんが、そこでの観測データに合わせるようにシミュレータを同化させて、南極上でどのような風が吹いているか、どのように物質が運ばれるかを予測しています。いろいろなリスク予測、災害予測等で非常にピンポイントな予測が期待されています。ダウンスケーリングというものは、まるごと全部シミュレーションしているのではなくて、ある程度の時間、空間スケールで計算したものを、それに合うような形でまた部分的時間・

空間中のより精密なシミュレーションを同化させていくということを入れ子的でやっていきます。上から与えるインフォメーションとしてのシミュレーションの結果は 1 個だけではなく、アンサンブルで与えると、いわゆるダウンスケーリングした結果のいろいろなリスク予測もアンサンブルになりますが、このような予測手法も大きな研究テーマになっています。

データ同化はどんな分野でも適用できます。我々はデータ同化の手法を統計数理の観点からきちんと定式化し、また新しいアルゴリズムを開発し、気象・海洋以外の分野にもいろいろ適用する研究を行ってきました。いろいろな偏微分方程式が実際時間、空間的にも差分化されています。今、津波のシミュレーションを考えてみましょう。その計算にはいろいろな物理量が表れてきます。そのようなある空間グリッド上の物理量を全部、ある時刻のものを縦に並べた巨大なベクトルを状態ベクトルと言います。そうすると、時間発展するシミュレータというのは、一時刻前の状態ベクトルから現時刻の状態ベクトルへの更新式といった、図 5 の最下段にある方程式で書けます。

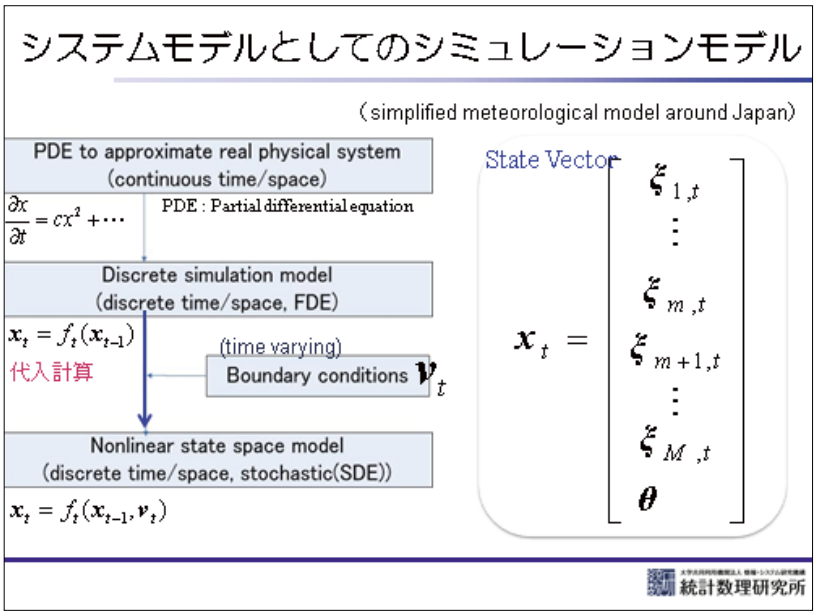


図 5

データ同化の重要な点は、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を明示的に用意することです。

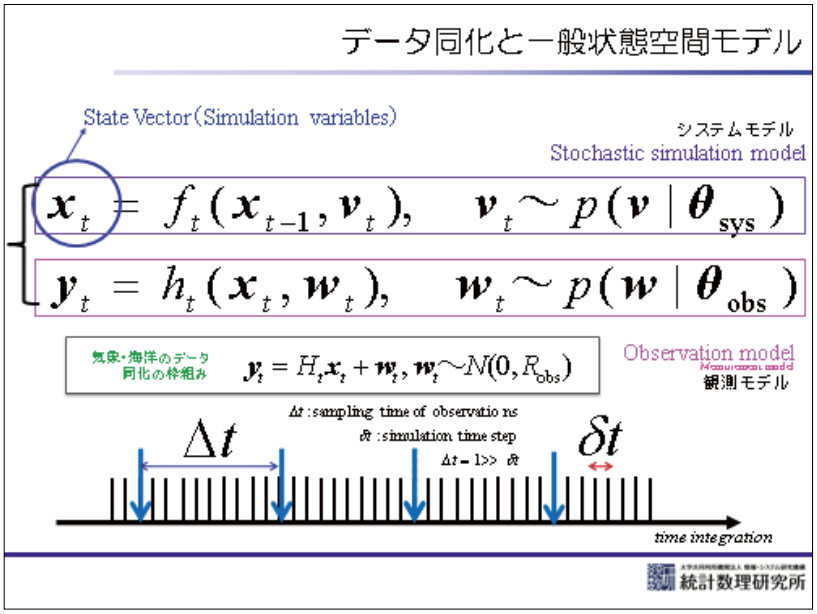


図 6

観測モデルの構成にもいろいろな工夫が必要です。図 6 に示したシステムモデルと観測モデルを組み合わせた形式は、統計や制御理論で大昔から研究されていた一般状態空間モデルというものです。ただ、大きな違いは、状態ベクトルの次元が 100 万次元、あるいは数百万次元になることで、それらをどう扱うかが大きな研究テーマです。

実際の計算では、状態ベクトルの次元の高さとモデルの非線形性から、シミュレーションの値を確率分布として表現できないので、アンサンブルでなんらかの情報を表現して、ある程度確率的な表現ができると仮定します。それでシミュレーションを行うと、アンサンブルメンバが時間発展します。そこにデータが入ってくると、データにあったアンサンブルメンバが変わる。再度シミュレーションを行いデータが入ってくると変わる。これを繰り返すのは逐次データ同化というものです。もう 1 つのデータ同化のやり方は、ベストなパスを事後的に求める方法で、気象庁等々の現業分野で採用されているものです。気象庁等の使っている予測シミュレータは巨大なシミュレーションが必要ですので、アンサンブルメンバを多数持つことが計算リソースの観点からそもそも不可能です。したがって、いちばんいい解だけを求めるのです。逐次データ同化の場合、アンサンブルを利用しますので、計算機上に載せられるシミュレーションの規模は自然と小さくなります。

最後に、データ同化の材料物質科学分野への応用可能性に触れたいと思いますまず 1 つは先端計測装置の開発にいろいろ寄与できるのではないかと考えています。たとえば、最近蛍光物質を使って細胞の中身にどのように流れがあるのかというのを可視化できるようになってきています。細胞の外側にどのような力がかかっているかは計測できませんが、流れの可視化データやその他の部分情報を細胞内流動シミュレーションに同化させることによって、細胞の外側にかかる力といった計測できないものを間接的に測定するようなことができるのではないかということです。また、人工衛星の計測装置は、非常に非線形性の高い測定装置になっていますが、その中には実験室外でのオフセット調整が必要ないろ

いろなパラメータが存在します。再実験ができない環境でのオフセット値等々を推定する逆問題解法にデータ同化の手法を使って、より良いオフセット値推定ができるのではないか、それでもっていろいろな高性能の計測ができるのではないか期待されます。

我々が共同研究で取り扱った神戸空港の人工地盤構築工事の例ですが、そこではどのようにして埋め立てていったらいいかが非常に難しい問題です。これは時間発展の計算法ではなくて境界要素法の問題ですが、シミュレータ内には、どれくらい水を含むか等々の未知あるいは不確実なパラメータがたくさんあります。そこで、シミュレーションを動かしながら、いろいろな測定装置で計測し、計測値で同化させることによって不確実性を低減しながら次のアクションをとるといった、アクティブな工事作業を実現することができます。

【質疑応答】

寺倉：ある分野では非常に強力な武器になっているものですから、こういうものが材料物質科学の中にうまく取りこめるような仕組みを工夫すれば非常に有効ではないかといつも思っています。

質問者：かつて工業的に水素を製造するためのある装置を開発する必要があって、その構造パラメータを決定していく必要があったのです。ところが、試作品をたくさん作るわけにはいかないので、1つ2つ試作品がある。その実験データと、構造パラメータ、流動解析をしないといけなかったのですが、CMT 計算にすごく時間がかかるのです。精度を上げようとするとメッシュを細かく切らなければいけない。それで流動、CMT 計算も最小にして、どういうパラメータで計算するかですが、その実験データと CMT 計算の計算値、シミュレーションデータを組み合わせて、メタモデルを作るわけです。それで十分に事足りるということがあったのです。

先ほど内挿と外挿の話がありましたが、今の話は内挿、外挿と関係ないのですが、材料開発をする場合に、割と関心のあるところのデータはたくさんとると思うのです。ところが、本当に目指す物性のところになかなか到達しなくて、そこはデータがないという場合があると思うのです。実験するのもお金がかかるという場合に、シミュレーションデータでも外挿領域を全部埋めてしまうという考えもありえるでしょう。内挿の部分についていくつかだけ計算をして、その違いがどの程度かを確認した上で一緒にしてしまってモデルを作ってしまうと。そうするとデータの疎密がある程度埋まり、外挿領域とされたところもそれなりにデータの密度があるので、予測の精度は上がるに違いないと。それに近いお話を先ほどの内挿と外挿と捉えてよろしいですか。

樋口:そのお話はデータ同化そのものだと思います。それを統計数理の観点から定式化し、結果として状態ベクトルは巨大な次元になるかも知れませんが、アンサンブルでいろいろ計算します。いわゆる、最初から1つの点、1つの量を求めるのではなくて、データ同化の基本となっているのはベイズ統計ですので、既知と取り扱われているようないろいろなパラメータにも不確実が入ってきます。そこが大きな違いです。

質問者：私も物性推算する場合はモデルをたくさん作って、アンサンブル予測をするようにしているのですが、そういうお話ですね。

樋口：パラメータ空間にもアンサンブルはもちろん入っていますし、特に気象海洋では状態ベクトルのほうも不確実性をもちますので、シミュレーションの計算結果は1個のパ

スではありません。今航空業界でホットなのは、CFD と EFD をどう結合するのかという問題です。航空業界の CFD は乱流がなければものすごく精度が高いですが、エンジン開発等々はものすごくコストがかかりますので、そこでもデータ同化研究がスタートし始めています。計算するときのコストが非常にかかる、あるいは実験そのものも非常にコストがかかるときに、どのように効率よくやるのかというので、広く言えば大きな枠組は全部データ同化でやっていきましょう、というのが大きなトレンドではないかと私は思います。

寺倉：データ同化をうまく使えれば非常に計算コストが下がるということはありうると思っていて、それがどうやったら実際の問題にデモンストレーションできるかというのはこれからの課題だと思います。

「企業における事例紹介及び課題とアカデミアへの期待」
射場英紀、信原邦啓（トヨタ自動車）

私は電池の研究のマネジメントをしています。昨日最終回の NHK の「メイドインジャパン」というドラマはまさに電池の負極材料の黒鉛をシリコンに換える技術を日本のタクミ電気の技術者が中国に特許と一緒に持ち出したという話で、国際競争を背景に材料の要素技術が課題になっているという事例でした。今日は、電池イコールほぼ材料開発というところで我々取組みを紹介するとともに、我々が考える課題をお話しします。

最初に、実験と計算の連携を図 1 に示します。計算だけで材料ができないのは当然ですが、タイミング的に計算が実験の先回りをできるのが理想的な形かなと思います。計算をする際にはずいぶんいろいろなことを考察する必要がありますが、実験の担当者も実験をやる前にそういうふうによく考えるというのは非常に意義があると思います。

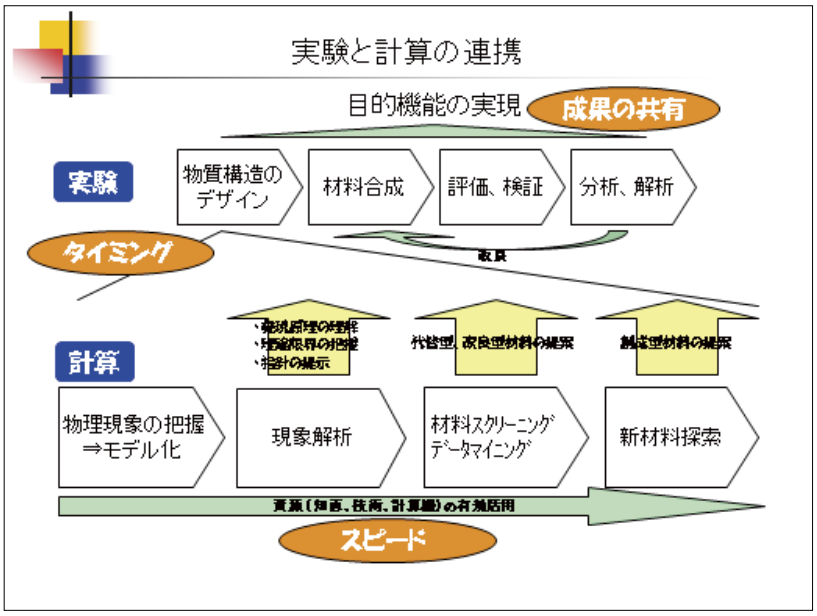


図 1

次に、実験と計算は実験計画のところで同時に始めるのですが、計算には時間がかかって、実験の結果が出た頃にしか計算結果が出てこないというケースもあります。それでも、次の実験計画を考察をする上では役に立つのですが、その実験に寄与するためには計算のスピードがやはり課題の 1 つです。

あと、私が最大の課題と思っているのは、成果の共有です。計算で実験計画を立ててその結果いい材料ができたとして、ともすれば実験だけでできたかのようなアピールの仕方になり、計算が裏方になってしまうケースが我々の会社の中では散見されます。いろいろな部分で計算が寄与していますので、それは実験と計算の両者の成果だということをお互いによく理解することが必要ですし、我々マネジャーサイドもよく見ていかなければいけないと思います。

こういう「タイミング」、「スピード」、「成果の共有」という 3 つの課題に対して、マ

ネジメントサイドの私がしたことは１つだけです。我々の電池研究部は全体で 60 人くらいで、電池の種類によってグループが４つありますが、以前はこの４つのグループと別に計算のグループがあり、同じフロアにいますから密に情報交換をしながらやっていました。しかし１つの試行として計算グループを解体して、各々のメンバーに各電池のグループに入ってもらったのです。そうすることによって計算の担当者も材料開発を自分の仕事としてやるようになり、情報は更に密に交換するようになって、もちろん成果も共有できるということになりました。他方、計算の手法開発の部分が少しできなくなった部分もあるので、あるところに来たらまた元に戻すことも必要かなと考えてもいます。この計算の担当者の中で今日来ている信原はこの新電池のグループチームのチームリーダですが、まさに彼の中で計算と材料開発が融合しているということです。

なお、計算の検討をこれ以上やるのか、もうやめてしまうのかという判断については、やはり実験の人と、これ以上やって意味があるのかどうかとか、ここはもっと深堀りしてくれとか、これはもう実験ですでにわかったからやめたほうがいいとか、そういうことを密にやりとりしながらできるだけ素早く判断する必要があるかと思います。

実験と計算の連携で必要なこと（図 2）としては、まず、実験サイドから見ると、できるだけ実験データと比較できるような電池特性が計算結果として得られ、それを見てすぐにアクションができるようなことです。

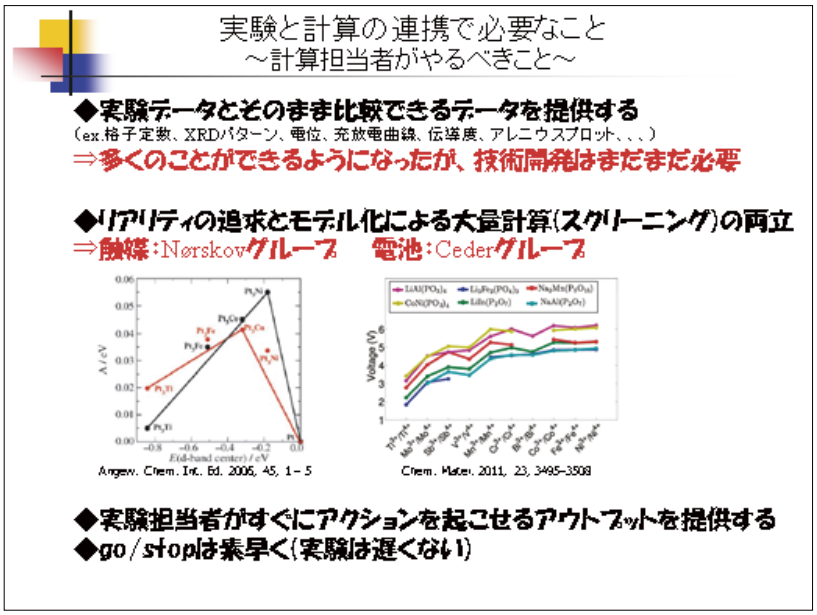


図 2

また、できるだけリアルな形で材料のスクリーニングができれば、計算、実験開発の加速につながっていきます。

電池の計算の取り組みでは、図 3 に示すようなナノのオーダーのところは割とやれていますし、電池の形になると、これはこれで従来の FM 等々でやれています。ところが、ちょうど中間あたりの、メソからマクロに移行するあたりが弱くて、電池の安全とか特性を決

める要因はどうもこの辺りが多いかな、というのが私の印象です。

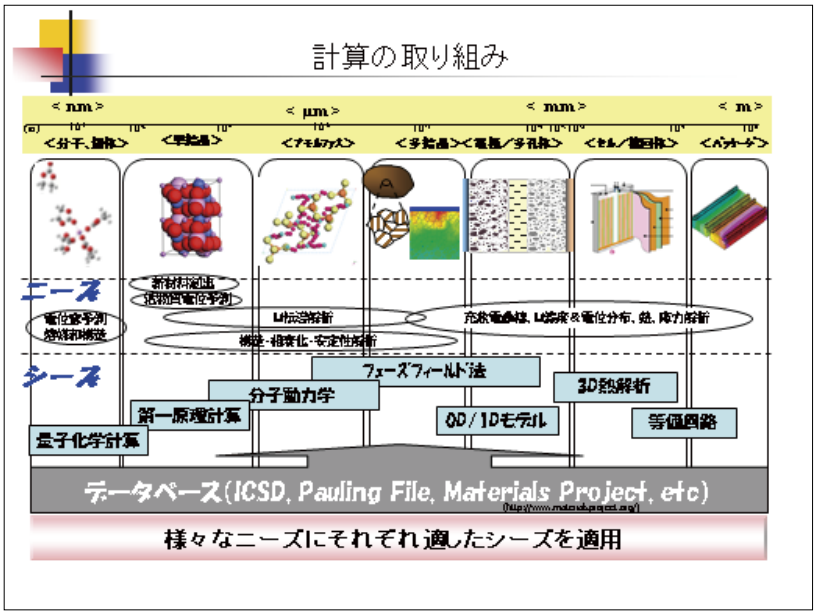


図 3

また、よく計算の方から「マルチスケールで全体を網羅して計算だけで電池を説明しましょう」というお話も伺うのですが、それはあまり意味がないと私は思います。いろいろな電池がありますし、使っている材料もまちまちで、機能特性もいっぱいあるので、その電池、材料、機能特性の各々でどのオーダーの問題がいちばん大きいかということを実験の担当者とよく議論して、そのいちばんキーとなるようなところの計算解析をやるというのがあるべき姿です。たとえば 2 つの計算手法を組み合わせてみるということは意味がありますが、全体を一度にやるというのは、あまり現実的ではないかな、と思います。

次に、事例紹介に移ります。最初はコンビナトリアル・ケミストリーでイギリスのイリカ社と取り組んでいる事例です。固体電解質の LLTO、これは文献でどの組成が、いちばんイオン伝導度が出るということはよく知られているのですが、コンビナトリアル・ケミストリーで最適化した組成がその文献の組成と一致するかどうかということを練習でやってみた結果です。薄膜でたくさん組成や構造やイオン伝導度を測定して、進化的アルゴリズムで解析して計算で伝導度を出しました。従来の文献で知られているのと組成は一致したのですが、エラーバーが大きくて、絶対値もずいぶん計算のほうが大きくなりました。なぜそうなるかを検討するため、構造解析、電顕で見ますと、リチウムイオン伝導の袋小路になるような、もう少しマクロな構造が組成、結晶構造とは別にあることがわかりました。ネガティブに働くものとこういうポジティブに働くものと両方あって、今後はこういう構造も最適化して取り組まないとなかなかトータルとしてのイオン伝導度が得られてこないという結果が得られたことがわかりました。ここでスピーカーが信原に替わります。

ここから少しデータを活用した例、計算を活用した例に関して紹介します。最初の例は、最近リチウムの次の電池ということで、ナトリウム系電池の開発が活発化されていますが、

その中で活物質として画期的に高電位の材料を実際に発見したという例になります（図4）。

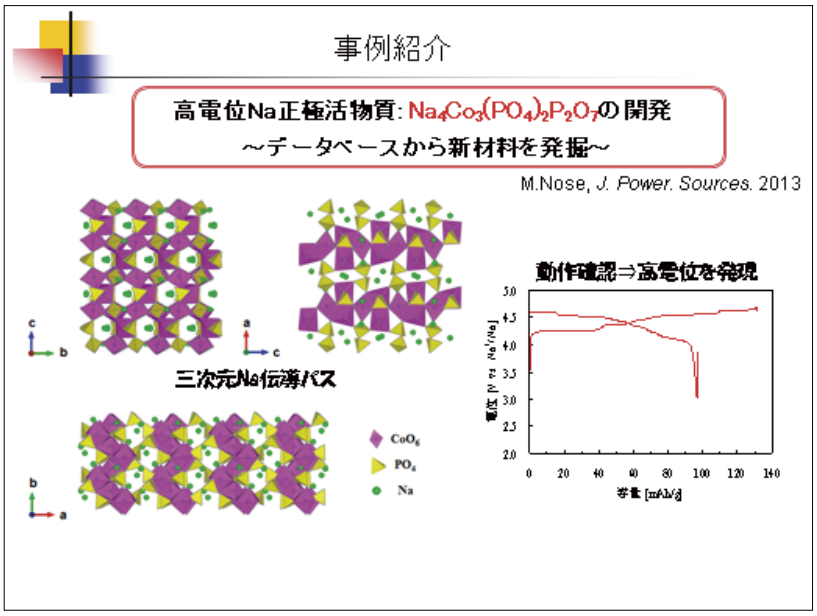


図 4

これは ICSD などを眺めることは重要であるという実例と言えます。開発者はデータベースを眺めて、絵を描いて、電池に要求されるものは容量や電位ではありますが、そもそも電池の基本機能としてナトリウムが動かないと意味がないということで、構造を眺めて三次元的にナトリウム伝導パスがあることをきっかけにこれで動かしてみようとしたわけです。なかなか予想できていなかった高電位が出たという結果が得られました。

続きまして、図 5 は第一原理計算と進化論的アルゴリズムを融合させて新材料の創出をするという事例です。結晶構造のデータがありますと第一原理計算でいろいろな物性、性能を予測するというのはもうすでにいろいろな結果も出ているのですが、これがないとなかなか何もできないというのがありました。この事例では、組成の情報からの結晶構造の予測が可能になります。例として、たとえばリチウム、バナジ、POK で組成を入れて安定構造を抽出すると。一応生成熱で評価しまして、プラスαで、convex hull の考え方で近くに安定な組成があれば、やはりそちらに引っ張られるので、そういう既存のデータの生成熱も計算して、本当にこれが安定に存在しうるかどうかというところまでして、有力なものを実際の合成に回すということをやっています。

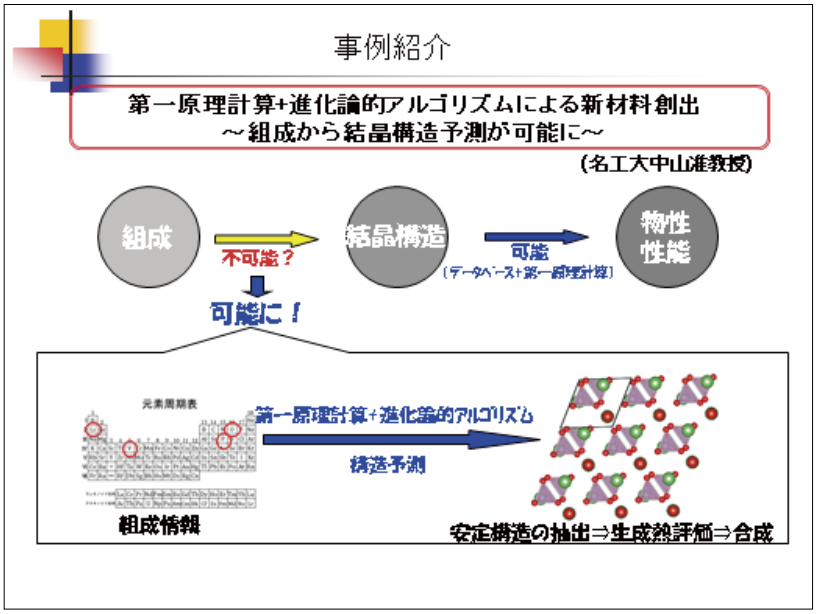


図 5

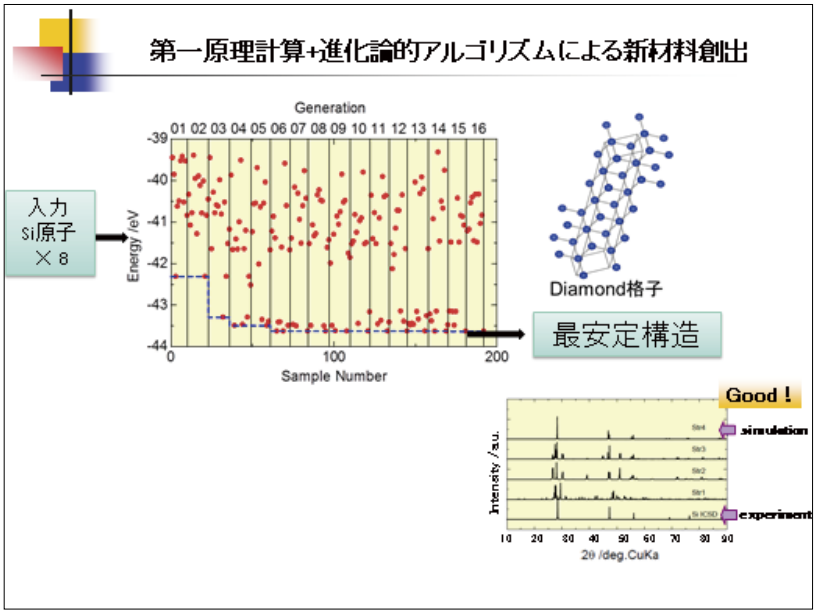


図 6

次の図 6 の事例は我々のオリジナルではないのですが、まず、構造、組成を与えて、ランダムに結晶構造をいくつか発生させます。次に第一原理計算で構造最適化計算、局所構造緩和をすべてさせて、出てきた構造（縦軸は第一原理計算から求まるエネルギー）を基に次の子供を発生させます。このときに、交叉や置換や突然変異といった進化論的アルゴリズムを使っています。最初の頃はこれが本当に使えるものなのかどうかと言われましたが、このすでに報告されている例では、たとえばシリコン原子 8 個を入力してアルゴリズムを回しますと、だいたい 200 世代までそれを生存競争させていって最終的に出てきたものについて、右下の XRD のパターンで示すようにシミュレーションと実験が一致しました。こういうことをいくつかの材料で確認し、今新たな材料創出に使っています。

次がデータマイニングによる新材料発掘ということで、これは東大の岩田名誉教授と一緒にやらせていただいたものです。岩田先生がディレクターになって作成されたポーリングファイルというデータベースを用いまして、電池では報告されていない新材料を発掘して実際に動かしてみたという例です。性能に関しては、画期的なものが出たというわけではないのですが、データマイニングから電池材料を抽出したという 1 つの例になっています。こういうことをやるときにやはりデータが不足するという問題が生じたため、実際にやりたいことはできなかったのですが、いくつかのものには目をつぶって材料を抽出したということになります。

次はきわめて単純な例で、グラファイトというのはリチウム系の電池の負極としては非常に有力な材料で一般的に使われているのですが、ナトリウム系ではどうもグラファイトが使えないという実験結果が出ています。これを単純なモデルで第一原理計算しますと、吸蔵エネルギーがナトリウムの場合グラファイトに入れていくと単純にもう安定ではないという結果が出ましたので、これを基に実験側にグラファイトの改良指針の提案を行ったという例です。

次が、MD 計算による固体電解質のイオン伝導機構解析から材料設計です（図 7）。

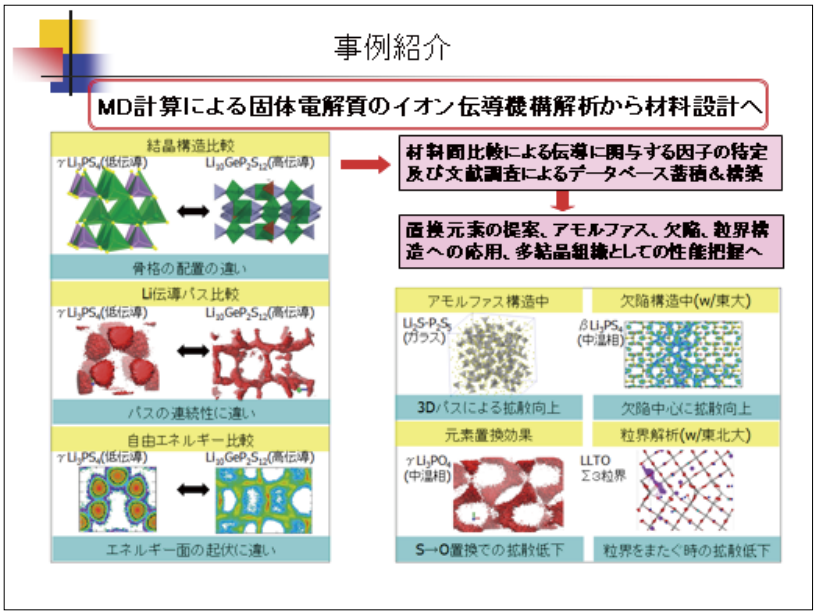


図 7

低伝導と報告されている LGPS といわれる非常に高伝導の材料に関して、MD でリチウム伝導パスを比較してみると、確かに高伝導のほうはよくパスがつながっています。自由エネルギーを比較して、エネルギー面の凹凸が少ないというところまで求め、これ以降は計算の中でアモルファスにしてみたり、欠陥を入れてみたり、元素置換をしてみたり、硫化での伝導はどうなっているかということを計算側で網羅的にやって、材料開発のほうにおとしているということです。

フェーズフィールド法についても電池材料への応用を今やり始めています。これまでもこのような放電曲線などというのは透過回路とかで一般的にやられていたのですが、もっと踏み込んで組織のところの情報を組み入れていけるということでフェーズフィールドを使ってこのような実験と比較できるデータというので実験屋さんと議論していくということをしています。同様のフェーズフィールドをインピーダンスの解析にも今使おうとしているところです。

次は Ceder の例ですが、実験と比較できるデータで議論することが非常に重要であるということを紹介させていただきます（図 8）。

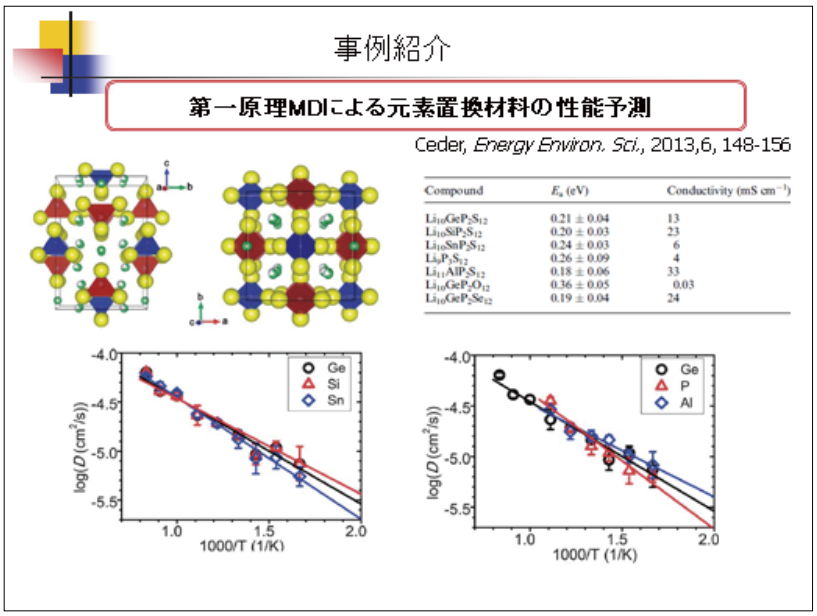


図 8

LGPS でゲルマのところにいろいろなものを置換していったときの伝導度ですが、実験結果と言われても不思議はないほどです。要は計算から実験結果があって、そのまま比較できるようなデータで議論していくと。置換したときに伝導度がどう変わりますか、活性化エネルギーがどう変わります、ということを議論していくと、実験屋さんにとっては非常にインパクトがあり、実験屋が動きやすいということになります。

最後になりますが、課題とアカデミアへの期待ということで、企業の中で計算をやっている私が感じている課題ということで、計算側としては、タイミングとスピードと成果を共用するということが非常に重要です。計算手法が社内にないとか、世の中にもないという問題もあるのですが、計算機の不足という課題を感じることはあります。そのときに、私は産学連携で相手先を探しにいくわけなのですが、アカデミアへの期待としては、いろいろなことを自由にやっていってもらいたいが、アプリケーション意識は当然ほしいというのと、技術発信を積極的にしていってほしいということです。またスパコン「京」も含め、大型計算機については、これまでできなかった大規模系の計算が対象の中心になるような傾向があるようですが、大量スクリーニングというテーマも非常に重要かな、と考え

ています。また、これは今いろいろなところでやられていると思うのですが、学 - 学の間での実験 - 計算連携を活発化も期待します。

【質疑応答】
質問者：私も企業とも共同研究していますが、どういうデータがありますか、という話になると、知らないと。そのときにうまくいかなかったデータ、記録にないとか言われることが多いのです。でも、あとであのときのあのデータもあると、たとえば考察を進めていく上で裾が広くなるわけです。そのときはネガティブデータだったのだけど、データの裾野が広いほうがモデルを作る上では考えやすいということがあるのですよね。ですから、データをどのように社内でも、大学でもそうですが、蓄積をしていくかというのは、すごく大事で、それがあると、別に大学の中だけではなくて、一緒にあわせることができますよね。

質問者：いわゆるマクロスコピックなイオン伝導については、こういった計算でマクロスコピックなディフュージョンコンスタントはなかなか計算できない、合わないということを聞いたことがあって、それはずいぶん長い間の難解な問題だと聞いています。物理的な解釈として、単なるイオンが個別にディフュージョンするのではなくて、1 つ動き始めるとあとはほかのやつがくっついてくると。キャタピラーモードということを聞いたことがあるのですが、そういった物理現象自身の非常に細かい、本質的なところもこういった計算でわかるのですか。

信原：今マクロとミクロのところ、イメージしているスケールが一致しているかどうかわからないのですが､それぞれのスケールのところでの課題に落として､それに適したシミュレーションをやる必要があると思っています。当然格子内の話と､あと硫化を含めた話は､硫化が律速しているのか、硫化があるからこそ動きやすいことになっているのかというのは、やはりそれぞれのスケールでやることが必要です。格子内のリチウムの動きやすさというのは、ミクロのシミュレーションで抑えていく。またマクロが問題のところはマクロスケールのシミュレーションで抑えていく。そこで結果を照らし合わせて、最終的にはものを考えるということをやっていくと。
質問者：計算ではいわゆるイオン間のコリレーションは自動的に計算に入ってきているのですか。
信原：そうです。

