

俯瞰ワークショップ報告書

ナノテクノロジー・材料分野 区分別分科会

「プロセスインフォマティクスの俯瞰
～材料合成プロセスへのデータ科学適用の現状と展望～」

2020年2月11日(火)、7月29日(水)開催



エグゼクティブサマリー

本報告書は、国立研究開発法人科学技術振興機構（JST）研究開発戦略センター（CRDS）が令和2年2月11日および7月29日に開催した『俯瞰ワークショップ ナノテクノロジー・材料分野 区分別分科会「プロセスインフォマティクスの俯瞰～材料合成プロセスへのデータ科学適用の現状と展望～」』に関するものである。

CRDSでは、ナノテクノロジー・材料分野の俯瞰報告書を2年毎に発行しており、直近では平成31年3月に「研究開発の俯瞰報告書 ナノテクノロジー・材料分野（2019年）」（CRDS-FY2018-FR-03）を発行した。これに続く2021年版の検討に当たり、前回の俯瞰報告書で欠けていた部分、研究動向の変化・進展のきざしから新たに加えるべき部分を考慮して、「光技術」と「プロセスインフォマティクス」の2つの領域の調査を目的にワークショップを開催し、ナノテクノロジー・材料分野の俯瞰を充実させている。

本ワークショップの主題である「プロセスインフォマティクス」は、必ずしも明確な定義がある用語ではないが、広くは、材料合成プロセス開発にデータ科学の手法を導入し、これまでにできなかった材料の合成や、既存の材料の特性向上を可能にするサイエンス／テクノロジーとして理解されている。

材料探索にデータ科学を適用する試みとして、先行するマテリアルズインフォマティクスにおいては、すでにいくつかの成果が得られており、新材料開発に一定の役割を果たし始めているが、多くの場合、候補となる新材料の組成・構造が予測されているだけで、その材料が「実際に作れるのか」、「どう作るのか」までが示された例はほとんどない。

これは、物性物理学の豊富な知見の蓄積により材料の組成・結晶構造からその物性をかなり正確に予測できるようになっているのに対し、特定の製造プロセスにより材料がどのように合成されていくかに関する予測可能性は非常に限定されていることに起因している。材料の組成・構造を決めるパラメータ空間よりも、製造プロセスを特徴づけるパラメータ空間の方がはるかに広大であることが一つの要因であると思われる。

また、昨今の先端材料に求められる特性には、従来以上に高度な性能や、相反する性能を併せ持つことが求められるようになっており、これまで行われてきた材料合成手段を超えた合成手段、準安定状態や非平衡状態を積極的に利用するような、新たな物質合成スキームを検討する必要がある。このような、新合成プロセスの構築には従来よりもはるかに精緻なプロセスパラメータの制御が必要になると予想されるため、プロセス開発の手法自体を革新していくことが求められている。

プロセスインフォマティクスは、従来の試行錯誤的手法では手に負えないほどの、膨大な数のプロセス探索を実験とデータ科学の融合により加速し、最適プロセス構築を可能にすると期待される。

このような状況を踏まえ、プロセスインフォマティクスの現状を把握し、将来の発展の可能性、解決すべき課題などに関する議論を深める目的でワークショップを開催した。なお、材料合成プロセスは、対象とする材料の種類（有機／無機／金属）や合成環境（気相／液相／固相）などにより、非常に異なったものになるため、共通点が多いと思われる材料の分野ごとに2回に分けてワークショップを開催した。

第1回ワークショップ「無機結晶・多結晶」分野（令和2年2月11日開催）においては、「新材料開発のスマート化の潮流とプロセスインフォマティクスの重要性」「結晶成長モデルベース開発（SiC 溶液法における大口径化を例に）」「結晶成長とインフォマティクス：温故知新」「多結晶インフォマティクスの現在地」「地球科学におけるプロセスインフォマティクス」に関する最新の成果と研究動向が紹介された。

第2回ワークショップ「触媒・有機材料」分野（令和2年7月29日開催）においては、「触媒インフォマティクス：材料組成とプロセス条件を考慮した触媒設計」「ハイスループット実験を基盤としたデータ駆動型材料科学研究」「反応経路のデータベースと機械学習を用いる触媒・発光材料の理解・分子設計」「マイクロフロー合成とプロセスインフォマティクスとの融合と展開」「NEDO 超超プロジェクト概要と高分子 MI に対する取り組み」に関する最新の成果と研究動向が紹介された。

また、第1回、第2回ともに、それぞれ総合討論を実施し、①プロセスインフォマティクスの現状：何ができて何が難しいのか、②プロセスインフォマティクスの将来像：どこまでのことが出来るか、③プロセスインフォマティクス研究開発の課題：技術的課題／制度体制的課題について討論を行った。主な討論ポイントは以下の通りである。

【データ】

- ・産業界の持つデータは開示されないと考えた方がよい。ただし、産学連携は重要。
- ・プロセス内で起きていることのその場計測やシミュレーションとの連携が必要。
- ・装置の経時変化や、実験条件以外の環境因子（室温や外気圧など）も考慮する必要がある。
- ・予見なく網羅的にかつ質の揃ったデータを大量に生み出すという意味で、ハイスループット実験データは重要。

【実験データとシミュレーション】

- ・プロセスインフォマティクスに関しては、最初から実験データを中心にするべき。
- ・実験データ取得には膨大な時間とコストがかかる。シミュレーション技術を上手に活用するべき。

【拠点】

- ・反応タイプごとにプロセス・センターを用意し、そこで実験することでデータも集まるような仕組みが有効ではないか。
- ・合成装置と評価装置と、データを蓄積するサーバ、データ科学者がいる形がよい。

【人材】

- ・マルチディシプリンな人材が必要。マテリアルを理解した上で、機械学習にも取り組むタイプのキャリアパスが現在、日本にはない。
- ・データ科学者にとっては、プロセスインフォマティクスの拠点に入る魅力が少ない。応用統計科学者の方が向いているかもしれない。

【その他】

- ・将来的には、例えばデータ取引のように新たなビジネスモデルの可能性がある。

これらの結果は、CRDS でさらに検討を加えて 2020 年度末に発行予定の「研究開発の俯瞰報告書 ナノテクノロジー・材料分野（2021 年）」に反映させるとともに、今後の戦略提言にも活用していく予定である。

目 次

エグゼクティブサマリー

1. 趣旨説明

曽根純一、眞子隆志、福井弘行（JST-CRDS）	1
--------------------------	---

2. 第1回ワークショップ（無機結晶・多結晶分野）

2.1 新材料開発のスマート化の潮流とプロセスインフォマティクスの重要性 知京豊裕（物質・材料研究機構）	4
2.2 結晶成長モデルベース開発（SiC 溶液成長における大口径化技術を例に） 宇治原徹（名古屋大学）	10
2.3 結晶成長とインフォマティクス：温故知新 柿本浩一（九州大学）	18
2.4 多結晶インフォマティクスの現在地 宇佐美徳隆（名古屋大学）	24
2.5 地球科学におけるプロセスインフォマティクス 桑谷立（海洋研究開発機構）	33
2.6 第1回 総合討論	41

3. 第2回ワークショップ（触媒・有機材料分野）

3.1 触媒インフォマティクス：材料組織とプロセス条件を考慮した触媒設計 高橋啓介（北海道大学）	54
3.2 ハイスループット実験を基盤としたデータ駆動型材料科学研究 谷池俊明（北陸先端科学技術大学院大学）	58
3.3 反応経路のデータベースと機械学習を用いる触媒・発光材料の理解・分子設計 畑中美穂（慶應義塾大学）	64
3.4 マイクロフロー合成とプロセスインフォマティクスとの融合と展開 永木愛一郎（京都大学）	73
3.5 NEDO 超超プロジェクト概要と高分子 MI に対する取り組み 青柳岳司（産業技術総合研究所）	81
3.6 機械学習から見た関心と課題 瀧川一学（理化学研究所，北海道大学）	87
3.7 第2回 総合討論	91

4. 総合討論（まとめ） 99

付録

付録 1. 開催趣旨・プログラム	101
付録 2. 参加者一覧	105

1. 趣旨説明

曽根純一、眞子隆志、福井弘行（JST-CRDS）

近年のグローバル化の動きの中で、国際的地位の低下が懸念される日本にとって、お家芸ともいえる材料技術を世界トップレベルに維持し続けることは、国際産業競争力の見地から特に重要である。このことは、材料そのものの性能が、それが使われるデバイス・システムの機能に事実上の上限を与えてしまうという点からも明らかで、最高性能の材料とそれを生み出す開発技術を有することは、技術立国としての日本の地位向上に大きく貢献することにつながる。

様々な応用分野で材料に求められる性能は、最近になってますます厳しいものとなってきている。2019年にCRDSが発行した戦略プロポーザル「未来材料開拓イニシアチブ～多様な安定相のエンジニアリング～」(CRDS-FY2019-SP-02)の中で述べたように、場合によっては相反しているとさえいえるような複数の物性、例えば、固体電解質における高速Liイオン伝導と広い電位窓、蓄電池正極材料における高Li出入り量と安定性、磁石材料における高保磁力と高飽和磁化、触媒材料における高い触媒活性と安定性・低コストなどを、高いレベルで両立することが求められてきている。上記プロポーザルでは、こうした高いレベルの要求に応えるために、材料の多元化や準安定状態を含む未利用安定相の活用を通して、材料探索の範囲を従来よりも広げ、これまで未開拓だった材料創生空間にアクセスすることを提案した。

しかし、材料創生空間を拡大するということは、今までに使われていなかったような条件で材料合成をするということを意味し、多くの場合、プロセスパラメータを今までより

ワークショップの趣旨

- 2019年版の俯瞰報告書で欠けていた部分、新たに加えるべき部分を補完するために、**光技術、プロセス・インフォマティクスの2つの領域を調査しワークショップを開催**することを計画した。今回のワークショップは**プロセス・インフォマティクス**に関するもので、以下に述べる観点から、議論を行う。
- 「プロセスインフォマティクス」は、材料合成プロセス開発にもデータ科学の手法を導入し、新材料の合成や、既存材料の飛躍的特性向上を可能にするサイエンス/テクノロジーとして期待されている。
- データ科学と材料科学を組み合わせたケモインフォマティクスやマテリアルズインフォマティクスにおいては、すでにいくつかの成果が得られ始めているが、ほとんどが、目的とする物性を持った物質の**組成**や**結晶構造**を見つけ出すものであり、その材料を**いかにして作るか**の指針を与える例はあまりない。
- プロセスインフォマティクスの**現状**と明らかになっている**課題**を把握するとともに、同研究分野の**発展加速**に必要な事柄を抽出する。

Copyright © 2018 CRDS All Rights Reserved.

図 1.1 ワークショップの趣旨

増やしたり、より精密に制御したりすることが必要となる。結晶構造・組成と言った材料パラメータ空間は、それ自体非常に広大なものであるが、このプロセスパラメータ空間は、それよりもさらに広大であることが容易に理解され、これまでのような試行錯誤や勘と経験に頼るやり方で、その空間内から求めるパラメータを探索することは著しく困難であることもまた容易に想像される。最適なプロセスを探索するためのサイエンス／テクノロジーを構築していくことが強く求められている。

一方で、近年、材料科学とデータ科学を組み合わせたマテリアルズインフォマティクスが、新材料探索や材料機能向上に効果を発揮する例が現れ始めている。これらの中には、実際に新規材料・高機能材料が発見されたり、新しい材料設計指針が見出されたりするだけではなく、これまであまり交流することがなかった材料系研究者とデータ科学者の共創による、材料科学の新たな潮流が生まれた効果も含まれている。

現状では、マテリアルズインフォマティクス、特に、第一原理計算を併用する手法を通じて発見された新材料・新物質については、多くの場合、候補となる新材料の組成・構造が予測されているだけで、その材料が「実際に作れるのか」、「どう作るのか」、まが示された例はほとんどない。これは、物性物理学の豊富な知見の蓄積により、材料の組成・結晶構造からその物性をかなり正確に予測できるようになっているのに対し、特定の製造プロセスにより材料がどのように合成されていくかに関する予測可能性は非常に限定されていることに起因している。データ科学をプロセス発見に繋げるためには、マテリアルズインフォマティクスで確立された手法をさらに高度なものに発展させていく必要があると思われる。

このワークショップでは、材料合成プロセスの発見にデータ科学を取り入れるプロセスインフォマティクスについて、材料合成の専門家の先生方に話題を提供いただき、その後プロセスインフォマティクスの現状、将来像、研究開発の課題について総合的に討論する。



総合討論について

話題提供終了後、下記の観点から議論を行う。

- プロセスインフォマティクスの現状
 - 期待と実力の差異はどのくらいあるのか
 - 現状の実用性(ほかの手段と比べての)
- プロセスインフォマティクスの将来像
 - 適用先, 予測性能, 有用性はどこまで伸びるか
 - 他の領域(他の材料, 他の学問領域)への拡張は可能か
- プロセスインフォマティクス研究開発の課題
 - 進展のための技術上の課題は何か
 - 制度・施策上の課題はあるか(規制, ファンド, 拠点…)

Copyright © 2018 CRDS All Rights Reserved.

図 1.2 総合討論の内容

材料合成プロセスは、対象とする材料の種類（有機／無機／金属）や合成環境（気相／液相／固相）などにより、非常に異なったものになるため、あまりに異なる分野同士の議論を同時に進めると、議論が全く噛合わなくなる可能性がある。そこで、ワークショップを2回に分け、初回に無機半導体の単結晶／多結晶成長プロセスについて、第2回に有機材料合成・触媒プロセスを取り扱うこととした。分野の異なる材料の合成プロセスを俯瞰することで、プロセスインフォマティクスの全体像を把握したい。

2. 第1回ワークショップ（無機結晶・多結晶分野）

2.1 新材料開発のスマート化の潮流とプロセスインフォマティクスの重要性

知京豊裕（物質・材料研究機構）

本日は「新材料開発のスマート化の潮流とプロセスインフォマティクスの重要性」というタイトルで、本分野の動向に関して私の認識を中心にお話したい。

今、マテリアルズインフォマティクスは様々な場面で注目を集めている。2019年2月には、人工知能（AI）を用いることで職人技よりも効率的に材料合成が行われたという新聞記事が掲載されていたように、マテリアルズインフォマティクスの有用性を示す成果が出始めている。

マテリアルズインフォマティクスという言葉が生まれた当初はどうやって使っていくのかというところから始まり、いくつかの有用な例が出てきて、最近では少なくとも材料のスクリーニングには有効であることが認識されている。

ご存知の通り、NIMSでは2015年からMPIというJSTの支援による拠点事業が運営されてきたが、その中でもマテリアルズインフォマティクスの次にはどのような概念が重要になるだろうと考えてきていた。2015年に福岡で開催された国際ナノテク会議（INC11）において、欧州の発表の中で「バリューチェーンを作るんだ」という話題があった。サプライチェーンではなくバリューチェーンであることについて、当時はよくわかっていなかった。その後、欧州 Horizon 2020 の中で NOMAD と呼ばれるマテリアルズインフォマティクスのためのデータを集めるためのプロジェクトが存在していることを知り、バリューチェーンの中にマテリアルズインフォマティクスも含まれていることに気づいた。つまり、欧州の中で材料から製品までを一つのコンセプトで見ようとする動きであった。結果的に NOMAD は米国の Materials Project の中に含まれてしまうことになるが、欧州は2015年の段階で既に欧州圏内のバリューチェーンを構築しようとしていたことが強く印象に残っている。

同じ時期にもう一つ印象的であったものとして「MAKERS」という本がある。2012年に発刊されたこの本の中には、小規模でありながらグローバル、職人肌でありながら革新的、ハイテクでありながら低コストといった「MAKERS」と呼ばれる小さなファウンドリーのような組織が将来的には実現していくことが書かれていた。

具体的な例としては、当時話題になっていた3Dプリンティングのことであるが、AIやオープンハードウェアや自動化やクラウドマニュファクチャリングといったものを用いた新しいものづくりが行われるようになると書かれている。AIと言っているが、コンピュータを用いて材料設計をすることも書かれており、マテリアルズインフォマティクスで材料を作ることと等価である。

このような流れは既に世界中で主流になりつつある。1つ目の例として、データ科学とロボティクスを融合した材料開発の自動化が挙げられる。例えば、かつて Clean Energy Project の中で分子材料を自動計算科学で100万件くらいのデータを作った Aspuru-Guzik 氏が、最近ではものづくりをやろうとしている。また、欧州では伝統的に強い触媒開発に

対して、ワークフローを自動設計して実際に合成してみようという動きもある。同様のコンセプトを無機材料に適用しようとしているのが東京工業大学の一杉先生である。

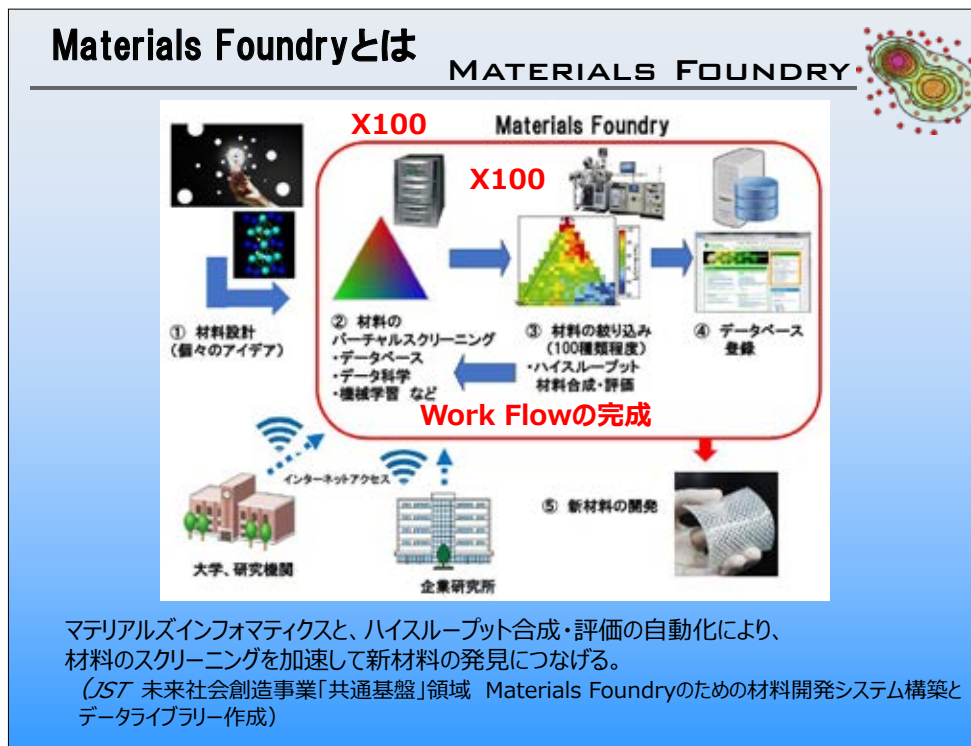


図 2.1.1 マテリアルファウンダリー

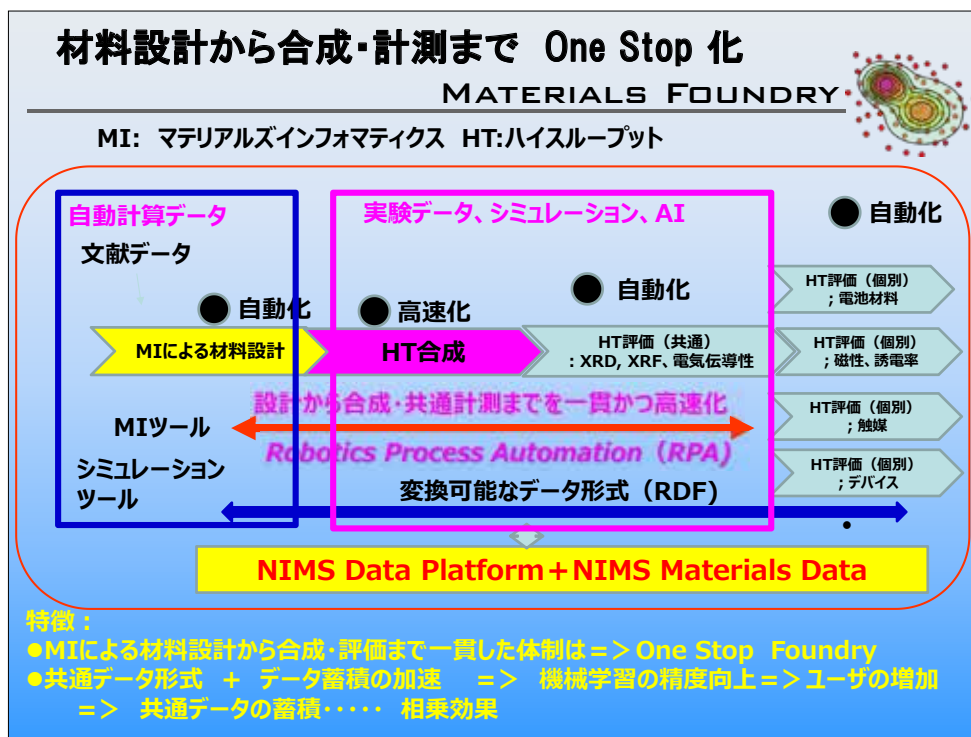


図 2.1.2 材料設計から合成・計測まで One Stop 化

一方、我々は JST 未来社会創造事業の中でマテリアルファウンダリーというものを提案している（図 2.1.1）。これは材料設計から合成までを自動化して、必要な材料のデータを自動的に集める体制ができないだろうかということで、フィージビリティスタディを進めているところである。その中心にはマテリアルズインフォマティクスがある。もちろんプロセスも重要なのでプロセスインフォマティクスも深く関係している。