質問者：リチウム電池の二次電池は、極端なこと言うと Goodenough 一人ですよね。NASICON のときから発想が何も変わっていないですよね。リン酸ジルコニウム型、あれで全部制覇されているわけですよね。それを一生懸命最適化しているなんて、いくらなんでも少し材料屋がだらしないのではないですかね。新しい材料が層間じゃなくて三次元の NASICON 構造、リン酸ジルコニウム型をリチウム、鉄、PO_4 だとか、皆 1 つの構造のバリエーションでしょう。名前が Goodenough だからって全部 Goodenough はちょっといくらなんでも。シリコンを使うというのは、あの辺からブレイクスルーが出ないと。
射場：おっしゃるとおりで、基本的には正極の酸化物というのは、層状酸化物の層間にリ

チウムが入るというモデルが主流ではあります。だけど、いろいろ新しい材料にもチャレンジはしています。たとえば金属空気電池などはそういう材料の層間に入るみたいな形ではなくて、完全に材料の酸化還元を使って、それと金属との組み合わせ、そういうものも我々は取組をしています。電池メーカーの研究は基本的には Goodenough モデルの延長線上ではあるのですが、それでは自動車のニーズにあう材料は出てこないということで、我々は新しい原理ということをあれこれ提唱して皆さんにもお願いしているということです。

質問者:最後に大学､学への期待ということで示されていた大規模系の計算で大量スクリーニングのほうが必要だということでしたが、その大量スクリーニングというのは、材料開発という点で具体的にどういうイメージなのですか。大規模系というのは実際の電池系なら電池系を意味しているのだと思いますが。
信原：少し規模は小さいかもしれませんが、たとえば Ceder のグループが精力的にやっていることとして、ナトリウム電池の研究開発が活発になりだした頃に、ナトリウム化合物の電位とかというのを網羅的にざっと計算して発表したのです。それが実験屋からすれば、非常にありがたいデータというか、もうすでにわかっているものであればそこからやればいいし、という、そういうイメージです。網羅的に元素などをふって、ある程度第一原理で予測できるデータは一気に出してしまっておいて、それを実験屋さんに使ってもらうと。だからデータベースを作るというイメージかもしれないです。

質問者：最後のところでスピードが問題という話がありましたよね。私も 30 年ぐらいこういう計算を企業でやっていて、大抵の場合ほとんど役に立たないわけです。だいたい開発が終わってから計算結果が出るから。それに関して非常にソフトな言い方をされたのですが、もう少しアグレッシブに何かを考えるという手はあるのでしょうか。単純に計算機を速くしろとかそういうことではなくて、たとえば今日の主題のようなビッグデータをもう少し真面目に利用するとか、何かそういうお考えがあるのかどうかということです。
信原：スピードの問題の中にはいろいろな問題があって、手法がないとか、人が足りないとかという問題もあるのですが、それくらいのレベルのイメージですが。
射場：補足すると、計算屋さんは実験屋の適当なニーズに対してなんとか応えようとしてがんばるのです。そうすると問題はすごく時間のかかるケースが大きくて、やはり実験屋はかなり歩み寄ってブレークダウンした課題を出せば、計算も速く結果が出る。そこが、今日は少し計算屋視点の話になっていますが、実験サイドも努力することによって全体のスピードは上がるのかな、と思います。

質問者：私も企業の中で計算をずっとやってきた人間で、すばらしいと思ったのですが、外部に研究者を出されたりしていますよね。特に国外国内問わず、計算屋に関してどのようなフィロソフィーでやっておられますか。計算屋に限らなくても結構ですが。
射場：弊社の電池の研究開発は大きく研究と開発で分けていまして、研究フェーズは私の責任のところなのですが、それは基本オープンでやりたいと考えている。知財に関しても学会発表に関しても基本はオープンスタンスで進めるということにチャレンジしています。そこから開発フェーズにフェーズアップした後はクローズにします。

「米国 Materials Genome Initiative（MGI）の概要について」
中本信也（科学技術振興機構 研究開発戦略センター）

Materials Genome Initiative（MGI）は関係省庁のアドホック会議が作成し、National Science and Technology Council（NSTC）の承認を得て、2011 年の 6 月に公表されたものです。科学的・技術的な詳細にふれるのではなく、政策的に目指す方向についておもに述べられています。1980 年代以降、技術革新や経済成長においていわゆる先進材料が非常に大きなウエイトを占めていることを指摘したうえで、先進材料が発見されてから社会で実際に利用されるまでに至るには 10 ー 20 年を要することを課題としてとりあげています。MGI は、先進材料の発見と実用化を加速するためのイニシアチブであって、材料イノベーションン基盤を整備して市場投入までの時間を 50%以上削減することを目標にするとしています。

MGI が材料イノベーション基盤としているのは、「計算ツール」、「実験ツール」、「Digital Data」の 3 つです。

「計算ツール」はこのイニシアチブのメインとなるものですが、そこではまず材料挙動の精確なモデルを作り、実験データによりきちっとバリデーションを行うことが重要であるとしています。また、いろいろなステークホルダー、関係者が容易に使用・維持できるようなオープン･プラットフォームを作ること、幅広いユーザー･コミュニティがそういった基盤の便益を享受できるようにソフトウェアもモジュール性をもった使いやすいものにしていくこと、が重要だとしています。

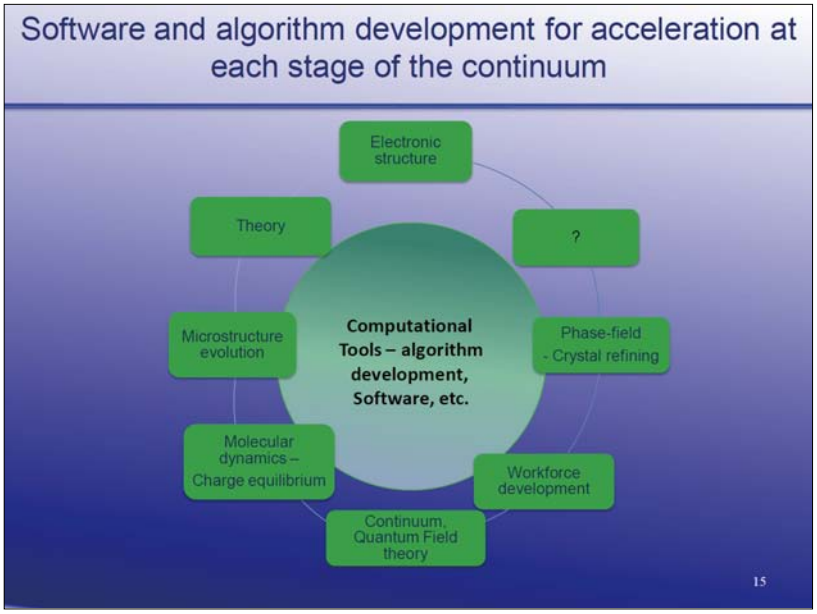


図 1

「実験ツール」については、このイニシアチブが計算やインフォメーションに重点を置いているため、その観点から実験ツールに何を期待するかという表現になっています。まず、合成・加工過程や実使用条件・環境下で諸特性を定量評価するという技法の重要性を指摘しています。こういった in situ の評価技術を計算ツールと組み合わせて、材料、反

応条件、あるいは加工条件を高速にスクリーニングする、そういったことをやっていくのが基盤としての「実験ツール」の役割であるとしています。

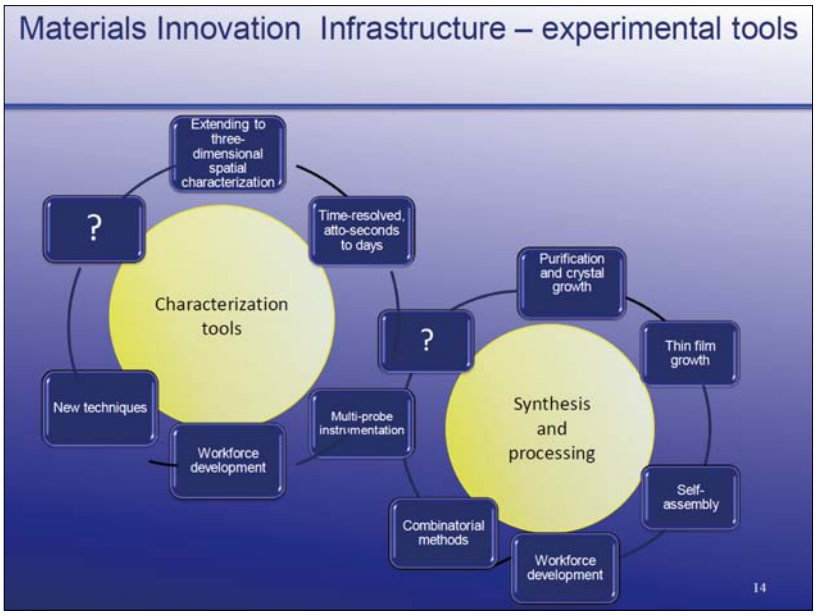


図 2

「計算ツール」の側からの要請として、モデルのパラメータの決定に必要な実験結果の提供をあげています。また、先ほど申しあげたバリデーションとの関係もありますが、有効性･信頼性の向上に補完的に作用するような実験結果の提供を期待しています。さらに、「計算ツール」におけるいろいろなモデルが、マルチスケールを扱っていく上では不十分であるので、スケール間をつなぐ上で非常に実験のデータが重要であるといった表現がイニシアチブの中に見られます。

「Digital Data」では、さまざまな手段でとられたデータへのアクセスビリティを向上させるために、材料研究者・開発者のためのデータ・ストレジシステム、データ転送システムを確立することが必要だとしています。また、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きちんと支援して材料データをもっとも有効に活用していこうではないかということを述べています。最後に、材料開発の一連のプロセスにおいて高度なデータを共有していくことが非常に重要だということを述べています。

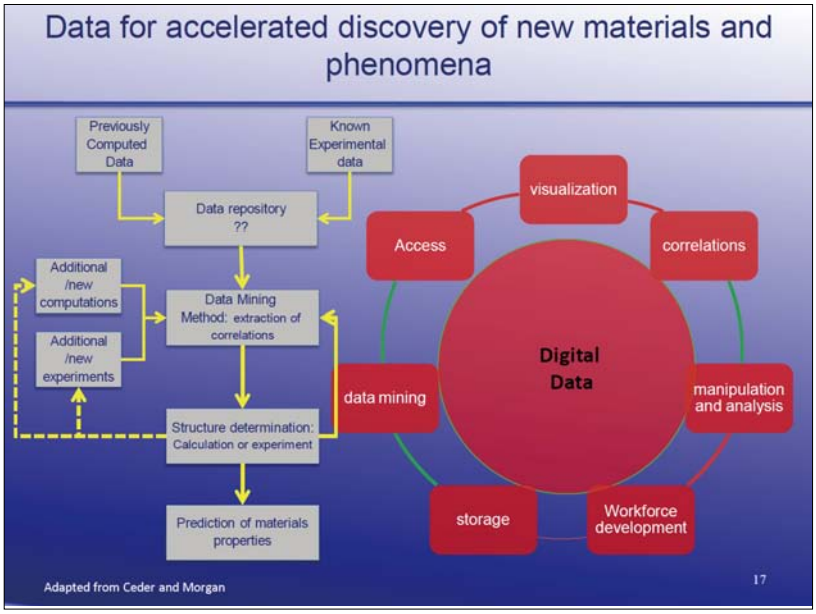


図 3

以上が MGI における 3 つの材料イノベーション基盤ですが、MRS　Fall Meeting で NSF の方が MGI の説明で使われていたスライドに“All Hands on Deck”（総員甲板集合）という表現がみられます。MGI の目指すところは、アメリカにとっては国内の製造業を強くしていくために、関係者が力をあわせてやっていきたいということを表しているのではないかと思います。そこでいちばん重要なのは open paradigms ということで、オープンにやっていけるところはオープンに、シェアできるものはシェアしていきたいといったことが MGI 全体としては述べられています。また、最後の部分でカルチャーを変えていくことが重要なのだといったことも述べられています。

MGI と同時期にオバマ大統領が発表した「先端製造パートナーシップ」は、米国の国際競争力、特に製造業分野で国際競争力を強化して雇用等も創出していきたいという話なのですが、その中で MGI が 4 つの施策の 1 つとして位置づけられています。そのほか、MGI に関連するイニシアチブとしては、ビッグデータに関するイニシアチブや、ナノテクノロジー関係では Nanotechnology Signature Initiative のひとつとして“Nanotechnology Knowledge Infrastructure”が MGI と前後して公表されています。“Nanotechnology Knowledge Infrastructure”の中でも実験データの共有によるモデルの高度化や、テータ・ソフトウェアの共有などが必要であると指摘されています。

２－４　コメンテータ総評

寺倉：今日はマテリアルインフォマティクスに関してかなり広い範囲に渡ったプレゼンテーションをしていただいたので、ある意味では全容を概観していただけたのではないかと思います。これから、いくつか議論すべき問題点を少し整理したものを出しますので、それを参考にしながら、コメンテータの方から今日のワークショップに出ておられて、こういうところが問題ではないかとか、などについてご意見を伺いたいと思います。

最初はデータを用いたインフォマティクスで何を狙うかで、1 つ目はデータを用いていかにハイスループット材料探索を行うか､というのは津田さんのいう予測という側面です。もう一つは、データから新たな物理・化学法則を探る、これは原因究明です。この 2 つに大きく分かれると思います。物質科学の開発に関わっているものは両方とも渾然一体となっていると思うのですが、この 2 つはごっちゃにしないで、それぞれ別の話だとして取り扱えということでした。

2 番目の研究の推進方策、異分野連携・融合に関しては、実験と計算の連携というのはインフォマティクスという意味では特に本質的ではないのかもしれませんが、計算結果の検証という点では重要ですし、物質科学全体としては広い分野の連携・融合が必要です。そこに軌道放射や J-PARC が絡んでくるということも、ユニークなアクティビティにつながるために非常に大切ではないかと思っています。

特に物質材料科学者と情報科学、あるいは数学との連携の問題というのは、特にこのインフォマティクスの場合に非常に重要かと思います。また産と学の連携・役割分担、これも必ずしもインフォマティクスとは関係ないかもしれませんが、動機として非常に重要ではないかと思います。

データベースのあり方は津田さんから RDF という一般化の問題の指摘がありました。

今後の展望、課題は、アメリカやヨーロッパ、韓国などでもうすでに動いていまして、日本はデータを活用したアクティビティというのが非常に低調になのですが、これはできるだけ早い時期になんらかの形で実行に移せるような仕組みを考えないといけないと思っています。

あとはスケール間のギャップというのは今日何回か議論がありましたが、これは大きな問題ですが、これ自身インフォマティクスと直接関係があるところもありますし、関係ないところもありますね。

それぞれのコメンテータの方でこの中で特にご意見のある部分について、お話しいただけないかと思っていますので、よろしくお願いします。

鷲尾隆（大阪大学）

私は特にデータからいかに新たな物理・化学法則を発見するか、ということに関しては、もしかして別の観点があるのかと思っています。単に原因究明ということではなくて、やはり実験をやっている方というのは測定がすごく重要なのです。実際に実験をやってものがきちんと見えないといけない。うちの研究所に材料の先生方もいますので、そういうことを悩んでいる先生はいっぱいいます。測定の精度をあげるとか、生のデータでは見えない背後の現象をきちんと統計的に推定して、判別して、現象が起きている、起きていないということを調べることは、実はデータマイニングや機械学習が非常に得意な話で、そこから新しい物理や化学の法則というのがわかってくるのではないかと思います。

先ほど海外の動向のお話でもわかるのですが、決してビッグデータという話ではないと思っています。むしろそういう足りないデータも含めて、貴重なデータだからこそ、いかにきちんとデータベース化して皆が利用できるようにすべきか、というところがすごく問題意識としてあって、ああいった研究が進められているのかな、と個人的には思いました。そういう意味でデータが貴重だからきちんとデータベースを整理しようという観点で進めることが大事ではないのかという気がします。

逆に言うとそもそもデータが足りないのだという場面が多いので、そういったこともある程度前提にして、どういった形で情報の技術を適用していくのか。逆に、情報系では非常に少数サンプルのデータからいろいろなものを推定するという研究もたくさんやられていますので、そういった観点はすごく大事で、そういったところともきちんとコミュニケーションを密にして進めていくほうがいいのではないかという気がします。

あと特にデータベースを整備しましょう、プログラムのツールを整備しましょう、と皆さんよく言うのですが、これは交通整理する人も必要だと思います。皆が勝手に自分の思いで、ではこういうツールというふうに勝手に登録しても、ソフトウェアがばんばんとネットワークに並ぶだけで、ほかの人が使えないと。だからどういうインプット、アウトプットで標準化してどういうツールとして公開するかということを、きちんと本来交通整理するような役目の人、我々情報系はそういうことをオントロジーという言葉で言うのですが、そういう人がついて研究者と一緒に整備していくという体制が必要なのではないかという気がします。そういうことからいうと、もし何か進めるのであればある程度トップダウン的に課題を示した上で、それについてそれぞれの研究者がいろいろな発想で取り組むという体制がいいのではないかと思いました。

松宮徹（新日鐵住金）

日頃、こういうプロセスをとったらどういう材料の構造ができるか。また、こういう構造をもっているものはこういう性質、こういう機能を示すという関係を捉えるべく実験をしているわけですが、離散的なので、それに網をかけて普遍的に法則を見出して説明したいと。しかも、機構に立脚した法則を立てることによって普遍性が増す、外挿もきくということで取り組んでいるわけですが、その場面で理論あるいは計算機シミュレーションがキーとなっているのです。その 2 つの関係を結びつけるのに 1 つの法則で足りるときはその法則に従って設計すればよいわけですが、いくつかの現象、あるいは素過程が絡んでいるときには計算機シミュレーションをもってその関係を結びつけざるをえないわけです。それぞれの法則が正しいかどうかというのは、やはり実験で検証しないと始まりませんので、そういう意味で計測とか観察技術と計算結果の対比が非常に重要になってくると思います。

また、実際に設計を行う場合には、法則あるいは計算機シミュレーションを動かさないといけないわけですが、そのときに必要な基礎物性がデータとしてないときには計算で補わなければいけませんので、ある程度多重計算が必要になってくると考えています。その多重計算の中に、今までやったことのない未知の材料も含めて計算を行うことによって、新しい革新的な材料が開発できる可能性も含まれていると思います。

また、インフォマティクスをどのように活用するかというところで、やはり法則を見出すときには機構を頭に描いてその機構に立脚したモデリングを行うわけですが、インフォマティクスの記述子を選ぶときにどのような選び方があるのか。やはり機構が頭に描かれていないと記述子が選べないのか。あるいは統計的に妥当な記述子が選べてくるのか。その辺がまだ少し理解できないです。リニアコンビネーションではなくていろいろな非線形の記述子を選んでくる理由が、あるターゲットの性質に対してどのように考えて選ばれるのかというのが 1 つよくわからないのですが。

津田：たとえば第一原理計算などをして、それを記述子として使うのですが、第一原理計算なんてものすごく非線形なものですよね。非線形な計算をして、その上で線形の予測をするということになっているわけです。だから、非線形な記述子を作ってさらに非線形な予測をすることもできます。そうすると、ただ、最後は線形な予測にしておかないと、あとで解釈ができなくなるのです。それで記述子はどのように作るかべきかということですが、これはやはりコミュニティワークであって、いろいろな人がいろいろなアイデアを出す。たとえば学会とかがあって、そこでフィルタリングされてくる。そういう形だと思います。

松宮：我々が物理的な背景をよくわかっている人たちといろいろディスカッションしていると、何をとってくるべきかというのが明らかになってくるのですが、スタティスティックにやはりこういうやつをとったほうがいいというような意見も出てくるわけですか。

津田：そうです。基本的にはたとえば物理というのは完全にあるわけですから、その物理的な発想でいろいろな記述子を作ってみると。ただ、実際どんなデータでもうまくいくものは存在しなくて、この記述子はこういうタスクに対してはうまくいくけど、こういうタスクに対してはうまくいかないみたいな、結果をどんどん積み重ねていった結果、最終的

に今音声認識などができているのは、そういう長いフィルタリングの結果、すばらしい記述子ができてきたからこそできるわけです。

寺倉：私の理解では、たとえば津田さんの発表の中にガーステインのポリアリルエチルスルフォンか何かの解析がありますね。それでいくつかコエフィシェントというのでどのパラメータがどの程度貢献しているかというのは、主なパラメータがいくつか決まりますね。それをベースにまた物理的な直観でそのコンビネーションか何かで、ベターな記述子ができれば、依存度がそのパラメータにしたらもっと上がると思うのですね。そこは最初には決まらないけれど、解析を工夫することによってそこに物理的な勘を入れて解明していけば、どのパラメータの組み合わせにしたときにそれがいちばんよく記述できるかということがわかってくるのではないかという印象をもっているのです。
津田:そういうこともあります。だから、記述子をいろいろ用意しておいて、結果を見て、また、いや、実はこれはこういうふうにしておいたらよかったのではないかと少し改善して、ということを繰り返していくと、より物理的な法則が明らかになるのではないかと。
寺倉：いちばん最初に示した鉄系超伝導体の、頂角みたいなものは、あの辺のパラメータを適当に入れれば、何かその辺が絡んだようなものが、T_c 決めるのに重要なパラメータになるというふうに出てくるのだと思うのです。そこから頂角にするか高さにするかは物理的なイメージとか、それをベースにして理論解析にしたときに、どれがいちばん効いているかということがわかってくるのではないか。

小谷：この間数学と他分野の連携ということでワークショップがあったのですが、そのときにお医者さんが癌の診断に数学が使えるというお話がありまして、お医者さんがおっしゃったのは、提示されるモデルが、我々が今までやってきた、直感的かもしれないけれども、その判断基準に従った形のものを数理モデルにして、効率よくできるのであれば使いたいのだけども、そうでないとなかなか使いたくないとはっきりおっしゃられたのです。それで質問は、もちろん皆さん物理的な法則もよくご存じだし、経験もおありで直感もあると。確かにこのパラメータだったら現象を記述しているな、というような指標を使ったデータ解析でないといけないのか、それともなんだかわからないけど指標が出てきてうまい相関があるということであれば、それは逆にそこからそこにあるなんらかの法則を見つけていく方向まで好奇心をもっていただけるのかということをお聞きしたいのです。
寺倉：そこの 2 つは区別しなさい、というのが津田さんのご意見だったのですが、一応そうは言いながら原因究明をやりたいとしたら、その予測はすごくうまくできた場合に、どのパラメータがいちばん効いているかということを、原因究明の方法を使ったら、どれがウエイトが高いかわかるわけです。
津田:そうなのです。ただ両立が難しいという現実があるということを少し知ってほしかっただけです。つまり、解釈しやすいモデルなど、今までの経験に基づいたものだけで予測しようとすると、うまく精度があがらないことが多いのです。それよりもいろいろな記述子を入れて、線形識別機でやるとあまりうまくいかなくて SVM でやるとうまくいっていましたよね。SVM というのは理解不能なので、非線形 SVM をやってしまうとまったく理解ができなくなるのです。でも、理解はできないけれども精度は高くなる。だから、それはどちらがいいということではなくて、だからこういうモデルが社会に受容されていく

ためにはその辺のバランスをうまくとりながらいかないといけないということだと思うのです。

細野：私は、ケミストリー出身なのですが、ケミストリーで役に立ったことは2つしかなくて、電気陰性度とイオン半径です。電気陰性度というのは物理の人に言わせるとあんな曖昧な概念はないと。でも、あれはすごく役に立つのです。あれでほとんどのケミストリーの人が仕事をしているのです。ですから、ああいうある意味ではコンプロマイズ、畳み込まれた概念ですが、ああいうものが出てこないかという期待はあります。

船津公人（東京大学）

私はケモインフォマティックス分野で 30 年ほど研究を進めてまいりました。企業との共同研究も多数行ってきた関係でインフォマティックスを利用すると言っても多種多様な工夫が求められることを多く経験してきました。

パラメータを多数使ってモデル作成を行うと、確かにモデル自身の「精度」、フィティングは良くなるのですが、モデルを作成するために用いたトレーニングデータにしか適用できないというオーバーフィッティングの問題が生じます。新しいデータに対する予測能力の低下と言う問題が当然起こるのです。それを避けるための工夫が必要となります。変数選択はそのための有効な手法の一つです。SVR（Support vector regression）は汎化能力に優れた非線形モデル化手法です。SVR は確かにモデルの中身が線形モデルのようにわからないのですが、変数選択をしながら精度および予測性能の高い R モデルを作ることはできます。たとえば説明変数を 100 用意して、最適化手法の一つである遺伝的アルゴリズムを用いて変数選択をしながら SVR モデルを作成すると、10 個程度の少ない説明変数で非線形モデルを作るということができる場合があります。そこで選ばれた説明変数を使って線形モデルを作り各変数の重要性を粗いながらも評価することは可能です。

そういうテクニカルな面ではいろいろ対応法はあり、最終的には精度、予測性能ともに高く、かつ使われている変数も物質特性や現象の説明に納得のいく形になるのが良いのは言うまでもありません。ではそのときに初期の説明変数集合としてどのようなものを準備すれば良いのかということになります。説明変数(記述子)については当然解析対象によって変わるものです。どういう記述子を最初に準備するか、基本的に化学者のセンスに依存するところは大きいですね。

たとえば新規機能性高分子材料開発する場合、ポリマーの原料となるモノマーの種類とその組成､各モノマーのいろいろな物性､そしてプロセス条件は必須となります。したがってモデル構築をする場合でも、具体的に設計に役立つ指針を得るためにこれらの情報を記述子、つまり変数としてモデルに織り込みことが求められます。ここで補足ですが、この記述子には、理論化学計算から得られる数値を利用することも積極的に行うべきと思います。また、構造設計を行うのであれば、構造につながる記述子を具体的に利用することも考えるべきでしょう。その上で、変数が多くなってきた場合には、モデル構築にあたって必須な変数は選択対象から外し必ずモデルに反映させ、残りの変数は遺伝的アルゴリズムなどで変数選択させることで、精度および予測能力の高いモデルができる可能性は結構あります。こういうことはこれまでの企業との共同研究などを通して何度も経験してきています。

変数選択と精度の追及は別々に行うべきとのご意見がありましたが、もちろん分けて対応することも可能ですが、ここでご説明したように両方同時に追求することが可能ということをお伝えしておきたいと思います。

また、実験と計算の連携ですが、インフォマティクスの場合に特に大切なことは、デー

タ密度が高い領域の予測は内挿ですので、ある程度の精度で、たとえば物性の予測ができます。また、モデルを用いた逆設計（inverse analysis）を行い提案された、例えば材料候補の実現性やその候補が持つであろう物性の信頼性はかなり高いものになります。しかし、外挿域、つまりモデル構築にあたってデータ密度が低い、あるいはデータの存在しない領域での設計結果はプアーなのは当然です。その領域からの提案候補が計算値と実験値と合わないのは当然ありうることでそれはそれで結構なのです。大切なことは、その外挿領域の実験データを織り込んで、改めてモデルを作り直すことです。常にモデルを更新することが必要なのです。一方で、データの無い領域のデータを補てんする実験を敢えて行う努力が必要です。回り道のように見えますが、結局はそれがモデルの質を高め、設計の信頼性を高めることにつながるのです。そういう意味でインフォマティクスにおける実験と計算の連携が必須なことを理解するべきでしょう。

データベースのあり方、これは非常に大事な事柄ですね。ところが、各実験研究室、あるいは研究者というのは、自分の生み出したデータやその近い範囲のデータは持っているのだけど、その周辺、裾野にあるようなデータ、またあったほうがよさそうなデータについては持ち合わせないことがほとんどと言ってよい。あるいはそのようなデータは眼中にないという場合もあるかもしれません。しかし、先ほど少しふれましたが、例えばガラス転移温度（TG）を予測するだけに限っても、ある研究室の温度範囲の狭いデータだけでモデルを作ると精度も予測性能も全くない誤差の大きなモデルになってしまうことがあります。ところが、関連の研究をしている研究室のデータや、目標とする機能性高分子関係から少しずれたフィールドのデータを集めてくると、共通の記述子でモデルを構築できる利点もあって、ダイナミックレンジの広い範囲で性能の良い一つのグローバルモデルを構築できることがあります。これについては経験したことがあります。ある研究者が僕はこの辺りの材料の TG を予測したいとなった場合、その研究者の研究室のデータも活きますが、ほかの研究室のデータの助けがあった始めて予測性の良いモデルができ、研究が進展することになります。この意味で関連分野のデータの共有化が図られる、つまりデータベースの整備は化学研究・開発にとって非常に重要な位置を占めていることが分かるのです。

まずデータを共有する、その手立てを考える必要があるでしょう。研究者は、論文は出しているわけですから、論文を出す際に、論文に掲載されたデータがデータベースとして収集する、論文に掲載されデータについても、公開対象を限定するなりして収集する。アメリカでも収集したすべてのデータを公開しているわけではなくて、たとえば半分は、研究機関外には非公開となっているものもあります。

データを提供しましょう、その際に基本的な記述子についても記載しておきましょう、あるいは関連データを別のデータベースから持ってくることができるかもしれないので、そこに対してリンクを張っておきましょう、それに加えて自分が本当に必要だと考える記述子については、いくつかの計算ツールをデータベースのルールオプションとして用意しておいて、それらのツールを使って各材料に対して欠落している記述子を埋めてください、というやり方もあると思うのです。

いずれにしても日本の研究者は海外の学術雑誌に研究成果とともにデータも記載して投

稿していきます。学術雑誌を発行している機関はそのデータをデータベース化して世界に対して販売しているわけです。日本人はそれを高額で購入して利用している。言ってみれば、国費で研究した結果が海外に出て行き、海外の研究者はその恩恵を受けているけれども、日本人はあまりその辺りの問題についてほとんどに目が向いていないという現実があるように思います。提案としては、論文にならなかったデータもたくさんあると思うのですが、それらも含めて公開データ非公開データを国内でどのようにオーガナイズしていくかは、今後のマテリアルズ・インフォマティックスを考える上でとても大切な視点ではないかと考えています。このあたりをJSTの情報事業部などで真剣に検討して頂く意義はあると考えます。

材料開発は様々な分野に関連しています。ある同じデータセットが自分の研究分野だけではなくて、別の研究分野および業種で利用されることもあるので、その辺りはよく議論して、どことどこがどのように使う可能性があるかといった広い視点でデータ供給と収集の仕掛けを考えてみるという視点もあると思います。ある機関だけが一所懸命データベースを作って供給するというのはもはや現実的に困難なのです。ですから、データを広く集めて、それをフィードバックする仕掛けと、企業なりに発信する仕掛け、そういったものをこの際しっかり作って、データ基盤を作っていく必要があると考えています。

それさえできれば、あとはデータの特性とモデリングのツールの特性がうまくマッチしていれば、精度の良い予測性のあるモデルはできると思います。そしてそのモデルを用いた逆解析、つまり精度の高い候補材料設計も可能となってきます。ただし、ここで考えなければならないことは、具体的に構造設計やプロセス設計などを行うため必要な記述子を織り込んだ統計モデルが構築でき、モデル解釈も行い、逆解析により提案された結果の信頼性も当然相応にあったとしても、それはあくまで材料開発を大幅に促進するための道具を利用していることを忘れてはいけません。いろいろな理論計算ツールがあるわけですから、材料候補が提案され種々の評価を通してそれらの候補が絞られてきたら、これらの理論計算ツールによってきちんと計算することも大切だと考えます。車の両輪のようなイメージですね。

ここで改めて申し上げるとすれば、なぜいまインフォマティックスなのかということです。材料開発、特に無機分野では第一原理計算に注目が集まった時代がありましたが、結局は産業上の実用に大きく寄与できなかった背景があることを忘れてはならないと思います。現場の材料開発部署では、社会のから要請に対していかに迅速に合理性をもって対応するかが求められます。これは世界の競争の中ではさらに深刻です。演繹的な手法もメリットはあるのですが、データを活用した帰納的な手法にこういう意味で世界が期待をしているのはいうまでもありません。帰納的な手法を利用する場合に意識しなければならないのは、対象をどういう記述子で表すかです。これについては先ほども述べましたが、解析と設計の物理的な意味を保持するために、理論計算から得られるパラメータセットも他の記述子と併せて用いることは有機低分子、薬物開発、機能性高分子開発でも日常的に行われます。材料、無機関係でもこのあたりを意識して帰納的手法のメリットを最大限に活用するように、マテリアルズ･インフォマティックスに関わる者全員が意識改革をしなければ、

我が国のマテリアルズ・インフォマティックスは真の意味で進展もせず定着もしないと考えます。ある程度の理論的根拠を以ってインフォマティックス手法により材料探索を行い、その候補の中で有望そうなものに対してのみ第一原理計算を行うくらいの大胆な役割分担を考えることが必要と思いますね。第一原理でうまくいかないからという理由で中途半端にインフォマティックスを持ち出したり、その逆にインフォマティックスでうまくいきそうにないからといって深い取り組みもせずに安易に第一原理計算を持ち出したりするのでは、これまでの失敗を繰り返すだけだと思います。この点は良く良く反省すべきことかと思います。

材料開発には、原料そのものも大切ですが、同じ原料でもプロセスが変わると全く特性の異なるものができます。原料、プロセスなどのスケールの異なる情報を同時に利用してモデル化し、材料開発のための設計が行えるポテンシャルを持っているのがインフォマティックスです。聞きかじりの手法や付焼刃的な対応ではなく、本腰を入れた対応を取らなければ手法とデータとのミスマッチも起こり、誤った方向で設計が行われることもあります。また、データベースについてですが、データベースを作ることだけを目的にするのではなく、利用する者がどういう情報項目を求めているかもはっきりとさせなければなりません。もともとデータベースは、その分野の研究者が、自分の研究を進展させる目的で構築してきたのが始まりです。この原点を忘れ、作る側と利用する側とが乖離してきたことが、気が付けば中途半端なサイズでしかも必要な項目が取り込まれていないために、結局誰も利用しないデータベースの山を築き本当に必要なデータベースがほとんど無いという現実を導いたのです。データベースは国家戦略に通じるとはアメリカのデータベース開発責任者の言葉です。その方が構築に関わってきたデータベース群はいまでも多くの人から利用されています。データベースは一朝一夕には構築できない。脇を固め、国家戦略の意味することをこの際良く考えて必要なことに取り組むことが求められていると思います。

福山秀敏（東京理科大学）

今日の議論では具体的な対象として、いろいろなものが一緒になって議論されているが区別して考える必要がある。基本的に物質の対象としては固体と分子系、つまりソフトな系、バイオ多自由度系の 2 つだと思います。この 2 つで状況がまったく違います。以下固体にフォーカスしたいのですが、固体でも物質に着目するか、材料に着目するかでこれもまったく違う。材料というのは私の定義では役に立つ物質という意味です。

物質と材料の間を見事にお一人で行ったり来たりしておられるのが細野さんで、新物質を見つけてそれ新材料にするという実績をお持ちです。その方がデータベースどうお使いになっているか、これをお聞きするだけで、今日のテーマ、つきているだろうと。加えて CREST の領域総括をやっていて、その中では今日ここにお見えの野原さんが非常にユニークな新物質を開発されている。野原さんのお話も少し伺ったのですが、やはり非常に印象的でした。

固体、これを物質としてみたときに当然物性が焦点になるわけですが、その観点で今新元素戦略が動いていて、磁石、電池、機能性材料、構造材料があるわけですが、それを具体的に考えるだけでデータベースがどういう位置づけになるかはかなりはっきりする。たとえば新しい磁石を作るときにどういうデータベースがあってどう使えるか？電池も同じです。

物質と材料の間にある大きなギャップをどう乗り越えるかというのが非常に大きな視点だと思いますが、物性は結局電子状態が関係していて、電子状態が関係したものも全部機能性材料と言っていいと思うのですが、これの予測というのは本当に難しい。物性のいちばんの基本は、金属と絶縁体の違い、相転移があるかないかの 2 つです。この一番基本にある金属と絶縁体の違いさえきちんと予測できていません。DFT（Density Functional Theory：密度汎関数理論）である程度できるけれども、たとえば高温超伝導母物質の銅酸化物、モット絶縁体、これに対してはできていない。また電荷秩序系もできない。できないから DFT が無力だと言っているわけではないのですが、そういう状況だということをまず意識して議論を進めなければいけないだろうと思います。相転移が関与している磁石、超伝導、これはもっと難しい。データベースがあったからといって、それですぐできるようだったら誰も苦労はしない。

データをどう使うかという点で、実際実験をやっている細野さんや野原さんのお話を聞けたのが、個人的には教訓的でした。両方の先生方は実際ある程度ケミストリーのデータから、マテリアルファミリーの見当をつけています。それがたとえば超伝導になるか、ならないかということに関しては、伺った範囲では NIMS の MatNavi が非常にいいレファレンスになっている。周辺のマテリアルがどういう物性をもっているかということを見るには確かにずいぶん役に立つだろうと思います。ただ元素式が並んでいるではなくて、その系がどういう物性値を持ち、どんな温度依存性をするのかまで含めて、こういうデータの 1 つとして非常にいい例で、日本としては大事だと思います。

材料にはプロセスが入り、例えば磁石などでは構成元素は同じでもどういう条件で作成するかで性質が変わってしまうわけで、これは別の種類の問題だと思います、これもやは

り分けなければいけない。大分けして固体とソフト系､固体の中でも新物質に着目するか、あるいは物質はわかっているけれど材料、プロセスを工夫するかという、ステージそれぞれ分けてそれぞれに適した情報が何かという分類をして、それにあったデータの整理が必要なのだろうと思っています。

寺倉：そういう意味で、このアクティビティに関しては田中先生が材料関係を担当してくださって、私は物理関係を担当して、二人三脚というか、細野先生にお尻を叩かれながらやろうかと思っています。野原先生には、ここでお話を伺ったほうがタイミング的としていいだろうと思うので、コメントをいただけますか。

野原実（岡山大学）

私は新超伝導体の物質開発を行っています。この研究では、NIMS の SuperCon という超伝導材料データベースと ICSD の結晶構造のデータベース、この 2 つがなくてはならないデータベースです。

新しい超伝導体を開発するとき、1 つのポイントとしてソフトフォノンに着目します。フォノン、つまり格子が軟らかくなると超伝導転移温度 T_c が高くなるという BCS 理論に基づいて物質を探す訳です。ではそんなもの（フォノンの情報）が載っているデータベースがあるかというと、それはありません。しかし、ICSD の結晶構造データベースから予測することが可能です。ICSD データベースの中には、記述子というのでしょうか、重要な情報として空間群が書いてあります。その空間群を見ると、構造相転移が起きて 1 つ対称性が落ちたらどうなるかな、という事が予測できます。ですから空間群を見比べると、そのデータベースの中からこの化合物はきっと構造相転移しているな、とか、構造相転移の直近にある化合物だなというように、ソフトフォノンがありそうな化合物を抽出することができます。1 つの例として、空間群 P-3m1 から C2/m に対称性が変化する化合物と抽出していくと、イリジウムとテルルの化合物で $IrTe_2$ が引っかかります。そこで最近この化合物の研究を始めました。

SuperCon を使うと、即座にそれが超伝導になるかどうか検索できます。SuperCon の、さらにすばらしいところは、その化合物が 0.3 K まで冷やされたけど超伝導にはなりませんでした、というネガティブデータまで網羅されているのです。実際 IrTe2 というのは対称性が落ちていて、0.3 K まで超伝導にならない、と書いてあるわけです。そこで、作業仮説をたてます。すなわち、対称性が落ちているから超伝導にならない、元の高対称性の相にうまく戻せれば、そこではソフトフォノンが使えて超伝導になるのではないか。ここからがデータベースに依存しない物質開発になるのですが、白金を少し混ぜると実際にもとの対称性が回復してそこで超伝導が起こるというように、実際にこの化合物を超伝導化することができました。

最近、新しい高温超伝導メカニズムとして、バレンススキッパーメカニズムが注目されています。例えば白金（Pt）等のバレンススキッパー（Pt^{2+}, Pt^{4+} だけが可能、Pt^{3+} のバレンス状態はスキップされる）と呼ばれるイオンを含む化合物では、局所的な「有効引力相互作用」による超伝導の発現が可能である、というのがあります。ICSD では、「白金の 2 価と白金の 4 価を含む化合物」という検索が可能です。そういう化合物を抽出し、それに化学ドープして超伝導にならないか、今そういう研究を進めています。ICSD で欠けているのは、「その化合物の電子状態がどうなっているのか」というのが見えてこないところです。カルシウム（Ca）の化合物で高温超伝導になるものがあります。Ca 自体は 100 GPa まで圧力をかけると 25K くらいのすごく高い T_c が出ます。また、グラファイトの層間にはさむと、CaC_6 というのが出来て 11K くらいの超伝導になります。第一原理計算によって、いずれも Ca の 3d 電子がフェルミレベルにかかると超伝導になると言われています。このように Ca の 3d 電子がフェルミレベルにかかる化合物を抽出したいのですが、それは今のところデータベースからは抽出しようがない。

今日、田中先生と竹内先生のお話の中で、米国のマテリアルズプロジェクトの中に ICSD に載っている化合物のバンド計算を入れこんだデータベースができている事を知り

ました。岡山に帰って、検索をかけて、その中でCa 3d軌道がフェルミレベルにかかる物質が引っかかったら、即刻合成したいと思っています。「この情報を検索したい」というのが具体的であればあるほど、きっといいデータベース、インフォマティクスのプロジェクトができるだろうと思います。

細野：今、野原先生が言われたことは非常に教訓的で、超伝導というのはこのくらい競争が厳しい。そういうところで予想もつかないのです。そうするとこれは一種のバイオに近いのです。バイオに近いから、それこそインフォマティクスでも活用しなくてはならないとも思うわけですが、そう簡単でもない。実は、超伝導が出る物質を見つけるのはそんなに難しくはないです。ただ、秋光さんの言う「松竹梅」の中の「竹」（T_c:77K以上）クラスの超伝導体を見つけるのが大変なのです。なぜか、MgFeGeは超伝導体LiFeAsの電子構造と似ているのに超伝導体にならない…。
寺倉：その違いが電子状態のどこから出てくるのか調べるのは、非常に教訓的な問題ですね。

中村振一郎（理化学研究所・三菱化学）

私は長いこと民間企業社で計算科学を用いた研究開発に関わってきました。いま、いちばん気になっているのは日本、その産業界つまり会社です。そこの仲間たちの顔が浮かびます。今日の、第4の新しいパラダイムが、会社を元気にするパラダイムになるかもしれない、そこでダイナミズムが生まれるといいな、と聞きながらずっと思っていました。ダイナミズムという意味は、データを集めれば、それを用いて、目的とする成果の予測ができる静的なプロセスではなく、それは瞬間的には起こるかもしれませんが、次の瞬間、データは古くなってしまうわけで、常にデータの新陳代謝が継続するという込み入った動的なプロセスの実現こそが重要です。このパラダイムがそういう運動を生むことを期待します。

今日の産業界では量子化学、分子科学を使うことは、もう当たり前になっています。私の所でも最初は1人だったのですが､今は平均すると15ー16人でしょうか。他の所でも、ある程度の規模の会社は大体同じ位のサイズの体制を持っています。これと同じように、今までなかった第4のパラダイムが会社に人を吸収する（ポジションを生む）ように発展してゆけばいいなと願っています。本当に役にたつなら、産業界のほうが、官学より貪欲でしょう。

最後に、これはいつも多くの所で、多くの方に申し上げていることなのですが、産業界の研究開発現場には、教科書にはない新しいサイエンスの芽とでもいうべき面白い現象、難問、課題がごろごろと転がっています。それらのほとんどは、いわば目的とする機能のスペックは決まっていて、それを実現する構成成分がわからない、という逆問題です。私の経験も大半がそれとの格闘でした。どこでもそのような問題には経験論で挑むしかないのが現実です。今日の第4のアプローチがこの打開策の一つになる予感がします。

常行真司（東京大学）

実験と計算の連携について少しコメントしたいのですが、計算機でできること、特に第一原理計算でできることというのは、限界が結構はっきりしているところです。できないことがいっぱいあるわけです。一方で、ミクロの情報に関しては非常にたくさんのことがわかって、たとえばアモルファスをやるというときには、非常に細かいことはよくわかる。そういう第一原理計算のデータをどのように扱うのか。これを実験でどう補っていくのか。これは大きな問題として残っています。これがもしDFT自体の問題である場合には、たとえばデータ同化の手法というのは使えないと思いますので、これをどうしたものかというのは大きな問題です。一方で、たとえば複雑な系の熱平衡の物性とか非平衡の定常状態の物性とか、そういうサンプリングが第一の問題であるような、計算機シミュレーションの限界を作っているようなケースに関しては、これはもしかするとデータ同化の手法が使えるかもしれないということで期待をします。

データベースのあり方の問題で、何をデータベースとするかというところでコメントです。たとえばSPring-8やJ-PARCなどで、あるいは先ほどの4Dの非常に綺麗な構造データがありましたが、ああいう最新の実験装置で膨大な実験データが出てきていて、それと、マテリアルインフォマティクスの手法というのを直接つなげることは非常に難しいだろうと思います。ですから、そこに何かしら意味のある二次データをマイニングするプロセスが必要に思うのですが、もしかするとそこに原子論的な計算機シミュレーションの重要な役割があるのかもしれないという印象をもちました。もう1点は、界面や欠陥、あるいはマルチスケールの構造、粒界のようなもの、もっと複雑な材料の物性値をどうやってデータベースにするのか。あるいはデータとして使えるのか。そこが今現在、私にはよくわかりません。これは考えていくべきことかもしれません。また、ネガティブデータに関して、NIMSのデータベースにそういうものが入っているというのは知らなかったのですが、ネガティブデータをどうやって集めればいいのかというのは、これは非常に重要なデータであるわけで、それを集める手法は何か考えるといいかもしれません。

今後の展望でマテリアルインフォマティクスの研究を誰がやるのかというところ。たとえば計算機シミュレーションをやっている人たちでこういう手法に乗り出そうという人がどれくらいいるかというところが未知数です。非常に大事だとは思うのですが、これが若い人たちに参加する意欲をもたらすことができるかどうかというところで、予算措置をするのがいちばん簡単なのかもしれませんが、そこの方策を何か考える必要があるかと思いました。企業でこれに対してどれくらい前向きに取り組まれるのか。企業の方がどう思われているかというところは知りたいと思いました。

寺倉：界面などの複雑な系のデータベースは例の画像データをデジタル化するという足立先生のお話がありました。いろいろなスケールの組織の画像をデジタル化するというのと結構似ているかな、と思うのですが、ともかく膨大なものになりそうなので、確かにそれをどうあわせるのかというのは難しい問題かもしれないですね。

曽根純一（物質・材料研究機構）

今日はたくさん示唆に富むお話を伺えて、いろいろな言葉が頭の中に残ったのですが、「源平合戦の時代ではなくて、現代流の戦いをしないといけない」、そうだと思います。いろいろな研究の手法、あるいは研究から生まれたいろいろな情報、知識化された情報というのは年代とともにリニアに進展してきていると思う。一方で、いろいろな計測手法はべき乗での進歩だろうし、コンピューティングのパワーは指数関数的増大です。ですから、我々はコンピュータの本当の力をもっともっとうまく使えるのではないかと、まだ利用しきれてないのではないか。そのときに、ただこれはツールであって、いろいろな新しいものを創造するのは人間であって、そこのインスピレーションがどう働くかがキーになると思う。人間のいろいろな物事を考える能力というのは、我々の脳のほとんどのパワーは、目から入ってくる画像を情報処理することに使われている。たぶんそれによってインスピレーションが働くのだと思います。

コンピュータの１つの理論で、モデリングしていろいろなメカニズムを理論的に理解するということもあるのですが、もう１つはやはり視覚化というのですか、いろいろな現象を非常にわかりやすく絵にできる。そこからいろいろな想像が働いていろいろな新しい物質につながっていくだろうし、現象の発見にもなっていくだろうと思う。同じような意味で、データベースもやはりそれを使って全体を俯瞰できるというのが大きなパワーになるのではと思う。いろいろなデータが、ポイントではなくて全体で見える、それによって人間がそれにインスパイアされて新しい発見をする。また、SVM みたいに機械学習を使って、いろいろな新しいパラメータを介して知識を発見することもあると思います。そのときにやはりデータベースによって非常に俯瞰的に情報が集まってくるというのが、非常に重要なのではないのか、と思いました。

あと、実験と理論ですが、物質あるいはデバイスを理論で全部理解するというのはほとんど不可能で、現実の物質というのはものすごく複雑でいろいろなことが起きている。ところが、計測というのはある意味である現象のワンショットの情報ですよね。だがワンショットの情報を使って１つのモデルなり、自分のアイデアなり、あるいは理論が正しいのか検証できる。計算のすばらしさは、その後、横展開すぐにできるわけです。検証されたモデルを用いてパラメータを動かしたり、何が起きているのか分析して想像することもできるし、あるいはそれを可視化して、いろいろなツールとして展開することができる。

NIMS は物質材料のデータベースを、MatNavi という形で世の中に提供しています。これはオープン・プラットフォームで、ユーザー登録していただければ誰でも利用できます。毎月百数十万件のアクセスがあります。ユーザー登録もものすごい勢いで伸びていて、今 7 万人の方が登録しているという状況です。このデータをどうやって作っているかというと、有機材料は発表論文のデータから新しい情報を抽出してデータベースを作っている。また、もう１つ重要なデータベースとして我々は、金材研の時代から、クリープですとか、腐食ですとか、疲労、そういう金属で重要な情報をデータベース化して、それも我々自身で実験のデータをとって、それを毎年データベース集として発行しています。ただ、これらを全部開放していいのか、ある部分プロテクトすべきではないかなど、日本の産業競争力ということもありますので、国益との視点で、今議論になっています。また、

課金などをどう考えていくのか。そういう難しい問題があって、今それは議論中という状況です。ただ、データベースはNIMSの中でエンジニアの人たちの大変な努力でこれを作成しているのですが､こういう作業というのは地道であるが故になかなか評価されない。こういった作業は、我々国研の任務だと思っているし、研究コミュニティへの支援としてやっているのですが、ぜひこういうものを積極的に活用して、こういう使い方があるのだということを研究者の方々にアピールしていただければ、いろいろなアプリケーションを開発していけると思います。

緒形俊夫（物質・材料研究機構）

補足させていただきます。今、いちばんアクセスが大きく、月 70 万件近いアクセスがある無機材料データベースは格子定数や状態図、また物性値を出していますが、10 年前のデータで留まっているということで、ユーザーの方々からはぜひ更新してくれ、という話があるのですが、なかなか更新がうまくいきません。1 つには予算的な面もありますし、また、コピーライトの問題もあるので、その辺はまず予算あっての上で更新かと思っています。さらに高分子のほうも抽出していますが、スタッフの高齢化という課題もあって、今後どのように継承していくかということについて課題を抱えていることも 1 つ報告させていただきたいと思います。

ただ、今後さらにこれをマテリアルインフォマティクスとして日本の技術の継承、データの継承を図っていかなければいけない、と思っています。

あとは先週オーストラリアのマテリアルインフォマティクスの会議に参加したときに、講演の中にペタフロップ級のコンピュータを各都市に置いて、ペタバイトのサーバをつないで、マテリアルサイエンスの高速ネットワーク、インフラストラクチャーを作りなさいと、ということで、今後 5 年間で 100 億円の予算をつけるような話もしていたので、今後日本としても何かそういうインフラの下に結集して、いろいろな材料の分野について京を活用するような材料選択なり特性の推算をするようなプロジェクトをぜひ考えていただければ、と思いました。

小谷元子（東北大学）

今日は、数学がどのように関われるのか、という観点でずっと話を聞いていました。数学の中で、ここ 20 年、30 年で非常に進歩した分野があります。その一つが樋口先生のお話しされたデータの中に隠れた構造を見出す手法に関わる分野です。もう 1 つは先ほど足立先生がお話しされたように、数学は複雑な形や現象からそれを表す簡単な指標、もしくは数値化することが得意です。先ほど例としてあがっていた曲率とかトポロジーとかは、いろいろな形を数値化する概念として幾何学が開発してきたものです。曲率やトポロジーなどの概念は 19 世紀、20 世紀前半に生まれ発達したものですが、扱われてきた対象は、滑らかで連続な、いわばマクロな現象、マクロな形でした。それがここ 20 年、30 年で、たとえば特異点がある形や現象、ミクロな視点でみる離散的な図形などを扱う数学、すなわち離散幾何学、が急激に発達しました。以前は扱えなかったものの曲率やトポロジーを取り出せるようになりました。材料のように複雑な対象に数学がどれくらい新しい視点を提供できるのかわからないのですが、私自身の興味は、最近開発された数学の道具がどのように役立つのか、更にデータベースなどにも、反映できるか、というところにあります。

数学は、少しずつ複雑な現象も扱えるように発展してきていますが、実際に現象を扱っている研究者と直接会話をすることによって、数学も更に発展するし、また現実の複雑な現象の解明にも貢献できるといいと思っています。1 カ月後に「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」という題目で、常行先生や塚田先生と一緒にワークショップを開催します。これは文科省の支援を受けている「数学と他分野の協働によるイノベーション創出のための研究推進プログラム」の一環です。数学と他分野が対話をすることによって何か新しいことができるのではないかという期待で、こういう企画を始めました。かなり面白い第一エンカウンターになると思うのですが、言葉が通じない可能性も非常に高く、そこはむしろそういうところから何か新しいものが生まれると信じて、活発な議論できればいいと思っています。

もう 1 つ、私が今いる材料科学の研究所（WPI-AIMR）は異分野融合を推進することがミッションの一つになっています。特に昨年度から数学と材料科学の連携、かなりチャレンジングなことと言われていますが、を強力に推進することになりました。その推進のために、インターフェース･ユニットという組織を作りました。インターフェース･ユニットは、理論物理や理論化学の若手研究者の集まりです。特徴は、若い研究者を一つの研究室に張りつけるのではなくて、独立研究者として、彼らが自発的に動けるようにしたことです。異分野融合推進においてコーディネーターが必要だということは必ず言われますが、コーディネーターという言葉の意味がどのように使われるかは大切で、特に若い研究者に自由に自発的なアイデアで動く機会を与えることは、新しい分野を開発するためにとても有効なのではないかと思います。

伊藤 聡（理化学研究所）

30 年くらい産業界でこういうシミュレーションをやっていて、最初の 3 年くらいはすごく役に立ちました。それはなぜかというと、半導体をやっていたからです。半導体はロードマップがしっかりしていて、ITRS があるので、3 年後に何が起こるかがわかるのです。つまり、ニーズがはっきりしているので、そこからやっていても間に合うのです。ところが、半導体を離れた瞬間にニーズは突然来るわけです。突然、こういう何か材料開発できませんか、触媒の活性があがりませんか、それを今年度中に出してほしいと言われて、だいたい間に合わないのです。だいたい何か状況がわかるのはプロジェクトが終わった後や、製品の開発が中止になった後などになってしまうわけです。ですので、産業界的に見ると、とにかくニーズドリブンで始まったときに使える仕組みがやはりいろいろ欲しいと思います。そのためにはもちろん物性研究に基づくような、そういう第一原理計算、ソフトウェアもきちんとする必要がもちろんあります。

しかし、京があってもできることはたかだか今の計算機の二桁か三桁くらいなのです。そうすると、もっと違う何か道具を使わないと画期的なことができないというのが私のすごく強い意識です。そこに今日お話があったデータ同化なり、あるいはインフォマティクスが使えたらぜひいいな、と思っているのが一番です。

もう 1 つはデータの充実に関しては、先ほどオープンリンクドデータ、RDF の話もあって、私もああいうのはぜひ進めていただきたいと思っているのですが、もう 1 つぜひ必要だと思っているのは、やはり成功事例です。産業界が見る成功事例という意味では、融点が見つかってもやはり少しインパクトが少ないのです。なので、細野先生が時々おっしゃるように、新しいものを見せろ、というのはそのとおりだと思っています。1 つでいいからそれができればいい。逆に言えば、そういう芽のあるところに皆で支援するようなことをしたほうがいいのではないかという気がしています。

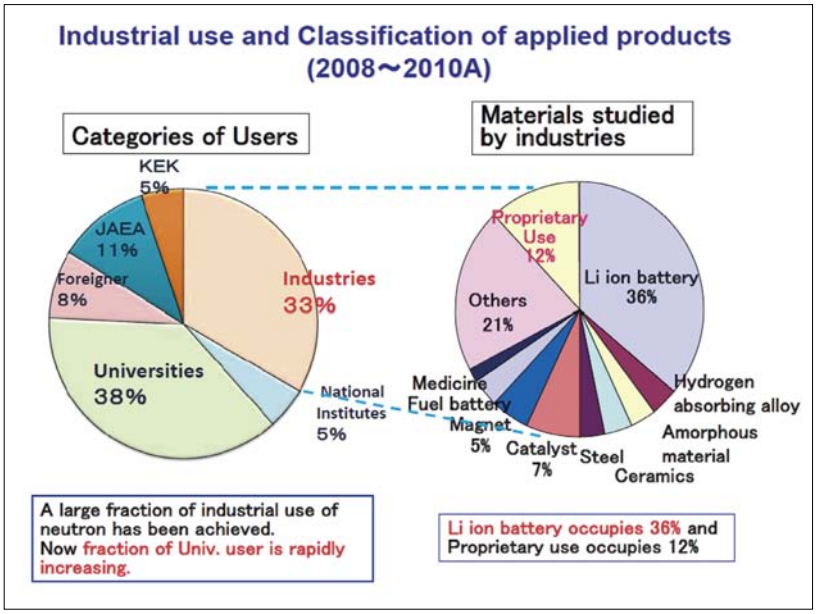
最後にもう 1 つ申しあげたいのは、全体のビジョンをやはり共有しないといけなくて、その意味ではビジョナリーをきちんともたないで、あちこちでいろいろなものがばらばら開発されたらものすごく無駄になると思うのです。果たしてトップダウンがいいのか、それともこういうコミュニティでビジョナリーをうまく語っていくというほうがいいのかというのはわからないのですが、いずれにしてもそういうものがないと、なかなか動かないのではないかと思いました。

とにかく、企業的な観点からすると、ツールはなんでも…もちろん私も物性をやってきましたから、記述子の意味は絶対知りたいと思いますし、それがなんだと言われたら、意味があるのだったら説明したいと思うのですが、その前にやはりものがきちんと早く作れるという仕組みがぜひ欲しいというのが切実な希望です。

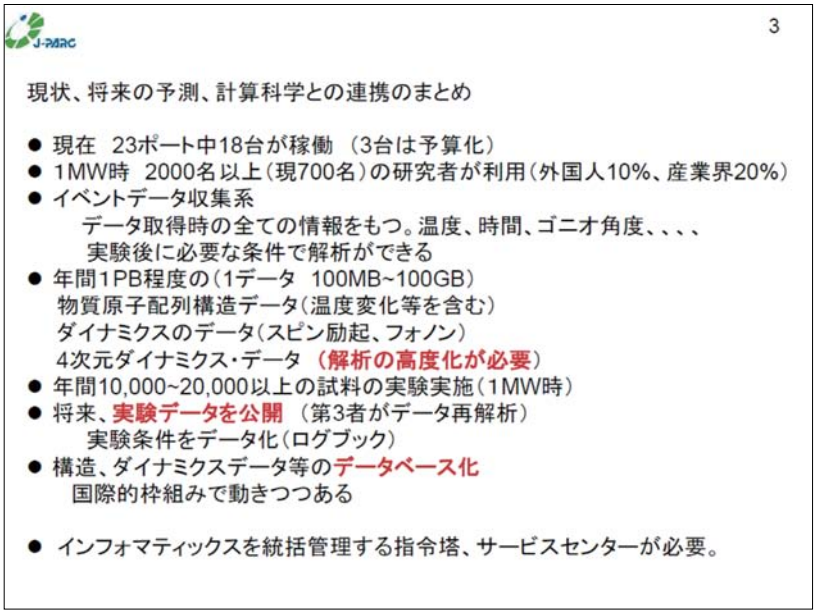
新井正敏（J-PARC センター）

今日は大型施設がいったい何を考えているのかということを少し話したいと思います。まず J-PARC の予算が認められた 2000 年に、当時の文部省と科技庁の間の、いわゆる事前の審査部会というのがありました。そこで、J-PARC は国際公共財です、と。これは世界中の皆さんにサイエンスの内容が良ければ無償で実験させましょう、ということが述べられました。これは、日本が国際社会に貢献することによって我が国が海外から認められることこそが国際競争力の一部なのだという考えに基づくものです。ですから、すでに大勢の外国人が J-PARC で研究を行っています。その典型例であるニュトリノ実験には現在ユーザーが 600 人いますが、そのうちの 400 人が外国人です。しかしながら、確かに海外に負けてしまっては困りますので、たとえば今回始まりました元素戦略を施設主導の研究プログラムに入れ込んで、その部分だけは優先利用してもらうような仕組みを作ることにより、国際貢献と我が国の国際競争力の推進が両立できるよう進める必要があります。

以下に 2 年前までの利用の統計を示します。中性子利用としては他に例がないほど産業界からの利用が大きいことがわかります。昔は大学の先生しか中性子は使いませんでした。ところが、今では日本を代表する約 50 社の企業が中性子やミュオンを使いにきているのです。



それらの企業が持ち込む材料の 40%が実にリチウムバッテリー関連の材料です。企業は自身の研究内容を表に出しませんから、もしかしましたら同じ材料を研究していることもあるかもしれません。施設で実施される研究課題は、申請をして審査を行って認められた課題です。この課題は実験が終わりましたら成果の公開が義務付けられています。非公開利用の課題は有償課題となります。その金額は 1 日 350 万円くらいです（現在は減額措置により 200 万円程度）。相当な額ですが、企業に聞きますと、研究内容を外部に出したくない場合この額は想定必要経費内とのことです。



J-PARC 物質・生命科学実験施設では 23 ポート中 18 台の中性子実験装置が稼働しており、大学、産業界、海外の研究機関の研究者に利用されています。さらにプラスして 3 台の予算化がなされました。実験室はほとんど満杯状態です。予想では、2 ー 3 年後には 2000 名くらいの方々が使いに来るものと考えられます。そのうち、10%以上は外国人です。あるいは更に増えるかもしれません。中性子に限ってのことですが、産業界からの利用は 20%以上と見ています。

材料研究で大事なことはその材料が使われる環境で状態を見ることです。従いまして、本施設ではその場環境での実験が行えるよう、設備、データ収集系を用意しています。その場測定ではイベントレコーディングという手法を使います。研究対象から散乱された中性子一発一発をその時点での試料の状態、たとえば温度や圧力などの情報とともに記録します。従いまして時々刻々状態が変わるような現象でも計測できます。また、中性子ですから非常に厚みのある、炉の中にある高熱の金属でも中性子が中まで入ってその状態を観測することができます。

さて、J-PARC 物質・生命科学実験施設において予想しているデータ量は年間 1PB 程度です。これは物性研究としては相当な量だと思います。1 データセット当たりで約 100MB ー 100GB です。この中には、物質の原子の配列構造のデータがたくさんあります。この種の実験を年間 1 万ー 2 万の試料について行うことが予想されています。ですから、この辺の情報が皆さんの目が届くような場所に置けるような仕組みさえできれば、相当活用されるのではないかと思いますし、あるいは先ほど野原さんが言っておられましたが、ダイナミクス、あるいはフォノンのデータなど、それは確かに世の中に整理した形で公開されていません。ですから、そういったものもきちんとデータベースに入れるということも考えらとよいと思います。

さらに、これまでのデータとは次元が異なるほど精度の高いデータが取れます。たとえばブリュアンゾーン中の三次元空間、更にエネルギー空間の四次元のデータが一測定で取れる時代になってきましたので、その複雑なデータを解析するには計算科学の人たちの力

を借りる必要があります。

第三者が実験データを公開するというのはどういうことかと申しますと、たとえばある実験が終わってから 3 年間後くらいに、第三者が同じデータを使って新しいアイデアのもとデータ解析して新発見をするような道筋も作りましょう、ということです。こういった大型施設のデータ自身は実はお金にすると相当な額です。一人の人がたとえば 1 週間実験に来るだけで中型科研費の総額に匹敵する額がかかります。従いまして、たとえば高温超伝導物質の動的構造因子の測定をして、スピンに興味があるからスピンの部分だけ解析して論文を出しました、というだけでは済まない時代ではないかと思います。同じデータの中にはフォノンの情報が入っていますから、これを解析して、たとえば BCS 理論に基づく超伝導機構について考えました、ということもありえるわけです。将来的にはそのような公開データの利用についての道筋を作る必要があるのではないかと考えています。このようなわけで、施設としては、解析された構造やあるいはダイナミクスのデータのデータベース化、あるいは元々の実験データの公開ということを今後は検討していきたいと考えています。

ただ、ここで大事なことは、大型施設のデータベース、あるいはインフォマティクスを進めるにあたり、管理すべきコンテント、記述子、範囲、あるいは方針をきちんと決める、司令塔あるいはサービスセンターを作る必要があるのではないかと思います。これがありませんと、国内にある多くの施設を巻き込んだ一体的な活用が行えません。

3. 第２回ワークショップ議事

３－１　挨拶

田中一宣（科学技術振興機構 研究開発戦略センター）

本日はお休みのところお集まりいただきまして、ありがとうございます。この問題は日を追うごとに、私自身、認識が随分甘かったと思うぐらい、進展が速くて、今日はこれ関して今後どうするかを、少し具体的に見通しをつけることができればいいと思っております。これに関して､寺倉先生のご指導の下にいろいろなアドバイスやご指示をいただいて、仮のプロポーザルを作るところまで来ております。

私自身は 4 月に、MRS でサンフランシスコに行きました。そこでガバメント･エージェンシー・フォーラムというのがあり、NIST で Materials Genome Initiative を担当し、NSTC が最初にレポートを出した時の特別委員会のメンバーの一人である James Warren の話を聞いてきました。2011 年 6 月からちょうど 2 年経つわけで、相当なプレッシャーを感じながら新たなメッセージを用意しようとしている情熱が感じられました。特にデータをデジタル化、標準化したり、そのためのデジタルデータ・コミュニティを設立するといったような動きがありました。

また 2 週間前にも、ベルリンでナノテクノロジー国際会議、これは日米欧の三極のものでしたが、その中でやはりアメリカが力を入れているのはマテリアルとマニファクチャリングです。これも同じように NIST や、Advanced Manufacturing National Program Office という新設されたところだと思うのですが、そこの Michael F. Molnar の話によれば国防総省が Digital Manufacturing and Design Innovation Institute を建てるとか、ネットワーキング化も含めて、相当な勢いでやっているということを、身をもって感じてきた次第です。

私自身はこの議論に関連するところとして一番重要だと思っているのは、システム的に物を考えるところです。インフォマティクスというのはすべての領域を越えて、系統的にデータを集めて、それを基にして最適化を行うというような作業だと思うのですが、そういう学際とか、あるいは連携とか融合とか、そういうことが非常に苦手な国民性があるわけです。コミュニケーションが非常に下手だということがあって、この分野だけどうも日本は勝てないというところがあるわけです。

私はこの今日の議論をどういうふうにプロジェクトとして展開していくかとか、国としてどうするべきかという議論の他に、中・長期的に考えますと、インフォマティクスだとかデータ駆動型の研究開発とか、こういうものは学校の必修科目のひとつにすべきだと思っているわけです。これは国語とか数学とかあるいは英語とか、それと一緒に、そういう物を若い時から入れていきませんと、おそらくいろいろな面で、今後、競争に負けていくのではないかという気がするわけです。これは、本日の会議と直接関係ないのですが、そんな印象をもっています。

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）

このワークショップは、実は2月の前に、12月にこじんまりしたものを一度やったので、個人的には今回で3度目なのですが、やる度にアメリカで何をやっているか等、いろいろな情報が入ってきます。確かにわれわれのコミュニティの場で、もうひとつ盛り上がらなくて、文科省からも「すぐにプロジェクトにしましょう」といった積極的な発言を頂けるレベルにまで至っていないのは、ひとえにコミュニティの盛り上がりがまだ不十分であるということだと思います。

これについては二つあって、ひとつはやはり、データをいかにして活用するかということに関して、アメリカは非常に進んでいて、これまでものすごく努力をしてこういうものを作り上げてきたのに対し、それに刺激されてわれわれは動いていて、もうひとつ、一番の根本のところからの積み重ねが足りていないのだろうと思っています。

一方で外の動きを見ながら、このままでは困るということで、できるだけ早くこういうプロジェクトを、何かの形で動かさなくてはという気持ちも非常に強いことがあって、京都大学の田中先生はかなり奮闘して来られたのですが、奮闘していてなかなか実らないと、だんだん疲れてくるので、是非、疲れ切らないうちに何か実りのあるものになんとかして繋げなければならない。そのためには、できるだけコミュニティを盛り上げていかなければと思うわけです。こういう会合を何回か重ねることで、少なくとも参加者が、こういう動きが明確に重要であると確信に至ってくれれば、かなり迫力の違った動きができてくると思うので、どこがこういう問題の本当の核心なのかということを、できるだけ把握していけるようなワークショップになればいいなと思っております。よろしくお願いします。

３－２　話題提供

「ベイズ推論と物性科学」
岡田真人（東京大学）

私は修士まで物性理論を研究していました。そのころ 1985 年ぐらいに神経、脳のモデルを作るとすごいコンピュータができる可能性があるということでニューロンブームがあり、そこで方法論として統計力学が脳科学に使えることがわかってきました。この分野は統計物理学者が活躍していました。このタイミングで研究の分野を脳科学とデータ・サイエンスに変更しました。脳科学というのは物質科学と異なり、第一原理から離れているので、理論と実験の融合が難しいのです。そのために、データ駆動型のアプローチを取り入れて、そこで一番成功したものの一つがニューロサイエンスではないかと思います。さらに今から約 10 年前に複雑理工というところに移って、物質科学も再スタートし、統計物理を使って、物性と脳と情報科学を研究しています。

今日、話題として話させていただくのがスペクトル分解です。マルチピークのスペクトルを、ガウス関数のようなシングルピーク関数に適切に分解するという問題です。こんな単純な問題がインフォマティクスかと言われるような内容ですが、何か分光をしたときに、マルチピークスペクトルが出てくる。それをシングル・ピーク・スペクトルに適切に分けるということで、物質科学にとっては非常に重要な問題です。適切にピークがわからないと誤った電子状態を推論します。だからいくらデータを取っても、シングル・ピーク・スペクトルに適切に分けることがきっちりとできないとダメなのです。当然ですが、シングルピークの数が多ければ多いほど、データにフィットします。でも例えば、n 個の点があって、n-1 次元の多項式でフィットできても、それはダメなわけです。実はこのスペクトル分解は数学的には非常に難しい問題で、代数幾何の広中平祐先生の特異点解消を議論しないと解決しない問題なのです。