NIMS 等で開発してきたシミュレーションツールや文献データなどを用いたマテリアルズインフォマティクスによる材料設計に加えて、自動化や高速化も含めてマテリアルファウンダリーを作ろうとしている（図 2.1.2）。その中で最近特に話題になっているのはロボティクスプロセスオートメーション（RPA）という考え方である。マテリアルズインフォマティクスとプロセスインフォマティクスを用いた事例として、今日は、我々のターゲットの一つであるハイエントロピー合金、いわゆる多元系金属材料で新しい磁性材料を見つけることができるかという挑戦について紹介する。

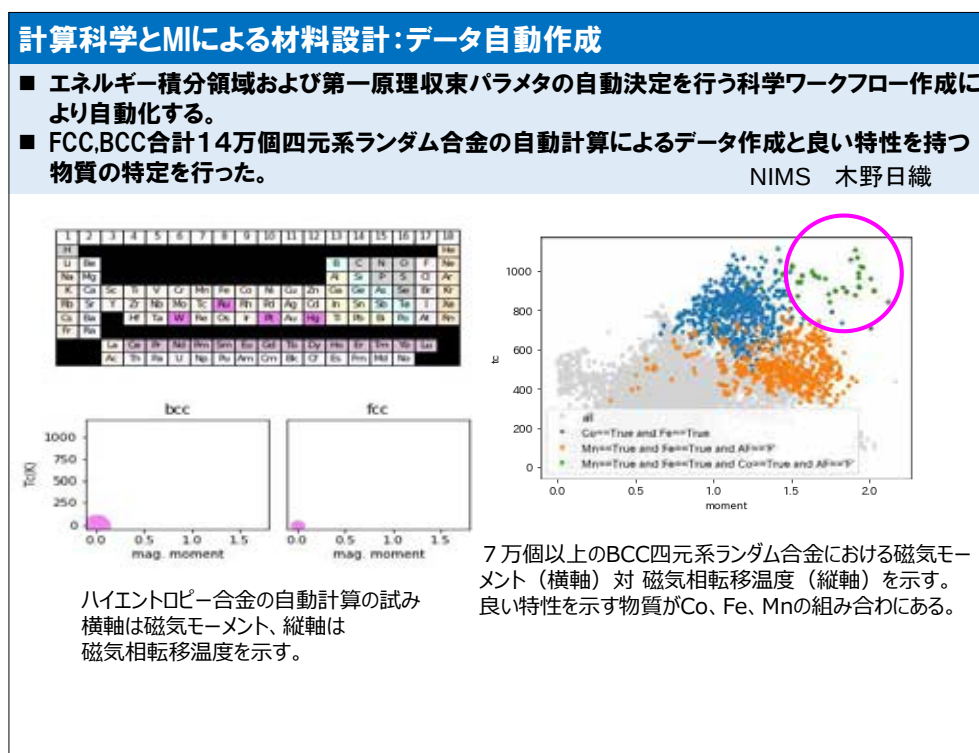


図 2.1.3 計算科学とMIによる材料設計

まず、自動計算に関しては、第一原理計算で一個一個計算していると時間がかかるため、第一原理計算の一つ KKR 法に基づいて 4 つの金属元素をあるアルゴリズムによって選択し、そのときの例えば温度や磁気モーメントを自動計算してデータを蓄積する。その結果を図 2.1.3 右側の図になるが、丸で囲った部分が磁気モーメントも大きく転移温度も高い材料である可能性が高いことがわかる。次に実際にその候補材料を合成してみようとなるが、合成については、かつて NIMS と東工大で合同開発したコンビナトリアル手法を用いて、さらに最近ではかなり自動化も進んでおり、例えば二元系であろうが三元系だろうがある程度は自動的に合成できるようになっている。

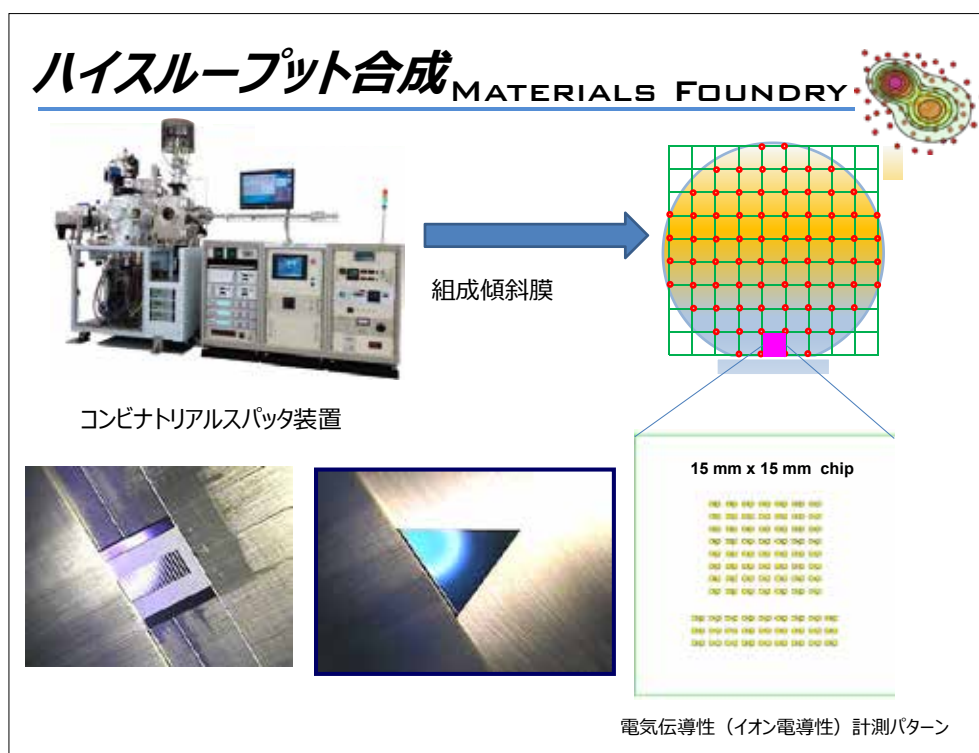


図2.1.4 ハイスループット合成

例えば薄膜合成では（図 2.1.4）、ある条件で組成が決まれば、1日あれば三元相図に相当する量の合成が自動で行えるレベルになっている。さらに、少し前までは単に組成傾斜だけを作っていたが、最近ではパターンを作った基板の上に薄膜を作製し、例えば各領域で数十個の電気伝導性や、あるいは電極をつけた上に製膜したものにさらに電極をつけてといった処理を行った後に電気測定や様々な特性測定を自動で行っている。その際、何℃で作るかなど、どういう条件にするかが問題である。これまでは例えば金属だったら 300℃ぐらいで酸素が抜けて比較的純度の高いものができるといった経験則に基づく部分が多かった。最近ではベイズ最適化を用いることで、いくつかの実験条件で作ったものを X 線で観測し、目的とする構造ができているかを確認した後、その結果をフィードバックすることでより最適な成長条件を見つけることが可能になっている。さらに、高速電子線回折や Phase Filed 法などのシミュレーションを統合して表面構造を予測することで、今現在できているバルクがどういうものか、目的とするものができているかどうかをある程度類推することも可能になっている。いわゆる機械学習を用いることでベイズ推定に要する時間をもっと短くして最適な成長条件を探索しようと試みている。

次にプロセスインフォマティクスの例として、薄膜成長で一般的に使われるスパッタ製膜に対して最近の我々の研究成果を紹介する。

スパッタ製膜は、いわゆる半導体プロセスをはじめ様々なところで使われる汎用性の高い製膜装置だが、そのプロセスは非常に複雑である。例えば、エッチングにプラズマを用いたり、ターゲットに関しても単結晶か化合物かによって異なっていたりする。ターゲットの組成がそのまま膜に反映されないため、経験的にターゲットの組成を操作して狙った組成になるよう合わせ込む必要がある。そのような作業を都度都度やっていると時間がか

かってしまうため、我々のグループではプラズマを使ったときにどれだけ材料がエッチングされるかを、機械学習を用いて予測する手法を考えた。スパッタではよく使われる山村・俵の式の中の重要なファクター（生成熱や質量など）を理論駆動記述子として、それに実際の実験データの中から抽出されたファクターを合わせたものをデータ駆動記述子と比較したところ、比較的よい一致を示した。

この記述子を用いて、スパッタリング効率を支配する記述子が何かを調べたところ、ターゲット材料の元素質量などいくつかの支配的な記述子を見つけることができた。つまり、様々なファクターがある中で、主となる記述子を特定することで、スパッタリング効率の大半が説明できるのである。これによって、スパッタ法のシミュレーションがほぼリアルタイムで行える可能性が出てきている。

今回の成果はスパッタリング効率のみに関する記述子を特定したに過ぎないが、この知見を活かしてプロセスをリアルタイムでシミュレートする、いわゆるデジタルツインが可能になるのではないかと考えている。実際にチャンバーの中で起こっていることを、リアルタイムにシミュレーションすることが可能になれば、ある条件を変えたときに膜質がどう変化するかをある程度予測することもできるようになるだろう。

次に計測の効率化を目指したマテリアルシーケンサーというものについて紹介する。

ここでは、ある試料に対して、試料そのものの画像データや、電気特性などの様々な特性を自動計測し、その結果を WiFi 機能付の SD カードに全て集めて、データを蓄積していく。その際、あとで機械学習に応用できるよう RDF というデータ形式にしておく。つまり、製膜のときと同様、合成から評価までを1つのデータ形式にしている。

SD カードを使っている理由は、計測装置が一般的にインターネットに繋がっていないからである。なぜなら OS をアップデートすると計測ソフトが動かなくなるからである。また、これまではデータを外に取り出すときに移動式ハードディスクを差し込んでデータを蓄積してきたが、最近の SD カードには WiFi 機能がついているため、OS にはタッチせずに、無線 LAN を用いてデータを蓄積することが可能になっている。このような装置を NIMS の中で試験しており、プロセスインフォマティクスに必要となる情報を集めている。

一方で、プロセスインフォマティクスを確立するにはまだまだたくさんのハードルがあるように感じている。

例えば、スパッタの例に限っても、ターゲットや雰囲気などいろいろな条件を最適化することが必要になる。スパッタリング効率に関しても、雰囲気によって、例えば反応性スパッタなのか単なる物理的なスパッタなのか、あるいは分圧によって膜の応力も変わる。基板によって極性や方位も異なり、表面の清浄化プロセスなど様々な条件に依存している。さらに、膜厚が変わればそれによって応力緩和も起こるため、全く同じものを作るわけではない。

さらに、いわゆる My Condition 問題もある。私の実験装置とあなたの実験装置の温度が全く同じとは限らないし、その他の条件も同じとは限らない。したがって、プロセスインフォマティクスは1つの実験装置、1つの場所でやる限りにおいては有効な手法になるが、他の人と共有しようとする途端に混乱が生じる。この問題をどうやって解決するかというのが今後の大きな課題になるのではないか。

それを解決する一つの方法として、拠点を作って、その中で様々なプロセスデータを蓄積する、いわゆるデファクトスタンダードを作っていく試みが必要ではないか。もし他の拠点と共有するのであれば、データとデータをつなぐための考え方やツールが必要になる。

その一例として、イギリスにあるマテリアルイノベーションファクトリーは、プロセスの自動化を図って、材料開発の拠点になろうとしている。また、半導体分野で有名な IMEC も最近では電池開発などに進出してきており、拠点化を目指している。拠点化することで少なくともその中のプロセス条件は統一感を出せる。

最後にまとめると、マテリアルズインフォマティクスに続く潮流として、自動計算や仮想スクリーニングのためのワークフローが重要になる。ハイスループット合成条件の探索や計測スペクトルの自動解析などが必要になり、それらを集めることによってプロセスインフォマティクスが確立されていくのではないかと考えている。

さらに、材料合成のためのデータ空間あるいはプロセスのデータ空間をどのように効率的に埋めるかも大きな課題である。そこには機械学習が必要になる。

【質疑応答】

Q:このような新しい取組はどんどん進めていくべきであると考える一方で、昔からの手法もある程度は残していく必要もあるように思うが、それらの割合はどうすべきか、ご意見を伺いたい。

A:感覚的なものになるが、3割くらいが先進的なことをやればいいのかではないか。例えば MPI では、データ科学者でありながら実験科学者とある程度議論ができるような人材を育成しようとしているが、そのような人たちが先駆者となってまずは有効であることを示していくことが重要ではないか。全員が同じ方向を向く必要はないが、新しい人材が中心になってすそ野を広げていく印象を持っているため正規分布のようなバランスが取ればいいと感じている。3割くらいの人材がマテリアルズインフォマティクスを使ったり、ハイスループット実験をしたり、自動計算を使って「このようにやると効率が上がるよ」ということを見せていく必要がある。ただ、そのときに重要なことはデータをどのように共有するかである。マテリアルズインフォマティクスのときには計算科学をベースにすることができたが、プロセスインフォマティクスでは My Condition 問題があるため、別々のところで同じようなデータがあったときにそれを同じものと判断できるかどうかは難しい問題である。

Q:例えば、多元系材料で太陽電池を作る際には、温度は一定ではなく、途中で上げたり下げたりすることで特性が上がるということがよくあるが、そのようなものも自動化が可能なのか。

A:今の装置は全てコンピュータコントロールになっているため、途中の温度設定も制御可能である。したがって、ここまでは何℃にして、ここからは何℃にする、という設定もでき、全て記録として残せるため、データとしても残すことは可能である。ただし、前工程と後工程によって特性が変わったときに一体何が起きているのかを理解するためには何らかの手法でフィードバックをかける必要があると思う。

2.2 結晶成長モデルベース開発（SiC 溶液成長における大口径化技術を例に）

宇治原徹（名古屋大学）

ここでは SiC の溶液法による結晶成長に関して述べる。ただし、最近では SiC に限らず、他の結晶の溶液成長や気相成長にも拡張している。

普通、SiC というのは昇華法、いわゆる気相成長で作られる。それに対して我々は長い間、溶液法をやってきた。一般的に平衡条件に近いプロセスほど品質の良いものができるので、溶液法の方が高品質と言われる。もう少し詳しく言うと、昇華法は下で原料を昇華させ上で再結晶化する方法であり、上下の温度差が結晶成長の駆動力になるため、必ず温度差をつけることになる。一方、溶液法で必要なのは濃度の差、すなわち過飽和度で、温度差は全体で 10℃ 以内くらいに抑えられている。昇華法は 100℃ から 50℃ ぐらいの温度差があり、それが原因で内部に熱歪みが生じ欠陥ができるが、我々は溶液法でそれを防いでいる。

昇華法で作られた市販結晶と、我々が溶液法で作成した結晶の X 線トポグラフィ像を図 2.2.1 に示す。影になっているのが転位で、市販の結晶には多くの転位が入っていることが分かる。それに対して我々のものは転位を探すのが難しいくらいにきれいな結晶であり、数字で表しても 2 桁少ない。

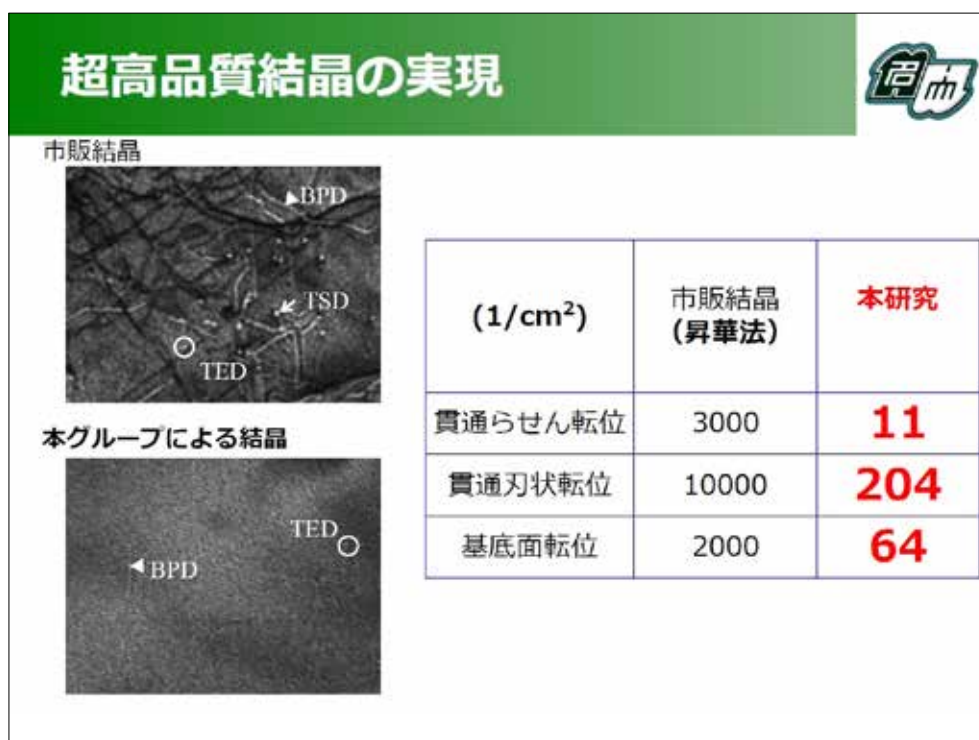


図 2.2.1 超高品質結晶の実現

このような結晶を実現するまでに実は相当な苦労があったが、以下ではインフォマティクスの話が中心なので、苦労の部分は割愛する。我々はきれいな結晶ができることを 2016 年～2017 年頃に発表したが、当時は 10mm × 10mm や 15mm × 15mm といった小さ

な結晶で、企業の人に見せると何インチができるのかと聞かれた。いくつかの会社がこの方法でチャレンジしてくれたが結局ギブアップしてしまい、これは我々が自ら作らなければならないと考えた。それが機械学習、インフォマティクスに我々が取り組むことになった一番のきっかけになっている。シリコンを例にすると、ネッキングという方法で無転位の結晶が作られたのは1958年頃。現在の300mmの大きさの結晶ができたのは、正確な記録は見られないが1995年～1996年頃と言われており、この間の40年弱が大口径化にかかった時間である。それに対して、SiCは現在すでに昇華法による6インチの結晶が市販されており、今は8インチの開発が始まっている段階にある。溶液法による10mm×10mmの結晶を8インチにするまで、40年もかけていられないので、我々は機械学習に取り組むことにした。

結晶成長は2,000℃程度でおこなうが、結晶成長装置の中は見るのが難しい(図2.2.2)。見た瞬間に温度が下がってしまうということもあるが、仮に見ることができたとしても見えるのはごく一部分だけで、ほぼブラックボックスになっている。そこで、シミュレーションをおこなうが、いくら簡単にしても数時間かかり、右図のように3次元でやろうとすると1日かかる。したがってシミュレーションで最適化するのには難しい。

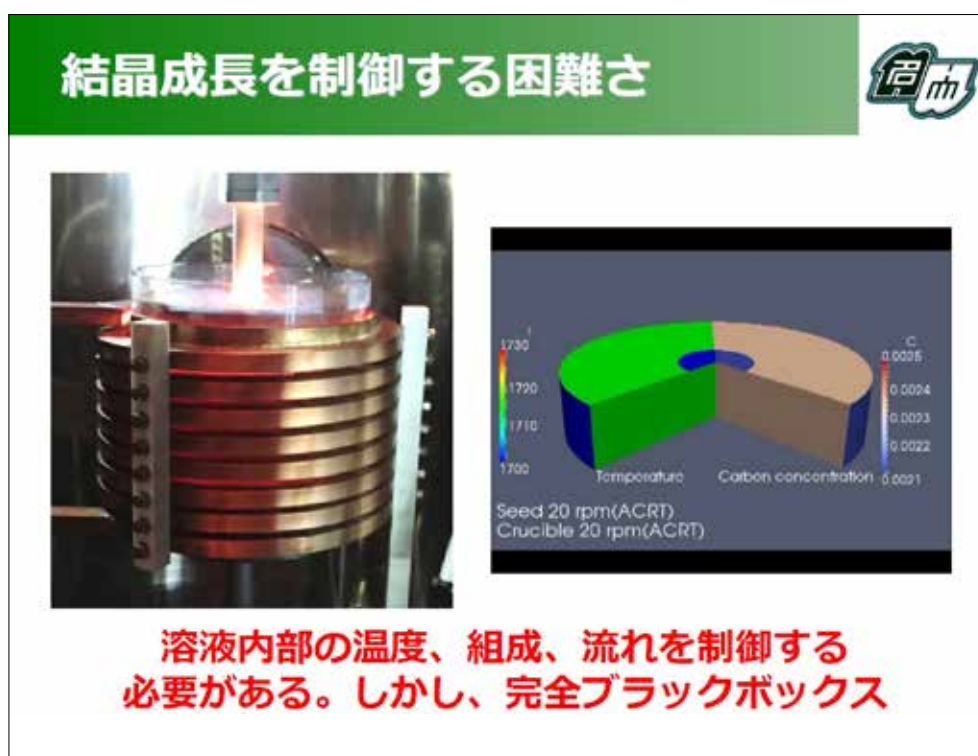


図 2.2.2 結晶成長を制御する困難さ

また、パラメータの数もかなり多い(図2.2.3)。るつぼの回転や温度のような数値を変えれば調整できるものだけではなく、るつぼの形状、断熱材の厚みなど、装置の設計に関わる場所までパラメータに入れる必要がある。これらは一度作ったら二度と変えたくないところなので、そこまでパラメータに含めるとなると途方に暮れてしまう。

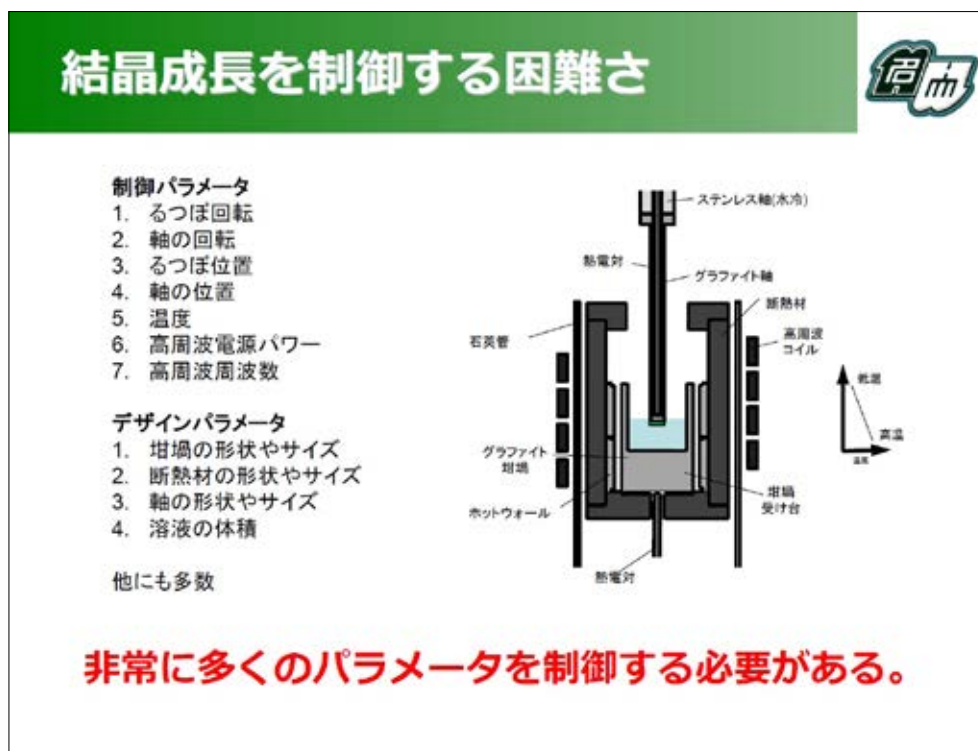


図 2.2.3 結晶成長を制御する困難さ

我々の夢は、自動車のようなモデルベース開発を結晶成長でやることである。モデルベース開発というのは、開発設計の半分をサイバー空間で、残り半分をフィジカル空間でおこない、両者を行き来しながら PDCA サイクルを回す。これがなぜ結晶成長でできないかというと、適当なモデルがないことが大きい。シミュレーションは厳密なモデルではあるが、スピードの遅さが問題になる。正確で早いモデルが必要であり、それを我々は機械学習で構築しようとしている。サイバー空間で実際とほぼ同じものをつくり、デジタルツインの中でモデリングして、最適化して実装することを目指している。

結論を言うと、我々はかなり良いモデルをつくっている。図 2.2.4 は、我々が実際に使っている装置の内部の構造を表したコンピュータグラフィックス (CG) である。結構リアルに表現しているが、見ていただきたいのは中央の虹色のところで、ここは $2,000^{\circ}\text{C}$ 近い溶液の内部の温度分布を表している。るつぼが動くとその温度分布もどんどん変わっていく。シミュレーションではこの 1 か所を計算するだけでも 1 時間や 2 時間かかるので、予め計算しておいて格好良い CG を作ったのかと言われてしまったりするが、そうではない。その場で読み取ったログデータから予測して、その場で CG に貼り付けて表示している。一種の擬似的なその場観察のようなことをやっているが、実際にはコンピュータの中に良いモデルをつくったということになる。後は、この中で最適化ができれば良い。

このモデルをどうやって作るかというと、まずシミュレーションで 150 ～ 500 個くらいのデータを作る。1 個あたり 3 時間くらいかかるため 500 個のデータを作ろうとすると 1,500 時間になるが、我々は 10 件程度のシミュレーションのプロセスを並行で動かして、なるべく 1 週間～ 10 日以内にデータを作る体制にしている。



図 2.2.4 高速予測モデルの構築

例えば、入力パラメータに対して、流れや過飽和度を予測するニューラルネットワークを作った。色々工夫したが、るつぼの形状を変えても1つのニューラルネットワークで予測できるようにしたことが一番の工夫である。シミュレーションで1時間程度かかるところ、ニューラルネットワークに入れると0.0003秒、ほぼ一瞬で予測ができる。(図 2.2.5)

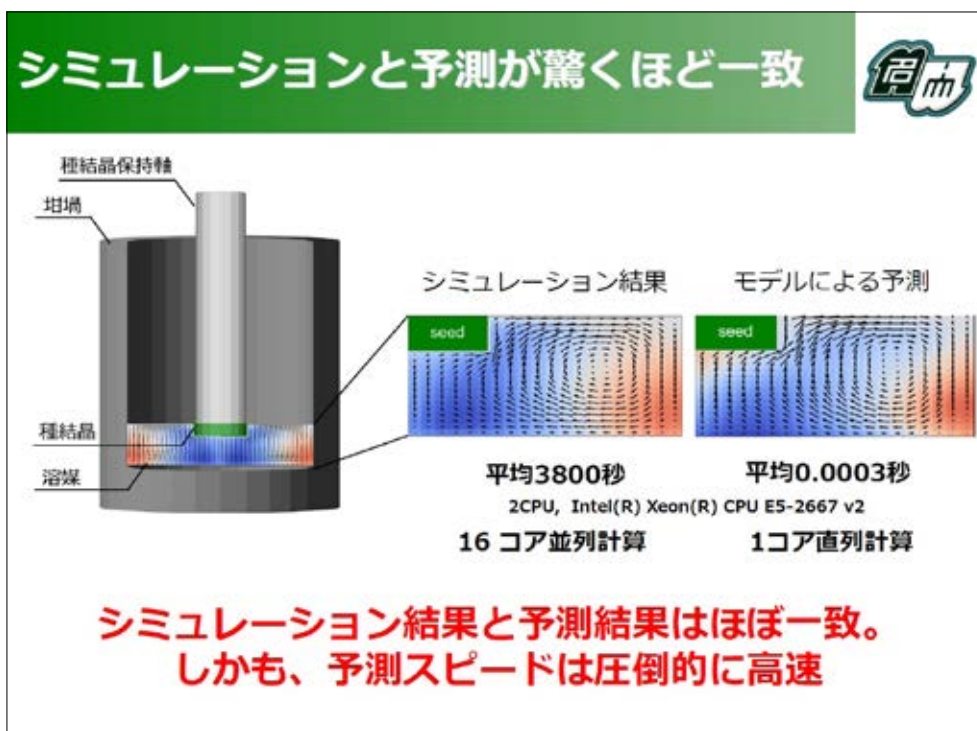


図 2.2.5 シミュレーションとモデルによる予測の一致

なお、これは高速で予測できることに意味があるのではなく、本当に意味があるのは微分可能な関数で表せたという点である。ニューラルネットワークで表現するというのは、複雑なシミュレーションを1個の複雑な関数で表すのとはほぼ同じで、そこに意味がある。それが一旦できると最適化が数学的にでき、遺伝的アルゴリズムを使うこともできる。例えば、液体の中の組成分布や流れ分布、温度分布を関数で表して、それを満たすような条件を探す。残念ながら今のところ流れや温度分布と結晶品質とを結びつける「良い結晶」という目的関数は書けていないが、まずは10mm × 10mmの結晶についての我々の重要な知見から、それを6インチにしたらどうなるか、12インチにしたらどうなるか、という風に進めている。我々の目的関数はここでは示さないが、目的関数を作るところには非常にタフな研究が必要で、そこには我々の10年の苦労の全てが含まれている。1回そのタフな研究をして成功すれば、その研究者自身が応用や拡張をする上でこのツールは非常に有効である。

この方法によって我々は2～3回の実験で3インチができるようになった(図2.2.6)。10mm × 10mmについて論文発表した翌年の2018年に3インチ、さらに、2019年には6インチができた。10mm × 10mmで実現した条件を6インチ全体で充たすように最適化を図ったが、それには1つの同じ条件で結晶成長するのではなく、シーケンスを途中で何度も変えること、全部で5つくらいのシーケンスの組合せが必要だった。このような条件を人間が見つめるのは不可能である。実は、6インチは2019年9月頃に1回できたが、品質が悪かったため中の断熱材、るつぼ、コイルのサイズなど、構成を全部やり直した。発注して組み直し、2回目で上手くいったが、このようなダイナミックなことをやっても我々が意図した通りのものができるようになった点がすごいところである。



図 2.2.6 3インチ超の結晶成長

もう一つ、この方法の仕掛けを説明する。最適化してもモデルが間違っていたら何の意味もない。間違える可能性があるとしたらどこかというところ、1つはモデルそのものだが、これは我々の場合はほぼないと考えている。良くあるのは物性値がそもそも違っていること。これは、2,000℃の溶液成長をやっているのだから大いにあり得る。そこで実際の実験結果とモデルで予測したものを比較して、どのような物性値を入れると実験に合うかという予測を機械学習でおこなっている。例として結晶成長速度を比較し、融液の物性値の予測をおこなったものを示す（図 2.2.7）。10 種類くらいの物性値を示したが、実験結果に合う値の組合せは無数にあり、それを 1 万通りくらいヒストグラムで表示してみると、ほとんどのものは平らな形になる。つまり、どの数字を入れても大体実験結果を再現するということになってしまうが、この中でも、例えば結晶と溶液界面の反応速度定数にはピークが見られる。しかも、この値は実験ではなかなか求められない。こういうものを 1 つ決めてどんどんやっていくと、他にもピークが出始め、全体として何かしらの推定値というものが出てくる。この推定値は文献値と比較すると合っているものもあれば間違っているものもあるが、実験で測定しにくい値が分かるということには、非常に大きな意味がある。

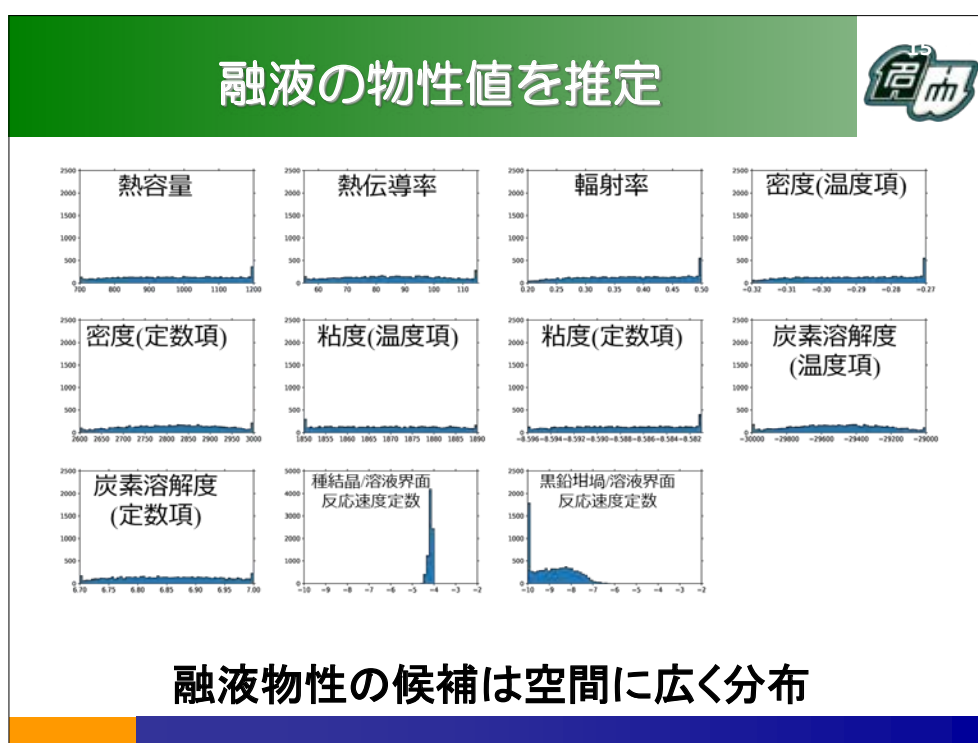


図 2.2.7 機械学習による物性値の推定

また、これにはもう一つ意味がある。ここでやっていることは、モデルが実際の実験と合うように物性値をパラメータに使うだけなので、正しいかどうかは検証しなければならない。ただ、これはまさに一つ前の講演で言われていた My Condition で、今、目の前にあるその装置にぴったり合ったモデルをつくることにはほかならない。なぜ我々が 2～3 回だけの実験で大口径化に成功できたかというと、自分たちのモデルを自分の装置にぴたりと合うように作り直したからである。

さらに、この方法の良いところは、一旦ニューラルネットワークで関数を決めてしまえ

ば、あとは実験データが1つか2つあればそこにぴったり合わせることでパラメータが自動的に決められこと。つまり、たった2～3回の実験であったとしても、現実のものを表すモデルが作れることを意味している。我々実験屋からしてみると、実験でビッグデータを取るのは大変なので、これは嬉しい。

最後に、最近の研究について紹介すると、液面の一部の温度分布から全体の温度分布の予測をおこなっている。また、シミュレーションと機械学習を併せた方法で、気相成長、例えば窒化ガリウムの結晶成長やCVDを扱っている。イオン注入などの半導体プロセスのモデル化や、熱回路の最適化にも取り組んでいる。さらに、ベイズ最適化を使って少ない実験データからの予測というのもやっている。例えば、シリコンのCVDでハイスループットを実現したり、半導体の基板結晶の品質を維持したまま5倍のスピードで加工すること、気相からの物質合成のスループットを2倍にすることなどに成功している。

【質疑応答】

Q: 6インチのSiC結晶作成については、るつぼ等の色々なものを作り変えて、2回目に非常に高品質で大きな結晶ができたという話だった。1回目の時には、るつぼなどの条件が機械学習の結果が出した最適解と異なっていたので、それに合わせて作り直したから2回目は上手くいったということか。

A: その通りである。実はそれまで4インチの実験をずっとやっていたので、4インチ用のまま、るつぼのサイズ等を変えずに無理やり6インチを作った。それなりに良いものはできたが、まだ不十分ということで、6インチ用にシミュレーションをやり直して全部作り直した。

Q: CVDに関心があるが、例えばALDのようにガスフローと表面反応で物がなくなっていく、あるいは脱離していくという場合にも、モデルである程度シミュレーションして機械学習にかけるということは可能か。

A: 白石教授（名古屋大学）による窒化ガリウムの気相反応、表面反応のシミュレーションがあり、一度それを使って試してみたところ、結構良い近似のことができることが分かった。恐らく反応定数などは間違っていると思うが、実際の実験と合わせれば反応定数がどれだけずれているかも分かるので、それで問題ないと思われる。

Q: 予測モデルを用いて最適条件を見出すとあったが、シーケンス、昇温温度や何℃で保持するかなどの組合せは無限にあると思われる。何百通りか何千通りか分からないが、それらの中から可能だと思われる実験条件を振って見つけたという理解で良いか。

A: 完全な非定常の計算というのは、ここではやっていない。都度、定常が成立すると仮定して、少し時間が変化したときのちょっとした変化、例えば、少し成長するとか、変なところからるつぼが融け始めるとか、そうしたものが分かるので、そのときどきの最適解というのを例えば10分おきぐらいに出す、そういう方法でやっている。

Q: 結晶成長の専門家というのは、何十回もやった中から経験的に良いものができる条件を見つけ出していると聞いた。プロセスインフォマティクスを使うとそれよりも広い範囲がシミュレーションやモデルでできて、より良いものができるという方向性であると思われるが、今回の話はその手前の段階ということか。

A: 物によると思うが、我々が扱っている系はそんなに反応が速くなく、全部定常で考えてもそれなりに合う。現実には産業で使うなら、この程度できれば十分と考えられる。パラメータ 10 個の組合せの最適値を見つけるのは人間ではほぼ不可能であるので、かなり役に立つと思う。

Q: SiC をシリコンの上につけたいというニーズがあると思うが、溶液成長法で多形の制御もできるか。

A: 多形の制御は全く問題なくできる。今は 4H-SiC を目的にしている、全部 4H-SiC になっている。

Q: 6H-SiC も可能か。

A: 少し条件を変える必要はあるが、6H-SiC が成長しやすい条件はだいたい分かっているので、やろうと思えばやれる。

Q: 非常に競争の激しい領域でベンチャーを作られたが、実用化に向けた技術的課題はどこか。

A: 最大の問題は、長時間にわたって安定に、かつ、速い成長速度で作るということ。長時間・安定の方は、途中でモニタリングをすとか、長時間のシミュレーション、変化のシミュレーションができるようになってきたので、大分クリアになってきた。一方で、成長速度を上げるところは、今回話した中にはまだヒントがない。溶媒の組成をいじるといったインフォマティクスのようなことをやる必要があり、そこが課題として残っている。とはいえ、昇華法に比べて半分ぐらいのスピードまで来ているので、そんなに非現実的な話ではない。

Q: 今の 6 インチでグレードが昇華法よりも圧倒的に良いということであれば、成長速度が半分でもそんなにデメリットにならないと思われるが。

A: どれだけきれいでも世の中からはコストと言われてしまう。ただ、種結晶で欲しいという要望も、いくつか受けている。また、我々は既に 8 インチを作るための炉を設計・発注しようとしているが、シリコンの専門家から、シリコンは 450mm ができたからこそ 300mm が安定にできるようになった、8 インチを作るなら 12 インチを目指すべきだと言われた。そこで、我々のヒーターは 12 インチが成長できるように設計してある。もし 12 インチができれば世界初である。

2.3 結晶成長とインフォマティクス：温故知新

柿本浩一（九州大学）

本日お話しする内容の1番目は、ある実験結果は他の実験結果と整合性があるのか。その点を考える際の解決策は、複数物理量を予測することである、ということについて。2番目は、逆問題について。3番目は、結晶成長というのは、非線形プロセスであり、必ずバイファケーションの落とし穴がある。要はローカルミニマムに陥る可能性があり、いかに回避するか、について。4番目は、実験と数値計算の協力連携が非常に重要、ということについて。数値計算は定量予測が必須である。5番目は、高温物性の予測について。私のグループでは、今日紹介する数値計算は全て自分で計算コードを書いた。25年かけてつくり上げた。実験装置も全部改造した。私の研究所では、買ってきたものをそのまま使うなど言っている。

まず、複数のデータの予測。太陽電池をつくるときの凝固の仕方については、必ず3次元で時間依存を計算する。冷やしたときにどんな結晶ができるか、温度だけではなく欠陥の種類とか、例えばこういう結晶をつくったときに、室温に下ろしてスライスしたときにどういう転位が分布するか、を定量的に予測した。X線で観測したものと比較すると、おおそ一致していることが分かる。

転位密度の次は、もう一個の物理量、残留歪みについて。実験では、光の偏波を使って結晶内の残留応力を測った。実験と計算結果と比較すると、合っているといえ合っている、合っていないといえ合っていない。

3つ目の物理量、不純物濃度について、冷却している間でも拡散が起こり、凝固界面に不純物の鉄が分布する。このような不純物があるところは太陽電池の変換効率が悪い。数値計算と、実験で得られたライフタイム、変換効率を比較すると、大体合っている。この数値計算では、凝固のときに体積が10%増加するという点を入れていない。

このように、様々な実験をやって非常にきれいな凝固、シリコン本来の凝固した結晶ができて変換効率も上がる。これはNIMSとの共同研究である。

2番目は、逆問題。必要なものだけやりましょう、という話について。温度をプログラムして「こういう温度にしたい」と、シリコン結晶下部の温度を設定する。それに見合うように、逆問題として、数値計算で2つのヒーターの温度を予測することで、デザインした温度変化が作り出せる、と考える。ところが、数値計算は、デザインした温度変化と後半部分が合わない。これは実験でよく経験する、温調はどんどん下がるが、実際の温度は下がり切っていない。これは熱容量が原因であるが、このようなふるまいが再現できる。

もう一つの結果は、1989年に私がベルギーにいたときに一緒に開発していたルーヴァン大での成果である。デザインした結晶を、逆問題を解いて作る、という研究である。この時点で、逆にヒーターパワーを時間とともにどう変化させるべきか、こういう予測ができるようになっていく。その計算で、どれくらいの容量のトランスを買えばいいか予測できる。オーバースペックのものは不要になる。すなわち、必要なものだけつくるという意味での逆問題。同様のアプローチは、化学工学、機械工学、航空宇宙工学などで随分やられている。

3 番目は、バイファークーションについて。これも非線形プロセスに特有だが、昇華法の炉での SiC 結晶の成長をシミュレーションで予測する。内部構造が分からないので、結晶がどうできていくか、について、数式に基づいて自分たちでプログラムコードを作り、実験との突き合わせを行う。

これは室温にまで下げた時の転位密度である。クリーンが行った実験と、我々のシミュレーションを比較すると、完全には合わないが、オーダーは合うようになってきている。

転位とか欠陥が入る問題は非線形問題である。スペインのグループが高温での応力歪み線図を計測しており、その結果に合うように、数々のパラメータを私の研究室で予測した多変数解析で決定すると、結果が合ってくる。