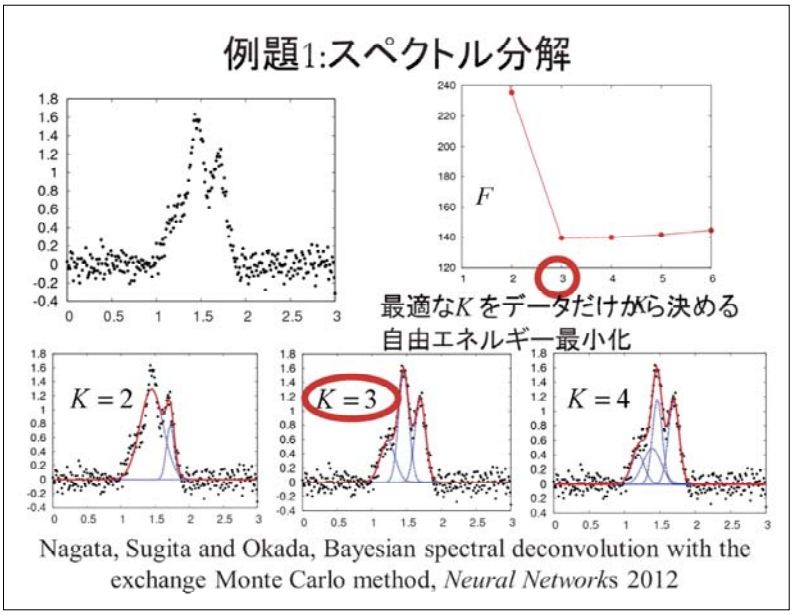


図 1

この例のように、スペクトルを分解するときに、この解像度だと二個のピークだとわかる。でも例えばシンクロトロン放射光でとれているスペクトルを、研究室内の装置で観測すると、ノイズが多くて二個のピークだと判断するのは難しい。だからコストをかけてシンクロトロン放射光を使うわけです。でも、ここで示すように、自分のラボでとったようなXPSでも、二個のピークと判断することができます。そういう限界を知ることも重要ですし、もっと重要なのはいろいろな人がとったデータがあったときに、どれを信用して、どれを信用しないかということをインフォマティクスを使って判断したい。実は開口時間について、この境界である種の相転移が起こって、シャッターの開口時間を4 msまで落としてもOKだけど、それ以上はダメだという判断が出来ます。これは非常に単純な問題設定を使って、データの質がどこまで保証されるかということを、数学的にきちんと定義する、という話の良い例になります。そういう話で得意なのは実はベイズ推論です。ここで何をしているかというと、電子状態を知りたい時にX線を当てて、その運動エネルギーを測ると電子状態が見えるけど、それがボヤっとしているので、それを適切な個数のシングルピーク関数に分離したいという問題です。

物質開発をするときに、レアアースの代替品が欲しいときは、強相関電子系を創らなければならないはずです。そのときに、第一原理計算だけでは不十分だということはわかっています。そこでわれわれはモデル・ハミルトニアンというものを作ります。いろいろな励起状態をきちんと求めないといけないときは、第一原理計算では、今の計算機でも無理です。だから物理的直感でモデル・ハミルトニアンを用意して、パラメータ・フィットして、本質を抽出するということは、今も続いているはずです。私の指導教官だった小谷先生はコアからd準位に電子が上がると、ある種の相互作用をアンダーソン・ハミルトニアンというのに付加しないといけないという定式化をおこないました。私は修士時代、このモデルについて一所懸命パラメータ・フィットをしていました。

そこから、系を記述する記述子やパラメータを最適に選ぶことをオートマティックにやる必要がある。つまり、与えられたデータだけからきちっとした情報をどうやって獲得するかという系統的処方をわれわれが持っている必要があると感じました。いくら大きなデータベースからたくさんのデータをダウンロードしても、やっていることの信頼性が担保される必要があります。データ駆動型のアプローチでもいろいろな立場があります。多分、マテリアルズ・インフォマティクスが一番効くのはベイズ推論だと私は思います。

新たな枠組みを作るために重要なのは、良い例題だと思います。データベースが大規模にあってデータマイニングしたという場合に、これまでの手法で物質開発をしている人にとっては、「本当にそうか」と信じられない気持ちが出てきます。ここでの例題は、マルチピークスペクトルをGaussianでフィットして、そのパラメータを最適化する。単にそれだけです。ただし、深い問題があって、実は局所解という問題があります。それを解決しないと、いくらデータを与えても最適なものは得られないという問題があります。

われわれはどうするかというと、確率の世界で物事を考えます。これがモデルです。われわれは観測しているのであるということを確率的に定式化します。それで最適化するのですが、ひとつの方法はレプリカ交換法です。どうするかというと、これは実際の焼きな

ましに対応するシミュレーティド・アニーリングと呼ばれる方法の双方向的な拡張で、ヒーティングとアニーリングを並列で行います。この手法でスピングラスの基底状態など、熱力学的な状態をいろいろ計算できます。どういう仕組みでご利益があるかというと、こうやって交換をしていくと最終的にいろいろな温度での最適値をとり、基底状態を探せたりします。

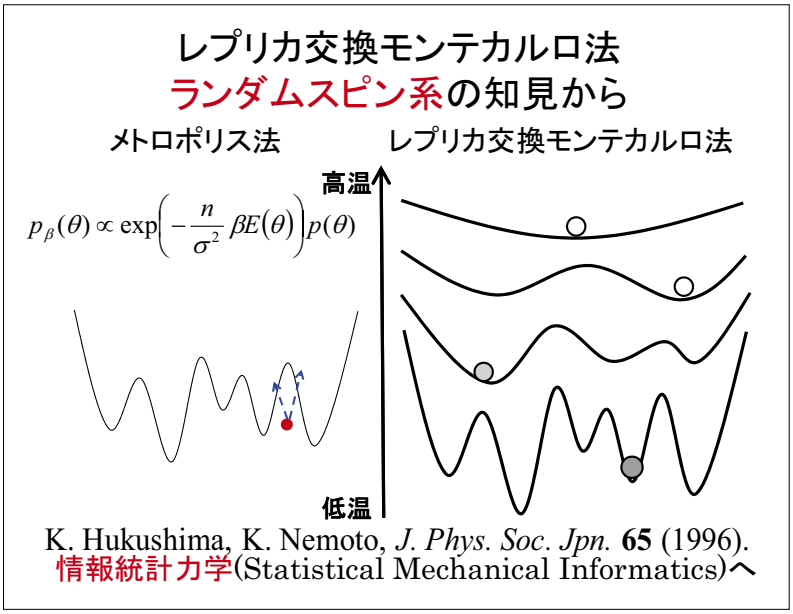


図 2

マテリアルズ・インフォマティクスをやる場合は、情報統計力学という分野は重要だと思います。これは 10 年程前から特定領域が走っています。実はここで話すベイズ推論と統計力学、特に分配関数の計算は実は数学的には同じことをやっています。統計力学の手法を使うことでベイズ推論の計算量爆発をできるだけ回避するということは、もう 10 年以上前から始まっています。もうひとつマテリアルフィンフォマティクスにとって重要なことは、物理がわかって情報がわかる人材が絶対に必要なのです。そういう人材が多いのも情報統計力学です。

データ駆動というと何かすごく新しいことを言うような気がしますが、実は第一原理もしくはモデルを組み込んでいます。つまり、あるパラメータがあったときにデータ自身が確率モデルというもので記述されるという書き方をします。この式は原因が与えられる確率、つまり条件付き確率です。われわれが知りたいのは、データがわかったときに、原因であるパラメータがどうわかるかということです。これは科学がやっていることそのものです。これを系統的にやろうと思うとベイズ推論を使うのが実は素直です。何がいいかというと、つまり知っているモデルをデータ解析に組み込めます。つまり、第一原理計算すらデータ駆動に入れることは可能です。難しい問題もいろいろな方法で最適化できますし、さらに、これは確率を計算しているのでデータからそれぞれのパラメータがどういう確率分布になるかということも自動的に計算できています。

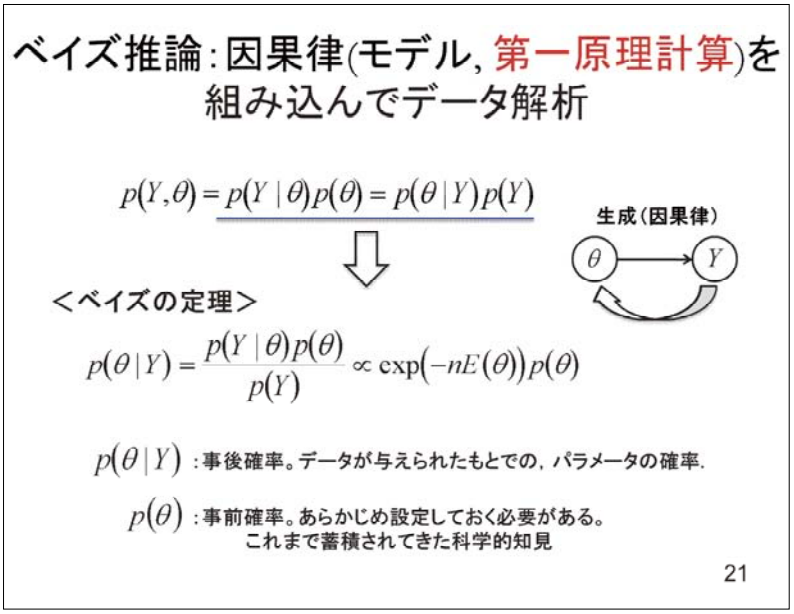


図 3

データがどれくらい精度が高いかということをきちっとわからないかぎり、人のデータベースは使えない訳です。そういうことが系統的にできる方法が､マテリアルズ･インフォマティクスで一番重要だと思います。

例えば Gaussian の中心とか Gaussian の幅が決まるというのは何となくわかるわけですが、先に Gaussian の数を決めないといけない訳です。欲しいのはデータが与えられた時の Gaussian の数です。われわれは一応全部モデルを持っているので、全ての確率分布を定式化できます。ここで重要なのは、考慮したくない自由度の系統的消去ができることです。つまり、ここを全部積分することで、実はデータと Gaussian の個数の関係を自動抽出できる。これがベイズ推論の強力なところです。その時に何が問題かというと、計算量爆発が起こるわけです。つまり、全てのパラメータ、サーチ空間をサーチしないといけないので、例えばパラメータ・フィッティングが 5 次元の時、各パラメータ 10 通りであれれば 10 の 5 乗のサーチをするわけです。このようなときでも、スピングラスの方法を使ってうまくサンプリングすることによって、この積分をかなり数値的に高速に計算できます。だからそれを積極的に使いましょうというのが情報統計力学です。

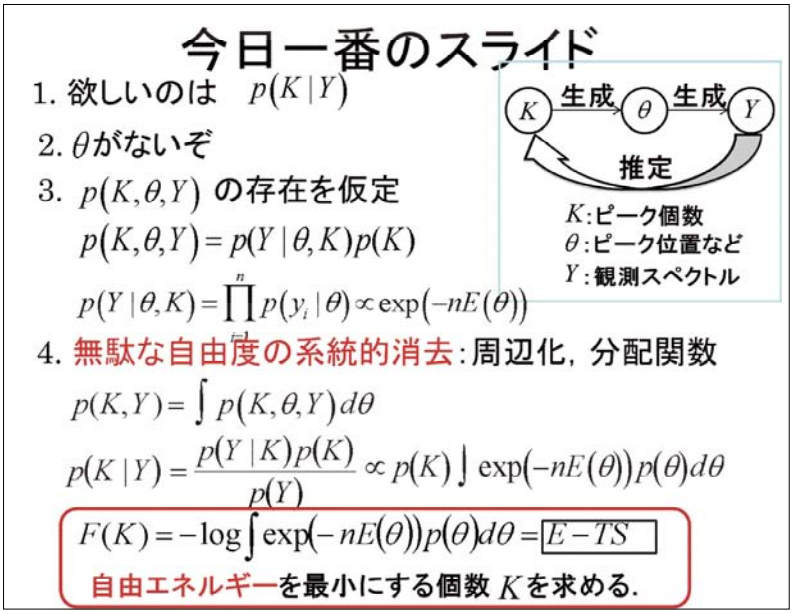


図 4

具体的には、実は自由エネルギーというのが出てきて、E というのがデータフィットの項です。だから、どれだけ誤差を減らすかということで、こちらがコンプレクシティもしくはエントロピー。つまり複雑なことをやって、誤差を下げてこういう罰金項が出てくるのです。その罰金項を実は自動的に抽出できます。罰金とデータフィットをどうやって混ぜたらいいのだという問題もありますが、実はそれを求めることが可能です。

さらに自由エネルギーを計算するテクニックはいろいろあって、サンプリングして自由エネルギーを求めるのはすごく難しいのですけど、これは大学の時に学ぶ自由エネルギーからエネルギーを計算する式というのがあって、内部エネルギーは実はサンプリングで簡単に求まるのです。だから、内部エネルギーをサンプリングしておいて、積分することで自由エネルギーを求めるということができます。実はレプリカ交換法ですべての温度で熱平均を取っているので、これはもうシミュレーションが終わった時点で、モデル選択に必要な情報は得られています。Gaussian での分解を考えたのは、一番簡単でパワフルで、誰もその適用範囲の広さに対して反論できないからです。例えばハミルトニアンをいれて対角化して、フェルミのゴールデンルールを入れてそれでパラメータがフィットすれば、私がやりたかった問題は明日からでもできます。ただしそれだけやっても、「ああ、それは強相関電子系の話だね」って言われるだけなので、一番広範囲で適用可能な例を研究しました。

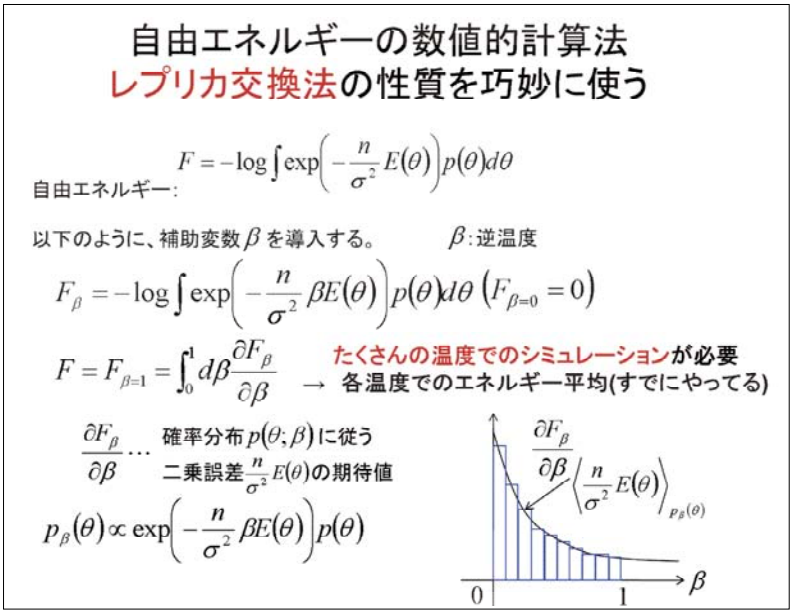


図 5

さらに、この問題は数学的にも非常に深い問題で、実はさっきの問題というのは特異性を持ち、自由エネルギーの回りの展開がガウス展開できないケースになっております。よく使われる隠れマルコフモデルとか、多くの確率モデルは特異構造を持っています。つまりガウス積分をして自由エネルギーを求めるということは不可能です。そこで広中平祐先生の特異点解消を使って、ゼータ関数を求め、自由エネルギーを計算したりします。一見、データ駆動型科学に関係ないと思われている代数幾何が関係しています。枠組の一般性としては、ガウス分布のところをポアソン分布に書き直せば、先ほどのように量子化ノイズの話ができます。それで、パラメータに関する確率分布も計算することができます。これは先ほどの確率分布を可視化したもので、ダブルピークだと、こういうふうに出るのですが、ピークの場所がどれくらい揺れるかということもわかります。ここまでは赤と青が分かれているということは、ピーク分離が可能なのです。この境界で、多分、相転移が起こっています。その相転移ポイントをピックアップすることで、そのデータ解析の結果が使えるか、使えないかがわかります。

このスライドがデータ統合を表します。今は一個の結果だけですが、ひとつの物質に関していろいろな測定方法があれば、いろいろな生成モデルがあるので、それを確率だと思ってベイズ推論をすれば、ノイズが大きいデータからの情報は少し取り入れて、少ないデータからの情報は多めに入れることが可能です。さらに情報を持ってないデータを落とすということすらできると思っています。

例えばこの枠組は STM にも使えます。表面のパターン形成を議論する時には、隣同士の原子が相互作用して、例えば原子が付けやすいか付けにくいかというのを取り入れます。例えば金森先生は研究されていた DAS モデルもその一種です。その相互作用をどこまで入れなければいけないかということも、先ほどと同じ枠組で議論できます。

まとめです。計測データとモデルを比較する。このモデルをこの中にビルトインして全

部、確率分布で書いて、それをベイズ推論でひっくり返しているだけなので、これは皆様がやっていることを系統的にやり直しているだけです。そこに数学的な手続きが入っている。これがひとつ目のメッセージです。

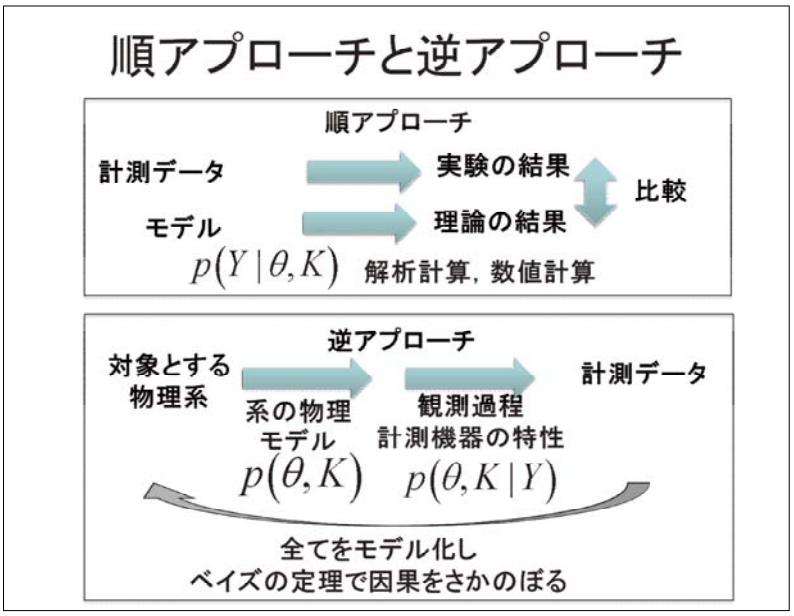


図 6

後、モデルがないということもあるわけです。この時はスパースモデリングをやってください。記述子の自動抽出ができます。スパースモデリングとは、いったい何かというとブラックホールが撮像できたり、MRI がいろいろ見えたり、いろいろあるのですが、画像という物は、どういう仕組みで作られているかわからないです。音声もわからないわけです。でも、例えば画像は圧縮できます。つまり画像がウェーブレットで圧縮できるということは､画像のある種の統計的な特徴を表しているわけです。音も高調波が出ますから、必ず倍音とか 3 倍音は出てくるわけです。そういうことを全然知らずに元の観測データをスパースに、つまり 0 ベクトルが多いような状況に表現する。基底を自動抽出するという問題を解きます。つまり何も知らなくても、答えはベクトルの表現がコンパクトになるだけということを課すと、実は自動的にウェーブレットや音声が出てくるということがわかっています。これは脳の研究でも示唆されていて、脳の視覚野の最初の段階には、実はウェーブレットに対応するような処理があります。聴覚野にも、和音を抽出するものがあります。つまりスパースモデリングというのは計測データを解析する手法ではありますが、実は脳がどうやってセンサー入力を処理しているかということにも使えます。何故ならわれわれの頭の中に数式が入っているわけではないので、なんらかの形で学習によって外界をセンスしないかぎり、物事はうまくいかない。

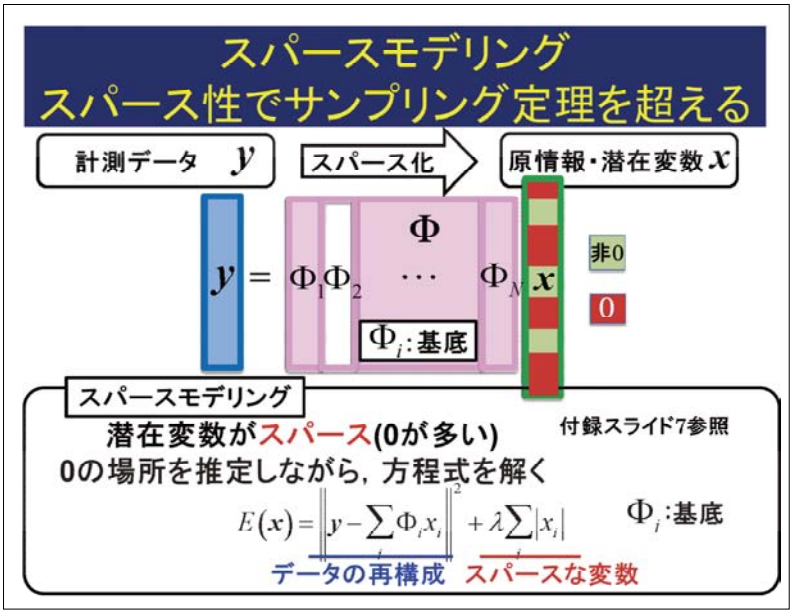


図 7

動機は、モデル・ハミルトニアンから派生しています。エフェクティブ・パラメータを自動抽出するということは、実は機械学習をきちんと使わなければならないということになります。そういう枠組の中には数学的に非常に深いものが入っています。全てのモデルを確率的に書けるので、どんどん高度化することができます。モデルがある程度わかっているところはベイズを使う。全然わからないところはスパースモデリングを使う。最終的にそのスパースモデリングとベイズ推論を融合することは可能なので、そういう指針でマテリアルフィンフォマティクの一部を推進することができると思います。

【質疑応答】
質問者: 一番初めに挙げられた例で言えば、パラメータを決めるのとモデルを決めるのと、両方が絡んでいます。モデルを決めるところは直感で決めるのですか。
岡田：はい、決めます。ただしモデルを大きくすることは可能です。例えば候補として、ハバードモデルとかアンダーソンモデル、全部、入れられるのです。しかもそこに付加パラメータをいろいろ入れられる。それぞれの自由エネルギーを計算することによって、実はその系はハバードモデルの方が的確かどうかが、自動決定できるということです。これまでは、それを人がやっているわけです。モデル計算をやって、いろいろな電子状態を計算して、その結果を実験の結果とあわせて、「よし、これだ」って言っているわけです。そのあわせる過程で、先ほどの E の誤差を計算しています。普通の場合は物理的直感で、こういう効果があるからこれぐらい入れて、モデル計算して、実験の結果をあわせて終わりな訳です。その点、ベイズ推論を使うとモデルハミルトニアンとして何が的確かを原理的には決定することが可能だと思います。

質問者：今回、多分、物性科学の物性というのは予測という問題に適用されて、それが例題だと言われていると思うのですが、例えばマテリアルを作ろうと思ったときに、そこで出てきたパラメータを入れて何か新しいことをできるのですか。

岡田：それは、ここにはそのまま入っていないです。ただし、例えばモデル・ハミルトニアンをやるときは何をやるかというと、例えばレアアースに関して連続的にパラメータをフィットして、その内挿で物事を予測するわけです。だから、いろいろな典型的なマテリアルに関して全部やっておいて、そこを少なくとも内挿することは可能なのです。では、内挿に対応する実際の物は何かというと、それはこの枠組みではわからないです。何でもできると言っているわけではありません。少なくともこういうツールがあれば、ディスカッションする時も系統的に議論することができます。そうなってくると、例えばいろいろなデータベースを統合するときも、それぞれのデータベースのクオリティ自身を、計算機を使って評価することも可能かもしれません。このベイズ推論だけで、物質開発が可能になるというわけではありません。ただし、データの持つ不確定要素を系統的に取り扱うことはできます。

質問者：データの無い範囲のところを外挿はできるのでしょうか。
岡田：それは基本的には無いと思ってください。だいたい人の思考はほとんどが内挿なんです。外挿するときも、内挿の状況を考えて、外挿しているので、基本的には外挿も内挿だと思います。その前提で機械学習がひとつできることは、内挿のジャンプを予測することです。例えば、全ての物質は混ぜたら連続的なんてことはなく、相転移が起こるので、どこかに飛ぶのです。その時の飛びをデータベースから自動的に抽出することは可能かもしれません。これは、どういう方法を使うかというと、画像の中から視覚物体を抽出する方法です。ほとんどの画像は連続的ですが、どこかジャンプするわけですね。境界線の抽出は、不連続な点は連続に繋がっているという性質を使います。現実的な観点で重要なのは、外挿よりは境界線を見つけることだと思います。

質問者:モデルが優秀であれば外挿もできそうな気がするのですが、そうでもないですか。
岡田：それは、外挿してみるだけです。これほど素晴らしいモデルなので、勇気を持って外挿しようかとする、これは人の意思です。コンピュータは判断しません。

「材料設計とデータベース」
及川勝成（東北大学）

私はどちらかというと物質発見というよりは物質開発とか、本当の実用化まで進めるような材料の開発をやっています。専門は計算熱力学であり、実用的な材料の熱力学関数をデータベース化していくというのも仕事になっています。金属材料がメインなのですが、金属材料ではミクロ組織が特に重要視されます。そのミクロ組織がどういうメカニズムで形成されるかというために、相平衡というのを研究するのですが、その辺を計算熱力学、また計算熱力学をするためのモデリング、あとはそれらを使ってデータベースを作る。あとは計算熱力学のモデルを作るために、その妥当性を検証するために実験等もやっています。

どういう材料を開発してきたかというと、快削鋼、これは削りやすい鋼です、あとは削りやすい銅とか、磁気記録媒体用、もう最近ハードディスクも段々なくなってきているのですが、盛り上がっていた時にはハードディスクの研究もしていました。あとは強磁性の形状記憶合金、コバルト基超合金、高強度亜鉛合金等の材料の開発をしていました。その他として最近は、基本、相図を書くための熱力学関数を求めているということなので、フェーズフィールド法や伝熱流体という計算をする。鉄鋼材料屋さんは、1 トンとか 2 トン、材料を造るわけです。そうすると、その中が均一であるわけがない。そういうマクロスケールな偏析がどうでるかを予想するために、こういうものと連携して計算する。こういうデータベースを作るときには、やはり実験だけではなかなか予測できないところもあるので、ab-initio 計算や MD と連携させて、実験からでは予測できなかった状態図のデータベースを作っていくということをやっています。

最初に鉛フリー快削鋼の話をしたいと思います。鉄も用途によって様々な特性を持たせます。この快削鋼は、とにかく削りやすさを改善するというのが目標の鋼です。ただ、鋼本来の性質である強度は持たなければならない。加工性もそのままです。ただ、必ずどこかドリルで穴をあけたり削り出しの加工をしたりします。その時の加工費用を下げるために、通常、鉄鋼メーカーでは鉛を入れていきます。鉛を入れるとどうなるかというと、鉄中に鉛が溶け合わずに単体で出ているのです。それが、液体潤滑作用みたいなものを起こして、削りやすくしていたというのが快削鋼の特徴です。ただ最近、RoHS 指令等の問題で、鉛はもう使えないと。そこでその代替物質の開発が望まれていました。

我々がこの時注目したのが、Ti4C2S2 という層状化合物で、潤滑性が強いものを考えました。先ほど言ったように脆くしてはだめで本来の鋼の性質はキープしなければならない。かつ、よく知られている固体潤滑剤ですね、モリブデン・サルファイドもあるのですが、それは鉄の中で析出しない。モリブデン・サルファイドを少量入れただけで析出すればいいのですが、そうでなければ使えないということで、計算の結果、モリブデン・サルファイドは、鉄中には生成できない。そういう計算を CALPHAD 法でやります。合金設計の指針として、ステンレス鋼をターゲットとしています。鉄、クロム、マンガン、シリコンは必須元素で、だいたいの組成が決まってしまっている。そこにカーボンも必ず入っている。それでサルファーも不純物なのですが必ず入っています。その中で、何かを少量入れて切削性を改善する物質が欲しいということになります。そこで注目したのがチタン

だということになります。実際いろいろ最適組成を決める時には、まずソルビリティ（溶解度）がほとんど無いことが重要で、ただ、Ti4C2S2 だけをこの鋼中に析出することは難しくて、有害な析出物、例えば TiC が出ると非常に固いのでだめです。クロムの炭化物はステンレスの耐食性を落とすのでだめです。マンガン・サルファイドも、ステンレスの耐食性を落とすのでだめです。そこで、ターゲットとなるのは、この辺の組成です。こういう相図の計算をすれば、だいたいわかってくる。ただ、実際作り込みのときには、必ずしもこの Ti4C2S2 が簡単に出なくて、これは、TiS と TiC を混ぜただけの自由エネルギーよりも、ほんの少し安定なだけの化合物なので、下手をすると TiS と TiC だけしか出ていないということもあるので、プロセスには少し工夫がいりました。

だいたい材料屋の仕事としては、ここまでで終わりかなと思うのですが、会社のエンジニアと二人で「でき上がったので造ってください」と言ったら、会社の組織はやはりいろいろあって、実際に造る人は考えが全然違って、ここのチタンがコンマ 55、こんな鋼、連続鋳造で造れるわけじゃないと一蹴されるわけです。チタンはノズル詰まりというのが起きるので、何百トンと溶かす溶解炉で、何十トンもお釈迦にしたら誰が責任を取るんだ、ということになるわけです。材料開発というのはフェーズによって、求められるもの、コントロールしなければならないものが違ってくるということになります。実際は物ができたのですが、どうしてノズル詰まりが起きなかったのかというのは、チタンも入っているのですが、カーボン、サルファーを高くしているので酸化しにくいということで決着しています。

次は、この強磁性形状記憶合金です、これは磁性を持ちながら、熱弾性型のマルテンサイト変態をします。従来にはない物性というのは磁場をかけると 10%近い歪みが生じ、今、開発段階というか実用化されている中で、一番大きな磁歪を出すのはターフェノール D と呼ばれるもので、0.1%ぐらい。それが一桁オーダー近い高い歪みを出すということで、センサーなどに使えないか、あるいはアクチュエーター材料等に使えないかということで、開発をしています。そういう意味ではこれはまだ発見開発段階の材料です。我々の研究グループでも、こういう材料がどうしてマルテンサイト変態しながら磁性も持つかというのは、何となく 3 元形状のデータを眺めながら、この辺怪しいぞというのでスクリーニングして見つけていくという手法をとっていました。

ただ、デバイス化の材料の段階に入ってくると、何が求められるかというと、単結晶化や、ある程度の大きさ、形です。単結晶でも単なる棒だと駄目ですと。ワイヤー状であるとか、いろいろな形にしないと材料にはなりませんよ、と言われるのですが、単結晶化するためには、液相と母相が接触しなければいけないという問題が出てきます。あとは単結晶ではなくてもよいと言われても、この相はけっこうボロボロで形を与えるのは大変なので、塑性加工できるように、何とかしなければならないということになって、その時にはこういうのに添加元素を加えていって材料開発をすることになります。その時に必要なのが多元系の状態図だとか、結晶構造の情報が必要になってきます。

次はコバルト基超合金の話になるのですが、超合金というのは基本的に高温での強度が優れているという材料で、ニッケル基超合金というのがよく使われます。ジェットエンジンや発電のガスタービンによく使われるのですが、こちらに示すように、ニッケルのマト

リクスに Ni3Al という $L1_2$ 構造の結晶が規則正しく析出しているというのが、非常に重要な組織になっています。このニッケル基超合金というのはだいぶ前に開発され、どんどん改良が進んできています。ただ基本的には、どんどん改良するために多元系化する。それで、入れる元素もだんだん、非常に高価な元素を添加しているということがあります。我々はそういう中でコバルト基超合金を発見しました。これはデータベースを構築しているときに、たまたまγ’（$L1_2$）がコバルトと平衡していました。実はコバルトは、周期律表だとニッケルの隣なのですが、2 元系では $L1_2$ が現れる系がほとんどありません。ただ、このデータベースを作っているときに、3 元系になればこいつが出るかもしれないということで、実験をして、コバルト基で初めてガンマプライムの組織を見つけました。これは、高温での特性もけっこうすぐれていました。ただ、高温特性がニッケル基超合金よりもよいだけではなく、他の特性もすぐれなければならない。まず求められるのが高温時の耐酸化性です。やっかいなことにタングステンが入っていると、耐酸化性が良くないということで、クロムを入れなければならない。あとは耐疲労性やクリープ強度等も求められます。

あとは、プロセス面です。こういうのは鋳造、つまり一度、金属を溶かして流し込んで造るということで、鋳造がちゃんとできますか、大型化して大型のインゴットができますかといった問題があります。あとは鍛造です。鍛造というのは金属を潰して造っていくのですけど、それができるか。それで、ある一定の温度に温度を高くしてフォージングするのですが、それができるかどうか。その時は多元系状態図、格子定数、密度などが必要になってきます。これは凝固のマクロ偏析で、インゴットの中は均一でないというのが、X 線 CT イメージでえられます、こういうのをシミュレーションできるようにやっています。シミュレーターを作って、会社に「こんなシミュレーターを作れたので、使わないですか」と言ったら「本数が違う」とか、会社からはそういうことまで要求されるのです。そうすると何が重要になってくるかというと、物性値の正しさです。モデルはだいたいいいとなると、今度は入れる物性値の正しさが重要になってきて、そのためのデータベースが必要になってきます。ここで重要なのが、状態図、熱力学データのラインナップなら CALPHAD 法でいいのですけれども、さらに粘性や熱伝導と、あとは密度です。密度の組成依存性と温度依存性です。後は熱伝達係数です。これは少し実験的なものになって、そういうものを求められます。

後はフェーズフィールド法と絡めてコバルトクロム基の二相組織が出る原因を明らかにしました。こういうのがわかれば、こういう組織になると、磁気記録媒体として優れているよということだったので、他の系にも展開できる。それでコバルト・モリブデンやコバルト・タングステンとか、そういうものを発見しています。この時にも重要になるのが、熱力学情報だけではなくて、拡散係数だとか、界面エネルギーだとか、そういうものが重要になってきます。

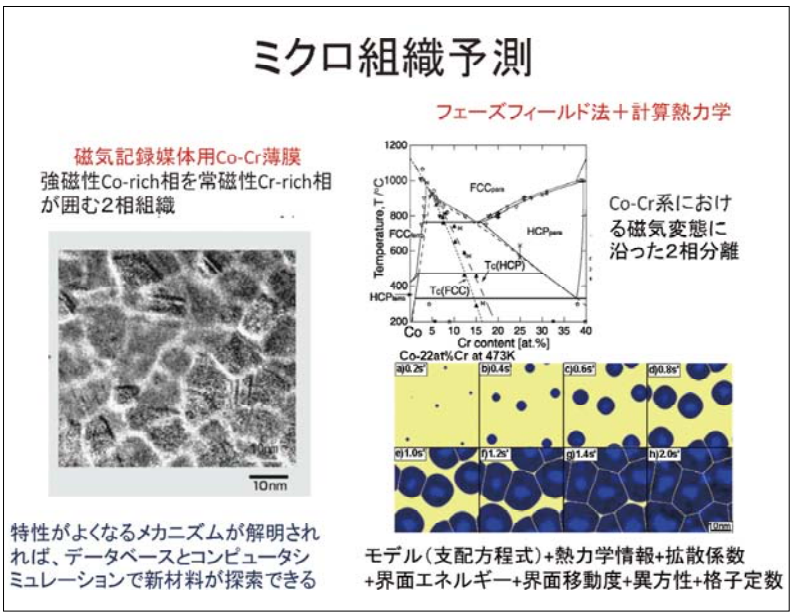


図 1

材料開発の流れで一番重要なのはどういうことかというと、先ほども言ったように、発見、開発です。あとは組成の最適化。特に大学の先生というのは、この辺まで多くの研究の対象にしているのかと思っています。ただ､実際会社で使っていただくときというのは、生産プロセスをどうするかというのも非常に問題になってきます。大学だと、とりあえずデバイス化までやっている先生もけっこう多いかとは思います。そのデバイス化、システム化の時に何か問題があれば、もう一度こちらに戻って最適化をしなければならない。そこまで終わって、信頼性もあるということになっても、では生産プロセスはあるか、既存の生産プロセスに載せるなら、今の生産プロセスを壊さないようにちゃんと設計をしなければならない。あとは、既存の生産プロセスでできないものを造るのですといったときには、新しいプロセスを作ることで量産化になるということで、やはりプロセス設計のための材料設計が重要になるかと思います。