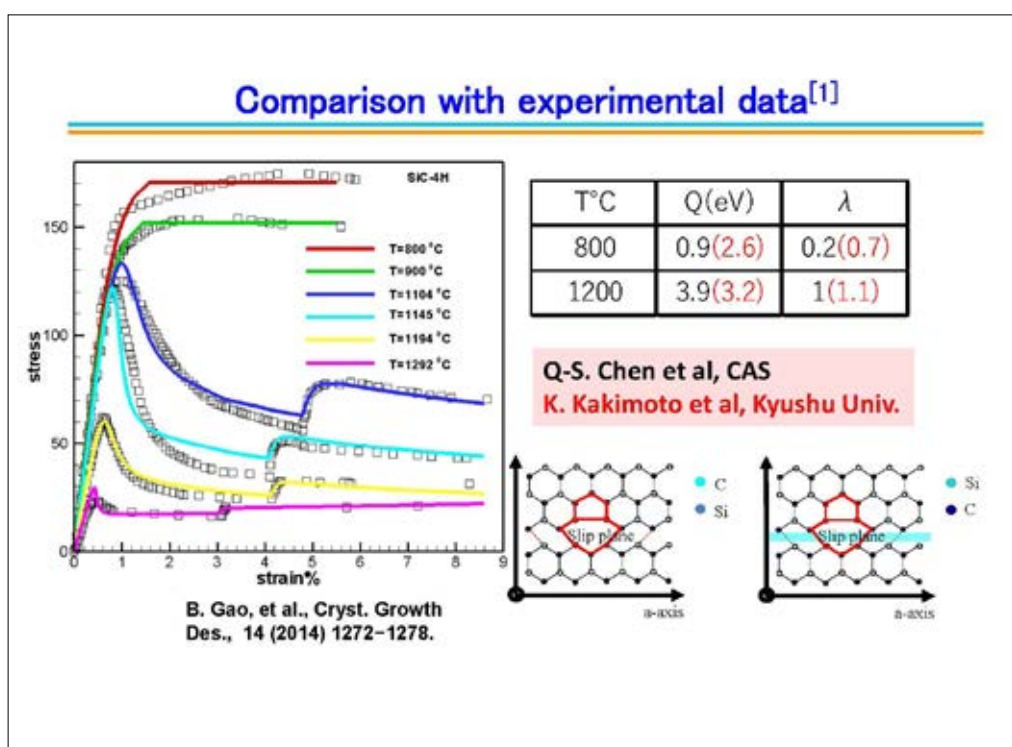


図2.3.1 実験結果と数値計算の比較

最近、中国科学院のグループが、我々のやり方を真似て「他の解があるよ」と言ってきた。彼らの結果がどれくらい違うかと思い、自分で計算したところ、中国のグループのパラメータで計算すると全く転位は入らない。これは間違っている。バイファークーションによる落とし穴であり、ローカルミニマムを求めてしまっている。チェックが必要である。

昇華法において、結晶多形というのはどんな要因で決まっているのかが重要である。二次元核生成が起こるとした場合の、表面エネルギーを第一原理計算から求めた。第一原理計算コードは SIESTA を使った。ヘキサゴナリティ、スタッキングごとに表面エネルギーを求め、計算のインプットとしている。

4 番目は、実験と数値計算の協力連携が非常に重要、ということについて。これは私の研究室にある 8 インチ Si 結晶をつくる装置である。これは、200 ミリの無転位の Si 結晶を作っていて、自動車用のパワーデバイス、IGBT を作っている。大体 10^9 から 10^{10} の

不純物のときに、市販されているウェハよりも純度が半桁高いものを作れる。

そこで、数値計算するために、炉全体を含める方針で、我々は開発している。下図は、液面の温度分布の計算結果である。数値計算の精度はおおむね良いが、まだ無欠陥状態。いわゆる空洞欠陥といわれるベイカンシーの凝集率を制御したいが、それを完全には予測できない状況である。

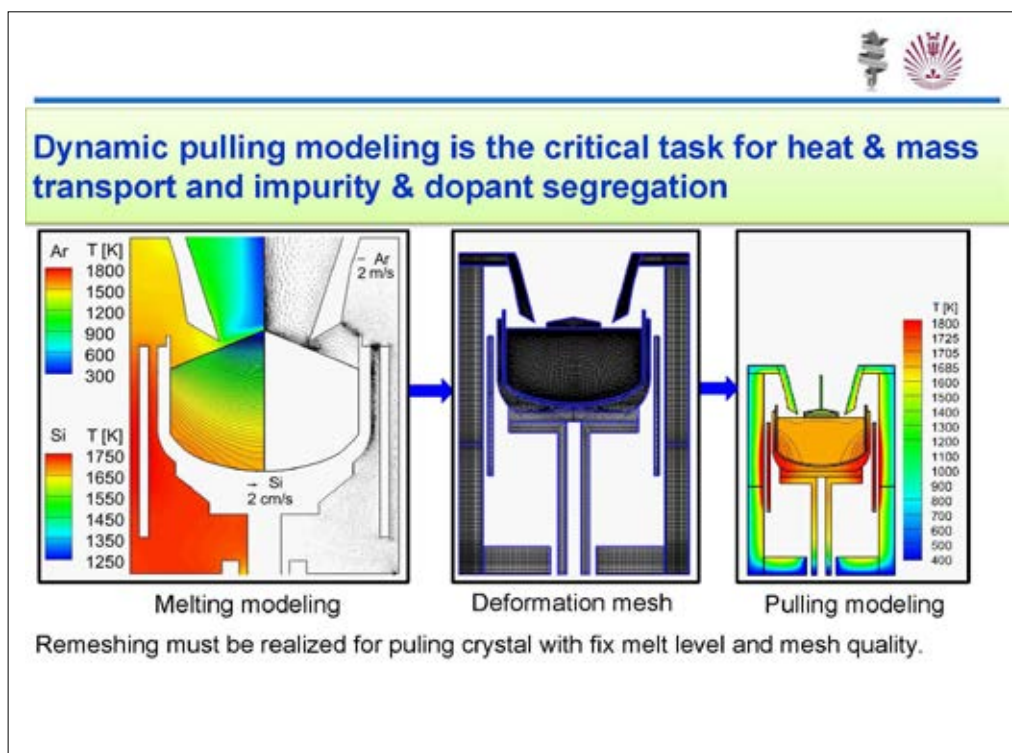


図2.3.2 8インチSi結晶用装置の数値計算結果

理由は、装置の大きさにある。10メートルぐらいあって、重要な部分はグリッドサイズを0.01ミリメートルより小さくしているが、それでも足りない。非常に薄い層を電流が流れていくため、きちんと計算しようとする膨大な計算量になる。現状、完全にはできていないが、それなりの精度で計算できる。

原料を溶かして、成長して、冷やし、最後にどういう結晶ができたか、までを数値計算で予測できないと、デバイス屋の要求には答えられない。原料を溶かす過程、結晶成長している過程で、本当に計算が合っているのか、を調べるために、炉内のCOガスのガス分析を行った。まだ完全ではないが、炉内の圧力に対してCOガスがどれくらいあるかを定量的に予測できるようになった。

結果として、ライフタイムが10ミリ秒、上から下まで均一な結晶をつくることができた。これをデバイスにしてパワーデバイスのプロセスにかけると、我々の結晶は、市販品のようにライフタイムがすぐに低下することはない。

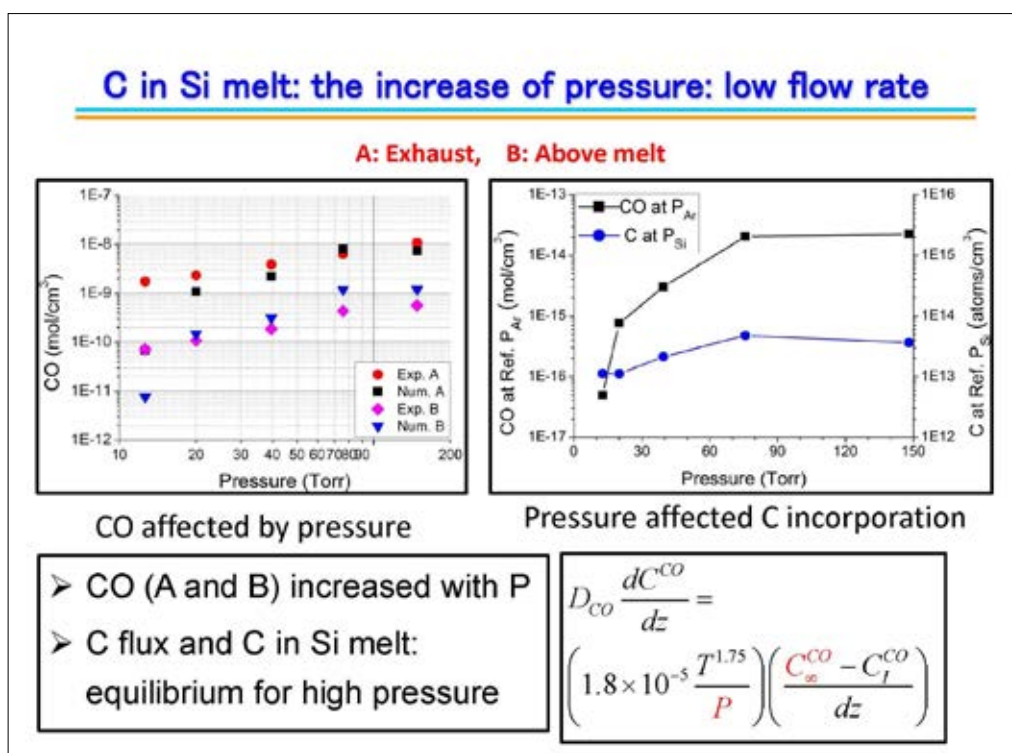


図 2.3.3 CO ガス圧力の実験と数値計算の比較

5 番目に、高温物性について。高温の物性、それから欠陥があるときの物性が必要になってくる。熱力学的な平衡のシミュレーションを、グリーン・クボの式を用いて予測する。熱伝導度が熱流束の自己相関関数の畳み込み積分でできるので、SiCの熱伝導度が、置換型原子、不純物、ベイカンシー等の濃度でどのように変化するかを予測できる。

量子コンピュータ用のシリコン同位体の熱伝導度についても数値計算した。シリコンは3つ安定同位体がある。熱電素子としては、電気は流すけれども熱は流さないのが良い。そういった物性を出すには、3つの同位体の混合比に近いものが良い。これはフォノン散乱が一番大きいので、熱伝導は小さいけれどもシリコンのバンド構造はそのままになる。こういう物性値が予測できる。

最後に、分野発展のための方策について。内容として、やはり物理、化学、数学、電気電子、機械、化学工学、経済学、全部が必要である。数値計算では、全て必要である。さらに、応用数学と工学の融合が必要である。工業数学の教育、数式による会話、これが減少している。それから、複数分野横断型のリーダーの養成。これは、例えば物理屋さんにはeV単位で話すが、化学屋さんはkcal単位で話す。この単位変換が可能なような、両方のバックグラウンドを持った人などが必要である。加えて、定量性の確保。これは社会実装には必須である。

また、膨大なプロセスパラメータがあるが、何を変数に選択するか、これも重要な視点である。さらには、誰にデータを渡すのか、という視点。R&Dなのか、製造なのか、解決策を提案できているか。単にシミュレーションで実験を再現しただけでは、シミュレーションは存在意義がない。

今後は、三次元、非定常予測法の確立が必要である。我々は今までもアルゴリズムを開発してきたが、今後必要である。それから、高温物理の推進。

安価で高品質の製造方法が重要である。日本ではあまりこういうプロジェクトをやらないうと思うが、これが一番重要である。アメリカがガリウムヒ素で現在勝っている理由の1つはこれである。結晶成長と結晶加工に科学を膨大に導入することが重要である。

結晶成長とインフォマティクス

分野発展のための方策

内容

1. 物理、化学、数学、電気電子、機械、化学工学、経済学等の融合
2. 応用数学と工学の融合: 工業数学の教育強化、数式による会話！！
3. 複数分野横断型リーダーの養成！！
4. 定量性の確保: 社会実装には必須！！
5. 膨大なプロセスパラメータ: 何を変数に選択するか？
6. データをだれに渡すか？ R&D? 製造? 解決策提案型！！
7. 3次元、非定常予測法の確立: アルゴリズム開発！！
8. 高温物理の推進！！
9. 安価で高品質の製造方法: 科学の導入！！

図2.3.4 分野発展のための方策

【質疑応答】

Q: 我々は新しいものをつくっていきいたいとき、特に多元系になったときに、様々な相があるが、その制御まで含めてやっていくのに使える基本原則のようなものはあるか。

A: 多元系だと、ある程度物性値が必要になってくる。温度、圧力等はある程度予測がつくが、熱物性などの物性値をある程度入手できないと難しい。実際の実験で信頼できるデータを取っていることが非常に重要である。

Q: 結晶成長過程は非線形に起源する、という話であるが、それを数学的に記述することはまだできていない、ということか。

A: 数学的に記述はできている。ただ、複数の極小点があり、確率的にどちらかに落ちる。あとは、確率論である。

Q: 逆問題というのは解析的に何かを解決できるということか。それとも試行錯誤的に「これが近かった」というレベルになるのか。

A: 数値解でやっている。具体的には、共役勾配法という多変数最適化問題用のマトリクス解法を使っている。

Q:最後の分野発展のための方策は、方向性としては納得するが、その実現のためにやるべきことは何か。

A:例えば、学部を出た後、同じところのマスタードクターに行かない、というルールを作ること。そうすると、自動的に異分野融合になる。

C:1つの研究室に複数の学科から来た学生を一緒にしてしまう、というやり方もある。

2.4 多結晶インフォマティクスの現在地

宇佐美徳隆（名古屋大学）

CREST「革新材料開発」（研究総括：細野秀雄）で進めている研究内容を中心にインフォマティクスの活用について紹介する。対象としている材料は単純なシリコン（Si）であるが、多結晶であるということが複雑であり、また組織、不純物、欠陥等から構成される複雑なシステムと考えている。まず単純なシリコンで学問をつくって、それを様々な材料に横展開していきたいと考えている。

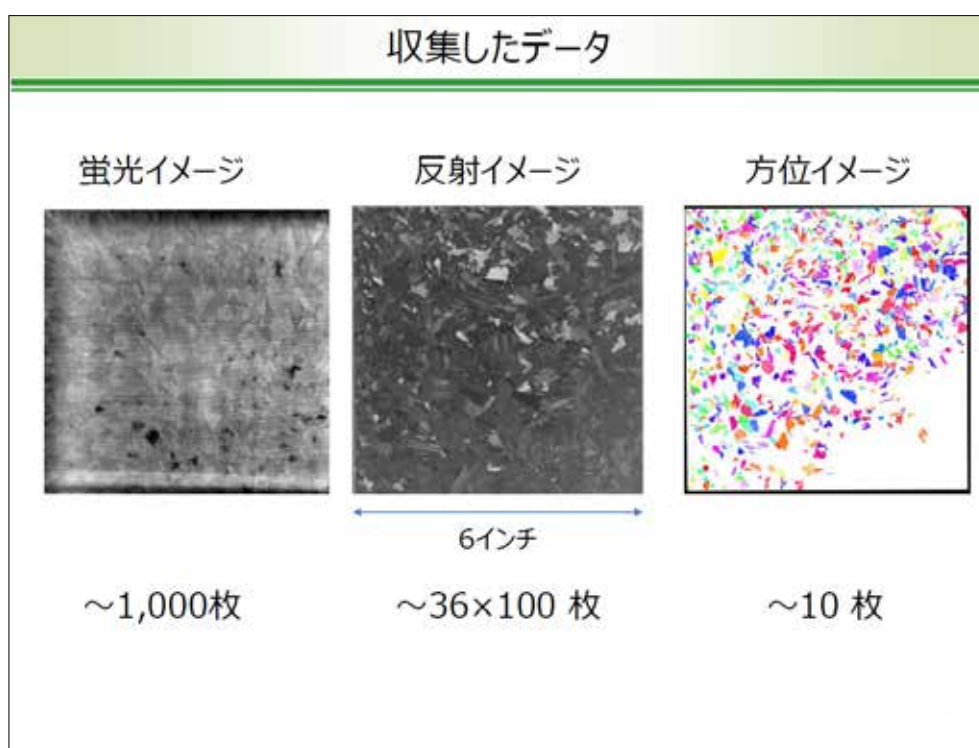


図 2.4.1 収集したデータ

図 2.4.1 の左側は実際に使われている 156 ミリ角の太陽電池用の多結晶シリコンウェハの蛍光イメージである。結晶粒が 1,000 個以上あるようなウェハであり、1 個の結晶粒界を表すだけでも少なくとも 5つのパラメータが必要で、非常に複雑である。これを網羅的に取り扱うに当たっては、ある現象を抽出するだけでもインフォマティクスを使うのが有用と考えている。

研究で狙っているものは、まずはどんなものをつくるかというマテリアルデザイン的な部分であり、それをどのようにつくるかというプロセスの部分まで踏み込みたいと思っている。多結晶材料の普遍的な高性能化指針を明らかにし、太陽電池や機能材料だけでなく、構造材料等にまで応用したいと思っている。

プロジェクトの中では実際に多くのデータを集めるところから始めている。簡単に大量に取れるデータや、少数だが最先端の計測技術を使ったデータ、など多くのデータを集めている。また、機械学習やシミュレーション等も使いながら、多結晶中に頻繁に表れる双

晶粒界のようなものだけでなく、一般的な粒界の構造や物性が分かるところまで開拓していきたいと思っている。さらに、実際のものづくりまで成し遂げて、このような方法の有用性を示したいと思っている。

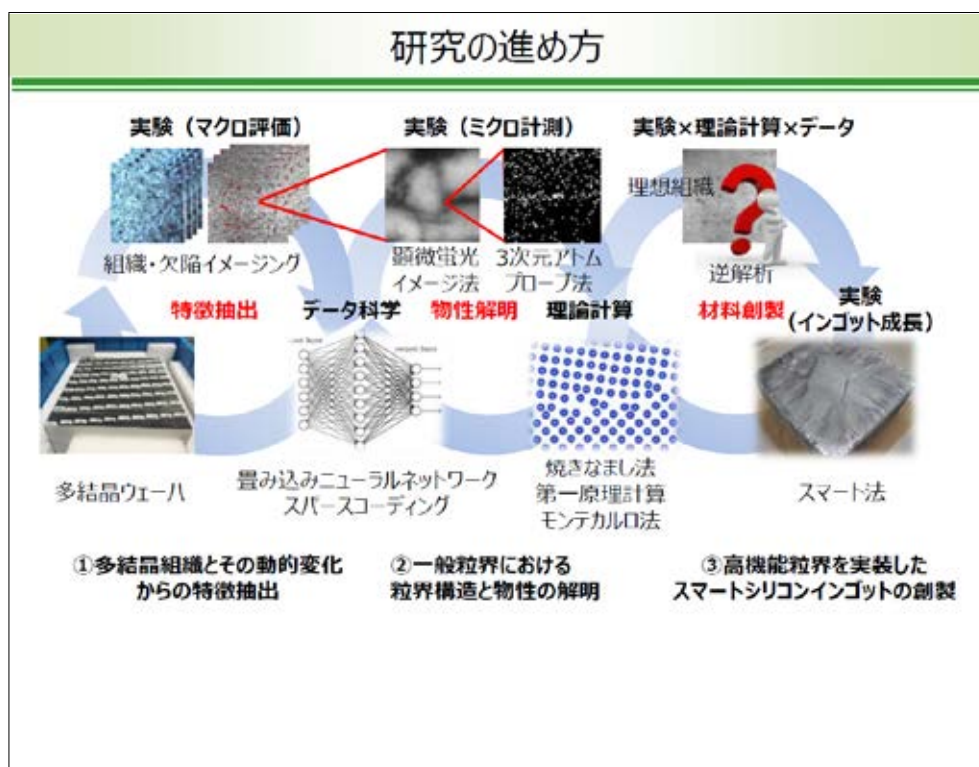


図 2.4.2 研究の進め方

図 2.4.2 に研究の進め方を示す。3つの研究項目がある。1つ目は多結晶組織とその動的変化からの特徴抽出である。インゴットの中から切り出した1,000枚程度のウェハの写真や蛍光イメージ、反射イメージといった多くのデータを取り、画像処理や機械学習的な手法を使ってマクロな特性に影響を及ぼしている特徴、あるいは結晶学的に転位の発生する場所を抽出するような研究である。

2つ目は一般粒界における粒界構造と物性の解明である。特徴的なところがどんな原子構造をしてどんな物性を持っているのかといったことを、シミュレーションの中に機械学習的な手法を取り入れ、高精度に速く多結晶の粒界を計算する手法も考えながらやっている。

3つ目は高機能粒界を実装したスマートシリコンインゴットの創製である。これはどうやってつくったらいいのかというプロセスの部分である。今はまだ踏み込めていないが、これまでの成果を少し紹介する。

ウェハから取ってくるデータとしては、図 2.4.1 に示すように大きく分けて3つある。1つ目は蛍光イメージであり、レーザーを当てて蛍光を画像化するものである。これは1枚を1秒程度で取れるので大量に撮ることができ、インゴット全体をスライスして全数検査するように実際の生産現場でも使われている。2つ目は反射イメージである。1枚のウェハに対して様々な角度から光を照射すると、方位の情報が取れ予測できるのではないかと

期待して集めている。3つ目は実際に結晶粒の方位を個々に測定して画像化するものである。これはとても時間がかかるので、多くを取ることはできないが、機械学習の教師データとして委託して取っている。手元にあるそれぞれのデータの数としては、図 2.4.1 に示した程度の数になっている。