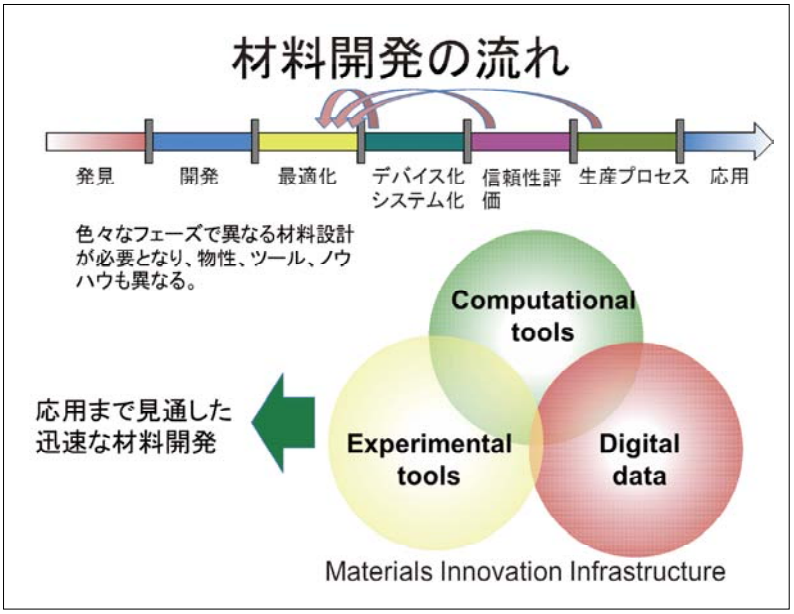


図 2

材料設計といっても、ここで必要なツールとこちらで必要になってくるデータベース、その辺がちょっと違うのかなと思います。迅速に、応用まで見通した迅速な材料開発をしましょうというのが、多分 NIST の James、私はそこの研究室グループに一時期いたのですが、彼らが言っているのはこの辺のことだと思います。

CALPHAD 法なのですが、基本は状態図と熱力学を組み合わせて状態図を計算しましょうという話です。まず最初に重要になってくるのは、実験状態図です。これはアメリカの金属学会等が編集しているのですが、その中にいろいろな状態図のデータがあります。それをまずデジタル化する。次にギブスエネルギーを組成と温度の関数でモデリング化してやる。その中に最適化するためのパラメータが出てきます。それで、モデリングが最適であれば、ギブスエネルギーと相図は一致するべきであるということでギブスエネルギーのパラメータから相図が計算できます。それで問題になるのが、実験データはいっぱいある、そのクオリティはどうなのだということが重要になってきます。その時に、このモデリングが正しければ、不正確なデータははじきながら、状態図が計算できるようになります。だからここで、データのクオリティを保証するというのと、ギブスエネルギーの組成と温度の最適化が行われます。

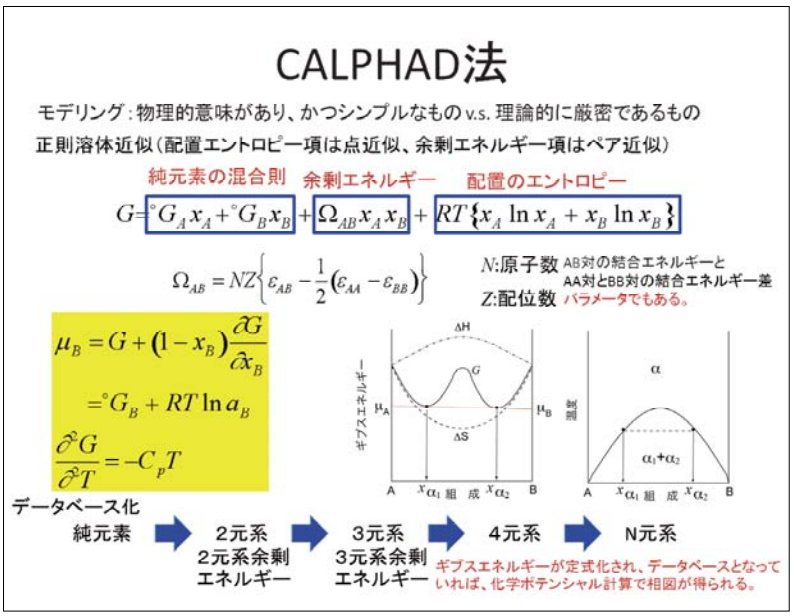


図 3

それを使って相図を計算する。自由エネルギーが組成と温度の関数で書かれているので任意の温度､組成で計算できる｡ここで関数をデータベース化するのと状態図をデータベース化することを行います。具体的にはどういうことかというと、モデリングは基本的にはCALPHAD 法ではシンプルで物理的にある程度意味のあるものをやりたい。ただ、物理屋さんの方としては、理論的に厳密なものをやりたいということで、この辺が、工学に近い人と物理屋さんでだいぶ違ってきます。一番シンプルなモデリングとしては、工学屋は統計力学の教科書に出てくる言葉で言えば、正則溶体モデル、配置のエントロピーは点近似で、余剰エネルギー項はペア近似ですということをやりますと。そうすると自由エネルギーが近似でき、ここは純元素の混合則になる。いわゆるセグレゲーション・リミットとか､リファレンスとか我々は呼びます。配置のエントロピーは組成だけで決まってきます。そうすると、この余剰エネルギー項を決めれば、ギブスエネルギーが決定するということになります。この余剰エネルギー項の書き方としては、ペア近似で書けば、こういうふうに書ける。AA ペア、BB ペア、AB ペアの値さえわかれば基本的には計算できるはず。このモデルをもっと複雑にして、第一原理から求めるということも可能なのですけど、エイヤッとこの辺をひとつにして、これを工学的な実験値に当てはめてやりましょうというのが、CALPHAD 法になります。いわゆる実験データに合わせ込んで決定していく。それをどうデータベース化していくかというと、まず純元素の G が決まっており、次に 2 元系ですが、2 元系はこいつを決まれば表現できる。2 元系を 3 元系に拡張すると、必ず 2 元系のデータは取り込んできて、3 元系で合わない、3 元系が 2 元系からずれる余剰項というのが必ず出てきます。それも、まあ、これに、ABC と書いてあって組成が 3 回かかる積になります。

次、4 元系をどうしますかというと ABCD になって組成が 4 回掛かる。こういうふうになるのですけど、組成が 4 回掛かると 1 以下の項が 4 回かかるということで、この Ω の数字がすごく大きな数字にならないと自由エネルギーにコントリビュートしなくなります。4 元系以上は無視して、3 元系までを組み合わせてデータベース化して多元系を計算

しましょうというのがストラテジーになってきます。このデータベースさえあれば、あとはギブスエネルギーミニマム式を解いて、状態図を得られる。

最後に CALPHAD 法の展開として、これまでは CALPHAD 法というのは熱力学的なものだけだったのですが、最近はこういうモル密度、モル体積とか熱膨張係数といったものを定式化して、CALPHAD 法と似たもので混合法＋余剰項という概念で定式化してあげて高圧下での相安定性も計算できるようにしましょう、というのもやっています。あとは、我々材料屋としては実験値が欲しいというよりは、信頼性のある実験値の温度・組成依存性が欲しいということになります。それで、その時にデータのクオリティということですが、拡散係数ですとか粘性とか熱伝導など、バラエティが富みすぎている。ワンオーダー違っても当たり前で、それをどう評価していくかというのが、CALPHAD 法をやっていこうとしたときに問題になっていきます。

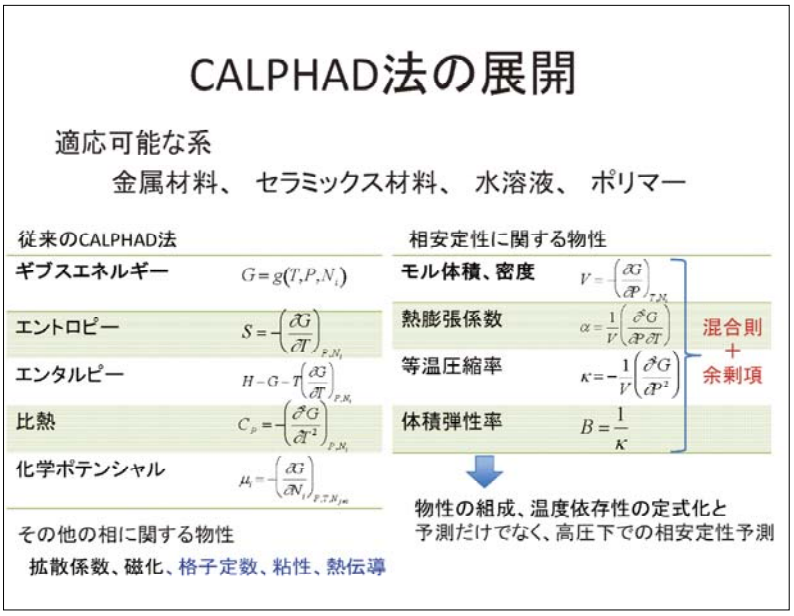


図 4

あとは、従来は実験データからの外挿だけだったのですが、それだと 2 元系には無い相を、突然 3 元系で求めるのが難しく、特にこういうのは 2 元系ではものすごくスタビリティが低い。実線が安定系状態図で、こういうところを実験では出せないわけです。それでも 3 元系を求めるためには 2 元系が必要になるので、2 元系で実験で求められないところを ab-initio を使いながらやっていくということをやっています。

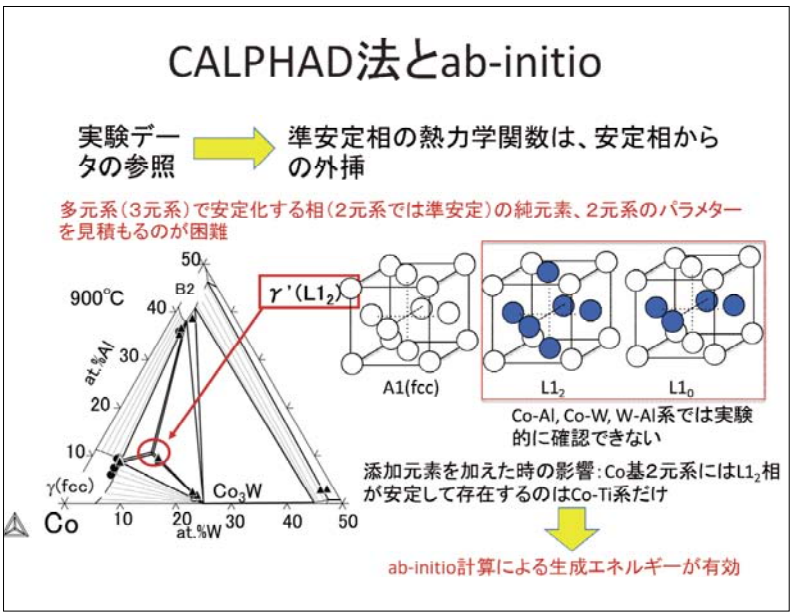


図 5

もうひとつ、こういうことをやるときにひとつ重要になるのが、ボルンのスタビリティクライテリア、いわゆる安定性です。フォノン分散にイマジナリティスタビリティの項が表れるというのと多分、同じことだと思うのですが、立方晶系が安定になるためには、弾性定数がこういうのを保証しなければいけない。こういうのを満たさない物は、第一原理で計算しても、我々が予測しているような比熱には乗ってこない物質になるので、こういうのを除きながら進めなければならない。また、第一原理だけでもかなり分配定数等は予測されます。

材料設計の時には最初、発明、発見というのは、データを眺めながら、やはり非経験的な予測で、材料特性を発見しなければいけないというのは、確かにあると思うのですが、従来の知見を応用しながら、こういうプロセスモデルやミクロ組織予測をやって、材料最適化ツールなんてものができたらいいなと思っています。

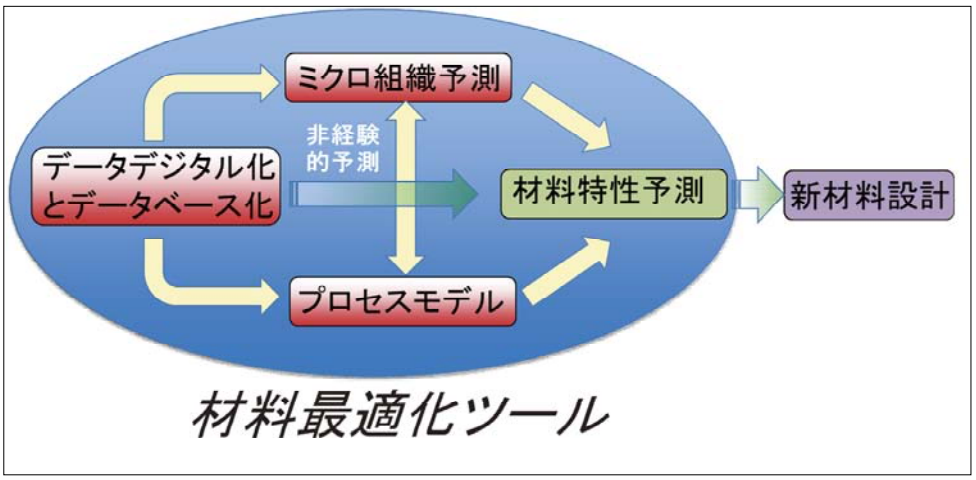


図 6

その中でCALPHAD法はどういう役割を果たせるかというと、実験、あるいはab-initio、アトミック・スケールのデータからデータを解析し、それを温度、組成、圧力の関数で戻してあげてデータベースを構築する。あるいは計算結果のデータベースの構築に役立つのかなと思います。

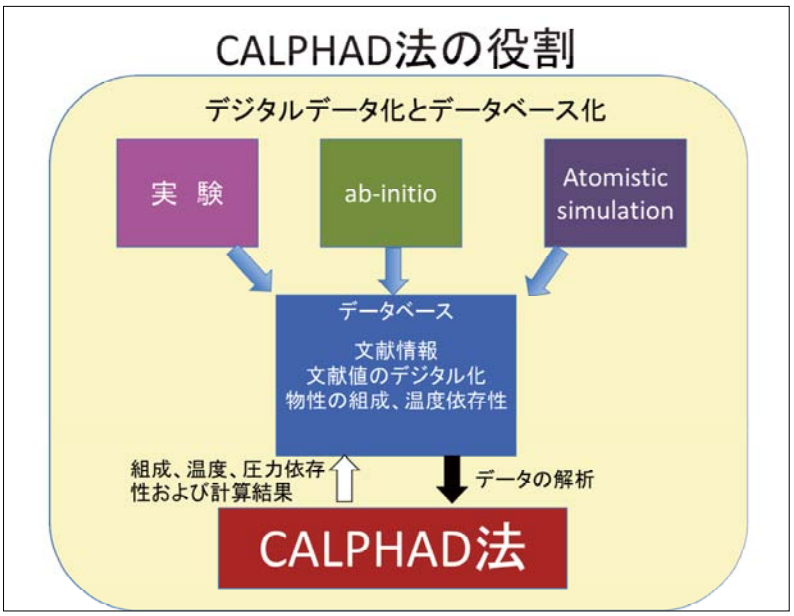


図 7

【質疑応答】

質問者：鋼に関しては素人ですが、相平衡をまず仮定されているのですが、その相平衡条件は鋼の材料であれば、必ず満たされているというふうに、すでにウェル・ノウンなりリザルトというのがあって、CALPAHD法というのを使われていると解釈すればよろしいですか。というのは、CALPHAD法を使っているということは、熱力学関数を使っているので、基本的には平衡状態の理論を使っているわけです。つまり平衡状態に材料がなければならないのだと思うのですが。

及川：物理屋さん的なイメージだと、平衡状態からずれているというか、準安定状態で平衡計算できるのか、という話になると思うのですが、こういうのをやるときは、基本、まず最初に一番安定な相を取り除いて、その後、いわゆる平衡計算をする。だから、そういうのは我々、準安定状態と呼んで、そこまでも予測しながら計算すると。けっこうアニーリングをするので、ある程度は平衡に近い状態になっている。

質問者：それは現実の系でも起きている。

及川：そうです。ただ問題となるのは凝固の途中とかです。そういう時はやはり非平衡状態なんですが、我々が使うモデリングとしては、界面は平衡です。界面では分配係数が並行状態に近いと、あとは拡散方程式とか解いて、それがどんどん拡散していきます。局所平衡モデルと呼ぶのですが、局所的に界面は平衡に近い値を取っているというのを仮定します。

質問者: すごくうまくいっているように思うのですけど、困ることはどういうことですか。

及川：今まで合金でやっているときは、ギブスエネルギーのモデリングは、実はあまり問題にならなかったというか、ちょっと複雑にしたもので何とかなったのですが、最近、酸化物等も含めて、メタルから酸化物、絶縁体までいくモデルを計算しなければいけない時に、このシンプルさではちょっと無理です。
質問者：多分、これはけっこうランダムの現象のところにあって、フェーズトランスレーションが起こるところでは駄目なんですが、先ほども言ったように、境界をうまく求めて、別のモデルをフィットすると。しかも、別のモデルの境界線を系統的に求めると、うまくいくのではないかと思うのですが。
及川：多分、一番典型なのはチタンオキサイドだと思うのですが、チタンオキサイドというのは TiO の NaCl なのですけど、ベーカンシー、チタン側に、O のベーカンシーがいっぱいありながらずっと酸化物まで連続的に変化していくのです。でも、チタン側はメタルリッチなのです。オキサイドはオキサイドなのです。それで自由エネルギーがものすごく途中でとんがるのです。そこのとがり方を区切るのか、でもフェーズは繋がっています。
質問者：要するに連続パラメータに関してフェーズトランスレーションが起こっているだけなので、そこで境界を引いていいのではないでしょうか。そして、違う自由エネルギーがヒットすればいいということだと思うのですが。
及川：ただ、1 次変態ではない。相変態的に言うと、2 次変態とか 1 次変態とか、そういう変態は起きていない。

質問者：先ほど、発見から応用まで一種の製造プロセスみたいなので、いくつかの段階にわけて進化していくという話がありましたが、その中に最適化というのがあって、最適化という要素については、その後、デバイス化とか信頼性評価とか、あるいは生産プロセスとかそういうところから返ってきて最適化をやるという話がありましたが、その中の生産プロセスというものの中にはコストとかいうものが入っているのでしょうか。
及川：コストをどの段階で考えるかというのも確かに重要かと思います。たぶん工学的な人たちは、たぶんこの段階で、最適化の段階でコストも考える。デバイスのイメージがあまりない時には、コストを考えずにここでやってしまってもいいと思うのです。そのコストで使えるデバイスであればいいわけで、材料開発をやる段階で二通りあると思うのですが、あまりデバイスイメージを持たずに材料を開発する。とにかく面白い現象があって、こいつは何かのデバイスに使えるのではないか、というイメージで開発するものと、ニーズオリエンテッドに、鉄鋼メーカーから直接、「こういう材料を開発してくれ」という時にはやはり鉄鋼材料に貴金属は入れられませんね、こんな元素は入れられませんねという制約下では、最初の段階でコストは決めていきます。
質問者：でもその場合は、生産プロセスをよく研究をしてみると、最初の最適化の段階で想定していなかった低コストプロセスがありうるわけです。それからいくと、当然、一度こちらから帰るのがあってもいいのではないかという気がするのですが。
及川：そういうこともあると思います。ここまで来てから、もう一回生産プロセスのコストを考えて最適化しなおすこともあり得るかと思います。

質問者:データベースとかデータ・ハンドリングの問題で現在、困っている点や課題になっている点は何かありますか。

及川：一番問題になってきているのが、やはりモデルのところになってくるのですが、最近は段々みなさん似たようなモデルを使うようになってきているのですが、アカデミアとしてそれがいいのかどうかと。このモデルがシンプルすぎるので、理論的に厳密でありたいというのと、シンプルでありたいというので、いろいろなモデルが実はあって、データベースを作るときは、基本的には同じモデルでないと作ってはいけない。ただ、可能なら違うモデルもうまいことやってシームレスに繋げられれば、けっこう大きなデータベースが簡単に作れるかなと。データベースを作ろうとすると、自分が使っているモデルでどんどん積み重ねていかなければなりません。それだと、たまたま他の研究者が自分と同じモデルを使っていればいいのですが、そうではない場合には大きなデータベースを作るためには、全部自分でやらなければいけないということになってくるので、その辺はひとつの問題になるかと思います。データベースを作るという観点では､それが問題だと思います。

「オープン・データの潮流とデータの統合利用」
山口敦子（ライフサイエンス統合データベースセンター）

まず私たちが使っているライフサイエンスのデータの概観について簡単にお話して、そのあと統合利用について、お話をさせていただきたいと思います。ライフサイエンスのデータには様々なものがありまして、遺伝子データやタンパク質データ等の分子データもあれば、医療データ、薬剤データのようなもの、あるいは生物、ライフサイエンスに関係する文献データだったり、あるいは生態そのもの、個体のデータだったりと、いろいろな物が含まれてくるわけです。その大きな特徴をざっくり挙げますと、多様性、大規模性、データ間のリンクが非常に多いという三つを挙げることができると思います。

それぞれについてもう少し詳しくお話ししていくと、多様性について、何故そういういうことが起きてしまうかというと、それぞれ研究者が興味を持っている生物種がバラバラで、生物種というのは命名済みのもので 120 万種ぐらい知られていて、まだ知られていない推定生物種だと 870 万種あり、これからもどんどん増えていく可能性があるわけです。さらに研究者が興味を持つ対象も、低分子化合物だったり、生体のなかの高分子だったり、あるいは細胞や組織だったり疾患だったりと、いろいろあり、さらにその組み合わせがあるわけです。さらに記述方法も、見たいものによって、化学式だったり配列だったり、あるいは立体構造、数字、画像、さらには文献やリンクなど、興味によってばらばらです。さらに実験手法がどんどん新しくなっていくわけですけれど、実験手法が新しくなるとまた欲しいデータも変わってくるし、たぶん材料系の方々のところのように、数式できっちり表せるという世界ではないので、モデルがどんどん変わっていき、どんどん多様性が進んでいくというところもあります。

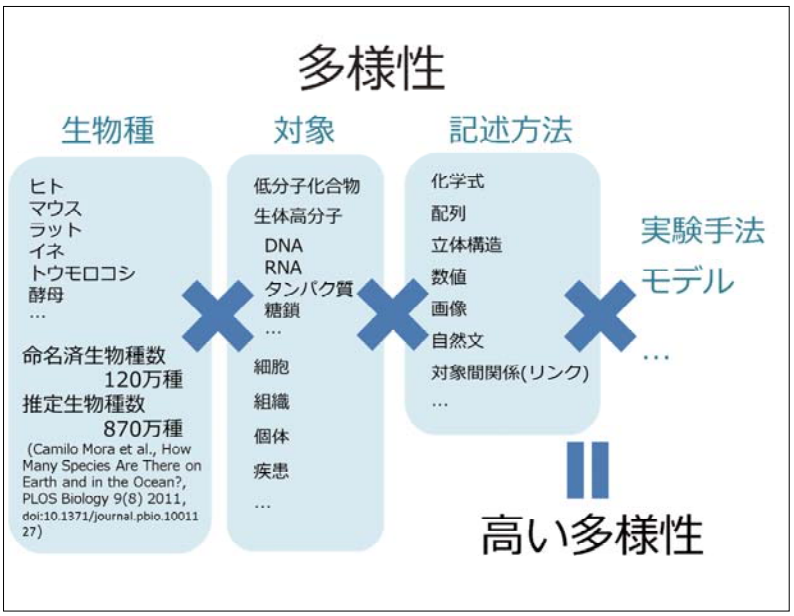


図 1

さらに、ひとつのデータも大規模になりがちという点もあります。これはDNAとかの核酸配列のデータベースですが、元々大きなデータベースであるにも関わらず、さらにどんどん大きくなっていることがわかると思います。これはシーケンサーの性能が上がるにつれて、計測のコストが下がっていくので、1研究室が出せるデータの量がどんどん増えているということもあります。

さらに、データの多様性にも関係するところなのですが、データの中にデータ間のリンクが非常に多いこともあります。例えばタンパク質のミオグロビンという例を見てみますと、これがどの遺伝子から作られるのか、というところに興味がある人は遺伝子データベースへのリンクを見ますし、それがどういうアミノ酸配列になっているかを見たい人は配列へのリンクを見るでしょうし、あるいは、それが実際に機能するときにどういう形をするかというのが見たい人は立体構造へのリンクを見ますし、あるいはタンパク質に低分子化合物がくっつくことで機能を出したりするのですが、そのどういう物がくっつくかということを見たい人は化合物へのリンクを見る。これらは本当に一部で、いろいろな見方があるわけです。研究者の興味によってデータの切り口が多様なため、データ内に多様なリンクができてしまうという傾向があります。例えばこれはタンパク質配列のデータベースで、多分最も有名なUniProtというデータベースですが、ここの青いところがリンクに当たるのですが、自分のデータベース内や他のデータベースへのリンクが数多く存在することが見て取れると思います。

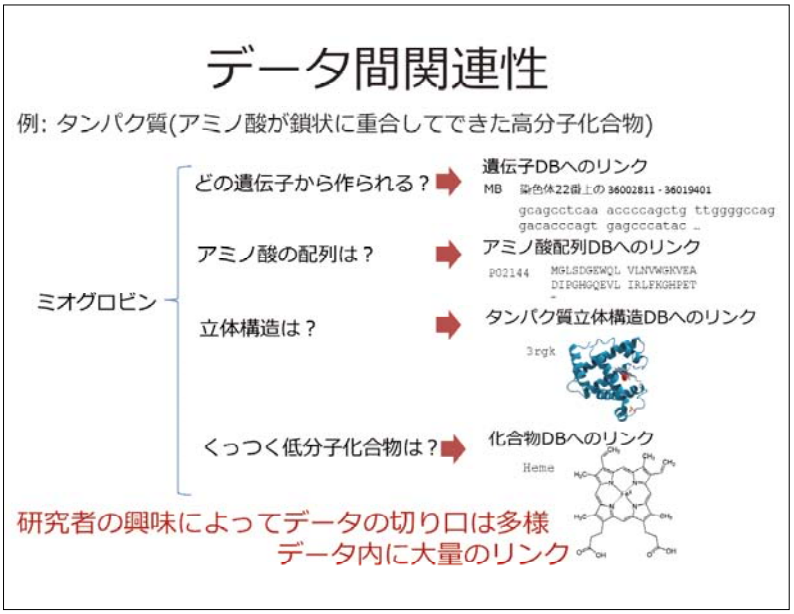


図2

こういうデータの性質を踏まえて、ライフサイエンスデータベースの現況を見てみますと、データベースが非常に乱立されやすいという状況にあります。データベースの重要性がわかってきて、データベースをひとつ作るとひとつ論文が書けるという状況にはなってきているのですが、それは反面、あるいくつかの生物種で、ある対象である新しいモデルで、というのをやると、ひとつが作れてしまいます。それで、小規模なデータベースが多

数、作成されるという状況にあります。また、国家プロジェクトの成果物がデータベースになることが多く、プロジェクトの数だけデータベースが作成されてしまうという状況があります。

では、それを１カ所にまとめて整理整頓してしまえばいいじゃないかと思われるかもしれませんが、これは非常に多様性が高い上に、それぞれの世界観が違う人達が作ったデータのために、整理整頓が困難という面もあります。さらに、データベースが元々大規模な上に測定機器の発展に伴うハイ・スループット化でどんどん大規模化されているという事実もあります。

大規模データで多種多様なデータベースを「自在に」利用できれば、これまで仮説・検証型を中心にしていた研究スタイルからデータ駆動型の研究にシフトできる状況にあると思います。しかし、「自在に」というところが非常に大事で、それができればより幅広い観点からの知識発見が可能になってきます。しかし、現状としては使いたいデータベースがどこにあるのか、あまりにも多種多様なものが乱立している状態なので、使いたいデータベースがどこにどんなふうにあるかという状況がわからない。また、例え利用したいものを見つけることができても、利用方法がわからなかったり、あるいはデータベース間でのフォーマットがバラバラで一元的に使えないという問題があります。

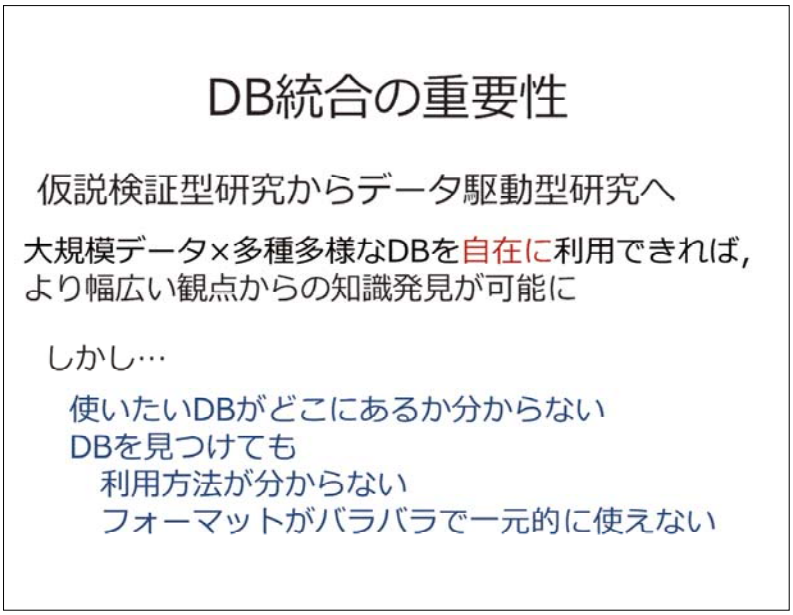


図３

例えば代謝経路、体内での反応経路ですが、その二つのうち、共通する遺伝子がどれで、その中で立体構造がわかっているものはどれか、という質問をしたいとすると、まずそれに関係するデータベースを探し、さらにプログラムが書けない人は、これに検索をかけてコピペして、さらにこっちに検索をかけてコピペして、さらに検索とコピペをして最終結果を得る。あるいはプログラムを書ける人であってもこの矢印の数だけプログラムを書かなければならない。さらに、同じ興味を持っている人はこのプログラムの再利用が可能か

もしれないのですが、少し変わると、もう使えないプログラムになってしまうので、書いたプログラムの再利用性は低い。つまり、けっこう無駄なプログラムを大量に書かなければいけない状況にあります。

こういう状況を何とかしたいということから、データベース統合を私たちは考えております。これは、NBDC のセンター長の高木先生が仰っている、統合のための三つのステップなのですが、私たちもこれに沿って統合を進めております。ひとつめとしては「データベースを網羅的に収集しメタデータを付与する」、二つ目として「それぞれのデータベースにおいてフォーマットと用語の統一を行う」、三つめのステップとして「複数のデータベースを再構築し使いやすいインターフェースにまとめあげる」。この三つのステップを通じて、使いやすいデータベースの統合を行いたいと考えています。

それぞれのステップについて、どのような取り組みを行っているかという話をさせていただきたいと思います。ひとつ目の取り組みについてですが、Integbio データベースカタログというのを JST の NBDC が中心になってサービスしておられて、これは文部科学省、厚生労働省、農林水産省、経済産業省の 4 つの省庁で集めているデータベースのカタログを統合的にこのサイトに集めて、それに名称や URL、生物種といったいろいろなメタデータを付与して検索ができるようにしたものです。同様の取り組みは、海外でもありまして、BioDBCore という取り組みがあります。これは、データベースを出すときには必ず必要なメタデータには何がありますかというのを議論しています。皆さん、これを付けましょうという取り組みです。こちらもデータベース名だったり URL だったり、連絡先だったりとか、これぐらいのものは、データベースを作るときには載せておきましょうと。これを検索できるようにしておきましょう、というのが BioDBCore の取り組みです。

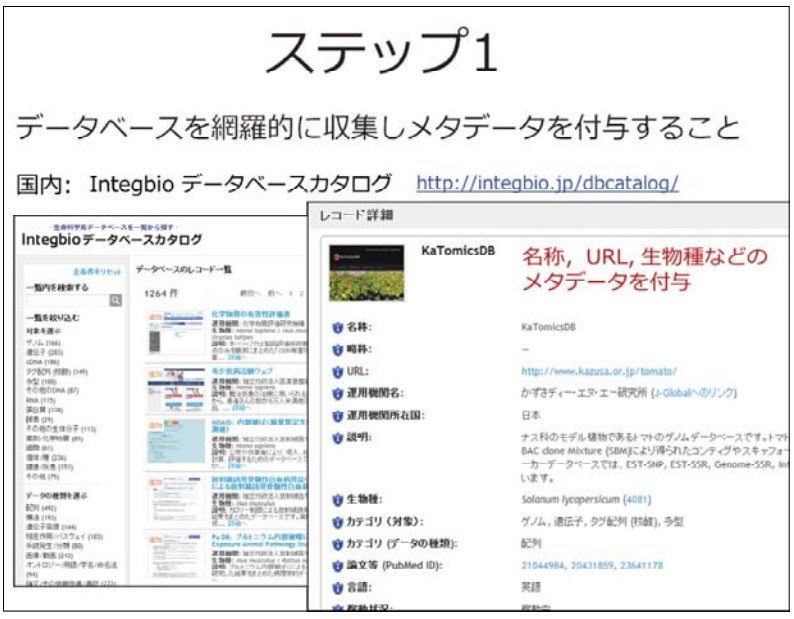


図 4

ステップ 1 がある程度できてきたという状況で、ステップ 2 の「それぞれのデータベー

スにおいてフォーマットと用語の統一を行う」ということについて、お話しいたします。先ほどから申し上げているように、ライフサイエンスにおいては多種多様な研究ニーズに応えることが可能であることが必要ですので、あらゆるデータベースの構造と独立に、シンプルかつ標準的な形で再構築する必要があります。それで私たちは何を選んだかというと、セマンティックウェブ技術にあるRDFとオントロジーを使おうとしています。フォーマットとしてRDF、用語の統一としてオントロジーを使おうとしています。そのそれぞれについて説明します。

図5

RDFというのは「データ記述・交換のための標準的枠組み」です。特徴のひとつ目として、主語、述語、目的語の三つ組で情報を表現します。例えばこういうAさんの身長が171で体重が65、Bさんの身長が156、体重が48みたいな、このような表があったときに、これを三つ組「Aさんの身長は171」「Bさんの身長は156」というように主語、述語、目的語の三つの組で情報を表現するということをします。これをやると何が良いかと言いますと、いろいろな表があっても、全て三つ組の形式に直すことができます。さらに、Aさんは上の表にも下の表にも出てきますが、同じものがラベルされた頂点のところを重ねていくと、このようなひとつのネットワークになっていくわけです。RDFというのは、こういうネットワーク構造を持っているデータだということも言えます。これで何が嬉しいかというと、ひとつのネットワーク構造に落とし込むことができるので、いろいろな興味に柔軟に対応することができます。例えばユーザー1がAさんの情報に興味を持っているとすると、Aさんに対応するデータだけを部分グラフとしてとってくれば必要なデータをとってくることができますし、別のユーザーが、例えばI社の社員の平均身長に興味を持ったとすると、I社に所属する人の身長のデータだけを、全体から切り取ってくることで対応することができます。

二つ目の特徴として、識別子としてURI（Universal Resource Identifier）を利用するところがあります。今、皆さんがわかりやすいようにAさん、所属I社と書いたのですが、