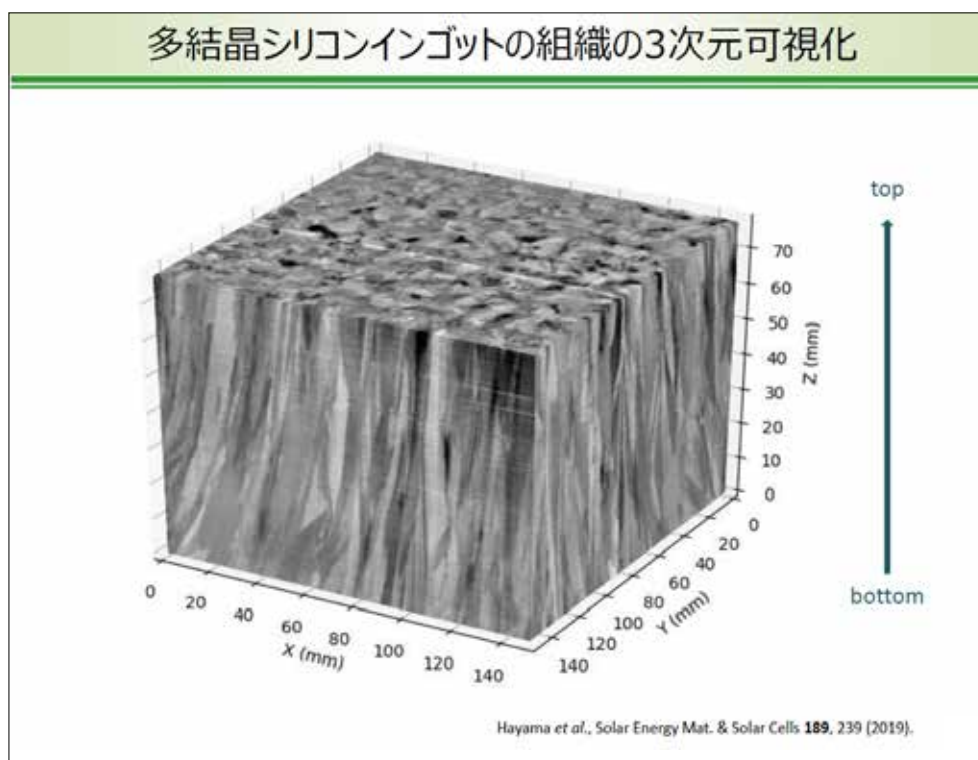


図 2.4.3 多結晶シリコンインゴットの組織の3次元可視化

図 2.4.3 は多結晶シリコンインゴットの組織の3次元像である。これは結晶成長のシミュレーションではなく、単純に下からウェアハの蛍光イメージを並べて三次元再構成し、どのような組織になっているかを三次元的に見せているものである。これは内部にもデータを持っているので、バーチャルにクロスセクションを取って、こういった組織をしているのか可視化できる。

これは蛍光イメージなので、暗いところは結晶欠陥に対応しており、そこを抜き出すと、実際につくったインゴット結晶の中で蛍光強度の低い転位クラスターがどのように分布しているのか三次元的に可視化できる。このため、成長過程で転位クラスターが発生し、ある部分で消滅したことや、分布に偏りがあること等が分かる。発生点を網羅的に取り出すようなこともでき、組織との対応等はこの範囲でできる。

ただし、蛍光イメージでは粒界が抜き出しにくいので、粒界抽出に関しては反射イメージを使っている。様々な角度から光を照射した多くの反射イメージを用いて、画像処理により粒界を抽出すると、結晶粒のセグメントの色分けができ、ここから様々な数値情報が得られる。例えば、図 2.4.4 に示すように、少し見にくいですが、粒界の三重点がどのように分布しているのかを網羅的に抜き出すことや、あるいは転位クラスターの分布で発生点がどの位置にあるのかを示すこと（左側の図中における赤い点）ができる。右側のグラフの

青色は、実験で得られた転位クラスターの発生点と一番近い粒界三重点の距離のヒストグラムである。オレンジ色はランダムに抽出した 4,000 点が最寄りの粒界三重点とどんな分布にあるのかを示したものである。赤い楕円で囲んだ部分を見ると、青色はオレンジ色と比較して距離の近いところにピークがあるので、転位クラスターと粒界三重点とは関係していそうだという知見を得ることができる。

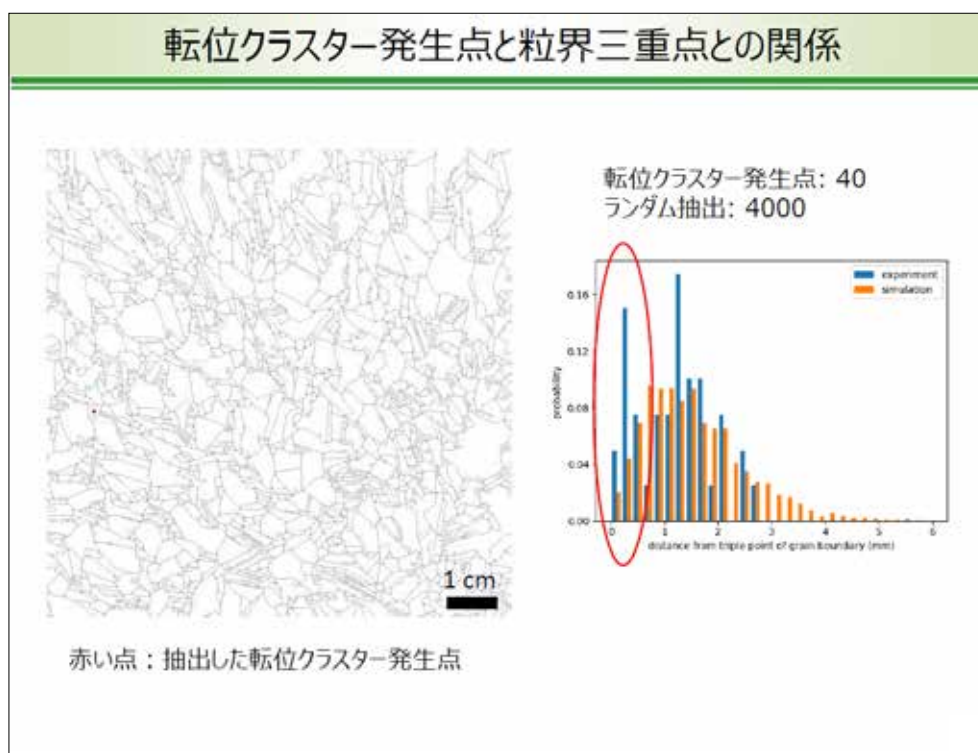


図 2.4.4 転位クラスター発生点と粒界三重点との関係

また、転位発生点がこういった画像的特徴を持っているのか、例えば情報学のコンボリューショナルニューラルネットワーク (CNN) の画像分類機の転移学習を使って抜き出すこともできる。人が見てもよく分からないが、転位の発生源となっているのではないかと機械が予測した画像を出すこともできる。

ただし、これらは蛍光イメージや反射イメージであり、まだ結晶学的な情報である方位や粒界などにまでは至っていないため、データとして多次元化していくことにも取り組んでいる。図 2.4.5 にニューラルネットワークを用いた結晶方位の推定方法を示す。ここで示すようなイメージを結晶粒ごとにクラスタリングして、その平均輝度の変化をデータ化していく。これには、1 枚につき 10° おきに照明角度を変えているので、36 次元ベクトルの入力が見られる。実際に測定で得られた方位を出力としてニューラルネットワークを構成すると、方位の予測機ができる。

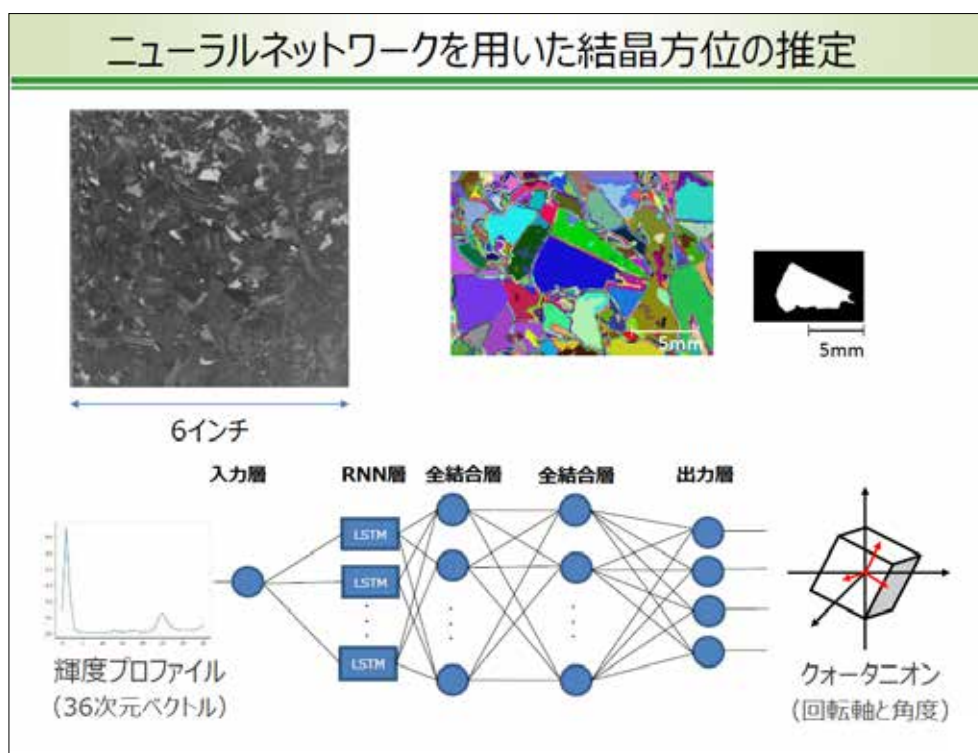


図 2.4.5 ニューラルネットワークを用いた結晶方位の推定

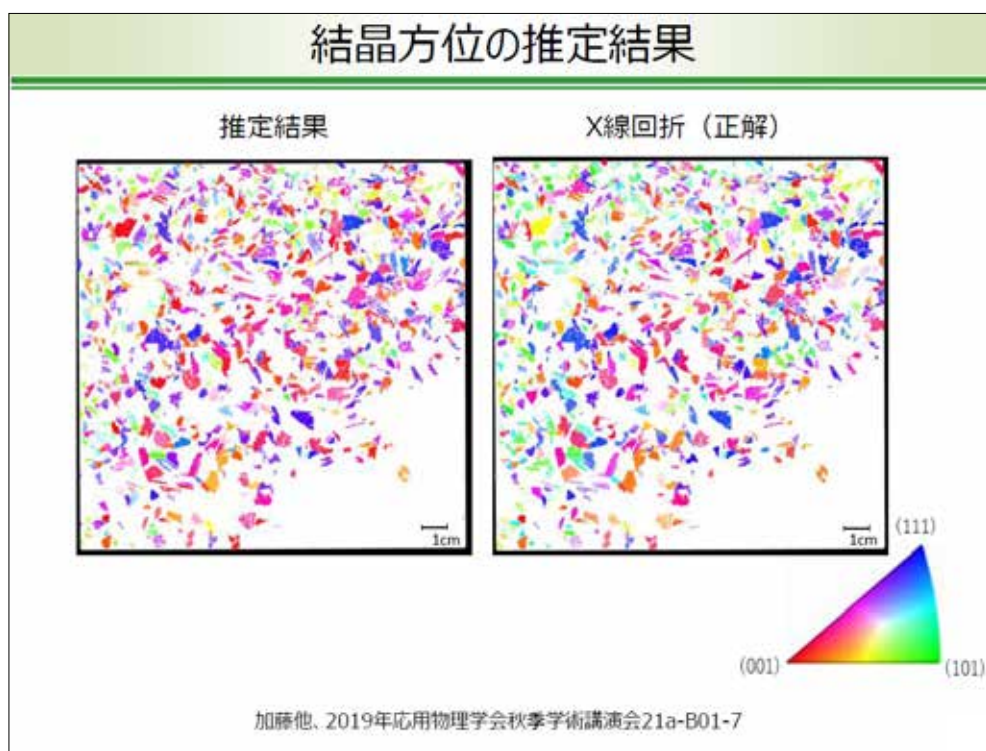


図 2.4.6 結晶方位の推定結果

実際に推定した結果の例を図 2.4.6 に示す。左側が推定結果、右側が X 線回折で得られた正解の方位分布である。方位の違いを色で表している。推定結果を見ると、平均の方位誤差としては十数°となっており、簡易的な解析には利用可能ではないかと思われる。

このように、簡単に撮れる写真から様々な情報が取れ、現状では結晶方位や転位クラスターなどの様々な結晶学的情報、電気的特性も含めた三次元分布の可視化ができており、欲しい情報を取ってこられるものができている。これとプロセスとの対応を取り、高品質な結晶をつくっていくことに答えを出したいと考えており、ここからプロセスに関連する話をする。

実際につくられる結晶は大きなインゴットであり、容器の中で溶かした融液を一方向に固めて作られている。CGSim というシミュレーションを使うと、マクロ的な熱流体解析によって三次元のシミュレーションができ、機械学習的な手法を使うとそれを速くすることもでき、そのパラメータの合わせ込みに様々な実験データを有効に使えるのではないかと考えている。例えば、意図的に成長過程で成長面に少し析出物が分布するような操作をやり、その分布を赤外で見て三次元可視化し、成長界面を多項式でフィッティングすることができる。成長界面の痕跡が見られることで、シミュレーションとの対応やパラメータ推定に使えると考えている。

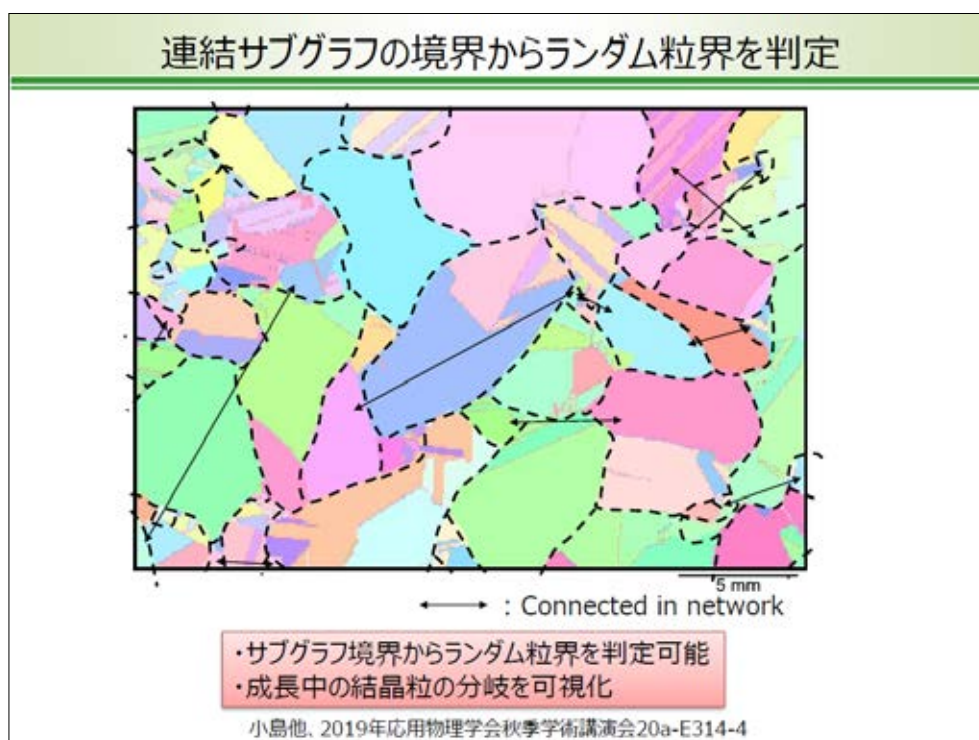


図 2.4.7 連結サブグラフの境界からランダム粒界を判定

多結晶の成長過程には $\Sigma 3$ という双晶が入って方位が変化するという過程がある。同じ結晶核からの成長過程で双晶が入ることによって方位が変わったと予測されるものは、結合サブグラフでつながっている。逆に、結合サブグラフでつながっていない領域の境界が、別の核から生成したものが衝突したことによるランダム粒界になると言える。このように、高次の対応粒界とランダム粒界を見分けるツールに使える。また、成長過程を遡って、組

組織の形成過程を予測するツールになると考えている。図 2.4.7 の中の点線で示したところが、本当のランダム粒界と予測しているものである。これを下まで遡っていき、最終的にるつぼの一番底を見ると、核形成したところにたどり着く。るつぼの底の組織を見て、ボロノイ図解析を応用して実際に実験で観測した結晶粒分布を再現するように核形成の位置とタイミングを考えていくことで、核がどこでできてどんなタイムラグで発生したかという核形成の予測もできると思っている。

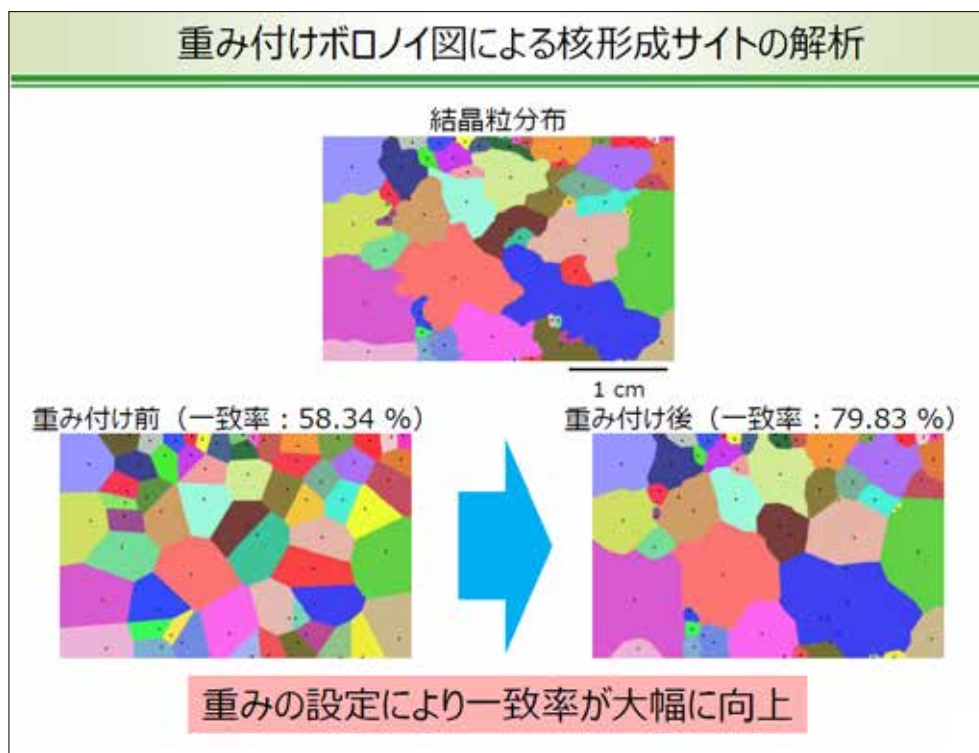


図 2.4.8 重み付けボロノイ図による核形成サイトの解析

図 2.4.8 は実際に組織に合わせ込むように、重み付けボロノイ解析を行ったものである。これで核ができるところや、組織が変化していくところをある程度予測できる。また、三次元的なマクロな熱流体解析もできるので、これらを組み合わせることで、転位の発生などマクロな特性に悪影響を及ぼすことがどのような物理に支配されているのかを考えることができる。粒界近傍だけでも原子一個一個を考えた凝固過程のシミュレーションを目指し、CREST のチームの中で、原子間ポテンシャルを経験的なものではなく機械学習型に置き換えて、分子シミュレーションの精度を保ちながら高速化する取組を進めている。経験的ポテンシャルは単結晶の中ではよく再現できるが、粒界や転位周りなど構造が大きく乱れたところの特徴は捉えられていないことが多い。例えば、有名なポテンシャルを使ってシリコンの多結晶の双晶を計算しようとする、エネルギーがバルクよりも小さくなってしまふことがある。そこで構造緩和をさせると、粒界が多く入ったものが安定になるという解が出てくることになる。そのため、多結晶を考える上では、構造が変わっているという特徴を捉えたポテンシャルをつくらないといけない。そこでやったことは、様々な構造に対する系のエネルギーと原子にかかる力を DFT 計算し、それを教師データに使う、ニューラルネットワークで構造を入れるとエネルギーや力が出てくるようなポテンシャル

をつくった。通常、分子シミュレーションをやるときには全ての原子に対して周りの様々な相互作用を取り込んだ方程式を解くのに多くの時間がかかるが、ニューラルネットワークで得られた原子間ポテンシャルで置き換えることでこの部分を高速化できるため、分子シミュレーション全体が高速化できる。

現状では、成長をシミュレーションするところまではできていない。例えば刃状転位のように少しだけ傾いたものはなかなか周期的にならないので計算が大変であるが、原子数13000程度の系で、構造緩和がきちんと計算できて、転位のコアの構造が計算できるようになっている。これを進めていくことで、大規模な凝固シミュレーションもできている。このようなものを使うことでマクロとミクロを組み合わせ、どのようなレシピで高品質な多結晶をつくったらよいか考えていきたい。

【質疑応答】

Q: デバイスの観点からすると、少数キャリアの寿命と粒界や欠陥などとの関係が欲しいが、その辺はどうなっているのか。

A: 今回は話さなかったが、それもまさにインフォマティクスを活用するものである。私たちは、蛍光イメージのプロファイルをシミュレーションでまず再現しておき、多くの教師データをつくっておく。そうすると、粒界の少数キャリア再結合速度と周りの結晶粒の拡散長に対して、得られるプロファイルをニューラルネットワークで結んであげると、逆にプロファイルから粒界再結合速度の推定ができる。そういったことができる、粒界の上における再結合速度分布の図も描けることになる。要は結晶粒の拡散長と粒界の再結合速度がウェハ全面にわたって得られるので、マクロな特性も予測できることになる。

Q: シリコンは多結晶でも太陽電池の発電効率が高くて使われている。一方、ガリウムヒ素 (GaAs) の多結晶太陽電池はつくられてないが、なぜなのか。

A: 多分、粒界の特性や結晶欠陥等に関係するのだと思う。今回紹介した原子間ポテンシャルについては、シリコンだけではなく、化合物半導体にも展開していこうとしている。ガリウムヒ素やカドミウムテルル (CdTe)、CIGS などに対してやろうとしているので、その中で答えが出てくると思う。