RDF にするためには全ての識別子に URI を割り当てることになっています。例えば、A さんだと、http で始まるいわゆる URI を割り当て、「所属する」という述語、「I 社」という目的語にも同様に URI を割り当てます。これによって、グローバルに一意の識別子を持って関係性を表現することができます。これは、機械が可読にするために、こういうことをしているわけです。別の普通の数字の ID でいいじゃないかと思われる方がいると思いますが、これの良いところは、HTTP 上の文書でもあるので、A さんについての情報をここに書いたり、あるいは「所属する」とはどういうことか、I 社とはどのような会社かといった情報をここに書いたりすることで、人間が見てもどういうものか理解することができるという点があります。

三つ目の特徴として、標準化された検索言語、SPARQL と呼ばれていますが、これが提供されています。これは、RDF の全体のネットワークから部分ネットワークを検索するための言語です。例えば A さんの情報に興味がある方は、A さんについて、述語が「？」で目的語が「？」のものを全部集めてきなさい、という検索のクエリを投げると、A さんに関する情報を持ってくることができます。また、I 社の社員の平均身長に興味がある人は、I 社に所属する変数の、身長の変数を全部持ってきなさいとすると、それにマッチするものを全部持ってくることができます。

このようにして、部分ネットワーク、同じ形を持っている部分ネットワークをとってくることで検索ができるようになっています。実際の SPARQL の例としては、こういうものです。これは実は、Integbio データベースカタログから哺乳類のデータを含むデータベースを求めよ、という SPARQL になっています。これだと別に RDF じゃなくてもいいじゃないか、と思う方もいらっしゃると思います。何故かというと、哺乳類のデータであれば、Integbio ではすでにメタデータが付いているので、それを検索すればいいのではないかと思われる方もいらっしゃると思いますが、何故、RDF とセマンティックウェブ技術を使おうとしたかというと、オントロジーと非常に相性がいいというところもあります。

今、よく使われているキーワード検索だと、ここの生物種のところがちゃんと哺乳類と明示的に書いてあるものは探すことができるのですが、ホモサピエンス、マウスというように書かれているものは探せないわけです。こういうものを探すためにどういう情報が必要かというと、ヒトは哺乳類です、とか、マウスは哺乳類です、といった情報が必要です。こうした概念を形式的に体系化したものをオントロジーと呼びます。

セマンティックウェブ記述ではオントロジー言語 OWL というものを使って概念を記述します。このオントロジー言語 OWL は、実はそれ自体が RDF、こういう三つ組のグラフになっています。必要な情報はこれなのですが、こういうものをいちいち作っていると、また今度、サルやリスがきたときにどうするの、ということになってしまいます。なので、普通、オントロジーの場合は概念を体系化して、それを OWL という言語で記述しますが、それをしておいて、さらにサブクラスのサブクラスは、サブクラスという推論を使って、例えば哺乳類と人の間には直接サブクラスというラインは引かれていないのですが、哺乳類のサブクラスにこの Theria というのがあり、そのサブクラスに Eutheria があり、といったように、サブクラスのサブクラスはサブクラスである、というのをどんどん推論していくことで、ヒトは哺乳類です、という推論ができ、他の上下関係も推論することでサブクラス関係がとれてくることがわかります。これによって、ひとつをきっちり書いておけば、

様々な上下関係について柔軟に必要な情報を取ることができるのです。

オントロジーはバイオの分野では、どんどん作られており、それを集めたサイトも存在します。BioPortal と呼ばれるサイトですが、ここには 351 のオントロジーが登録されています。ここはユーザーが勝手にオントロジーを登録するサイトなのですが、ここに最もたくさん登録している人達が OBO と呼ばれるサイトの人達で、Open Biological and Biomedical Ontologies というサイトで活動されているのですが、ここでは 8 つのオントロジーを正式にリリースしており、113 のオントロジー候補を利用できる状態になっています。

今まで話してきたように、バイオのデータを公開するときには、RDF 化してかつデータ内の概念をきっちりオントロジーで記述して公開することが望ましいわけです。しかし、既存のオントロジーではカバーできない概念がまず多数ある。さらに、だからといって自分でオントロジーを作って更新していくのは、非常にコストが高いわけです。新しい情報がくると、オントロジー自体もどんどん変わっていきますので、更新自体も大変だし、生物学者が自分の概念を数学的にきっちり表現してオントロジーに直していくのは、けっこう大変な作業になってしまうということもあり、オントロジーを作って RDF を公開することを考えるとすごい時間が掛かってしまいます。

図 6

ですので、RDF 化したデータを他のデータと繋いで公開するだけでも価値があるのではないかというムーブメントが起きており、それが Linked Open Data と呼ばれるものです。公開された多種多様なデータが、ひとつの大きなネットワークとなるために、オープン・データの評価指標があり、これは「5 ★ Open Data」と呼ばれています。1 つ目の星、一番低い段階としては、データをオープン・ライセンスで Web 上に公開する。フォーマットは不問で、表が画像になっていようが、PDF になっていようが、フォーマットは不問。とにかく、公開されていることが望ましい、というのが 1。2 番目はデータが機械

可読な形で公開することが望ましい。これはたとえばデータが Excel の表だったり、Word だったり何でもいいのですが、機械が読めるという形であることが望ましい。三つ目は機械可読。2 番目まではソフトウェアを持っている人ではないとできないので、データがオープンな形式、例えばタブ区切りやカンマ区切りのような形式で公開されていることが望ましい。四つ目としては、ちゃんと URI がついた RDF で公開する。5 つ目さらには RDF で他のデータと繋げて公開したもの。その段階までいったもので、実際はこの先にいくつか条件があるのですが、それを Linked Open Data と読んでいます。

現在、Linked Open Data のクラウドに含まれているデータセット数は 295 あり、三つ組の総数としては 310 億あります。外観を見ると、これは文献のデータで、これは新聞や放送局などのメディアなので、ここが最近お聞きになられたことがあると思いますが、オープンガバメントですね。政府系のデータで、ここのピンクのところがライフサイエンスなので、全体として大きな割合をライフサイエンスが占めていることがおわかりになると思います。

中央の、みんなからやたらリンクされているものは何かというと、これは DBPedia つまりざっくりいうと、Wikipedia を RDF にしたものです。これに繋げることで、異分野間を繋げることができています。

Linked Open Data を用いることで、先ほどのような複雑なクエリについても SPARQL で一個記述して SPARQL のエンドポイントに検索をかけると結果がでます。SPARQL を書くという手間は掛かりますが、そんなに複雑なことをしなくても検索ができる明るい未来が待っているかもしれないということです。ただまだステップ 2 についても課題がいっぱいあり、RDF 化したデータを扱うことができるようなデータベースシステムが未成熟であって、大規模データに関してはロードや検索に時間が掛かることが知られています。ですので、これに関しては動向を注目しつつ、継続的にベンチマーキングをしていきたいと思っています。また SPARQL を自在に扱うことには、かなり楽になったとはいえ、慣れが必要です。それは敷居が高いので、SPARQL の生成を助けるインターフェースなどが必要になってくる。これはステップ 3 のインターフェースのところにも関わってくる問題になります。

ステップ 3 の複数のデータベースを再構築し、使いやすいインターフェースにまとめあげるというところは、実際、私たちも開発中で、現在進行中の話になると思うのですが、例えば先ほどここに出てきたような自然文による質問文を入力すると、それを解釈して SPARQL を自動生成するシステムであったり、あるいは表形式のデータにユーザーのデータを、欲しいデータを、データベースをどういう内容か指定すると、SPARQL を自動的に書いて追加したデータにしてくれるシステムだったり、あるいはゲノムに特価してゲノム研究に典型的な可視化テンプレートを用意することで、その研究に特価した SPARQL のテンプレートを用意しておいて、それを使って検索を行うというような物を開発しています。

まとめとしては、ライフサイエンスにおけるセマンティックウェブ技術、主に RDF を用いたデータベース統合に向けた動向と取り組みを紹介しました。今後の取り組みとして

は、データベースの RDF 化をさらに進めるとともに、アプリケーションを見つめたオントロジーやインターフェースの開発を進めたいと思います。

【質疑応答】
質問者:ライフサイエンスの場合、データをとるためのレギュレーションやプラットフォームを同じにしても、たぶん違うデータがとれてしまうことがあると思います。例えば細胞とか発現とか、例えばレギュレーションを同じにしても再現できないような層はあります。それは構わずデータベースに入っていて、今のような統合データベースではそれを引っ張ってしまうわけです。バリデーションというか、どれが正しいかということはアプリケーション側に任されていて、データベース側、統合側では何もしないのですか。
山口：統合側では、生データをオープンに出してもらってそれらの統合を目指しつつ、さらにそれにメタデータですとか、信頼性がどうかとか、そういうものを付け加えていくことは、サードパーティーとして私たちのセンター内のチームでやっているところがあります。
質問者：ただ、データが膨大なので追いつかないです。それはどういうふうにやるのですか。
山口：クオリティ・コントロールのための評価基準を作ったり、あるいは論文化されているかというところを見て、ある程度、半自動的に評価をしているのが現状です。
質問者：では、検索のところでもそういうフィルターが掛かっている。
山口：検索のところと、データ基盤のところは、ワンステップ離れていて、データとしては入っているのですが、そのデータに三つ組として、このデータはどうです、ということは、別に後から付け加えることができるわけです。ひとつの大きなグラフなので、好きにデータは、構造を考えずに、どんどん付け加えることはできるので、そこにサードパーティーとしてデータが良いとか、どうだといった評価は、あとで付け加えることができる形です。

質問者:先ほど、データの更新に問題があると言われました。確かに、使いやすいインターフェースを整えるのがひとつのアプローチだと思うのですが、それで本当にうまくいくか、という感じが個人的にはします。知識をあのようなオントロジーでまとめようというのは、いわゆる AI の分野で昔からやられていて、やはりうまくいっていないのです。例えば IBM などは昔会社に行って、そういうエキスパートシステムを作りませんかという提案をしたのだと思うのですが、やはりまず、現場の人がそういう知識を入れられるのかという点でハードルが高い。さらに、やはり知識の更新です。まったく新しい概念が出てきた時にどうするのかというのは、やはり解けなくて。

私の理解では、最近のトレンドは、きっちりしたオントロジーみたいなものを作るのは諦めて、学習、というと言い過ぎですが、ある程度自動的に知識を引っかけるような仕組みを作る方に行かないと駄目なのかな、というのがあると思うのです。例えば、この生物の分野だと割とタクソノミーがはっきりしているからできるはずであるとか、あるいは他の分野ではどうかというところでコメントいただければと思います。
山口：まさにご指摘の通りで、ここに書いておりますが、オントロジーを作ることは本当にコストが高いので、セマンティックウェブ技術はいったん廃れかけたのですが、それが

Linked Open Data というムーブメントでまた復活しつつあるというのが現状です。ここではもう、オントロジーを作るのは基本的に諦めて、RDF 化したデータを他のデータと繋げるだけで価値を出していこうというのが、Linked Open Data の考え方です。

質問者:生命科学の分野で、オープン･データベースの中にデータを入れ込むインセンティブはどこにあるのでしょうか。それと知財の関係はどのように取り扱われているのでしょうか。
山口：オープン・データにしようという動機については、まず雑誌がデータをオープンにしていないと、それをアクセプトしない場合が多いというところがかなり大きくて、例えば自分で実験をしてさらにそれをこういう結果が出ましたと言っても、みんな実際に確かめられないので、データは必ずどこかで見えるようにしておいて、サプリメントデータという形で置いておかないと、ジャーナルがまずアクセプトしてくれない。論文を書かないと業績になりませんから､皆さんそれでどんどんデータを出す｡あるいはジャーナルによっては、ここのデータバンクに登録してくださいと指定しているところもあります。
　知財、ライセンスについては悩ましい問題なのですが、今私たちが進めているこういうライセンスを使いましょうというものには、クリエイティブ・コモン・ライセンスというのがありまして、今までライセンスにはいろいろなものがありましたが、それも統合的な形にして、例えばこのデータは本当に自由に使っていいです、というものから、改変は駄目ですとか、必ず著者名を入れてくださいとか、いくつかの制限を標準化することでわかりやすいライセンスにしてデータをオープンに提供してくださいという方向で進めています。

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）

データベース統合というと最近はオントロジーを使うのがはやりで、すでに何年か前に開発したものがあるので、その辺の経緯と、やはり材料物性のデータを Linked Open Data に載せる上での問題など話をさせていただきます。

なぜ今こういうことをやっているかというと、元々バックグラウンドは原子炉材料の研究室にいて、データベースなどをやっていたのですが、これからは分散型データベースもネットで繋ぐ時代かなと思って、D 論は複数のデータベースを繋ぐといった方向で書きました。その後、新日鐵の計算科学グループにいましたが、この時は先ほど話にでた CALPHAD 法とか計算流体力学などを組み合わせて、鉄鋼の精練プロセスを統合的にシミュレーションするというようなシステムの開発を松宮さんの下でやっておりました。その後、2000 年前後に東大の岩田教授が代表者になって、振興調整費で仮想実験というプロジェクトがあり、これは第一原理計算からミクロからマクロまでのシミュレーションを何とか繋ごうという話ですが、その中で私はどちらかといえば、情報系の寄与ということでモジュラー・シミュレーション・フレームワークというのを提案しました。これは RDF 等を使って、複数のシミュレーションの間のデータのやり取りをできるようにして、モジュール化して繋げられないかというものだったのですが、フレームワークは作ったのですが、中身を充実させるのは大変な話で、なかなか計算材料科学の専門家の協力も得られず、ある程度、サンプルを示して終わったということになると思います。ちょっと間が開きまして、今日、お話しする材料のデータプラットフォーム、これはオントロジーを使ってやっていたのですが、その頃からいろいろな知識工学的な、最近のはやりのセマンティックウェブ等を使って、どう材料科学に関する知識を書いていけるかなという話をやっていたので、その辺を話していきます。

これは 2006 年、7 年ぐらいに NEDO の知的基盤の受託研究、2 年間のプロジェクトでやったのですが、この時の目標は基本的にはエンジニアリングの材料に対するデータなのですが、プラットフォームのアーキテクチャーとして、マテリアルズ・インフォマティクスに必要なものはある程度含まれていると思うので少しお話しします。先ほどバイオの話にもありましたが、材料のデータベースというのはいろいろあるはずなのです。というのは、いろいろなプロジェクトで作られていますが、年が限られたプロジェクトで終わると、それ以降は更新されなくなる。小さなデータベースがたくさんあり、それらがうまく活用されていないのではないか、という問題意識で提案を作りました。かといって、データを全部くださいというと、なかなかそうはいかないので、どこかにプラットフォームを作ってそこに共通のフォーマットを用意して、そこからリンクできるような形にする、ということを提案しました。共通のデータフォーマットがあれば、計算といったところともデータの交換が可能であろうし、比較していくことも可能だろうと。

この時、当然データベースはいろいろな種類があるので、辞書が必要なのです。先ほどのオントロジーとは何かという話に、私の考えは、単なる階層化した辞書です。それ以上

のものはあまり望んでいない。しかしオントロジーなど、はやりだすといろいろなオープンソースで使える編集ツールやビジュアライゼーションツールが出てくるのです。そこで、オントロジーをベースにこういったデータ交換をするということを考えました。

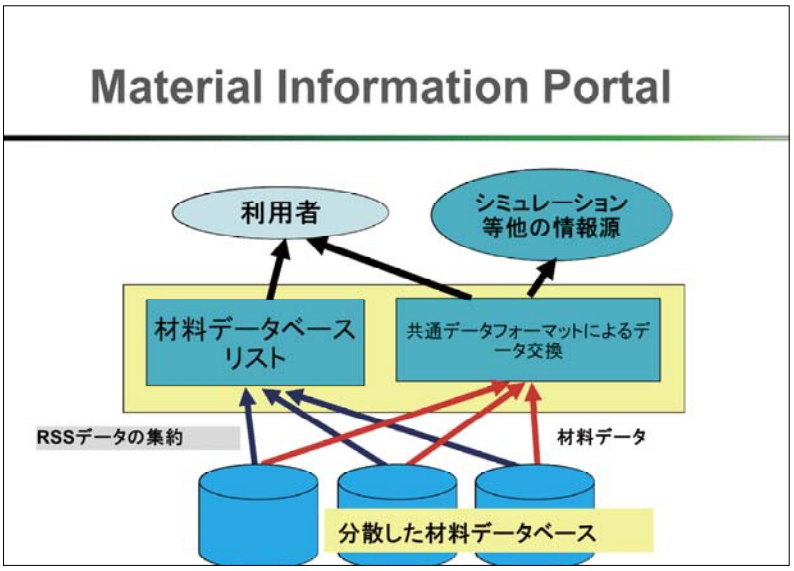


図 1

この時の組織ですが、まず物質・材料研究機構の材料情報ステーション、それから産業技術総合研究所の計測標準で、熱物性データベースをやっているグループ。それから高知工科大学は鉄鋼のデータベースをやっている方がいて、データあるいは共通フォーマットとデータと各個別のデータベースとのマッピングということを考えた。それを見て、共通のデータフォーマットを作るところは、私の東洋大学と東京大学のグループでやる。

CODATA（Committee on Data for Science and Technology）という組織があり、そこにマテリアル・データのタスクグループを作りましょうという提案して認められて、私とアメリカのローラ･バルトロ教授が協同でチェアをやっていましたので､そこに持っていって何とか国際的に広められないかと思ったのですが、メンバーの皆さんそれぞれ自分のデータベースプロジェクトでお忙しくて、そういったデータ交換まで頭が回らないということで、盛り上がらずに終わってしまいました。

材料は物性値が非常に多すぎて、辞書を作ろうとしてもとてもカバーしきれないとうことで、最終的に実際にわれわれが作ったのは、産業技術総合研究所と物質・材料研究機構が共通で持っているデータということで、熱物性をまずターゲットにしました。かといって、材料全体の共通の部分をおろそかにするわけにはいかないので、ここではどうしたかというと、まず材料データというのは、まず何らかの物がある。そこに環境とプロセス、これはテストメソッドも試験方法も含めてプロセスとここで呼んでいるのですが、それで何らかの環境下で何らかのプロセスが与えられることで特性値が出てくるのであろう、ということで、四つのコア・オントロジーを作って、さらに当時、物理乗数や単位等の定義もなかったので、ある程度用意しまして、トータルで 600 ぐらいのクラスを作った。こういった階層構造のビジュアライゼーションなどもオントロジー等セマンティック・ウェ

ブ技術に乗っかっておくとツールが開発されます。

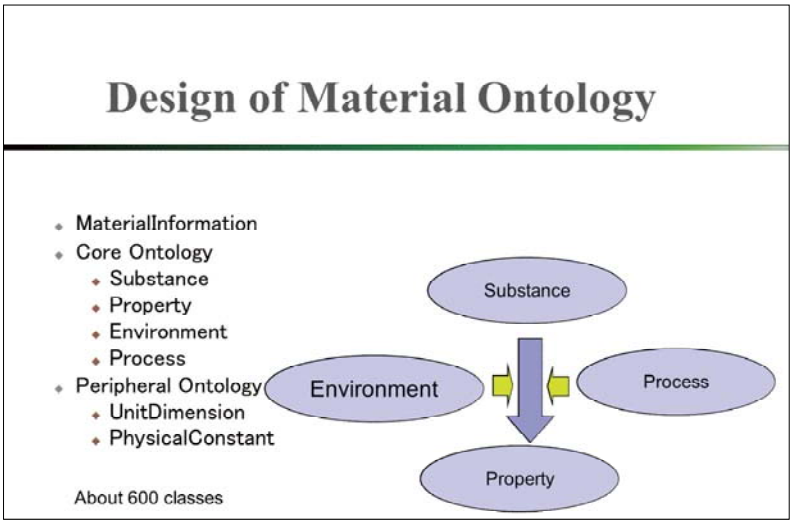


図 2

実際にデータ交換を行った例が図 3 になります。中段の黄色で示した部分がオントロジーで定義した辞書になります。上が物質・材料研究機構の熱物性の部分を取り出してきたスキーマ、下が産業技術総合研究所が作っているデータベースのスキーマ、ということでオントロジーを中心に上下に繋げて、これを経由して材料の熱物性のデータが交換できるでしょうというところまでをやりました。

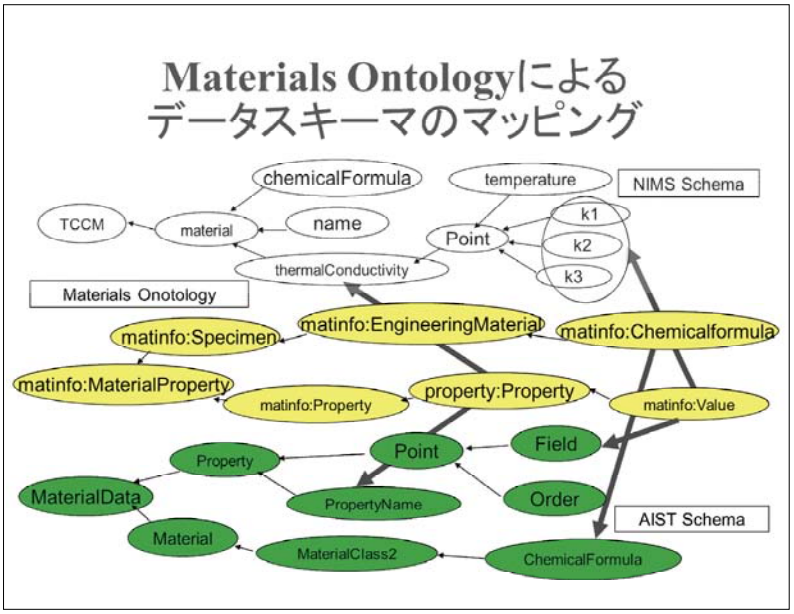


図 3

熱物性のデータは、いろいろな人工物の設計等に必要ですので、基本的にはエンジニアリングの部分がターゲットになっています。2008 年に終わって以降、データベースのプロジェクトでも時限が切れると続かないという問題がありますが、データベースのプラッ

トフォームのプロジェクトも年限が終わると切れてしまって、予算もなくなってなかなか続けられない。残念ながら材料データに関しては統合センターといったものもできなかった。生命科学の方では異種のデータベース連携に関連してオントロジー開発というのが随分続けられてきたわけですが、エンジニアリング分野でもここ数年少しずつ出始めているようで、ここ 1、2 年、海外の大学や研究者から「うちは今、エンジニアリングのオントロジーを考えているのだけれど、おたくのマテリアル・オントロジーというのはどうやったらダウンロードできるの」といった問い合わせが、ぼちぼち来るようになりました。エンジニアリング全般に関して辞書を作りたいのだが、マテリアルの部分はどうなっているということで、サーベイをしているところがあるようです。

国際的な動向ということで、やはり今ある程度、国際的に動いているのはエンジニアリングに関するもので、マテリアルサイエンスにしてもアメリカがマテリアル・ゲノム・イニシアティブとか言い始めたので、当然そのうち言い出すと思うのですが、先日、マテリアルサイエンスのオントロジーどうなるのという話をマテリアル・ゲノムを前からやっている Krishna Rajan 教授に聞いたのですが、あまりはっきりした答えが得られませんでした。

エンジニアリングに関するオントロジーで今、わりと活発にやっているは CEN（EU の標準規格委員会）で、ワークショップを何年か前からやっているのですが、去年からやっているのが WS/SERES で、これの前に WS/ELSSI-EMD というのがあり、これはオントロジーといっても包括的なものを作るのは大変なので諦めて、ISO の機械試験だとかすでに標準規格になっているものに関して、ちゃんとオントロジーのような形式を使って電子的にデータ交換できるフォーマットを作りましょうということで、これは EU で材料データ、原子力データ、原子炉材料のデータなどをやっているペテンのジョイントリサーチセンターの方が中心になって動かしています。もうひとつは、VAMAS（先進材料の標準化プロジェクト）に TWA35 というデータベースのグループが数年前にできましたが、ただこれはあまり動いていないようです。CEN の方は、こんど VAMAS の TWA35 を通じて作ったデータフォーマット、オントロジーを ISO のアペンディクスに載せていこうかという話をしているので、少しは活発になるのかと期待しています。

先ほどのプレゼンテーションにもあった Linked Open Data（LOD）ですが、これは去年から総務省と経産省がオープン・データ活用ということで、現在の LOD Cloud を見ていただくとバイオとか地理情報が中心で、よく見ると材料物性みたいなものがほとんど無いのです。NEDO のプロジェクトの分担者に入っていただいた、産業技術総合研究所の熱物性の方が、昨年度から担当されているので、どういうふうにしようかといった話をしているのですが、今の LOD Cloud には物性値としてちゃんとしたものが無いので、産業技術総合研究所がぜひ世界の LOD Cloud の物性値のコアになってくださいとお願いしているところです。

ただ、こういう LOD に材料データを載せる話をしていると、科学技術情報に関して、共通に使えるようなインフラの部分は意外にちゃんとしていないところがあるのです。例

えば物理をやっていると当然、単位やディメンジョンは大切なのですが、NASA が地球科学の情報を書く関係でこうした単位やディメンジョンの定義・辞書は用意しているのですが、ちゃんとディメンジョンとして数式でできていない。あるいは数式をどう書いていくか、例えば今、我々がお話ししているのは熱物性のデータですので、例えば熱伝導率も本来あれば定数ではないわけです。温度や圧力が変われば全部変わる。ですからそれに対して、経験式のような形で定義域を決めて、いろいろ経験式や実験でフィッティングしたパラメータが産総研のデータベースには収録されているのですが、関係型のデータモデルであっても LOD　であっても、このような数式をどう書いていいかという標準が無いのです。

そこで、OpenMath という数学の人々が作っているマークアップ・ランゲージを使ってこれを書くということを考えています。この場合、記述を二つに分ける必要があります。なぜかというと、物理的な知見などに基づいて作られた基本的な多項式の形式があって、そこに実験データから最小二乗法などでフィッティングしたパラメータが乗っているので、両方をデータベースとして独立して作らなくてはいけない。これを今、うまく何とかして、例えば熱物性データベースに入っている数式を、そのまま数式処理システムに食わせて、実験データを更新するとそれで最小二乗法をやる、パラメータをアセスメントしてくれる、そういう仕組みを作ることができないかなということを産業技術研究所と一緒にやっています。

ただ、このワークショップ基本的には材料開発が主眼なのだと思うのですが、今、お話ししたような LOD　等では、基本的にはリファレンスデータというか、統計処理を済ませて、ある程度信頼できるデータを対象としています。LOD の担当者には個別の実験データまで入れなくてもいいのではないかというお話をしましたが、マテリアルズ・インフォマティクスという意味では、ちゃんとやらなくてはいけないという気もします。

NEDO 知的基盤は、その時は単なるデータの交換ということでセマンティックウェブというような話はしなかったのですが、基本的には最初の設計の段階でオントロジーを使って意識して作ったので、このボキャブラリー辞書として、構造は別として、ボキャブラリーはそのまま LOD でのデータのリンクに使えるのではないか。標準資料や標準試験のデータは、当然 LOD に、オープン・データということで公開する。個別の実験データは、非常にいろいろなデータ形式があるので、標準化して全部一緒にするのは、あまり現実的ではないでしょうね、というようなお話をしていました。

今までお話ししてきたのは、主にエンジニアリング材料の話なので、新材料開発という意味で、こういう技術をうまく使うにはどうするのかなということを考えますと、基本的に材料開発に使うデータセットと、何か人工物を作るとか設計に使うようなデータセットは別物として考える必要があるのだろうと思います。というのは、いろいろな実験があり、新しい測定技術なども出てきます。小規模なデータセットでいろいろな物がでます。当然、それを全部取りまとめてひとつのモノリシックなデータベースを作るというのは、コスト的にかなり無理があるだろうと思っています。ただ、ある程度 Linked Open Data で代表的な部分だけでも決まったデータフォーマットを作っておけば、資源として活用できる。

最近私はケモインフォマティックスで、化学物質の計算をしている方とお話をすると、化合物のデータベースはオープンなものがあるのですが、実は構造に間違いがある。そうすると、そうしたものを計算と突き合わせて間違っているようなものと対照できるのではないかという話があったのですが、そういうところはある程度データ構造などを統一してお互いに検証することもできることは確かです。

人工物の設計に使われるデータは非常に重要な部分で、計算でマテリアルズ・インフォマティクスをやるにしても、リファレンス用に必要とされる信頼性の高い実験データというのは、物質・材料研究機構や産業技術総合研究所が作っている高度なデータベースは、さらに重要度を増してくると思います。ただデータの信頼性を確保するには時間もお金も掛かるので、多分、新材料開発のライフサイクルを考えると、間に合わない。そうすると、速報性の重要なデータはやはり別にしておかなければまずいのではないかと思っています。

そういった材料開発のためのデータをデータベース化するにはいろいろと問題がありまして、先ほどのバイオの場合、ゲノムなどはデータベースに登録してアクセッションナンバーを付けて投稿しないと論文が受け付けられないというような仕組みを作られたのですが、材料の場合は多分、論文を書くのはよくてもデータベース化するという労力まで掛ける人はあまりいない。最近はデータそのものが重要であるということで、Data Publishing であるとか Data Citation ということでデータセットそのものにも引用するための DOI を付けて、いろいろな論文からデータそのものを参照してもらおう、参照されたデータはいいデータだからそれ自体も研究者の評価にしようといった動きがあって、こういうものも何とか材料の研究をしている人がデータを出すインセンティブにならないかな、といったことを考えています。

【質疑応答】

質問者: Linked Open Data を産業技術総合研究所でやって世界のコアに、という話があったのですが、データベースとして使うには、データの量が大切になってきます。そうすると新しいデータだけではなく、古いデータも活用していかなければならない。当然、それをビジネスにしている人もいるわけで、有料性の問題もある。あるいは、この今、作ろうとしているデータベースを最新のデータに更新していくためには、やはり資金が必要です。その辺りはどう考えているのですか。

芦野：Linked Open Data 自体の考え方は、日本でもオープン・データ活用協議会ができたということですが、国が得たそういう統計データは、主に海外でオープンガバメントと呼んでいるのは国の統計データなので、毎年新しい物が出るのですが、そういう意味では物性値のような物はあまり例がない。要は持続可能性の問題だと思うのですが、それに関しては若干心配されるところではあります。今、オープン・データと言っているので国がお金を掛けて産業技術総合研究所のデータベースをオープンにしたと。それでプロジェクトが終わったらそれまでですねということでは困るので、永続的な URI をつけてくださいねというのは産業技術総合研究所の方に釘を刺していますが、プロジェクトが終われば打ち切られる可能性はあると思います。

ただ、産業技術総合研究所がやっているのは計測標準のグループです。そうすると、あ

そこは継続的にデータを上げて、標準を国内に提供するミッションがありますので、比較的、な持続的な体制になっていますので、なんとかなるかなという感じです。新しいデータをどんどん集めてやるというのは、継続性がやはり難しいと思います。

３－３　ディスカッション

寺倉：最初に、前回のワークショップやいろいろな欧米の情報をベースに、島津さんが提言をまとめてくださったので、骨子を簡単に説明していただきたいと思います。

島津（JST CRDS）：御手元に二つ、報告書の案をお配りしておりまして、今、ご紹介いただいた方が、青のカラーで「戦略プロポーザル」となっているものです。これまでの活動を踏まえて、今後、こういった提言を出していければと考えております。もうひとつお配りしているのは、「科学技術未来戦略ワークショップ」という資料ですが、前回のワークショップの議事録をまとめたものです。今回のワークショップも同じようにまとめて、一冊として報告書を出したいと思っています。

まずここで、今後の提言の方向性、中身を簡単に説明いたします。今日のワークショップの目的をもう一度確認すると、物質材料研究はこれまでデータ駆動型の実験科学が主流でしたが、最近になって実験と計算の連携という所がやっと回り始めました。これは国の元素戦略プロジェクトの貢献も大きいと思いますが、非常にコミュニケーションがとれるようになってきたところに、もう少し将来の事を考えると、やはりデータ駆動型科学の中でもデータ･サイエンス的なものを取り込んでいくべきだという所を､今日のワークショップで合意できればいいな、というところです。

これを物質材料のコミュニティに落とし込むと、こういった「物質・材料創成（プロセス)」のコミュニティ､「解析・評価」のコミュニティ､「モデリング・シュミレーション」のコミュニティといった三者の連携に加え､インフォマティックス（データマネジメント）が真ん中にあって、こういった関係がうまく回っていくことによって、科学技術手法として解析・予測だけでなく、いまだ方法が確立していない設計というところができるようになってくるのではないか。また、こういった原理駆動型の手法と、データ駆動型の手法がうまく併用されることで､研究が効果的に進むのではないかという問題認識がありました。

データ・インフォマティクスが何故必要、重要かというと、物質材料の設計、探索にはデータの蓄積が必要条件である。前回の細野先生や本日の及川先生の発表、また前回の岡山大学の野原先生のコメントにもあった通り、原理、物質材料探索に携わる先生方は、結晶状態や電子構造のデータベースを睨みながら、候補物質の探索や設計を行っているが、その割にはデータの共有やデータベースの整備が不十分なのではないか、というところが１点目。

２点目ですが、これまでに合成された物質の数と物性が膨大になっていて、従来の整理要素だけでは見通しが不十分になってきている。そういった見通しに加えて、バイオでもでもシークエンサーの話がありましたが、物質材料も同様で、計測技術の進歩によって短時間で大量のデータが取得可能になってきているというところ。さらに第一原理計算が手軽にできるインフラ環境の普及によって、構造と物性値のデータが容易に入手可能になっている等のことから､まさに今､研究の手法が変わりつつあるのではないかといったこと。

もう１点はマルチスケールの対応と書いてありますが、依然として物質物理と材料の間が離れすぎている。これはコミュニティという意味もありますが､離れすぎているため､科学の研究成果が実用に繋がるスピードが非常に遅い。そこで、この間を繋ぐのがモデリングであり、データであるのではないかという仮説、問題意識を持っています。

そういった時に、今後、物質・材料研究において、データのハンドリング、インフォマティクスがキーとなる時代が将来的に来るのではないか、少なくとも米国はその可能性に気付いてマテリアル・ゲノム・イニシアティブやナノテク・イニシアチブの中でも、データのイニシアティブを作っていると考えられます。

そういった状況の中で、データ群に潜在する蓋然性をコンピュータの能力、計算による網羅性や機械学習により予測、可視化といったところを、今の物質材料研究ではあまり有効に活用できていない。早い段階で計算、実験、データの研究者の連携を始めていく必要があるのではないかというところが問題意識となっています。

具体的に、物質・材料科学の飛躍的発展への課題は二つあると思っており、ひとつは今日、中心に話題提供をいただきましたが、データの取得と共有、データベースのあり方を、もう少し考える必要があるかなということ。もうひとつは、そういったデータを解析する手法です。これまで経験と勘でやっていた所を、もう少しデータ・サイエンスの取り組みを導入していく必要があるのではないかという二つです。

なかなか物質・材料研究では、データベースやこういったデータ解析という取り組みがピンと来ない状況ですが、データ共有を促進するための仕組みとインセンティブ、およびデータ駆動型研究の推進に向けて､コミュニティ文化､意識を変えていく強いリーダーシップが必要ということを書かせていただいています。

本日のワークショップの議論で話し合っていただきたいことですが、残りは「誰が進めるか」、そして「どうやって進めるか」ということだけだと思うのです。全体として、今まで説明したことが重要だと皆様に認識いただけるのなら、こういった仮説があるということで、物質・材料研究者とデータ科学者の連携を少しずつでもいいので進めていかなければならない。現状は意識が低い、プレーヤーが少ないといった問題があります。成功事例があれば非常にわかりやすいのですが、米国でもマテリアル・ゲノム・イニシアティブは開始されてまだ一年強で、こういった新興融合分野はすぐには成果が出ないという特徴があります。一方で米国等で成果が出てからでは遅いという指摘もあるので、国がサポートしていくことも必要であるのではないかと考えています。