Q: 化合物半導体の場合には、例えば結晶粒界にポテンシャルバリアができたりするが、そのようなものも予測できるのか。

A: それはできている。まずは構造を決めて、その電子状態の予測まで持っていくことは検討している。

Q: 今はポリシリコン (多結晶 Si) であるが、今後は様々な多結晶に対して応用できるようになると考えているのか。

A: 応用できる部分とできない部分があるとは思いますが、方法論としては使えるところは多くあると思う。

Q: 全く違った物性、例えば磁性体の粒界のところでの交換結合がどれだけ切れているか

といった話の場合にも、コアシェル構造がどのくらいになるか計算できるということか。

A:是非そのようなところに展開していきたいと思っている。CRESTの中でも多結晶を扱っている他のチームがあるので、チーム間連携をやりながら様々な多結晶材料に展開していきたいと思っている。

Q:粒の中は、形や大きさ、方位が異なっても、同じような物性、欠陥、密度、などを持つ均質なものと仮定しているのか。また、粒ごとの違いをどのように決めているのか。

A:粒界のキャリア再結合速度を求めるときに、粒内のキャリア拡散状況もパラメータとして出てくるが、それは1個の粒内は1つの値にしている。ニューラルネットワークで物性と出力の計測データとの関係をつくっておくと、計測結果から粒の電気的特性を表すようなパラメータも個別の粒ごとに出てくる。

Q:シリコンの粒界にポテンシャルができるということだが、そのポテンシャルは何で決まってくるのか。バンド構造のラインナップの違いだけではなく、ワークファンクション、結合などが考えられるのではないか。

A:私たちは粒界の原子構造を決めて、そこからその物性を決めている。全ては見られないので、実験と合わせる部分もあり、計算でやる部分もある。マクロな対応関係が同じでも、つくり方によって微妙に構造が違ったりすることもあり複雑ではあるが、基本的には原子構造をきちんと決めることをやっている。

2.5 地球科学におけるプロセスインフォマティクス

桑谷立（海洋研究開発機構）

岩石学について説明した後、プロセスインフォマティクスの研究例として2つ紹介する。「変成岩岩石学」という岩石学で私は学位を取ったが、当時はそれこそ崖を登って石を取ってくるのが生業だった。ポストドクとして東京大学の新領域創成科学研究科に在籍する頃に転機が訪れ、機械学習とその地球科学への応用という情報科学の研究を進めるようになった。その後、東北大学を経て、現在の海洋研究開発機構（JAMSTEC）に所属している。現在は情報科学と地球科学、岩石学、地球物理学、環境科学などのテーマに、インフォマティクスを応用した研究をしている。

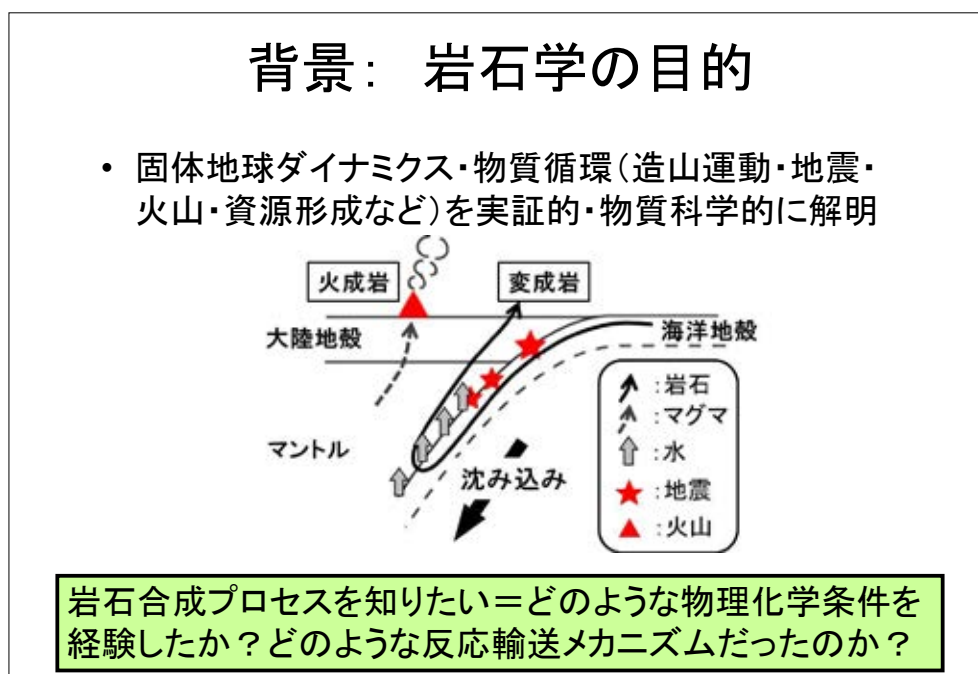


図2.5.1 岩石学の目的

そもそも岩石学とは、固体地球ダイナミクス・物質循環に関する学問であり、山がどうできたのかという造山や、地震や火山、資源がどのように形成されているか、そういったことを実証的に、物質科学的に解明する学問である。岩石は一度地下に沈み込み、そこから上がってきたものを我々は見ることができるのだが、岩石の形成プロセス、合成プロセスを知ることで、その岩石がどういう物理化学条件を経験したのか、どのような反応メカニズムで形成されたのか、そういったことを明らかにするのが岩石学である。岩石学の本質は、結果である岩石の何らかの分析データから、原因である過去の様々なプロセスを推定する、難解な逆問題といえる。一つのプロセスからではなく、多くのプロセスが重なり合って岩石は形成されるが、私達には最終状態である岩石だけしか情報がない。ここから何かを抽出しなければいけないが、例えば1辺1ミリくらいの岩石薄片を観察してみると、多成分・多相系で、10成分入っていてかつ10相が含まれていたりする。様々な鉱物が非常に複雑に入り組んで存在しているので、どうやって情報を読み取ればいいのかよく分か

らない。そういう絶望的な状況だったが、大量のデータを得ることができるようになってきた。化学組成や方位のデータなど様々あり、データを最大限活用すれば、困難な逆問題に対応できるかもしれないと考え、データ駆動科学を進めることにした。自然科学の逆問題を解くためのデータ駆動科学の方法論の一つとして、バイズの定理を使って結果から因果律を遡って原因を推定する枠組みがある。物理モデル、シミュレーションモデルを活用するという基本姿勢を持って進めている。

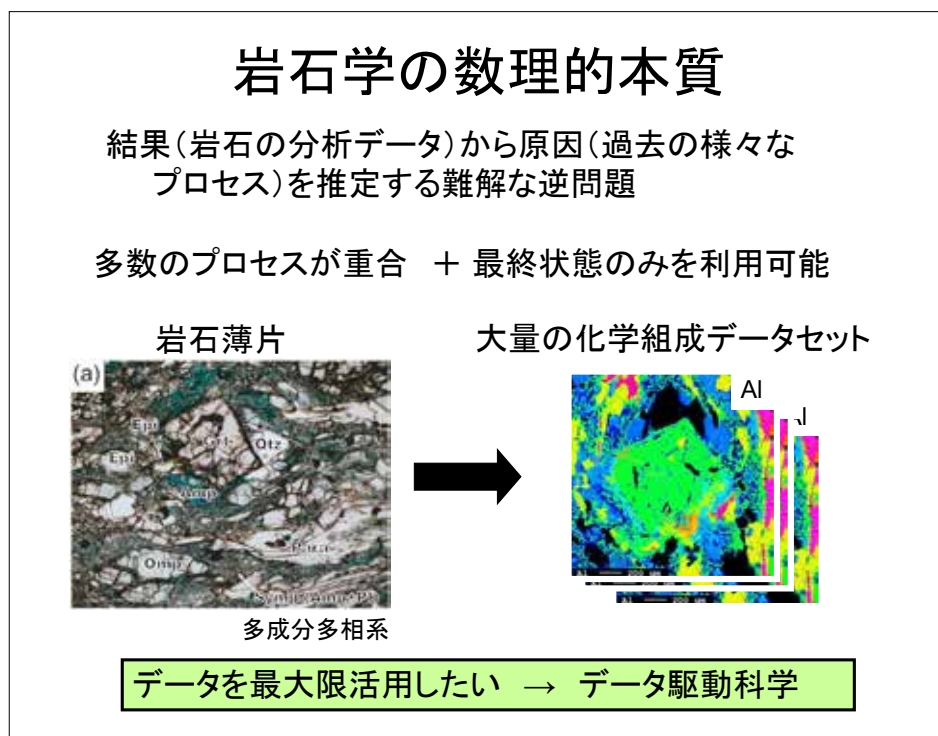


図2.5.2 岩石学の数理的本質

最初に、データ同化を用いた鉱物組成累帯構造からの温度圧力履歴推定について説明する。最終状態からプロセスの物理化学条件の履歴を推定する、という逆問題になる。組成累帯構造とは、鉱物内部粒子の同心円状の化学組成の空間変化のことである。地下深部の情報を記録する天然の記録メディアともいえる。図 2.5.3 は鉄の化学組成の元素マップだが、岩石の内部から外側にかけて化学組成のバリエーションが見える。これがどうやってできたかという、地表で核形成したものがどんどん雪だるまのように結晶成長していく。中心から外側にかかる縁辺部の化学組成は、鉱物が成長していくそのときの物理化学条件、この場合は温度圧力変化を記録している。地下内部の物理化学条件を知るために、この組成累帯構造を使って、温度圧力の変化を推定するという研究がある。熱力学的におよそ平衡であると仮定すると、化学組成は温度と圧力の関数になる。化学組成を中心から外側にかけて分析して、プロファイルを取り、熱力学の式を逆解析することで、中心から外側にかけてこの岩石が経た温度圧力のパスを推定する。これに確率論を使う研究をしてきた。この方法では温度と圧力の時間推移は分からないが、時間ごとに圧力がどう上がってきたかも知りたい。深さに対応して、どうやって沈み込んでいったのか、どのように外側の温度が上がっていったのか、等を推定する。時間軸的なものは、鉱物粒子の成長則を仮定する。鉱物粒子の半径の時間変化はアレニウス型で表現され、拡散律速のときに成り立つ成

長則になる。これによって半径の時間発展が分かり、さらに温度圧力から結晶成長時の化学組成が分かる。結果として、組成累帯構造をシミュレーションすることができることになる。

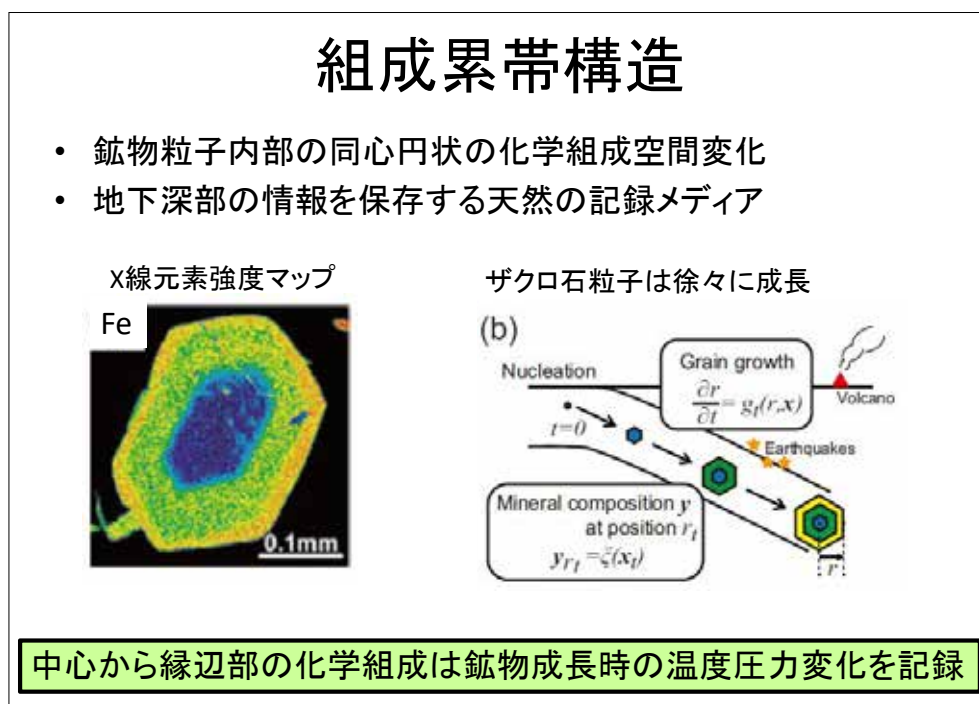


図2.5.3 組成累帯構造

この際に使うのが、ベイズ推論ごとのデータ同化になる。データ同化は、数値シミュレーションと観測データを融合する枠組みであり、この研究は恐らく岩石学で初めての適用例になる。データ同化は状態空間モデルというものを考え、時間ステップごとに数値計算をし、持っている観測データを合わせる。このように、数値計算のシステムモデルと、観測モデルの二つから成る。しかし、通常のデータ同化では歯が立たない。岩石学の問題は時系列のデータを得ることができないので、結局最後の空間データのみしかない。通常のデータ同化では、観測とシミュレーションが全部ではなくてもどこかで合うが、それが無理な問題であるため、新たなデータ同化の方法論が必要となった。そこで私達は、状態空間モデルを改良した。数値計算と空間的な観測データはまばらにしか取れないが、これらを融合するにはどうすればいいか。各数値計算の時間ステップに対応する時系列観測データをつくる必要があった。そこで、空間補間によって数値計算が観測データを追跡できるようにすることで、アルゴリズムを作成した。

モデルをつくったら、人工的につくった観測データから本当にこの P-T-t path (圧力・温度・時間) を推定できるかを確かめていく。たくさん数値シミュレーションを同時に何千個も走らせて、仮想的に鉱物粒を成長させる。観測データに合っているものは、その数値計算は生き残らせて、間違っているものはやめて別の計算を行う。この方法を使って一番よい P-T-t path を推定できる枠組みになっている。材料にもこういった組織があれば、適用しようと考えている。また、複数の粒子や二次元の化学組成マップにも応用しようとしている。

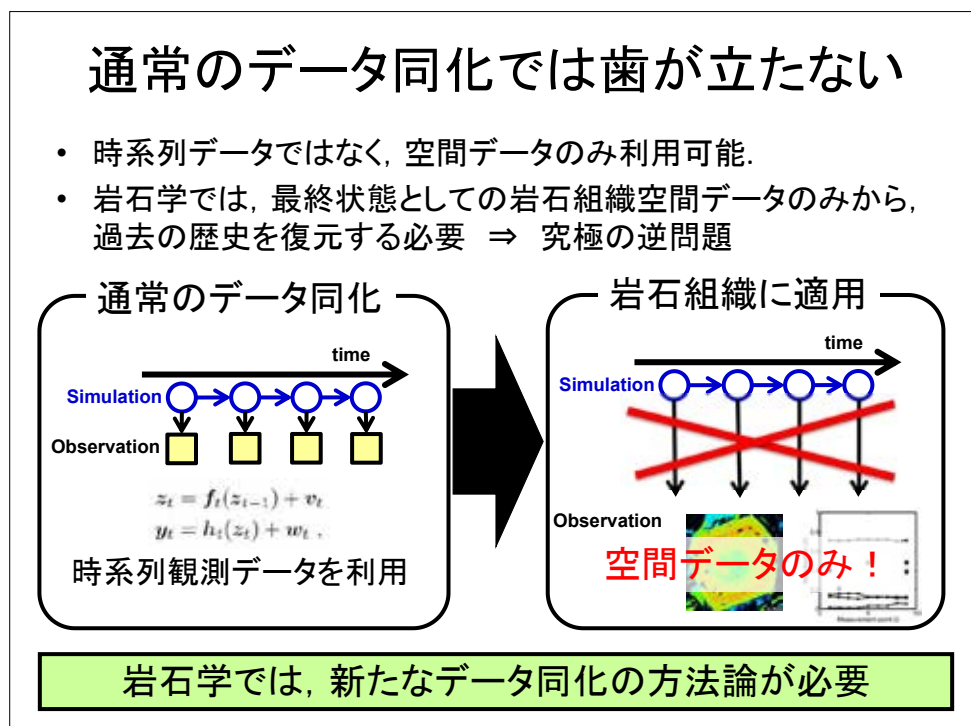


図2.5.4 通常のデータ同化では歯が立たない

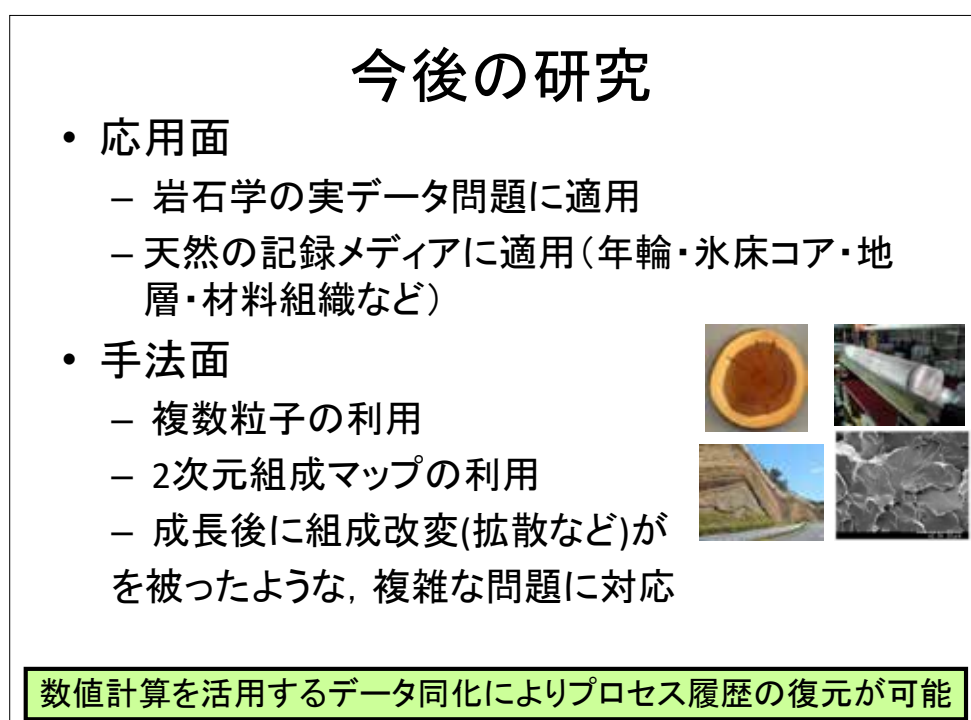


図2.5.5 今後の研究

もう一つの研究例は、スパースモデリングを用いた岩石流体反応の速度定数・経路推定である。最終状態からプロセスのメカニズム、法則を推定する逆問題になる。反応速度を得る通常の方法は、単純なシステムで実験を行って、溶液濃度の時系列データを測定して、そこからデータ同化によって反応速度を推定するような方法である。今回の実験は少し違

い、反応-拡散系の水熱実験で、かんらん石の粉末と石英の粉末を接して、ここに水を入れて 300℃、飽和蒸気圧の中で約 90 日間置く。反応後に内管の断面を観察して、実験結果として反応生成物の空間分布のデータを得る。かんらん石がどのように減って、かんらん石と石英と水が反応することで蛇紋石がどれだけの量が生成したかを、空間分布として得る。この空間データから、化学反応ダイナミクスを推定する問題になる。

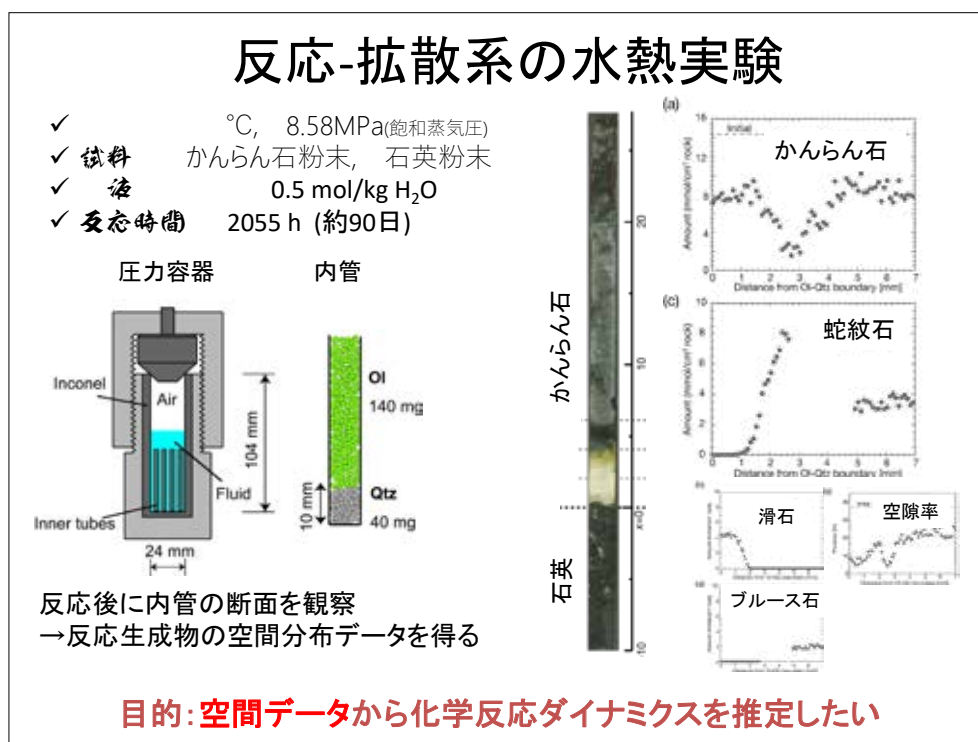


図2.5.6 反応-拡散系の水熱実験

濃度の時間変化が溶液中の一次元の拡散で、多数の鉱物の溶解析出反応から成るという反応拡散方程式になる。多成分・多相系になるので、反応のパスが多数出てきてしまう。最終的にはかんらん石から蛇紋石になるが、中間生成物があることでたくさんのパスが考えられる。このケースでは7つのパスが取られる。このような条件下で溶液の化学組成の変化と、空隙率の変化を数値計算する。さらにこの研究は、どの反応経路が使われて、どの反応経路が使われていなかったか、そういったことを明らかにしたいという研究である。ここでやはり人工データによる解析を行い、この方法がうまく働くかを推定するためにあらかじめ反応速度定数を仮定し、そのうちの幾つかは機能しないパスであるとして0を仮定し、人工観測データを測定する。データからこの反応経路及び反応拡散係数をきちんと推定できるか、そういった研究になる。