そうはいってもコミュニティの自主努力というか、冒頭、寺倉先生のお話しにもありましたが、どうやっていけばいいのかというところを考えていかなければならない。中核となるプレーヤーとしては、すでに計算と実験が連携する元素戦略という場がありますし、計算物質科学イニシアティブ（CMSI）のようなコミュニティがあって、NIMS の材料情報ステーションがあって、数学と材料科学が連携するような東北大学の WPI がある中で、

誰がどのようにやっていくのかという問題になります。

理想を言えば、データ基盤整備のプログラム、NIMS にある Mat Navi を中心としたデータベース網の拡充、改定のようなものと、データ駆動型の研究開発プログラム。はじめは FS 的な取り組みで良いと思うのですが、こういったものを一体的なガバナンスで進めることができれば非常にいいと。その際、今日もお話しいただいたようなバイオサイエンスデータベースセンターやケモインフォマティクスのような先行する取り組みがあるので、そういったノウハウを活用して進めていければいいのではないかといったことを提言の中には書かせていただいています。

これは提言の中には書いておりませんが、国として、こういった出口の分野を横ぐしで支える計算や計測、解析の各々は非常に充実してきていて、その間の連携も少しずつですがとれるようになってきています。そこに、将来的にはデータ、インフォマティクスといった横串の基盤があれば、なお一層研究が有効に進むのではないかと考えています。

寺倉：どういうふうに進めるか。次のプレーヤーが書いてありましたね。最初に話したように、いろいろなレポートや様子を聞いていると、我々の努力は必ずしも十分ではないことは良くわかっているのですが、悠長に構えている問題ではないのは明らかなので、具体的に今後どのように進めていけばよいのか、いろいろご意見をいただければと思います。まず、ここに書かれている、中核と成り得るプレーヤーは、最初の情報としては非常に有用で、ここで閉じるという意味ではないのですが、こういう所は中核になって、そこに新たに加わるのでもいいのですが。あるいは、そこで連携しながら、そういう形で動き出すべきことではないかなと思っています。それと同じ物を、最後の絵、概念的にはこれでいいのかなと思いますが、こういう物を参考にしながら、アクティビティに関する重要さとか、まだこういうことが足りていないのではないかといったご意見がありましたら、ご発言いただければと思います。

緒形（物質・材料研究機構）：私ども先週 NIST で VAMAS（新材料及び標準に関するベルサイユプロジェクト）の会議と合わせてデータベースのワークショップをやってきました。データベースのワークショップでも、マテリアルズ・インフォマティクスという語義に加えて、オープン・データ、これは 5 月 9 日にホワイトハウスで推進が表明されたものを、今後どのように進めるかという話もありました。またそのデータワークショップの中では、ひとつの議論の中に、例えばデータベースを使って開発した材料の利益を誰が享受するのだという話しがあって、NIST の方でもデータベースは提供するものであって、それに対して使った人からは、開発で得たものに対して報酬を取らないといった考えを示したりして、今後のデータベースのあり方、課題、例えばデータベースのソースのあり方についていろんな議論がでたと思っています。一応 NIMS と NIST は MOU を結ぶ予定でいますが、今後 NIST、米国としてもデータベースは国際的に連携していきたいと。ひとつの国、またはひとつの機関だけでは今後のデータベースは開発しきれない、維持できないという考えがあって、なるべくデータを共有できるような方針を今後探っていくような動きがありました。

今回、私もこのワークショップに参加して、先ほどあったように、各コアの機関が結局データベースをどのように継続性を維持するか、また対外的にも日本のデータベースを維持していくかということがあるので、これは個々の機関でやる時代は終わって、国家的な取り組みに何とかしてほしい。どこかが取りまとめて、継続性を維持して、その中でもコアである機関がその特徴を生かして進めると。

これまでのデータベースは、とにかくデータを入れてみましたと。今度はどのように使っていくか。前回のワークショップであったように、材料だけでなくて、組織とかそういう映像データもデジタル化して取り込むようなことも今後していくことが、対外的にも日本のデータベースの有利さにつながる。

データベースに関して、各国が関心を持っています。その中で日本の立場、これまでに得た蓄積を生かすようなデータベース策を講じていきたいなと。NIMS としても NIMS の特性だけではなく、実際の研究の蓄積も盛り込めるようなデータベースを開発していきたいのですが、NIMS だけでは、やっぱりできない。このような場で、いろいろな方々が知恵のセンスを持って進めないと、今後進まないのではないかなと。今、こういうデータベースをしっかりしておかないと、今まで得てきた技術の蓄積が消えてしまうのではないかと少し不安もあります。データベースは今後の材料開発のイノベーションを促進するものとして位置付けるとともに、技術の継承も兼ねて、是非データベースを継続的になるように、できるだけ国的な取り組みにして、海外的にも欧米と対等に情報交換できるようなデータベースにして頂けるとありがたいと思っています。

寺倉：マテリアルズ・インフォマティクスの場合には、データベースが確かにベースになるので、データベースのあり方とか、いろいろな検討の問題。もうひとつ日本で掲げているのは、物質に関して、データをいかに使っていわゆるデータマイニング等のアクティビティを定着させるかということがあると思うので、それをどうやって推進していくかということ。その二つをやらないといけないと思います。

最初にデータベースのことが出たので、しばらくはデータベースについて議論をして、その後でデータマイニング、データ･ハンドリング、機械学習等について、シミュレーションとモデリングということも含めて、どうしていくべきか議論したいと思います。

先々にこういうアクティビティに繋げるためには、データベースが基盤ですが、それに対する我々の認識は必ずしも十分でない所があったので、今日はそういう話を発表いただいたのですが、データベースそのものにいろいろな問題があって大変だということがよく分かりました。その話を補足したり、あるいはコメントがあったらご意見をいただけないでしょうか。

曽根（物質・材料研究機構）：今日、バイオの話を聞いて思ったのですが、2000 年の頃にヒトゲノムの解読が、クレイグ・ヴェンターのセレラ・ジェノミクス社によって IT、コンピュータをフルに活用して、実行されてしまった訳です。国際ゲノム機関よりも先に、商業化した。それを契機に、電気メーカーは、これからはバイオは製薬会社だけではなく、IT の会社も主力の一翼を担うべきだという、そんな雰囲気があった。アメリカなどでは、データベースのビジネスが、ベンチャーとしてガンガン立ち上がった。でも、ほとんどつ

ぶれちゃった訳です。つまり、データベースは投入する資金、用意する資金に比べて、ビジネスにしてインカムとして入る量はたかが知れているのです。

でも今回のこの議論の中でも、データベースが物質の世界でもこれから質を変えて重要な役割を果たしていく、そういう時代になってきたのだと思います。そのためには、普通のビジネスでは成立しない世界ですので、国がしっかりサポートする。それも、乱立させてもしょうがないので、どこかに集中させて、そこにインフォマティクスの最新の成果を取り込んで発展させていく。可視化もそうですし、機械学習もそうだと思うんですが、そういうものを使って、より魅力的で使いやすいもの作っていく事が重要と思います。今、NIMS にもデータベースはありますけれど、そういった最新のインフォマティクスはほとんど入っていません。それが、入ってきて、シミュレーション等、いろいろなサポートデータが入ってくると、非常に魅力的なものになる。それは ALL JAPAN でやっていかないとできない。今まさに、そういう時代になっているのではないかなと思います。

伊藤（理化学研究所）：昨日、東大医科研の宮野先生のお話を伺う機会があって、非常に面白いことをおしゃっていて、2003 年にヒューマンゲノムが全部出たと。その時にアメリカから「これから生物学は完全に ICT の世界だ」と言われたんです。確かにアメリカは、そういうふうに切ったわけです。アメリカではその時に、いろいろなソフトウェアベンチャーができ、ほとんど全部潰れたんですが、一方で今、アメリカがバイオで一番儲かっているのは、次世代シークエンサーのようなハードを作って、そこで解析するためのソフトウェアです。つまりデータベースそのものは全然儲からない。そこはむしろ国がきちんと持って、それを使った産業が 10 年間でものすごく儲かっている。

ところが日本では、それが全然できなかったので、宮野先生の話だと、この 10 年間、つまり 2003 年から 13 年までは失われた 10 年間であると。医科研だけで一所懸命やっても全くできない、というお話をされていました。ですので、やはり今言われたように、バイオのデータというのは材料よりも 10 年先にやっていたと思うので、データはどこかきちんと集中して持っていくというのが大事で、それをみんなで、日本なり世界で活用することで新しい産業ができるということは、明らかにバイオの方で実証されていると思います。

寺倉：ワシントンで「計算機を使ったいろいろな活動の、社会的インパクトが特に強いものが何か」という、ものすごい数のヒアリングをしたというのを、私も知りあいから聞いたことがあって、そういうものをベースにしてアメリカはすでに動いてきたんです。MGI のワークショップのレポートを、ザッと見ていたら、非常に徹底的に問題を細かく分析して、どこから取り掛かるかなどをプライオリティーに従って整理しているんです。その手のかなり緻密な積み重ねで向こうは動いていて、そういう意味でさっきから我々の努力はまだまだ足りないと思っているのですが、そうは言いながら、それだけ言って同じことをもう一度繰り返している時間は全然ないと思うので、今、言われたように、データベースについては基本的には確かに我々のところでいろいろできるような問題じゃないので、国の方からかなりサポートしていただかないといけないと思うのです。国からいったら、データベースはいったん始めると止められないから、なかなか始めるのが大変だというのを聞いたことがあって、両面があるのだろうと思うんですが。

一方でバイオやケミストの方でのいろいろな蓄積があるので、我々がやろうとしたときには、そういう所からかなりノウハウを吸収してやれるという、それはひとつのメリットだと思うのですが。よほど覚悟してやりださなければ、後々に何も残らないという懸念もあると思います。

鹿野（分子科学研究所）：今の件で、その際に誰がデータをインプットするのかということを少し議論していただきたいのです。なぜかというと、例えばここで議論されてデータベースを作りますとなったときに、担い手になるのは多分、僕の同世代。僕の同世代が今、何をやっているかというと、ポスドクをやっていたり、ドクターの学生をやっていたりする訳です。その人が先ほどの山口さんのお話しにもありましたが、データベースを一本立てれば論文一本になると。論文で評価されている以上は、基本的に論文を出したい訳です。そのために基本データ何か出してくださいよっていって、僕らに何かインセンティブがあるかというと、実際の担い手には無い。結局、いいデータベースがあれば、その後に今日のベイズ推定の話であったり、スパースモデリングの話であったり、そういうものが乗っかってくる訳です。例えば、それが今、多分ポスドクといいましたが、これが技術職員でもいいと思うんです。ただ技術職員みたいなのを今は技官さんと言って基本的には実験の技術職員になっていますけど、専門にそういうものを入れられるような技術職員を技術の中に置いていただいて、その人にガンガン、データを入れていただくのも、ひとつの案かなと思います。皆さんは、どの様にお考えなのでしょうか。

寺倉：具体的な話になると、段々、そういう問題が浮き彫りになってきますが、それは必ずしもデータばかりではなくて、計算科学、プログラミング等で散々、我々も経験したことです。要するに、最後の実を採るところではなく種を蒔くところ、仕込むところです。基本的に支援部隊に近いところだと思いますが、それは実は、非常に長いこと議論をしながら、何も進んでいないのだろうと思います。ずっと我々の頭の中にありながら、なかなか具体的に答えられない所で、ここで議論しようとしても、多分どうにもならないと思うのです。

鹿野：それを議論して解決しない限りは、多分、一歩も進まないんだと思うんです。例えば、データベースを本当に作るとなったときに、では本当に誰がやるんですかっていうのは、それはセットにして考えていただかないと、若い世代に押し付けられても困ってしまう。

寺倉：データベースを作るためだけにデータを作るという、そこだけに頼ったらそれは動かないと思うのです。具体的には、最近、我々が本を書いてもサプリメンタル・インフォメーションとか、エクスターナルな情報を載せるのです。我々は、実際に論文を書くときに、ものすごくたくさん計算しても、それを全部載せたら論文が長くなってしまうから、ポイントだけ載せて、後はもうサプリメンタルインフォメーションに回すわけです。だから、そういう物は活用できるだろうと思うのです。データベースを作るだけのためにデータを作るというのは、多分それは非常に困るので、それはもう借りてくればいいんです。

細野（東京工業大学）：バイオの場合は全然違うと思うのです。日本の情報分野は世界的に圧倒的に強いということはありませんよね。材料は世界で一番強いのです。世界で一番強いところが全部情報を出すことはないのです。それが日本の戦略です。アメリカは情報が無いのです。だから計算機で食うしかないのです。ヨーロッパだって、ほとんどまともな情報が出てきてない。それがなぜ日本が、強い産業をもっているところが、データを出さなきゃいけないのですか。これこそ国益に反します。だから、今のところは公開情報だけを、データ化する。日本語で書いた物は、絶対にデータ化しない方がいい。日本はそもそも英語でハンディを負っている。ハンディを利用しなければだめです。

寺倉：田中先生のところでは、実際にいくつかやっておられて、計算機でデータをいっぱい作っていくのは、ある程度機械的にやれるところもあるので、元になるデータを、どこかから借りてくるのですが、Mat Navi の問題で最近、思ったのは、Mat Navi のデータは電子化されていないのです。要するに、コンピュータで読み取れないのです。それで、例えば計算機で自動的に情報を読み込んで、物質の一連のものを調べようとしても、Mat Navi ではできないのです。

山崎（物質・材料研究機構）：Mat Navi は無機材料に関しては、クリスタルインフォメーションファイルはダウンロードできるようになっています。それで今、誰が作るかという問題があったのですが、Mat Navi の前身は、JST が 1995 年に高品質物質データベースの開発ということで、7 年間 JST でやられて年間 4 億円ずつ、約 30 億使って開発されたのを、NIMS が引き受けて 10 年間で 15 億円ぐらい。ですから、今の Mat Navi は 45 億ぐらいすでに資金が投入されているのです。やはりデータベースは、続けることに意義があって、継続こそ力といった所があるのです。では誰がそれをやってきたかというと、JST 時代から民間会社の研究者の OB が論文読んで公開されたデータを数値化してデータベースにするということをやっています。

それから今、いわゆる無機材料の結晶構造データベースですが、JST 時代にスイスの MPDS と Dr. Villars が、今、ポーリングファイルプロジェクトと言っていますが、そこに JST が資金を投入して、約 8 万件のデータを当初 7 年間で構築したわけです。その後は MPDS が資金繰りをいろいろやりながら、今 25 万件の結晶構造データを持っているわけですが、それを今、我々購入したいと思っているのですが、著作権などいろいろな関係があって、単にお金を払えば我々の方に移管できるという問題でもないのです。ただ今年度、24 年度文科省のご配慮で補正予算がつきましたので、今、我々、結晶構造データベースと電子構造データベースを統合する基盤整備を進めております。

緒形：どうやってデータベースを作るか、そのソースについては、NIST でも非常に議論があって、ソースを論文から自動的にサンプリングするという手法を今はやっている所もある、投稿したら必ずデータベースに入れるという学会もあるのだけど、実際データの信頼性をどう確保するかというのが、各データベースで皆さん、慎重にやっているところがあります。また NIMS でも、例えば有機のポリインフォのデータをとるにしても、論文を何十件か選び出して、さらに査読者がそこから選び出したデータについて年間 600 件のデータを登録している。それだけ選んでデータを登録しないことには、データの信頼性

を確保できない。そうすると、データベースの信頼性はNIMSではそういう査読者の見識で選んでいると。また、無機材料についてはVillarsという、何十年にもわたって世界的な結晶データを作ったところしか、たぶん今、もうコスト的に作れないのではないかというぐらい、非常に限られたデータソースになっていると。そういうところがあって、それに匹敵するデータベースを今後、どう作っていくかが重要になると思います。

岡田（東京大学）:まずインセンティブの話なのですが、ブレイン･マシン･インターフェースという分野があって、fMRIで脳をスキャンして、その人が見ている画像を再構成するという論文が、Neuronという雑誌で出版されています．このデータはアップロードしているのです。なぜかというと後で述べるように、最終的には引用が増えるからなのです。そういう意味で、データベース解析に関しても、データ解析が職人芸的なものだと思っていると、多分データをクローズにしたくなります。でも、ベイズ推論などの系統的手法で解析してあると、どういう手法でどうやったかということがクリアになっています。そうすると、いろいろなデータ解析手法を試そうというきっかけなって、そのデータはダウンロードされます。そういう状況では、引用という形でインセンティブは必ず最初の人にあるわけです。要するに、どんなに後で精密な事をやったって、最初に何かやった人が評価されるというのが学問なのです。そうなってくると､いいデータは勝手に自己組織化して、使いやすくなると思うのです。さっきの話で、いいなと思ったのは論文からデータに飛べるというのは、各自が勝手にやれるわけです。そこでいいデータで、実験者にとっても実験はきちっとやれば、誰が触ってもいいわけですし、データ解析もきちっとやれば、後で誰がやっても良くて、そして最初にやったってことで一番偉いということになります。そういう仕組みが必要で、そのためにはデータ解析、もちろん実験をどうやったかということをクリアにすることと、マイニングをどうやったかということをきちっと書けば、多分、うまくいくのではないかと考えています。

もうひとつは、どういうデータベースを作るかではなくて、データ獲得がミッション・オリエンティドであることも良い流れをつくると思います。私はJAMSTECの方と知り合いなのですが､DONET（地震･津波観測監視システム）というので､彼らはデータを取っているだけではなくて、シミュレーションもしていいます。そしてDONETのデータを、例えば井手さんとかが納得するような方法で第一ステップとして､きちっと解析しておく。そうすれば、より科学が進んで、最初の解析が不十分で、10年後には前のデータ解析結果と違うということであって、それはそれで良いのだと思います。

話しは飛びますが､さきほどのオントロジーを自動的に作るというのも良いと思います。多分、井手さん等がやっているようなテキストマイニングをすれば、自動的に階層構造が出てきたりするかもしれません。

寺倉：テキストマイニングというのはどこまで進んでいるのですか。

井手（日本アイ・ビー・エム）：必ずしも私はテキストの専門家ではないのですが、確かにポピュラーな研究テーマだと思います。特にバイオ系のMEDLINEだとか文献データベースをクロールして、何か知識を作ろうという試みはあるのだろうと思います。ただ、それがどこまで現場の研究者が嬉しいことなのかというのは、少し私の立場からはむしろ

よくわからないところがあるのです。むしろ逆というか。もちろん、テキストマイニングをやった人は、「こんなの出ました」と自慢するのでしょうが、ちょっとそれが本当の現場というか、実験をされる方からどう見えるのかというのは、何とも言えないところです。

岡田：評価してもらって、テキストマイニングの技術を上げればいいじゃないですか。要するに、やりっぱなしだから、だめなのだと思うのです。例えばテキストマイニングをそのまま使えないのであれば、ランキングをしてあげて、評価されるようにテキストマイニングのアルゴリズムを変えればいいだけであって、情報学者、物質学者、そういう人たちのループが大事だと。情報科学者の方には、物質サイドからのランキングから情報処理をどう変えるかという方法はいっぱいあるわけです。それで、テキストマイニングの方法を改良すれば、多分、情報科学もすごく面白いのだと思う。情報科学の人も、論文を読んだエキスパートが、どう考えるかということを自分にフィードバックするということもっとやるようになると良いと思います。私の意見は、こういう問題は、実は情報科学の人も交えて両方とも進むというふうにしないと、多分、うまく行かないと思います。

細野:バイオや天気予報とは違う。そこを間違ったら、いい相関があっても役に立たない。フィジカル・プリンシプルこそが大事であって、それがあってモデリングができる。ですからフィジカル・プリンシプルにつながるようなデータがほしいのです。そこから次にそれを解釈して論文を書くだけではだめで､材料というのはそれを､実際に新しいものを作って、それで製品にしていかなければならない。

マテリアルサイエンスというのは、マテリアルというと物質と一緒になってしまうが、材料は何かというとお金を稼げる分野なのです。それは、面白いからやるだけじゃだめなのです。そのためにはどうすればいいかというのは、そんな甘い問題ではないのです。ただ、それも今のマテリアルサイエンスのやり方だけでは、不十分なことはみんなわかっている訳です。ただ､そういうのを何とかもっとうまくやっていかないと、このままいくと、日本はまたうっかりすると、米国に全部スキームを作られてしまう。エレクトロニクスでアップルが何も作らなくて韓国と台湾に作らせているように、部素材分野も下手をするとそういうふうになりかねない。このまま行くと、世界でナンバー１だとか世界ナンバー２だとか下らん議論をしていて、実際はそうじゃないところで、大事なところを取られてしまう危機感があるのです。そのためには、マテリアルゲノムとか情報分野の取組みでは最初負けている。でも実動部隊である材料分野は強いわけです。それを両方うまく組み合わせて、米国と同じ所をフォローしないで、なおかつ勝てるという式を作らないといけないでしょう。

寺倉：細野先生もこれまでずっとこの会合にいつも出てきてくださっているのですが、その背景には、フィジカル・プリンシプルは非常にはっきりしているんだけど、直感だけではなかなかわからない物がいっぱい増えてきたので、機械学習などで要するにヒントが欲しいんです。ヒントまたは何かの先ほどの記述子を適当に選んで、データマイニングで原因を追及したとしますね。でもそれは最終ソリューションが全然わからないのです。ただ、そういうものをヒントにして次のステップに進んでいくという、そういうところに行けるのではないかというのが、私の今の印象なのです。最終的な答えがいきなりこれから出て

くるとは思ってはいません。

岡田：私もそうは思っていて、私のこれまでの論調はデータベースをどう考えるかということで話したいだけで、データマイニングでマテリアルサイエンスをどう変えるかというのは、まったく別に思っていまです。ひとつだけ言っておかねなければいけないのは、いきなりニュートンの運動方程式は出ないです。データから、よき現象論を出すというプロセスは非常に大事で、それには私が説明したスパースモデルが使えるわけです。例えば、T2 乗と R3 乗というのをイコールで結ぶというのは、これは直感で出せますけど、例えば T128 乗と何かを足してイコール 0 になるような関係がもしかしたら埋め込まれている可能性があるわけです。それを人は直感では出せないだろうと私は思う訳です。そういう事を積み重ねて、よき現象論を作ったあとで、そのフィジカル・プリンシプルが出てくるということだと考えています。

寺倉：田中先生に、ご自身のところでやっておられることを、今までの話を関連させて少し説明してくださいませんか。

田中功（京都大学）：データベースについては今、議論を聞かせていただいて、既存の実験データから、あるいはパブリッシュされたものから作っていくようなところと、例えば第一原理計算をして新しいデータを積み重ねていくところと、方法としても違うのじゃないかなと思っています。文献からとっていくことは、ものすごくいろいろなノウハウがあって、それはちゃんと年月のかかることですし、それは大事にしていかなければならないと思います。アメリカ等でもマテリアル・ゲノムでやっているのは、第一原理計算等を元にしてやっているのがひとつの中心。第一原理計算の部分は、自動的に計算したものを共通のフォーマットで積み重ねていくようなシステムを作れば、極端にいえば後は機械が勝手にやっていくようなところがありますので、そういうフォーマットのところをできるだけ早く作る。それも日本だけということではなくて、オープン・データというものと非常に相性がいいと思いますので、そういうものと組み合わせても、アメリカの連中は向こうのデータベースを使ってくれていいよということを言ってくれているので、そこのデータソースの部分では競合するのではなくて、世界中オープンにしてやると。

データを使う、そのノウハウのところでうまく独自性を出して、多分それは直接、産業のプロダクトを作るところにつながると思います。使うところは知財の問題なんかも関わってくるので、うまくそれぞれのところで隠しながら、といいますか、手の内を見せないでやっていくという、その使い分けが大事だと思います。

寺倉：実験のデータと、計算でやるのは違いますけど、そうはいいながら計算もたちどころに CMSI で考えた時に、そういうふうなアクティビティを積極的にやっているかというと、そうではないです。

鹿野：つまり第一原理計算は一個良いコードがあれば機械が勝手にやると。そうしたら、もう研究者はいらないのではないかという議論につながりかねないじゃないですか。

田中功：それだけをやるだけだったら、本当にいらないと思います。プログラムを自分でも作らずに、商用のプログラムを買ってきて、ピッとクリックしたら、論文ができてしまうのです。でもその人は多分、5 年後には仕事はないです。そんなことは機械が置き換わ

るんですから。ただ今、寺倉先生がおっしゃったように、今それ、インセンティブがないから、そこの所はなかなか誰もしたがらないですけど、それは国から後押しがあって、共通フォーマットなりでき上がって、入れていきましょうということになれば、勝手にやっていくこともできると思うので。それはかなり持続可能なのではないかと思います。

寺倉：マシンラーニングで、もうひとつ私の頭にずっとひっかかっているのは、特にマテリアルの時にマルチスケールということは、データ解析と関係させてうまくやらなければいけないのではないかと思うのです。そういうモデリングとかマルチスケールに関して、数学では何か議論がありますか。

小谷（東北大学）：常行先生、田中先生、鹿野先生を招いて、階層性をキーワードに研究会をやりました。研究会を通して、実際の材料科学研究に役立つマルチスケール解析にはまだ課題があることが認識できました。数学では階層を繋ぐ概念や技術が現在、急速に発展しているところですし、そのように新しく開発された数学概念で他分野では使われていないものがまだたくさんあります。まずは、そのような技術を使っていただきたい。　それによって新しいことができます。異分野融合の一番良い時期でインセンティブはあると思います。数学と材料は今までインタラクションがなかったので、この間の研究集会ではお互いの持つ技術や課題を提示して、「面白そうですね」、「それを使ったら何かできるかもしれませんね」というところで終わったのでないかと思います。常行先生などから事例をたくさん紹介していただき、数学の人間も、「ああ、それは数学のこういうところに関係しているかもしれないので、これからやっていきたい」という感想でした。非常に数学にとっては刺激的だったし､材料科学との連携に加わる動機を持てる集会だったようです。

寺倉：物質科学でいうと、日本があまり強くないのは、いわゆるマルチスケールのところがあまり強くないんです。そこは欧米の方が、泥臭くてもいいから手間を掛けて、そういうところにアタックしていこうという努力をずいぶんされてきたんだけど、それは基本的には基礎から応用につなぐときにひとつの大きなネックになっていると思うのです。例えば、及川さんが今日お話しをされましたが、いってみれば第一原理計算から相図を作るところだってマルチスケールの問題ですし、フェーズフィールドはマルチスケールで…ああいうところがもう少しきちっと対応できるようなアクティビティがないといけないのではないかと。実は数カ月前、JST で日欧の元素戦略の共同研究の審査にかかわりました。ほとんどの提案において日欧の役割分担は非常に極端で、日本は物質を直接扱うと、欧米は計算をして、その計算もほとんどマルチスケールの計算をするということで、そういう組み合わせの提案が 90％ぐらいだったと思うんです。その時私は、非常に日本の計算が蔑ろにされているような印象を受けてがっかりしたんです。データ解析、データマイニングというのは、今のようなマルチスケールの解析のところでも、多分ずっと使える可能性がずいぶんあるし、モデリングは、物理的原理がないといけないんですけど、データの受け取りとか、そういう所を含めて、マルチスケールの問題というのは、非常に重要な課題として残っているのではないかと思うのですが。

中川（新日鐵住金）：マルチスケールと言った場合に、ミクロとメソですか、メソとマク

ロですか。
寺倉：まずはミクロとメソの方です。
中川：恐らくミクロとメソのところは、多分ミクロがいわゆる原子格子の、いわゆる離散系です。メソは連続系なので、非常に難しい問題だと思います、そこは。メソとマクロはある程度、均一化コードがありますので、ある程度できていると思うんです。今後はそのミクロの、メソのところに、いわゆる物理原理を入れながら、どういうふうにするかというのが問題。

小谷：原子分子材料科学高等研究機構は、まさにそこにフォーカスをしてやっています。若い人がすごく関心を持っています。過去 5 年間で AIMR がやってきたことは、原子、分子を観測し制御し、それがどのような現象に繋がるかを解明することです。昨年、そこに数学とか理論物理の研究者が加わり、階層間の関係を見つけようとしています。本当に、皆がすごく盛り上がっているという印象です。数学と材料科学連携が始まり、更にそこにシミュレーションが入ると、議論がスムーズになりますね。

寺倉：他にデータマイニングや、機械学習等、インフォマティクスの基本的な問題に関してコメントはありますか。

小谷：私はデータベースの事はよくわかっていないのですが、データベースを使った側の情報を収集することは、許されるでしょうか。例えば Google とかは、どの人はどういうふうに情報を使って、更に、その次にどういう情報を使うか等、そのような使用者の情報を解析しています。データの価値も上がりますし、さらに材料科学の開発において、無意識に研究者がやっている材料の開発のアイデアを、そこから取り出すこともできると思うのです。

曽根：先ほど細野先生が国営の話しもされていましたが、NIMS がどうなっているかというとデータベースにアクセスする場合は、まず登録をしてもらう形になっていて、どこの国の人が、どの機関の人が何にアクセスしているか。これは見ようによっては恐ろしいことで、だいたい月 100 万から 150 万件ぐらいアクセスがあるわけで、分かってしまうわけです。世界がどういうふうに動いているか、何にどこの機関が、どこに集中的にいろいろな事を調べようとしているのか。だから、ある国に対しては制限をかけるとか、国益という意味ではいろいろな事ができると思います。ただ、そのためにはそのデータベースが世界を席巻するようなボリュームのものでなければならず、まさにインターネットがそうですが、まさに Winners take all です。

細野：こういう方向が進むと、知財が大変なことになりますね。データのハンドリングについて、上位の概念で捉えられちゃうと。実際にはアメリカとヨーロッパは、そこは考えていると思います。

田中一宣（JST CRDS）：ちょっと観点が変わるのですが、今の、国がある程度ファンディングして世界に冠たるデータベースを作るということを含めて、インフォマティクスに対

する企業の危機感はいかがでしょうか。学術的な面から見れば、これは必要だというのはわかるんです。しかし、インセンティブはないけど、国がお金を付けてくれというのは、なかなか難しいんです。実際、そういう一面ではなくて、今の話にあったように、Google はすべての情報を集めています。それで例えば GDP も、一番信頼性のあるものはいずれ Google が出すだろうと言われるぐらい全てを読まれるわけです。中・長期の視点で見て、今、こういう所にファンディングしないと大変なことになると思っておられるのかどうか。あるいはファンディングしていただければ、それはそれでありがたいねという程度なのか。

伊藤：今回の議論はものすごくサイエンスに偏っていて、工学という観点がなくて、この提言も科学研究になっているんです。しかしそうではなくて、私が最初にこの分野に絶対に必要だと思ったのは、とにかく産業界はこういう取組みを進めたいわけです。私もずっとシミュレーションをやっていましたから、だいたい何かの開発を一緒にやります。実験部隊のやることに、全然シミュレーションは追いつかない訳です。これをやりましょうと言って、解析しても全然間に合わない。それでプロジェクトが終わった頃に計算結果が出て、「お前ら何やっていたんだ」と、必ず言われるわけです。それは何故かというと、やはり物理・化学だけで頑張っても間に合わないっていうところがある。では、他に何か使えないかというところに、最近データベースが出てきて、特にバイオインフォマティクスを見ていると、実験をやらないのに、あいつら何かいろいろな事、パスウェイ解析か何かやって、出ているじゃないかと。それができるなら是非やったほうがいい。

多分、シリコンとかかなり成熟しているところでは、インフォマティクスはいらないんです。私が後半、特にやっていたのは、蛍光体だとか燃料電池だとか、何だかよくわからない物を 3 年間で物にしろとか言うから、シミュレーションはほとんど間に合わないんです。そこで、産業界的にはそういう意識はものすごく強い。極端に言えば、物を開発するのに物理・化学である必要はないんです。最後はそれがわからないといけないことはよくわかっているんです。だけど、開発している時は物理・化学である必要はまったくなくて、答えに早く到達すれば何でもいい。だから、その意識というのは、どこのメーカーも強いと思います。

塚本（昭和電工）：企業のひとりとして非常に気になっているのは、ずっとお聞きしていると、データベースだとかオープン化だとか、あるいは機械で処理して出力しようと。それは言い方を変えると全部、デジタル化していくことです。それで我々企業側から今までの産業の歴史を見ると、デジタル化した途端に、世界の競争力を失うんです。液晶がしかり、電池もそうなり始めているんです。非常に大事なのは、日本の素材産業がこれまで強かったのは、アナログの世界で勝負してきたんです。もちろん、物理学的な原理原則はそれは当然いるとして、多くがやはり勘と経験の世界でいろいろやっている。それで今、私はたまたま化学会社ですが、化学の世界もペトロケミカルはまったく競争力がないんですが、ファインケミカルは非常に競争力を持っています。それは何故かというと、やはり勘と経験で整合性の問題とかいろいろなことを「アッ」と思う人がある物を合成して、思わぬ物ができるという世界なんです。

今、金属やセラミクスだけではなくて、ファインケミカルの分野でもインフォマティクスは非常に使い道があると思うんです。理想的に言えば、分子構造を完璧に並べれば鉄よ

り強くなるわけです。理屈の上ではそんな世界もあるわけですけれども、一方でこの議論が私非常にリスクがあるなと思うのは、ではデータベース化しましょう、共有化しましょう、オープン化しましょうといったら、まさしく日本の国力をいっきに外国に教えるようなものだと。アメリカは材料世界で遅れていることはわかっていますから、今、やっきなんです。

日本が下手にデータベースを公開することに私は大反対で、どうやって国の中に閉じこめてセキュリティを守るかということを一方で考えないと、アルゴリズムまでオープンにされ、論文が出、あ、そうかこのデータベース作ってこういうアルゴリズムでやれば、こういう実験せずとも答えがでるのか、というと、あっという間に日本の競争力がなくなります。

もちろん一方、こういうことをやろうとすれば、人材がどんどん育って、すそ野が広がってということはある、それやるためにはデータベースのオープン化というのは必要なんですが、どういう仕掛けで、必要なところ、肝のところはクローズにするかと。我々の会社の事業でいけば、例えば電動用のカーボンですとか、ハードディスクのメディアですとか、これはいっさい特許を出さないのです。何故かというと、出した途端にデジタル化して、海外に漏れるわけです。なので、すべてノウハウで担保する。それだけが相変わらず収益の源泉になっています。そこらのところはよくご理解いただかないと。

学術の世界でどっと行っていただく。これは中身としては非常に私も是非、やっていただきたいと思っているのですが、国の産業競争力ということで言うと、どんどんアナログがデジタル化して、デジタル化であっという間に国境を技術が越えますから、そこのリスクをどうヘッジするかというのが非常に大事だと思います。

田中一：今の塚本さんのお話で私が一番気になるのは、国（日本）にそういう情報を閉じこめるということができるでしょうか。国民性としてできないと思っているんです。必ず漏れるってことを前提として戦略を立てないと、日本は戦略にならない。

細野：やはり一番危険なのは、上位の概念で知財を押さえられてしまうということです。各論ではなくて上位の概念で捉えてしまう。だから日本はやはりこの取組みをきちっとやらなければならないと思うのです。本気でやらないといけない。情報は漏れるんですけど、ゆっくり漏らしゃいいんです。上位の概念はサイエンスでは関係ないと言いますが、やっぱり世界的にはどこかでサイエンスが担保されない限り、嘘のつき放題になってしまいます。だからやはり、ある程度のサイエンスがあって、ある程度の客観性があって、個々の物質名にとらわれないコンセプトで知財が成立する時代になっちゃうと思うんです。日本はこのまま行ったら、戦艦大和と同じで、また大事な部分がなくなってしまいます。好きとか嫌いとかに関係なく、もうやらなきゃ仕方がないと思うのです。ただ、日本流にやらなければいけない。アメリカの後を追っても勝てないです。

寺倉：多分、一番基本的なことは、データを活用するような研究の進め方というのを、どのくらい重視してやっていくかという、そこだと思います。あとはまあ、気にしなくてはならない事はいろいろあると思うんですが、少なくとも我々のコミュニティの中で、データを非常に重視しながら研究を進めていくという、そういうムーブメントが無かったとい

うことだと思っていて、これをうまくスタートしなきゃならんというのが一番の動機です。だからあとの細かいいろいろな問題があることはわかっているのですが、それはその過程で整備していく以外になくて、いろいろな提言のためにデータを活用していくことに関して、あまりネガティブにならないような進め方を是非とも早くしなければならないのかなと、そういうふうに思っています。あとは具体的にどうやるかという話なのですが。

中川：異分野融合という視点から物を言わせていただきますと、よい事例があれば良い人が集まります。データも良い事例になるようなデータをいただければ、そういう事に関心のある人を集めて、物理法則も含めて、それをどういうふうに表現するかというモデリングや、数学手法も含めて、そういうことができるかなと思います。だから初めは良い事例を小さく始めて、それで具体的な事例ができるのを議論していくことができれば良いと思います。もしデータをいただく場合に、どなたに相談すればいいか、ですとかね。そういうふうな、窓口を示していただければ、まずアプローチして、小さな事例でまずは具体的にこつこつ始めるというふうな…。

寺倉：ええ、大規模にやる程、実はコミュニティがボリュームはないので、その現状からいっても、始めるにしても少なくとも最初はこじんまりしたところから始める以外にないと思います。ただ、どこにも種がなければ費用の割りようもないので、種をどこに植え付けていつ頃から始めるかという、そこが一番頭の中に問題点として残っているんです。すでに中核となるところは多分だいたい決まっていると思うんですが、そこが中核となるような具体的な動きをどうやって進めていくかという…。

伊藤：もちろんこういう方々が中心になってやられると思うんですけど、この前後にプロジェクトの立て方です。多分、どういう開発プログラムをやるかというところだと思うんですが、今のような方々でそれをやると、いわゆるシーズ・オリエンテッドなやり方になると思うんです。持っている技術から始まると。今回やはり、特に産業界の観点からすると、やはりニーズ・ドリブンであってもいいのではないか。元素戦略はそういうところが当然あって、レアアースを使わない、使わないのにも関わらず現実にはけっこうシーズ・ドリブンになっていますけども。本当にニーズ・ドリブンでやっていった時に、もちろんそこが物理・化学に落ちるというのは従来の考えだし、マルチスケールシミュレーションに落ちるかもしれないし、いきなり本当に情報科学に落っこちてしまうかもしれない。そういうやりかたがひとつや二つあって…。ニーズ・ドリブンは多分産業界の方々にデータがあるので、そういうことがひとつ、二つあってもいいのではないかという気がします。

寺倉：そうでしょうね。私がだいぶ前から元素戦略で引っかかっているのは、私としてはそれがニーズ・オリエンテッドだと思っていたんです。だから、データマイニングとか、インフォマティクスをともかくやってみて、遊びのままとまっているのでは困るので、何が本当に求められているか明確な所と連携しながらやると。それが多分、現実的には一番いいんじゃないかと。いきなりはっきりしないところから進めるよりは、その辺りを頭に入れながらやれるようにしてはどうかと思っています。今、伊藤さんがおっしゃったのはもう、少し広く考えるべきかもしれませんけども、発想は特には違っていないと思います。

だからこれでいうと、データ・インフォマティクスの新たなコミュニティを、情報系の人を交えて作らなくてはならないと思う。計測解析の所は、ああいう大型施設が使えますし、元素戦略のようなものを念頭に置きながら、何か具体的なアクティビティがセットアップできないかと思っているのですが、それが政策的にいいかどうかどうかはちょっとわかりませんが。

馬場（文部科学省）：こういうマテリアルインフォマティクスの話を重要だと認識しつつも、ではそれをどうやってやるかは悩ましいなと思っていたところです。もし国としてこういったプロジェクトをするときは、やはり目的を見失わないようにしていかなければならない。何が言いたいかというと、先ほどデータベースの拡充の話もありましたが、データベースをたくさん作ろうということだけが目的になってしまうと、恐らくニーズとかけ離れてしまうことになってしまうと思いますし、アメリカのマテリアル・ゲノムも材料開発が 30 年、それを半分にする等、いろいろな目標がある中で多分彼らもプロジェクト明記をしていると思うので、それをどこに置くかということをきちんと考えていかなえればならないだろうと思っています。

次にそれをどうやってやるかということなのですが、日本の国益ではないですが、やはりしたたかにやる必要があるんだろうなと、今のところは思っています。個人的な意見ですが、NIMS の Mat Navi といったものをまず中心に拡充するというのは、やり方としてひとつあると。歴史的なことを参考にご紹介すると、旧 JICST は元々情報整備のための法人があって、そこでいろいろデータベースを作ってきて、独法を制度設計するときに、その中で材料系については NIMS に移管したという話なのです。昔 JST を担当していて驚いたのが、旧 JICST 時代に本当に凄まじい数のデータベースを作っていたのです。それが今、どうなっているかというと、公費投入した分に対して活用されているかということは、ちょっとなかなか疑問なところがあったりはしていて、それはもしかしたら、データベース化自体が目的だったからそうなってしまったのではないかという気はしています。ただ、その一方で、こういったものを国として基盤としてちゃんと確立していかなければならないということで、改善・拡充していかなければならないとは思っているところです。

あと二つ目のデータ駆動型の部分について思うのは、元々は戦略目標などが念頭にあったかと思うのですが、やはりそれがメインになってしまうと、どうしてもまたそういった研究者と乖離したところで、ひとつの分野ができてしまうような気がしていて、やはり成功例のようなものをひとつ作っていくことが大事だと思っています。今であれば日本の中でも元素戦略であったり、東北大学の WPI であったり CMSI など、そういったところで取組みが進んでいるので、そういったところで成功例を出しつつ、それはクローズでやりつつ、それはそれとしつつ公表されるデータベースというのをやはり両輪とした方針というか方向性を出していくべきなのかなと、個人的には思っています。

もうひとつだけ言えることとして、他の分野のデータベース等は参考にすべきだと自分も思っています。今日、山口さんから説明があったライセンス統合データベースもそうなんですが、多分、国のひとつの反省として、この統合データベースができる前は各省庁でライフ系のデータベースが、文科省だけでなく農水省、経産省でもできて、それを CSTP で音頭を取っていただいて統合データベースを JST でやるという経緯があったところも

あります。そこで、もしこういったものを国として力を入れてやるからには、最初の段階でどうやっていくのかというところを当然省庁に限らず、企業の方々のニーズ、また関与も含めて設計した上で、検討していくというのが、まず大前提だし、そうなってくるのかなと思います。

今、個人的な意見を申し上げましたが、それをどういうふうに進めていくかは今日、いただいた議論も参考に政府の中でも検討させていただいて、引き続き議論を積み重ねていくことが必要だと思います。

本間（文部科学省）：ニーズ・オリエンテッドなデータベースとインフォマティクスというのは、非常に重要だと思います。データベースに関して、私は一応民間企業（新日鐵住金）出身なものですから、少なくとも構造材料に関しては歴史的伝統的に明らかにデータ駆動型でやってきたのだろうなと思います。それをコンピュータでやるかどうかだけの違いかなと思うわけです。例えば今日、及川先生が説明された CALPHAD をひとつとデータベースとして見れば、かなり成功した例だと思います。世界中にかなり普及しております。しかし一方で、CALPHAD は鹿野先生がおっしゃった通り、平衡状態のデータベースですから、現場で実現されることは絶対にありません。しかしながら現場は理想状態から乖離があるけれども、その乖離がどうあるかということを考えるわけです。それがノウハウになります。それは絶対に公開いたしません。

ただこれを新日本製鐵が何十年もかけて人のノウハウで培ってきたことで、今後それではもうもたないということで数理研究所のようなものができているわけです。この辺りのノウハウを、もし今日岡田先生がおっしゃったようなベイズ統計で対比ができていって、その手法を隠せるのであれば、これは非常に強い力になると思います。

細野先生が指摘された知財として上位概念を立ててどうするかというところは、非常に悩ましいところで、現実問題、Mat Navi には新日鉄が出しているデータもありますが、「これ出してもいいや」と、「維持するよりも出した方が安い」ということで、提出しているのがありまして、コストに見合うという物ですが、そこは戦いだと思いますので、とられる前に何を上位概念として日本が出せるかだと思います。守りではなく攻めなければいけないと思います。そういった意味で、どんどんチャレンジしていく必要があるかなと。

ただ本当に、データのためのデータベースを作るのは本当に意味がないことは弊社も経験していますので、とことんニーズ・オリエンテッドを追求していきたい。そういった意味で今回、数学者の方々がここまで材料のことに参画してくださって、非常にいい傾向だと思ってます。小さな成功例で言いますと、たぶん構造材料が一番手近かなという気がします。セラミックスの世界でもそうなのかもしれませんが、小さな例をひとつ。例えば中川さんと相談して出せる部分を少し紹介していただいたりしてもいいかなと思います。その中でもう一度データベースのありかた、それからデータの拡充の仕方です。無闇やたらに広げるのではなくて、こういうデータが必要だからこういう形で欲しいというのをやっていただけたら、きっと数学者の方々、理論物理の方々も納得していただけるデータベースの方に行けるのではないかと思っています。

最後にひとつだけ、私も日米欧のレアアースワークショップに出て参りましたが、レアアースにかこつけて日米欧の材料化学がお互いにどのように評価しあえるかということを議論したのが実態です。その中でアメリカのマテリアル・ゲノムの官僚側が「本当にこん

なことを始めて、良かったんだろうか」と悩んでいて、「日本から見てどう思う」と言うわけです。私がからかい気味に「あんなのダメだ」と言ったら「やっぱり止めようかな」とか言ったりするわけですけれど、逆にアメリカもまだまだ本当に日本に勝てるかと暗中模索です。データに関しては、日本のデータが一番質がいいということは、世界中の誰もが何十年間も認めていることで、逆にこれを逆手にとって世界中のインフォマティクスを牽引しない手はないと思います。その辺り多分、細野先生がおっしゃりたかったことではないかと思いますが、日本の強みをとことん抽出して、その中で世界から「やはり日本がいないとマテリアル・ゲノムはだめだ」と言ってもらえるようなインフォマティクスを作り上げていただきたいなと思います。

常行（東京大学）:どう進めるかという話を、一言だけさせていただきたいと思うのですが、データベースを作って維持管理するところと、そのデータを使ったりする所は切り離してできるとは思うのです。ただ、、データベースを作っているところ、例えばNIMSを中心に、データベースを作って、それをずっと維持管理して、それだけで終わってしまうと、NIMSとしては非常にやりにくい、メリットを感じられないだろうと思うんです。その意味では、最後の成功体験を何かつくるというところに、データベースを持っているところが積極的に絡めるような仕組みを、まずは作るのがいいかなと印象を持ちました。

それからもう一点は、基本的な話ですが、データマイニングという言葉がたくさん出てきましたが、今日の岡田先生の話を聞いていて思ったのは、実験データから情報を抜き出す、意味のあるデータを抜き出すところは、またちょっと物を作るとは別の話として非常に大事なところで、例えばJ-PARCとかSPring8で膨大なデータが出てきて、処理しきれないでいるわけです。そこからどういう情報を抜き出してデータベース化すればいいのか、一次データから二次データの意味のあるものを作るところが、多分ものすごく問題で、そこはそこで別枠で考えなければならないのかなという印象を受けました。

寺倉：一次データから二次データの問題点というのは、こういう事をやっていく時に計算結果のバリデーションとか、逆に実験結果を計算でバリデートするというのがあると思うんですが、そこの実験と計算やデータ解析は非常に密接に繋がっているので、そこも含めた議論を進めていきたいとは思っています。

細野：NIMSにそんなデータがあるなら売ればいいじゃないですか。別のベンチャーを作ればいいわけでしょう。売れないものをどんどんと作る必要はないです。第一、我々みんなICSD（結晶データベース）買っていますよ。ICSDがあって、それを越えられないならいらないです。そこはNIMSで真面目に考えてください。今あるものより良くて売れるものを作らなければ、国費を投資する意味はないです。だいたい、データベースでも、コンピュータもソフトウェアでも、日本発で売れているものはあるんですか。いつも我々、買っているだけです。計算のシミュレーションの人達も甘えていると思いますよ。結局、VASPでもWin2kでも皆海外製を使っているんです。データベースだって全部外国です。日本発のものが無いじゃないですか。

曽根：それは重要なポイントでNIMSでも議論しています。レアアースと同じで、デー

タベースも中国に握られたら恐ろしいことになります。そういう方向を目指していくということです。そうはいってもデータがぼこぼこ出てきても仕方がないところがあって、問題はニーズ・オリエンテッドでそれをどう活用していくかです。今、元素戦略のセンターができたわけで、あそこをうまく活用して、是非、こういう視点でいろんな元素戦略のプログラムをマテリアルインフォマティクスという視点もしっかり入れてドライブしていく、そういう仕組みも必要だと思います。ここで最先端のインフォマティクスの技術の人達との会合からいろんな新しいものが生まれてくるだろうなという期待はすごくあります。その両者がいろんなコミュニケーションして新しい物を作る、相当バリアがあると思いますが、これはやる価値があると。

細野：日本発のソフトウェアとか日本発のジャーナルも含めて何もないんです。日本のジャーナルって、ニーズで切ったらゼロになっちゃいますよ。それで、データベースに追加投資してキリがないです。我々はICSDがあればいい、それで間に合っているんです。研究現場ではNIMSのを使わなくても全く不自由していません。

曽根：今、相当の勢いで我々のところもアクセスが増えてきていて、月150万件ぐらいあります。

細野：そこはタダだから来ているんです。

曽根：インターネットのビジネスは、まずユーザーを確保して、それからお金を取る仕組みです。ユーザーを握ってしまえばいろんな仕組みを作れる。とにかく握っちゃわないと前へ進めない。そのあとならいろんな、細かいお金の落とし方はいくらでもできる。

細野：計算のソフトとデータベースのソフトと、全部、外国に握られたら、我々どこで勝つんですか。制海権握られて、制空権握られて、あとは地上戦ですか。そこを本当に真面目に考えないと、最後は言葉の問題に行き着いてしまう。東洋の島国は勝てないのかっていう、ずっとそれが根底にあるんです。だから田中一宜さんが昔言われたけど、日本人は欧米と競っていくには、2倍以上働かなければならない。今でもまったく同じことです。

緒形：売ることに関してはNIMSも十分に検討していますし、一部は売っているものもあります。あと、売るだけでなく、データベースやデータは資源ですから、先ほども言ったようにどの国家が握るか。計算は握られているかもしれないけど、データも海外にアクセスしに行くようになってしまったら、全部握られちゃいます。データベースをアクセスしていけなくなってしまったら、もうその材料開発ができなくなってしまう。だから独自のデータベースを持つということも、ひとつの考えとしてあると思うんです。もし外国が日本をシャットアウトしたら、見に行けなくなってしまうわけです。もしICSDのデータを全部どこかが握ってしまったらば、日本は高いお金を払わないと見に行けなくなってしまうわけです。いかに自前のデータを維持するかということも大事なデータベースの位置だと思います。

岡田：今の制空権を握られるという話はNMRもそういう面はあるそうです。圧縮センシングのソフトがバンドルされているのがヨーロッパ製らしいです。スイッチ一つだけで解析まで出来るのですごく売れるのです。計測精度は日本製がすごくいいと聞きますが、結局、ソフトで負けているようです。日本は、さらに情報科学と計測が融合していないとい

う話しは良く聞きます、ヨーロッパは既にそれをビジネスチャンスとしてちゃんとやっているということです。製品を作るマザーマシンを押さえるべきだと思います。ここでも高い計測装置を押さえて、そこにバンドルソフトをいれて、現場の人がスイッチだけ押せるようなマシンを売るというようなビジネスとしての戦略です。それが日本の一番弱い点だと言われています。ある研究費の提案で、ベイズ計測という枠組を以前に出したことがあります。ベイジアン・センシングです。センシングの所にベイズ推論を入れて、それで高度な計測装置の市場をおさえるという戦略が必要かと思います。そのパイロットマシンを共同研究として先端の研究者の人に使ってもらって、データをとれば、サイエンティフィックな面の向上も狙えます。

寺倉：細野先生のコメントで、だいぶピリッと締りましたが、それは本当に深刻に我々もとっています。日本は群雄割拠でいっぱいいろんな物を持っているんだけど、まとまった形では、VASP とかいくつかの物とくらべて売れるようなものは無いことは事実で、どうやって力を結集していくかというところを、いつも CMSI で努力はしています。非常に厳しい警告ですが、心してやらなければならないことだと思います。

4. サマリー（ワークショップからの気づき）

背景

- マテリアルの研究者の多くは結晶構造などのデータベースを睨みながら、候補物質を探索している。合成された物質の数と物性が膨大になっており、結晶構造や分子式など従来の整理要素だけでは見通しが不十分になってきている。
- 最先端のデバイスや触媒などの高効率化に資する理論の解明において残された問題は、非常に複雑な系（物理状態）にあり、従来の手法では効果が限定的。
- 実験と計算の間には、「タイミング」、「スピード」、「成果の共有」という 3 つの課題がある。考察に足る計算結果の創出に時間がかかりがちであるため、実験の途中に計算結果をフィードバックできない。
- デバイスや素材に必要な機能をいろいろな元素の組み合わせで実現しようという社会的な強い要請。
- 物質と材料（社会・産業の役に立つ物質）の間には大きなギャップがある。
- 計測技術の進歩により短時間で大量のデータを取得可能。特に、3 次元計測技術の進展が著しい。大規模計測施設では、これまでのデータとは次元の違うデータがとれる（たとえばブリュアンゾーンの中の三次元のベクトル方向、更にエネルギー方向の四次元のデータが一回に取れる）ため、今までのようなデータ解析をすれば済むような問題ではなくなってきている。
- 第一原理計算が手軽にできるインフラ環境の普及。
  ⇒電子状態を確認してから実験をやった方が効率的。従来は実験の傍証としての計算であったが、ここ数年でパラダイムシフトが起こっている。
  ⇒計算科学の長所として、1 点で実験との検証をすると、そのあとの横展開が容易。
  ⇒計算によって構造と物性値（一次情報）のデータを容易に入手可能。米国のマテリアルゲノムイニシアティブを受け、計算による電子・結晶構造のデータベースに進展あり。
  - ✓ MIT「MaterialsExplorer」、「Phase Diagram App」
  - ✓ デューク大学「AflowLib: Ab-initio Electronic Structure Library」
- 各スケールでの解析は行ってきたが、スケール間の連関が希薄。
- データマイニングや機械学習は、測定の精度をあげる、あるいは生のデータでは見えない背後の現象をきちんと統計的に推定、判別して、現象が起きている、起きていないということを調べることが非常に得意。また､非常に高次のデータをシンプルにし、因果関係を推論することや非常に少数サンプルのデータからいろいろなものを推定するという研究が進展している。

目的

- 科学研究手法として、解析⇒予測⇒設計の道筋の確立を目指す（設計は未だし）。
- 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）の発見、その活用による新材料の創出を図る。
- 膨大なデータから意味のある情報を抽出し、物理・化学的法則の発見につなげる（機

械学習、データマイニングによって、データから規則を割り出す。不確実性のあるデータから確実性のある結果の創出。機械学習（原因予測）によって、意味のある記述子を見出す）。
- 新物質の発見から新材料の実用化までの時間とコストを大幅に縮減することで、材料の側面から社会課題に対応し、産業競争力を強化する。

手段

近年進展してきたプロセス、計測、計算の研究者の密接な連携に加え、データ処理（情報科学、数理科学）の連携・統合により、科学技術研究の新しい流れを創る。
- 研究者個人の能力、セレンディピティに加え、シミュレーションによる大量データ、又は実験データベース（例えば、MatNavi）の有効活用によるコンピュータ（①機械学習による予測と②計算機の持つ網羅性）活用型の発見、材料創成へ。
- 演繹的手法と帰納的方法をうまく併用することにより研究開発のパラダイムシフトを起こす。

そのためには自分が実験等して得たデータと、パブリックなデータを統合して解析できることが理想である。また、バイオインフォマティクスやケモインフォマティクスなど先行する技術・ノウハウの活用が有効である。

➢ データ解析の方法・技術
- 物性物理や無機材料における QSPR（定量的構造物性相関）の確立
- 計算可能な物性値を記述子とする複雑特性の予測
- 第一原理計算とクリギングの組み合わせによる最適物質自動設計システムの構築
- シミュレーションとコンビナトリアル合成・評価の組み合わせによるハイスループット材料設計サイクルの構築
- セルフコンシステントな予測と検証
- データ同化の取込み
- 複雑な一次データの多面的な解析技術

➢ 定量データを通じた実験（計測）と計算の連携によるモデリングの高度化、マルチスケールの連結
- 計測、モデリング、プロセスを初めから組み合わせた材料開発
- 計量的な特徴値だけではなく、トポロジカルな特徴の記述

➢ データ整備（無機材料は 10 万ー 100 万物質。米国でもまだ数万レベル）
- 材料データの多面的な蓄積技術とデータベースとのリンク
- 計算 DB と実験 DB のリンク
- RDF などによる DB の整備
- マシンリーディングの活用、発展
- 構造材における組織データの 3 次元、4 次元のデジタルアーカイブ化
- データ取得に適した特殊装置の開発

障害

- 物質材料分野の研究者にデータをインテンシブに活用するという意識が希薄。材料コミュニティや産業界における多くの研究者は、その可能性やその存在さえも気づいていない。
- シミュレーションや理論という演繹的な考え方と経験や大量データから知識を産み出す帰納的な考え方は、まったく異なる。この異なる考え方をつなぐ頭の整理が難しく、科学者自身に戸惑いがある。
- 複雑世界では、予測可能性のために説明可能性を諦めなければならない（道具主義）という意識の醸成
- 実験による結晶構造や物性データベースはバラバラに存在し、パブリックにアクセスでき、二次利用できるものはほとんどない。
- 鉄鋼材などの構造材料において、プロセスと特性を結びつけるための組織のデータがいまだにアナログで、モデリングにつなげられない。

効果

科学的側面

- 広範な物質系（有機 / 無機 / 金属）の俯瞰方法の創出とそれを用いた物質科学の因数分解（物質・材料を記述子で張る空間で特徴づける）
- 未開拓領域の発見（対立軸による展開、空白領域）
- 大量のデータからの物理・化学量を制御するパラメータの発見
- 一次データの共有による多面的な科学の発見

実用的側面

- ハイスループットな物質設計、物質開発
- 物性値のマトリックスとしての設計（工学）
- 包括的候補物質選択（構造探索）
- 定量的な材料設計指針の確立
- テクノロジーギャップの解消
- プロセスパラメータの決定

具体例：

- Lee plot や Miedemarule のような経験則が導かれる → 経験則の物理的意味付け
- 物理的洞察に基づく記述子の選択（酸素還元反応での OH 吸着自由エネルギー）→ 有効なデータハンドリング、物質開発指針

付録

1．第 1 回プログラム

開催日時：平成 25 年 2 月 11 日（月）10 時 00 分ー 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 ー 10:30　オーガナイザー挨拶　田中 一宜
趣旨説明　寺倉 清之（北陸先端大、産総研）

10:30 ー 11:00
「機能に基づく材料の設計ー理論、実験、計算、データの統合への期待ー」
細野 秀雄（東京工業大学）
11:00 ー 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 ー 12:00
「『京』を用いたインシリコ創薬」
奥野 恭史（京都大学）

休憩（昼食）

12:40 ー 13:10
「第一原理計算を用いたマテリアルインフォマティクス」
田中 功（京都大学）
13:10 ー 13:50
「Combinatorial and High-Throughput Screening of Materials Libraries、及び MGI 等米国の動向」
竹内 一郎（メリーランド大学）
13:50 ー 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 ー 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 ー 15:40
「データ同化によるモデルの高度化ー物質材料研究への応用ー」
　樋口 知之（統計数理研究所）
15:40 ー 16:10
「企業における事例紹介及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 ー 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 ー 17:45　　閉会　田中 一宜（JST/CRDS）

コメンテーター : 新井 正敏（JPARC）、伊藤 聡（理化学研究所）、小谷 元子（東北大学）、曽根 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）

2. 第 1 回参加者リスト（敬称略）

発表者（発表順）

寺倉 清之	名誉リサーチャー (シニアプロフェッサー)	産業技術総合研究所 (北陸先端科学技術大学院大学)
細野 秀雄	教授	東京工業大学　フロンティア研究機構
津田 宏治	研究班長	産業技術総合研究所　生命情報工学研究センター
奥野 恭史	教授	京都大学　大学院薬学研究科
田中 功	教授	京都大学　大学院工学研究科
竹内 一郎	教授	メリーランド大学
知京 豊裕	ユニット長	物質・材料研究機構　MANA ナノエレクトロニクス材料ユニット
足立 吉隆	教授	鹿児島大学　大学院理工学研究科
樋口 知之	所長	情報・システム研究機構　統計数理研究所
射場 英紀	部長	トヨタ自動車株式会社　電池研究部
信原 邦啓	主任	トヨタ自動車 電池研究部

コメンテータ（50 音順）

新井 正敏	ディビジョン長	J-PARC センター　物質生命科学ディビジョン
伊藤 聡	コーディネーター	理化学研究所　計算科学研究機構
小谷 元子	機構長・教授	東北大学　原子分子材料科学高等研究機構
曽根 純一	理事 (特任フェロー)	物質・材料研究機構 (科学技術振興機構　研究開発戦略センター)
常行 真司	教授	東京大学　大学院理学系研究科
中村 振一郎	特別招聘研究員 (フェロー)	理化学研究所 (三菱化学)
福山 秀敏	副学長	東京理科大学
船津 公人	教授	東京大学　大学院工学系研究科
松宮 徹	顧問	新日鐵住金株式会社
鷲尾 隆	教授	大阪大学　産業科学研究所

オブザーバー（50 音順）

赤木 和人	准教授	東北大学　原子分子材料科学高等研究機構
浅井 美博	副部門長	産業技術総合研究所　ナノシステム研究部門
安藤 康伸	産総研特別研究員	産業技術総合研究所　ナノシステム研究部門
石田 豊和	主任研究員	産業技術総合研究所　ナノシステム研究部門
石橋 章司	研究グループ長	産業技術総合研究所　ナノシステム研究部門
井下 猛	特別研究員	物質・材料研究機構　材料情報ステーション
緒形 俊夫	ステーション長・ グループリーダー	物質・材料研究機構 中核機能部門　材料情報ステーション／環境・エネルギー材料部門

門平 卓也	主任エンジニア	物質・材料研究機構　つくばイノベーションアリーナ推進室
古宇田 光	プロジェクトマネージャ	東京大学　物性研究所
香山 正憲	上席研究員･グループ長	産業技術総合研究所　ユビキタスエネルギー研究部門
小西 優祐	産総研特別研究員	産業技術総合研究所　ナノシステム研究部門
佐々木 直哉	主管研究長	日立製作所　日立研究所
高田 章	主幹研究員	旭硝子　中央研究所
塚田 捷	事務部門長	東北大学　原子分子材料科学高等研究機構
塚本 建次	技術顧問 （副会長）	昭和電工株式会社 （ナノテクノロジービジネス推進協議会（NBCI））
藤堂 眞治	特任教授	東京大学　物性研究所
長岡 正隆	教授	名古屋大学 大学院情報科学研究科
中村 壮伸	助教	東北大学　原子分子材料科学高等研究機構
野原 実	教授	岡山大学　大学院自然科学研究科
溝口 照康	准教授	東京大学　生産技術研究所
宮本 良之	研究グループ長	産業技術総合研究所　ナノシステム研究部門
山崎 政義	特別研究員	物質・材料研究機構　材料情報ステーション
山下 晃一	教授	東京大学　工学系研究科　化学システム工学専攻
渡邉 聡	教授	東京大学　大学院工学系研究科
渡辺 啓正	マネージャー	HPC システムズ　HPC 事業部
グエン・ヴィエット・クーン	シニアエンジニア	HPC システムズ　HPC 事業部

関係府省

守屋 直文	政策企画調査官	内閣府 政策統括官付（科学技術政策・イノベーション担当）
永井 雅規	室長	文部科学省　研究振興局　基盤研究課　ナノテクノロジー・材料開発推進室
馬場 大輔	室長補佐	文部科学省　研究振興局　基盤研究課　ナノテクノロジー・材料開発推進室
河村 麻美	係長	文部科学省　研究振興局　基盤研究課　ナノテクノロジー・材料開発推進室
本間 穂高	調査員	文部科学省　研究振興局　基盤研究課　ナノテクノロジー・材料開発推進室
小笠原 敦	センター長	科学技術政策研究所 科学技術動向研究センター

科学技術振興機構

佐藤 一美	主任調査員	研究プロジェクト推進部
住本 研一	調査役	知識基盤情報部
辻 伸二	主任調査員	研究振興支援業務室／戦略研究推進部／科学技術イノベーション企画推進室
恒松 直幸	上席主任調査員	情報企画部
中山 智弘	参事役・調査役・エキスパート	科学技術イノベーション企画推進室／研究開発戦略センター／戦略推進室
古川 雅士	研究監	科学技術イノベーション企画推進室　ナノテクノロジー・材料分野
山本 摸	主任調査員	産学基礎基盤推進部　事業推進担当

科学技術振興機構 研究開発戦略センター

田中 一宜	上席フェロー	研究開発戦略センター ナノテクノロジー・材料ユニット
石原 聰	特任フェロー(研究所長)	研究開発戦略センター（ニューメディア総合研究所）
島津 博基	フェロー	研究開発戦略センター ナノテクノロジー・材料ユニット
鈴木 響子	フェロー	研究開発戦略センター ライフサイエンス・臨床ユニット
永野 智己	フェロー	研究開発戦略センター ナノテクノロジー・材料ユニット
中本 信也	フェロー	研究開発戦略センター ナノテクノロジー・材料ユニット
馬場 寿夫	フェロー	研究開発戦略センター ナノテクノロジー・材料ユニット
的場 正憲	フェロー	研究開発戦略センター 電子情報通信ユニット
宮下 永	フェロー	研究開発戦略センター ナノテクノロジー・材料ユニット

3. 第２回プログラム

開催日時：平成 25 年 6 月 1 日（土）13 時 00 分ー 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（予定、敬称略）
CMSI：寺倉 清之（産総研）、常行 真司（東大）、田中 功（京大）、伊藤 聡（理研）
元素戦略：細野 秀雄（東工大）、田中 功（京大・再掲）
NIMS 材料情報 S：曽根 純一、緒形 俊夫（NIMS）
東北大学 WPI：小谷 元子（東北大）
数学イノベーション：中川 淳一（新日鐵住金）、小谷 元子（東北大・再掲）
データ・統計：井手 剛（IBM）、鹿野 豊（分子研）
文科省（ナノ材室、数学ユニット）

Ⅰ　挨拶と趣旨説明（13 時 00 分ー 13 時 30 分）
13：00 ー 13：10
田中 一宜（JST 研究開発戦略センター）および寺倉 清之（産業技術総合研究所）
ご挨拶
13：10 ー 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 分ー 15 時 30 分）
13：30 ー 14：00
「ベイズ推論と物性科学」
岡田 真人（東京大学）
14：00 ー 14：30
「材料設計とデータベース」
及川 勝成（東北大学）
14：30 ー 15：00
「オープンデータの潮流とデータの統合利用ーライフサイエンスの例ー」
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 ー 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 分ー 17 時 45 分）

15：45 ー 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 ー 17：45

意見交換

4. 第２回参加者リスト（敬称略）

発表者（発表順）

岡田 真人	教授	東京大学 大学院新領域創成科学研究科
及川 勝成	教授	東北大学 大学院工学研究科
山口 敦子	特任准教授	ライフサイエンス統合データベースセンター
芦野 俊宏	教授	東洋大学　国際地域学部

討議者

寺倉 清之	名誉リサーチャー	産業技術総合研究所 （北陸先端科学技術大学院大学）
田中 功	教授	京都大学 大学院工学研究科
常行 真司	教授	東京大学 大学院理学系研究科
伊藤 聡	コーディネーター	理化学研究所　計算科学研究機構
細野 秀雄	教授	東京工業大学　フロンティア研究機構
曽根 純一	理事	物質・材料研究機構
緒形 俊夫	ステーション長	物質・材料研究機構　中核機能部門　材料情報ステーション
小谷 元子	機構長・教授	東北大学　原子分子材料科学高等研究機構
中川 淳一	上席主幹研究員	新日鐵住金　技術開発本部　先端技術研究所
井手 剛	担当部長・シニアリサーチャー	日本アイ・ビー・エム IBM 東京基礎研究所
鹿野 豊	特任准教授	分子科学研究所 理論・計算分子科学研究領域

オブザーバー（50 音順）

赤木 和人	准教授	東北大学　原子分子材料科学高等研究機構
池田 進	准教授	東北大学　原子分子材料科学高等研究機構
石原 聰	研究所長	ニューメディア総合研究所
井下 猛	特別研究員	物質・材料研究機構　材料情報ステーション
古宇田 光	プロジェクトマネージャー	東京大学　物性研究所
徐 一斌	主幹研究員	物質・材料研究機構　材料情報ステーション
塚本 建次	技術顧問	昭和電工（株）
西川 宜孝	シニアマネージャー	みずほ情報総研
三宅 隆	主任研究員	産業技術総合研究所　ナノシステム研究部門
山崎 政義	特別研究員	物質・材料研究機構　材料情報ステーション
渡邉 聡	教授	東京大学　大学院工学系研究科

関係府省庁

守屋 直文	政策企画調査官	内閣府　政策統括官付（科学技術政策・イノベーション担当）
馬場 大輔	室長補佐	文部科学省　研究振興局　基盤研究課　ナノテクノロジー・材料開発推進室
河村 麻美	係長	文部科学省　研究振興局　基盤研究課　ナノテクノロジー・材料開発推進室
本間 穂高	調査員	文部科学省　研究振興局　基盤研究課　ナノテクノロジー・材料開発推進室

JST

植田 秀史	副センター長	CRDS
田中 一宜	上席フェロー	CRDS ナノテクノロジー・材料ユニット
島津 博基	フェロー	CRDS ナノテクノロジー・材料ユニット
白木澤 佳子	室長	バイオサイエンスデータベースセンター（NBDC）
辻 伸二	主任調査員	科学技術イノベーション企画推進室
富川 弓子	フェロー	CRDS システム科学ユニット
永野 智己	フェロー	CRDS ナノテクノロジー・材料ユニット
中本 信也	フェロー	CRDS ナノテクノロジー・材料ユニット
古川 雅士	研究監	科学技術イノベーション企画推進室　ナノテクノロジー・材料分野
的場 正憲	フェロー	CRDS 電子情報通信ユニット
宮下 哲	主査	戦略研究推進部　グリーンイノベーショングループ
矢倉 信之	フェロー	CRDS　ライフサイエンス・臨床医学ユニット

■ワークショップ企画・報告書編纂メンバー■

田中　一宜	上席フェロー
島津　博基	フェロー
永野　智己	フェロー
中本　信也	フェロー
的場　正憲	フェロー
石原　　聰	特任フェロー　※ 25 年 3 月まで
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
宮下　　哲	主査（戦略研究推進部）

CRDS-FY2013-WR-03

科学技術未来戦略ワークショップ

データを活用した設計型物質・材料研究（マテリアルズ・インフォマティクス）ワークショップ報告書

平成 25 年 8 月

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
電　　話 03-5214-7481
http://jst.go.jp/crds/