逆解析の方法は、観測データと計算値のフィッティングを MCMC というサンプリング法で評価する。たくさんの反応経路があるなかで、どの反応経路が動いているか、動いていないか、7個のパスがあり、2の7乗、128通りの中から最もよい経路の組合せを確かめる。モデルの評価基準として、ベイズ自由エネルギーの近似値を与える WBIC というものを使って計算した。結果、最初に仮定した正しい組合せをちゃんと再現していることを確認した。

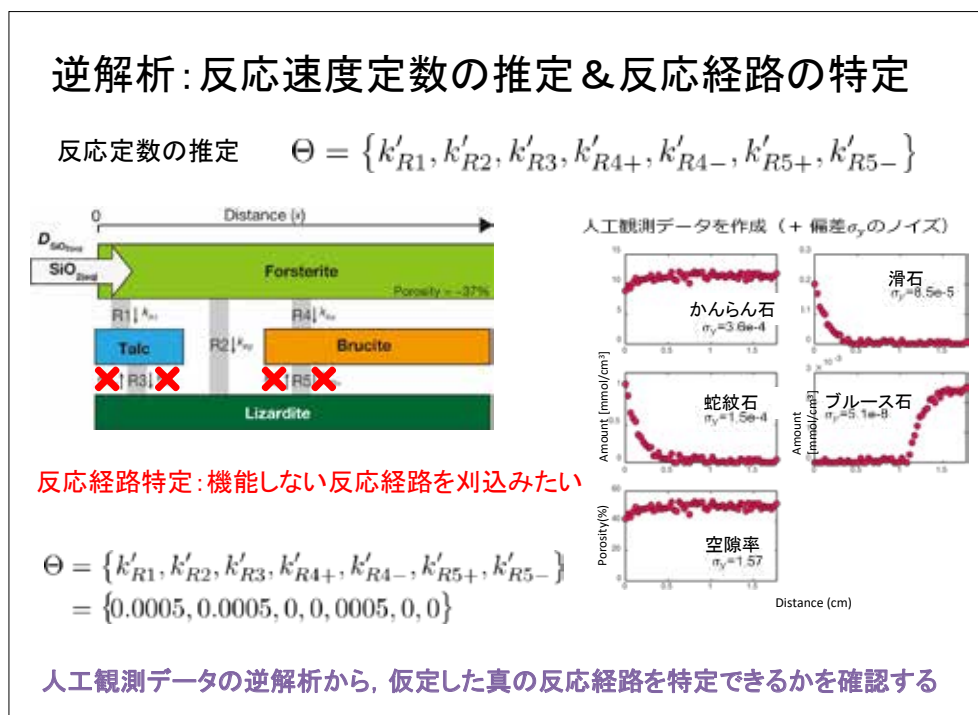


図2.5.7 反応-拡散系の水熱実験

このケースでは反応経路が7つしかなかったが、現実の系は多成分・多相系であり、反応経路は複雑で何十パターン、何百パターンになってしまう。実時間での計算は困難なことが多く、例えばスパースモデリングの代表的手法である L1 正則化を活用したり、アルゴリズムを工夫することで解決していく必要がある。その場合も、地球科学者だけでは無理であり、数理・情報科学者と連携することで、数理的に問題を解決していくことが重要になる。

議論

- 水熱実験の空間観測データから反応速度定数および経路を特定する方法を開発
 - 反応拡散方程式の数値計算と実験後空間データをFitting
 - 組み合わせ網羅的探索により, 反応経路を特定
- 現実の天然系(多成分多相系)の反応経路は複雑
 - ⇒ 組み合わせ爆発 (2^n)により, 実時間での計算は困難
 - ⇒ L1正則化(スパースモデリングの代表手法)を利用

$$E(\Theta) = \sum_i \frac{1}{2\sigma_i^2} \sum_j^{N_{obs}} (y_{i,j} - y_{i,j}(\Theta))^2 + \lambda \sum_{l=1}^7$$

L1正則化項

数理・情報科学者と連携することで, 数理的な課題を解決

図2.5.8 現実の複雑な反応経路には、数理・情報科学者との連携が必要

【質疑応答】

Q：岩石学や鉱物学でインフォマティクスを切り開いている。先端的計測技術等も活用して研究しているが、装置の使い方やデータの取得などのノウハウ等はどうやって獲得したのか？つまり、ある研究者が、「こういう新しいことをやってみたい」と思ったときに、どういう環境があればそういう研究が花開くようなところまで行くのか、このイメージ像は今後の制度設計に役立つので知見があれば伺いたい。

A：JSTの情報計測領域で、ベイズ推論などを多用して、高精度化する研究をしている。私自身、いろいろな地球科学者と共同研究することで、いろいろな対象、いろいろな計測でたくさん論文を書くという研究スタイルでやっている。情報科学等に少しでも興味があり、数式を展開するのが苦ではなくて、プログラミングも書ける人がいたらディスカッションをして仕事をする、そういうことをいつもやっている。

Q：岩石ができるときの温度は、どのようにして決められるのか？

A：温度を決めるのが一番大切。熱力学的平衡を仮定すると、化学組成は温度・圧力の関数になる。それによって、組成の測定結果から温度・圧力を導き出す。

Q：岩石の空間的な構造からいろいろなプロセスを推定しているが、なぜ空間的な構造ができてくるのか。

A：反応の進行とともに速度定数等が変わるので、パターンが形成する。温度圧力の履歴などによってパターンができる。非常に多成分・多相系なので、温度・圧力を変えるといろいろな鉱物が生成したり消滅したりする。そういうものを我々は見ている。最終的なパターンから、地球内部の情報をできるだけ抽出したいというのが我々の研究スタイル。鉱物の相の生成や消滅も導入したシミュレーションも一応できるが、今はまだ入れていない。今後は、それも入れようとしている。

Q：圧力－温度－時間で考えるとき、圧力と温度と組成の話は分かるが、その時間の軸、つまりそれが何億年なのか何千年なのかということをどのようにマッピングしているのか？

A：反応速度定数が分かっているものを仮定したので時間軸を求めることができる。私たちは常にある意味、特殊な状態を使っている。岩石の中でも非常にフェーズがたくさんあり、ギブズの相律的に自由度が2まで落とせるサンプルを使って、関係式を出している。基本的には、各瞬間、各瞬間は熱平衡状態であると仮定している。

Q：スパースモデリングでは、それをどういう仮定にするかによってどのパスが正しいかわかってしまう。ブラックホールの「絵」がよく取り上げられるが、あれも100%正しいと言っているわけではない。最も確からしいブラックホールの形はあれだと言っているだけで、それが8割正しいのか99%正しいのか私たちには分からない。岩石においても、ある仮定を置けば最も確からしい履歴はこうだという程度にしか解析でき

ないのではないか？

A：私が岩石学から情報科学に行ったのは、ベイズ推論に感動したためである。ベイズ推論では物事を確率論的に捉える。それまでは、決定論的に「地球の歴史はこうだ」と岩石1つから語っていたが、本当はやはり確率でしか分からない。その確率でしか分からないことは、このベイズ推論を使えば一応定量的に表現できる。そこがこのポイントであり、ベイズ推論を用いると、完璧ではないが、今、人類が最も真摯にやった答えが出てくる。私たちには科学的信念があり、世の中はシンプルであるだろうと考えている。そういう意味で、ゼロが多いことを使うスパースモデリングは科学者の信念と合致する普遍的なモデリング手法だと考えてこの新学術領域研究を立ち上げた経緯がある。

2.6 第1回 総合討論

ファシリテーター 伊藤聡 (NIMS、JST-CRDS)

【ファシリテーターによるトーンセッティング】

私自身はマテリアルズインフォマティクスのプロジェクト MI^2I のプロジェクトリーダーをやっているが、マテリアルズインフォマティクスは、狭い意味では以下のようなものだと考えている。すなわち、計算データや実験データから様々なアルゴリズムを使用して、機械学習モデルを実際につくり学習をさせる。その学習済みモデルを利用して、例えば「バンドギャップがこれこれの程度の四元系の化合物は何か」といった問いに対し、それを満たす化合物組成を求める。これが普通の意味でのマテリアルズインフォマティクスだと考える。

一つの実例を挙げると、私のプロジェクトで、熱伝導率の高い高分子材料をつくるために、高分子の構造と熱伝導率の間の機械学習モデルをつくっているが、ひとたび機械学習モデルができると、もういろいろな構造が次々と出てくる。そして、ここから現実のものをつくる時には、いくつかのフィルタリングをする必要がある。その際に、やはり一番大きな問題になるのが「それは合成できるのか？」という点である。

この問題に対しても機械学習を使うことはできる。具体的には、化学反応のデータベース（「化合物Aと化合物Bから化合物Cができる」という情報）としては、すでに膨大なものがあるので、そうしたデータベースを基にして、逆反応解析（「この物質をつくるためにはどういう化合物を持ってくればいいのか」）を機械学習的に導くことは、現在では可能になっている。しかし、得られた結果が実験的に検証できるかという点、実はこれがなかなか難しい。

高熱伝導率高分子の時には、合成のしやすさは、有機化学実験の専門家に候補分子の分子構造を見てもらって、できそうなものを探した。やはり、現状ではまだ、人間の知見のほうがはるかに優れており、「つくるにはどうしたらいいか」というところはマテリアルズインフォマティクスでは必ず問題になる。

材料開発者にとっては当然のことであるが、ものをつくるためには図 2.6.1 に示すように「何をつくるか」、「どうやってつくるか」、そして「何がつくられているか」、この3つをちゃんと回さないと新しい材料は開発できない。この「何をつくるか」というところが、いわゆるマテリアルズインフォマティクスで、「何がつくられているか」は、最近、計測インフォマティクスと呼ばれるようになってきている。それから、「どうやってつくるか」が今日の主題のプロセスインフォマティクスということになる。

マテリアルズインフォマティクスに関しては、 MI^2I 、NEDOの超超プロジェクト、さきがけなどの研究開発プロジェクトが実施されており、計測インフォマティクスに関しても、JSTの情報計測というプロジェクトもあり、学会としても、人工知能学会の計測インフォマティクス研究会がある。これに対して、現状では「どうやって作るか」に関する部分がちょっと手薄というか、アプローチがまだよく分かっていない状況のように見える。

ここから総合討論に入るが、その前に、もう一点だけ触れておきたい点がデータに関する問題である。

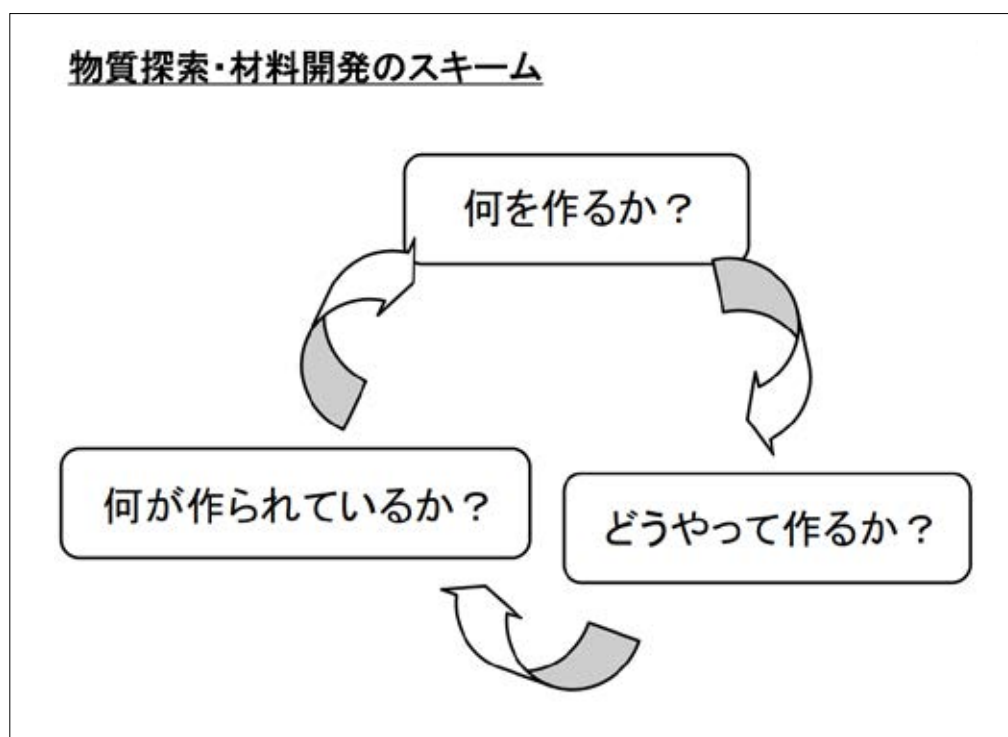


図2.6.1 物質探索・材料開発のスキーム

データに関しては、実験室で日々記録されているワーキングデータから、論文となり出版されるデータ、ハンドブックに収録される標準データに至るまでのヒエラルキーがあるが、最近では、この日々生み出されているワーキングデータもきちんとデータベースに収録しておくことが重要ではないかと言われるようになってきている。今までつくられていた材料データベースは、出版された論文のデータが多いが、そうではなくて日々生み出しているデータが重要ではないかという考え方である。そうすると、この日々創出されるデータに付随して、プロセスの条件、計測の条件を、きちんとデータとして取っておかなければいけないが、この部分に関しては、最初の知京先生の講演にも指摘があったような、様々な課題がある。

また、大量のプロセスデータが産業界に存在するという問題もある。産業界がプロセスデータを開示することはまれであり、この問題をどうするか。それから、プロセスデータの本質的な問題として、鉄鋼業などの産業で使われる製造装置は、多くが一点物で、この装置でのプロセスデータは取れるが、装置を変えたらプロセスデータが変わってくるという問題がある。この My Condition 問題をどうするか。

半導体のプロセスの場合、あらゆる工程で不良が出ないように、企業はしっかりとした工程管理をやっていて、そこでは様々な手法でいろいろなことをやっている。これはプロセスインフォマティクスそのものだといえるが、そのノウハウは絶対外には出てこない。それこそが企業の競争力なので、絶対に出てこないことになる。これをどうするかといった問題がある。

そういうことを踏まえて、「1. プロセスインフォマティクスの現状～何が出来て何が難しいのか?」、「2. プロセスインフォマティクスの将来像～どこまでのことができるのか?」、「3. プロセスインフォマティクス研究開発の課題～技術的課題／制度・体制的課

題」について議論を行っていききたい。

【プロセスインフォマティクスの現状～何が出来て何が難しいのか？】

- ・今、一番難しいと感じているのは、材料合成の全工程に関するシミュレーターをすべて揃えること。シミュレーションがあれば少ないデータでもうまくいくという話をしたが、現状ではまだ揃ってない。それができない間は、やはり人間の知恵を使った実験を常に裏でやっておかなければいけない。
- ・実験家自身が、シミュレーションを最終まで全部自分でつくことはできないため、全プロセスのシミュレーション化は、誰かと協力する必要がある。シミュレーションできない部分を経験で最適化しようとする一年二年では済まなくて、三年四年とかかってしまう。本当は、そこを短くしないと効率のいいものづくりはできないので、そこを何かで乗り越えないといけない。
- ・シミュレーションで物性値を全て正しく求める必要はないかもしれない。簡易なモデルか、関数形（経験式）さえ分かれば、実験と合わせることができるので、必ずしも完璧なシミュレーションを目指す必要はないと割り切れば、シミュレーターは効率よくつくれるのではないかな。
- ・炉の中でも、MBEの中でも、温度の分布推定は、そんなに難しくない。しかしここに流体が入ってくると極端に難しくなる。シミュレーションの正しさを検証するには、データを取りやすい産業界とサイエンスに特化した大学・国研の連携することが、サイエンスのチェック、テストができるいい機会になる。
- ・産業界と強い信頼関係があって、プロセスデータを開示してもらうことができるとすごく役立つと思うが、なかなか難しい。結局、自分でつくったものを評価して自分でシミュレーションしてみたいなことをやらなければならない。本当に産学が共同でやればもっと早く到達できそうなところを、結局何かちょっと遠回りしてしまって学だけでやっている。すごくパワフルで大きいものまで自分でつくってしまおうということはみんながみんなできるわけではない。
- ・これまでの議論で、プロセスインフォマティクスは、他のデータ駆動型以上に産学連携が重要なように見える。あまり企業における事業の機微に関わらないところで産学連携をする必要があると思う。
- ・データ以外の問題点としては、結局、機械学習でもシミュレーションでも、材料開発をやったことのある人が見ないと正しいかどうか分からないというのが一番の問題。データサイエンスの専門家にデータを送っても何もならないというのは、ほとんどの素材メーカーはよく分かっている。
- ・プロセスをやっている本人が機械学習を使えることが大事だと考える。また、そういったトレーニングができる環境も必要になる。サンドボックスというか、お試しできる場、ツールがあったりサーバがあったりして試せる環境が大事。

- ・ 学生さんは本当にカジュアルにプログラミングができ、高度な計算を Python とか Tensorflow 等で行って結果を出してくれる。その辺の敷居はすごく下がっているのも、若者と一緒にやると良い。
- ・ コマーシャルソフトウェアをユーザとして使う上では問題ないと思う。だが、海外のどこかで開発されたプログラムとかアルゴリズムとかが使えるだけで、自分用の機能を自分では付け加えることができないというので良いか。それを日本でやらなくていいのかどうか。例えば第一原理計算、昔は国内でみんな一生懸命つくっていたが、今はやめてしまっていて、全員 SIESTA とかそういうものになってしまっている。流体だと Fluent とか。本当にそれでいいのだろうか。
- ・ プログラムをすべて作るとなると、Gaussian のような大変複雑なものに至るまで、自分たちでつくらないといけないことになるのか。
- ・ そこは、HPC の分野の人の PR 不足だと思う。日本独自のアプリ、例えば Gaussian なんか全然足元にも及ばないような分子軌道法のプログラムは日本にもあるし、今でも使えるような状況になっている。他にも、そういうソフトは多数、国のプロジェクトで作られて、それはかなり使えるようになっている。ただ、それが分野外の人にはほとんど知られていない。
- ・ 専門家の作るプログラムは、複雑すぎて分野外の人に使いこなせないのではないかと思うが。
- ・ 国としては手を打っていて、アカデミアがつくったものに関してはソフトのベンチャーに渡して、そのベンチャーに GUI をつくらせて使えるようにするというのも幾つかある。ただ、アカデミアがつくったものすべてにはできないので、当然売れ筋のものになる。構造解析などのユーザが多いものは良いが、結晶成長関連のようなユーザが少ない分野ではベンチャーも成り立たない場合が多い。
- ・ Purdue 大の nanoHUB という拠点は、ウェブベースでいろいろな計算をやるサイトだが、そこに自分たちでつくったソフトも入れることができ、GUI も向こうのスタッフがつくってくれ、意見交換の場があったりもする。ソフト作成者とユーザがバーチャルでミーティングできるという形の出会いの場と、自分のつくったソフトのテストの場のようなものをプラットフォームとして持っていて、それが全世界でユーザが 150 万人ぐらいになっている。これは NSF が 10 年間にわたってサポートすることになっているが、サポート終了後、将来的に有料化しても、nanoHUB はスタートアップとして独立できると思われる。日本でも NSF のような形でサポートすればベンチャーも成り立つのではないか。
- ・ アメリカにはそういうものを面白がって参加してくる人が多いが、日本人はすごくまじめというか少し引いて見ている。「これは使い物になるかな」みたいに。特に産業界の人はその傾向があって、国民性の問題かもしれないが、同じことをやろうとするときに結構問題になるかもしれない。
- ・ データとソフトウェアの話が続いているがハードウェアに問題はないのか？ 現状では、一部だけをシミュレーションして残りを機械学習でやっているが、ハードが早ければ全部シミュレーションできるのではないか。

- ・ハードウェアが発展することはウェルカム。ただ、シミュレーションが早くなっても、機械学習はさらにそのシミュレーションの100倍速いものができるので、組み合わせによって、同じ時間により高度のことができるようになる。
- ・計算機以外のハードはどうか？例えばセンシングなどで課題はないのか？
- ・X線の構造解析とラマン散乱の測定を同じ試料に対して行って、何らかの対応がつけられないかの検討を行っている。構造上X線回折が行えない成長装置内にラマン測定系を構築して、プロセスパラメータをあつめられないかと検討している。
- ・実際に成長しているときにいろいろなことが、シミュレーションでしか現状見られないようなことがその場で計測できるとすごくいいと思うが、そうでない場合、シミュレーションにつながる何らかの計測を同時に行うのが良い。
- ・流体の計測に関しては、以前に結晶成長学会で精力的に行ったが、あれで数値計算が正しく行われていることの検証ができた。今は、もう数値計算のほうが正しい結果を与えらると思う。実際、航空機設計のためのデータは、現在では、風洞実験よりも計算の方が信頼できるようになっている。他のシミュレーションも、ちゃんと計算をやっていけば結構将来は明るいと思われる。実際の系の測定結果をシミュレーションにフィードバックをかけることで信頼性を高めることもできる。
- ・プロセスというのはいろいろなデータを取っておかないと、一点物の実験もあるので後で取り返せないということもある。可能な限りそのときに取ったほうがいい。そう考えると、無駄かもしれないけれども取っておくというのは、必要なのではないかなと思う。
- ・結晶成長をやっていると成長装置の部材の劣化（経年変化）、例えば断熱材の熱伝導率が変化していくようなことが起こる。装置パラメータを固定したシミュレーションは、それだけで全然合わなくなってしまうが、熱伝導率を推定できればシミュレーションの補正が可能である。熱伝導率そのもののモニタリングは難しいので、測定可能なデータから熱伝導率を推定するソフトセンサー的な発想が有効である。例えば、断熱材からの光の透け具合や、装置全体のサーモビューアー像とかもすべて取っておくことで熱伝導率の変化が推定できる。プロセスにおいては、このような経時変化を管理することはものすごく重要である。
- ・今、一番重要だと思っているような、ヒーターのパワーや炉の周辺の温度といったデータ以外のデータも取っておいたほうが、後々何かの役に立つかもしれないという話だが、実際実験とかプロセス、製造装置をつくるときにそこまでできるものか。どのぐらい取っておけば良いとなるのか。装置周りの全部の温度を取っておいたほうがいいのか……。
- ・それはリソースとアイデア次第。
- ・十分制御できていないデバイスプロセスなどで、「どうも雨の日は出来が悪い」などと後々気付くことがあって、何で湿度のデータを取っておかなかったのかと後になって思ったことがあるが、実験した時点では、「半導体プロセスをやる際に外部の湿度を取らなければいけない」という発想はなかったのではないかなと思う。逆に言うと、プロセスインフォマティクスをやるときに、「実験室の中のこことここ、ひょっとしたら床の振動が重要かもしれない」とか、そこまで取っておくというのは難しいのではないかな。学理とまではいかなくとも、「こういうものは測っておいた方が良い」というのは基本的にできるだろ

うか。

- ・過去何十年間でそういう知識が人間にあるわけで、まずそういうものを何かの形で顕在化というか共通化というか、そういう形にしておく、他の方々も参考にできる。それを一から学ぶのはしんどいし、それがあるかもしれないからといって可能性があるデータを全部取る、それは何かオーバースペックとしか思えないので、逆に言うとそういう、属人的な情報をどう形式知化するかは重要な点だと思う。
- ・自分の研究室での反省点は、学生が行った「これこれの条件での計測結果」というようなメタデータは、ノートや個人のエクセルにしかなかったりする。形式が全然標準化されていないため、過去のせっかくの成果が有効に活用できない。データベースになっていけばいろいろな解析ができると思うが、そういうところできていない。
- ・まずは、メタデータの標準化だけは最低限必要。

【プロセスインフォマティクスの将来像～どこまでのことができるのか？】

- ・これまでの発表の中には、非常に賢いシミュレーターができて、その中で自動的にいろいろなことができるという将来が見えたが、その先にはなにがあるのか。
- ・とにかくシミュレーション屋さんは、ずっと突き進んでもらうのが一番将来像に近いと思う。シミュレーションのスピードの遅さを補っているのが機械学習だというのが自分の認識なので。
- ・真空装置メーカーの人と話をすると、「プロセスインフォマティクスは自分の頭の中にしかない」といったことを言われる。パラメータが山ほどあって、そのノウハウの蓄積みたいなところがあるので、それをいかに標準化というか、汎用化していくかが重要なと感じている。薄膜の場合は、もう分圧の圧力1つで応力がどんどん変わるし、再現性を取るにも常に、いろいろなモニタリングをしていないと、1週間前と今日とで条件が変わったりするという人もいる。そういう中でデータを取っていく。やはり何でも取って関係あるものは全部デジタル化しておくことが必要。
- ・そういう意味で言うと、実際に作る場所から計測まで、ぐるっと回すというマテリアルファウンダリーが、いろいろな分野でできるのが一つの将来像になるのか。
- ・IMEC がそうなろうとしている。一生懸命ほかの分野を入れて。やはり拠点化というのはどの国でもある程度進めている。
- ・とはいえ、何でもできるということは何でもできないということの裏返しでもあるので、例えばバルクから液体から何でもというわけではなくて、例えば今、一番日本が必要としているようなところ、例えば薄膜なりチップコンデンサなり、例えばミクロンクラスの膜ができる装置とか、そういう幾つかのいいショーケースをつくってみる必要がある。
- ・そうなってくると、人力で行うよりは、東工大の一杉先生のようにロボットで、役に立つかどうか分からないけれども、とにかくデータだけは測定しておくというのは、やはりこれから必要なのではないか。それを人手で取っていると大変なので、やはりロボットで取ることがかなり重要で、労力をかけずにデータを取らないと無理になる。そうい

- うプロセスデータを取るところでロボティクス技術は非常に重要ということになる。
- ・ 今、ロボティクスは、10年前に比べて価格がすごく安くなっている。Python ベースでコントロールできるもので、昔は1,000万円したものが30万円ぐらいで動かせる。ハードウェアの価格が下がって使いやすくなっている。なので、今後の材料科学実験にロボティクスというのは重要だと思う。
 - ・ 今日の話は比較的要素の数も少ない単結晶の話が多いが、これが、「多元系の単結晶で、しかも固体だけれどもボイドのような高次構造もあって、それが機能の元になっている」といった材料にまでこのプロセスインフォマティクスの手法が使えるようになるのか。やはり、少しずつ難しい問題に挑戦していくことになるのか。
 - ・ 機械学習的には、ステップ・バイ・ステップで進んでいくわけではなく、最初からできそうもないところとできそうなところをくつつける。だから機械学習でやろうとなる。つながった後、人間が見て解釈できるものはしていく。ステップ・バイ・ステップでいくなら多分、インフォマティクスはあまり要らないという考え方もある。
 - ・ 似たような例で、フィルターをシミュレーションしようとしたとき、実際のフィルターの迷路のような複雑な流路構造をモデル化するのではなく、「フィルターのような構造」を自動生成するプログラムを作って、シミュレーションしたことがある。シミュレーションデータの自動生成というのも一つの方法になる。普通にやろうとすると、X線CT等で測ってそれをデータにすることになるが、時間のかかるX線CT測定や、X線CT用のサンプル作成に大変時間がかかる。そこで「きつとこういう形になるに違いない」という似たような構造をコンピュータの中で大量に自動生成する。似たようなサンプルを1万個つくってしまうみたいなのができれば、ボイドのような複雑な構造もその中に出てくる。
 - ・ 現在の、材料科学への機械学習の使い方というのは、まだちょっとプリミティブだと思っている。材料科学に機械学習を適用しようとして、10次元とかになるとパラメータの数ばかりやたら増えてしまっていて大変になってしまっているのが現状だが、自然言語処理の分野をみるともっとパラメータが多いのに、それをうまくパターン認識してやっている。自然言語処理では、結構距離の離れた前後の関係とか、すごくローカルな構造に特徴を持たせて解析したりしている。あそこら辺の発想をもうちょっと入れていくと、そういう多次元みたいなものにもっと適用できるはずだが、そういうアプローチは現状そんなに見たことがない。
 - ・ 今の話は非常に重要なところだと思う。指摘の通り、言語に関しては今、急速に機械学習の技術が進んでいて、ものすごく翻訳の精度が上がっているが、あれは別に構文解析なんてやっておらず、単にパターンマッチングをやっているだけ。だから、そういうものを使ってプロセスで「何が起きているのか」ということを学んでおいて、「こんなやつ」と言ったときに最適なプロセスを出してくれるというのも、一つの方向になると思う。
 - ・ ただ、複雑なものにいかうとすると、複雑なものをそのまま「どうモデル化するか」という技術をもっと高めていったほうが良い。

- ・複雑系をそのままステップ・バイ・ステップでやるのではなくて、まずはアーティフィシヤルなものでいいから解析できるものをつくる。それをどう実現するかは、それぞれ研究者のアイデア次第というその考えは、大変面白い。
- ・アメリカでマテリアルズインフォマティクスを引っ張っている Cedar さんが NIMS での講演で強調していたのも、くしくも自然言語処理。今、たくさん出ている論文を集めて、テキストマイニングでプロセスのデータベースをつくる。そうするとプロセスについていろいろな像が浮かび上がってくる。あとそれを引っ張り出すのはインフォマティクスだと述べていた。業界を引っ張っている人がそう言うんだから、ある意味で極論ではあっても、そういう方向もあるんだろうなと認識した。
- ・その一里塚的な例として、シュプリンガーから出た「Lithium-Ion Batteries」という本がある。ベータ・ライターという人が書いたとされているが、実は、リチウムイオンの電極に関する約 5,000 を超える論文を機械学習でマイニングして、それで機械が全部要約をつくって、例えば「リチウムコバルトオキシサイド」というと「こんなことが知られています」と半ページぐらいにまとめて書いている。これが、すごく分かりやすい。日本人が書くよりも全然よく分かる英語を書いている。そういう、カタログをつくるみたいなことは、今、機械学習の自然言語処理でできるような状況になっている。ベータ・ライターの「ライター」は書く「writer」で、もちろんペンネームというか、シャレというか、そういうもの。
- ・Cedar さんが言っていたことをさらに言うと、そのときに、最初の原料は何なのか、最終的な生成物は何なのか、途中でどんな副生成物が出てくるのか。それからプリカーサーみたいな途中の構造みたいなものも含めて、そういう情報を全部引っ張り出すんだ、そういう感じのスピーチだった。
- ・そういうことをやっている人たちは、材料を研究している人たちなのか、それとも、データ処理を研究している人たちなのか。いろいろな研究機関で、近年データアナリティクス専門にやっている人と、材料やマテリアル、デバイスをやっている人が、一緒にやっていたりもするが、その境界にいる人間のスキルが問われるところがある。材料屋は、データアナリティクスの手法自体は開発できない。データ科学者のやっているものは高度過ぎて材料屋には全くいじれるようなレベルではない。先ほどから出ている、機械学習の高度な利用は、何となく材料屋の領分を超えそうな気がするが、そうでもないのか。
- ・そこは、歴史的には図書館学の人がいる。最近の図書館学ではデジタル・ライブラリといって、図書館をデジタル化するという動きがあるのだが、そこに書誌学の人がそれぞれの分野ごとに入ってきていて、例えば、NIMS などでは、そういう図書館学をやってきた人が材料学を勉強して、材料学向けのデジタル・ライブラリをどうするかというところで、今のように企業のデータをどう構造化していくかを検討している。そういう風に、分野ごとにやっていくのが多分一番いいのではないか。
- ・図書館学の専門家と材料の専門家がいるようなところでそういうものが、フレームワークとしてつくられているわけか。
- ・NIMS の場合は、図書館学をやってきた人間が材料学を勉強して、材料の専門家といろいろ議論しながらそういうデータベース、こういうものがあつたら今、言ったグラフに

したほうが使いやすいよみたいな議論をする。

- ・それはさっきどこかで議論があったように、まさに今回、そういうマルチディシプリンな人材が求められている。Cederさんは決して情報学の人ではなくて、彼の講演の中に盛んに物質の名前が出てきて、バッテリーの話も出てきて、こんな材料を合成したと主張している。その人がプロセス設計の世界に入ってきて、情報科学と物質科学の両方を理解しながら進めている。

【プロセスインフォマティクス研究開発の課題 ～技術的課題／制度・体制的課題】

- ・ちょうど議論がそこに来たので、拠点の話をしたい。拠点が大事なのは、最初にも出たように、プロセスデータは一点物になりがちで、これを何とか標準なり汎用データにしたい。一点物が標準になるというのは、その拠点、例えばIMECのようなところでは、そこで一点物のデータを取っても、実はそれがみんな来ると皆が共有する標準データになっている。産業界は絶対データを出さないし、産業界が必要なのはデータではなく、データの解析の方法であろうから、1つは何かお手本を示すというか、拠点の内製データを使ってお手本を示して寺子屋風にやって、そこで普及させていくというのも当然あると思う。
- ・実験してデータを取るところから、プロセス結果のデータベースをつくるところまで、全部がその1つの拠点の中に入っているのか。
- ・IMECの場合、1回そこで実験をするとその実験データはIMECに落ちていく。それで、IMECを使ったんだから当然そのデータは次に来た人が参照してもいいねと。嫌だと言ったらIMECの使用料がバーンと上がるみたいな、そういうビジネスモデルはすぐに考えられると思う。
- ・これは、例えば半導体だったら半導体の結晶成長装置や成膜装置があったり、露光装置、エッチング装置があったり、他にもいろいろ揃えることになるし、他のプロセスならまた別の装置群が必要だし、色々考えると、工業地帯をつくるぐらいのレベルの巨大拠点にならないか。
- ・Chris Andersonのメーカーズ的な発想の、コンパクトな拠点を作れば良い。自分が材料を開発して、この材料が本当に電子デバイスとして使えるかどうかというときに、テストする場所が今のところないのが悩み。ナノテクプラットフォームでできることにも制限があって、完全にチップができるわけではない。そういうときに、ミニマルファブは結構魅力的に見える。ミニマルファブみたいな一つのメーカーズがあって、マテリアルファウンダリーのようなメーカーズがあって、その中には幾つかの装置と標準化されたようなものがあって、それらが、バーチャルなり、必要に応じて共同して物を作る。茨城の北に行くと、発注するといろいろな小さい会社が集まってバーチャルな会社をつくって対応する、ああいう形が面白いのではないかと思っている。
- ・電子デバイスだったらミニマルファブで組む、あるいはバッテリーだったらNIMSの電池拠点と組むとか、あるいは何かだったら産総研の何とかと組むとか、場合によっては他の会社のそういうところと組んでしまうとか、そういう小さいメーカーズがバーチャ

ルにつながっている。アルミなら結晶成長のメーカーズがあるというのが日本的でいいのかなと。あまりコストもかからない。

- ・プロセスインフォマティクスができるプラットフォームが置かれるような拠点があるとして、それを実際にデバイスなりにしていけないといけない。そこまでできないと駄目。電池だったりセンサだったりのデバイスにしていこうとすると、これは分野ごとによって、大分違うから、それぞれの応用分野に携わるところとアドホックにネットワークができるような仕組みが提案できるのではないかと思う。
- ・その、プロセスインフォマティクスの拠点には何が必要なのか。
- ・製膜装置と評価装置と、それからデータを蓄積するためのサーバ、そこにデータサイエンティストがいるという形かと思う。IMECのように、データをそこに落としていってくれる形を整えることが重要。IMECは、昔は、ナノテク関連の学会などに参加して、世界中に広告を打っていたが、今はやらなくなっている。今は、待っていればみんな「使わせてくれ」と来るから。だから、自分のところにプロセスデータがどんどん落ちていくまで発展していけばいいのではないか。
- ・どうも汎用のものを考えているようだが、やはり目的に合わせて装置はかなり違うのではないか。磁性体の強い磁石を作るのと半導体のデバイスと電池の正極とでは、用いる装置が違う。それを1つのものというのは無理ではないか。
- ・装置をどう整えるかは別の議論が必要。重要なのはみんなが使える装置がそこにあるということ。自分たちが何億円も出して装置を買うのではなくて、その拠点に行けばデータなりみんなが使った装置があって、コンタミもあるかもしれないけれども、取りあえずデータが取れる。そして、そのときのプロセスデータも取れる、そういう拠点が1つあるといいのではないかということ。
- ・それは1か所でやる必要はなくて、複数箇所でもいいと思う。ナノプラは、バーチャルネットワークを組んでやっている。ただ、プロセスインフォマティクスの拠点はすごく重要で、やはりそこにデータが集まってくることが大事。
- ・バーチャルネットワークは良いが、共通化してほしい。基板とか。ナノプラの最大の問題点は、基板サイズが違うということ。アメリカでは、バイオをやる人も半導体の人も、少なくとも何インチと決めて、装置間はシャトルでウェハが飛んでいった。あれが今、日本ではできない。ある大学は10ミリ角だしあるところは3インチを使うし、この装置は4インチだしというふうになっている。
- ・ポストナノプラを考えると、データを取るものと、それからインフォマティクスをセットにしたようなものを考えていくのが1つの考え方だろう。
- ・政策立案側の視点に立って、冷静な意見も聞いておきたい。個々の事例として「この人のところでは、このプロセスで、こういうことがちょっとできましたね」というのはできるような気がするが、やはり国家的視点からすると、「拠点化してもそんな簡単にいかない」のではと感じてしまう。何かこういう方向で頑張ればいいのではないかということころは割と言えるが、現実的にはそううまくいかないかもしれないとの危惧もある。
- ・誰かにデータを出せというのは、私はもう考えないほうがいいと思っている。少ないデー

タでもできることをしつこく言い続けたのは、もうデータを出すというのはほぼ不可能だと思っているということと、あと My Condition 問題で、データを出し合ったとしても結局使えないという答えになると思っているので。一番いいのは、やはり百個二百個のデータでもその技術が発展する方法をつくっておくことが一番冷静なやり方と思っており、その1つがシミュレーションだと信じている。まず少ないデータでシミュレーションを活用して、今までの10倍の効率をつくっておいて、それを見て誰かが「シミュレーションさえつくればうちの分野でもこれが可能になるのだ」ということを、ロードマップとして構築するのが一番美しいかなと思っているし、それなら今の自分でもできている。

- ・データがないというのはもう現実問題なので、データベースを探してもないケースが結構ある。特に特に多元系になるとほとんどゼロ。とするとデータは、つくるしかない。そのときに、やはり何かきっかけとして数千個でもいいデータを、例えば計算科学で加速してつくることが必要。そこは、やはりデータがないと機械学習も何も扱えないので、多くても少なくともデータは絶対必要なもので、自分が関心のあるところの周辺のデータをつくることが重要になる。
- ・自分の実体験では、「100個のデータで、何でできるのか」とよく聞かれる。「少な過ぎる」と。この感覚は、ビッグデータの人たちにはあまり持てないと思うが、100個に1個間違いのデータが入ると、機械学習をやっても精度が非常に落ちる。結局、材料研究者は100個を1,000個にすることはできないので、100個の中にとんでもない1個が含まれていないかという検証コードを書いていて、それを除いた途端に精度が著しく上がる。だから、そちらの発想のこともやっていかなければいけないかなと思う。
- ・今の話はデータの精査という問題で、とても大事。バイオでやられているデータ精査は統計学的にできるが、スモールデータの場合それができないので、新しい精査の方法を考えないといけない。これは非常に大きな問題。
- ・もう一つ議論しておきたいのは人材の話。先ほどから、いろいろな、マルチディシプリナリなような人材が必要だという議論があったが、とはいえ、最初から、若いときからマルチにすると、誰かが言っていたように、根無し草になって、「あなたの専門性は何」と言ったときに困ってしまうというのもまた問題。日本はなかなか、ダブルディグリーが定着しないとかそういう議論もあるが、そこに関して何かないか。
- ・それはすごく単純で、マルチディシプリナリなリーダーみたいな人をきちんと何人か見つけてきて、その人のところに置くのが一番いいと思う。実際そういうマルチなリーダーが、本当にダブルディグリーだったかというのと、決してそうではなくて、最初は何かの専門家なんだけれども、その人の個性もよるが、だんだんその人の興味でマルチになっていっている。だから私は別にダブルディグリーなんてやる必要ないと思っていて、やはりこの世の中には特殊な人がいて、何にでも興味を持つ人がいる。そこはもう本当に属人的で、そういう人をきちんと見つけてきて、そこにいろいろな根を持った人たちを集めてその人の真似をさせるというのが一番いい。
- ・卓越教授みたいなものを幾つもつくって、それに弟子入りさせるような場所を作る。そ

こに、やはりマルチのできる人を探してきて確実に置くということが重要。それは一人二人で十分かなという気がする。

- ・古典的な言い方になるが、やはり「背中を見て育つ」ものだから、まず背中を見せることが重要。普通の大学の1つの研究室に閉じ籠もっていたらその先生の背中しかないから、その先生がマルチでない人だったら絶対マルチにならない。
- ・その1つは、大講座制みたいにするのが大事ということになるか。
- ・特にインフォマティクスとか数理科学の人は、多分、緩い連携が必要。例えば、東北大学のWPI-AIMRでは、数学を使って物質科学をやっているが、あそこには、それぞれ物質科学研究分野はあるが、数学者はどこの分野にも属さないで所長直轄になっている。それでどの会議、どの研究会でもいいからぐるぐる回って、ぐるぐる回っているうちに面白いものにぶつかったらそこで研究をやれ、という仕組みでやっていて、それはそれで面白いやり方。
- ・日本の大学もできれば有機的な大講座というか、国研も含めて何かそういうものがあると、私はまだいけると思う。
- ・融合的なところをつくるのは基本大賛成だが、実際、大学でどの部局でも融合的な知識体がつくれるかという、やはりそうでもないと思う。特に学部というのは教育組織であり、そうすると、電気工学、機械工学、化学といったディシプリンが決まっている。学部学生には決まったディシプリンを教えないといけないので、そこをいきなりマルチにするのは難しいと思う。先ほどから話の出たWPIは教育組織ではなく、横串型の研究組織なのであれができています。大学も附置研なら教育組織ではないので、混ぜてしまえる。
- ・具体的には大学の附置研のような非教育組織、あと国研。これを色々混ぜてしまえばよい。昔、産総研の前身の工技院に融合研というのがあって、混ぜてしまえというのがあった。ただ、現状は「混ぜてしまえ」と言っても、まだそれを本格的に認めているという状況ではないし、大学でマルチにやってしまうと、例えば科研費1つとったって細分化していて「どのカテゴリーに入りますか」になってしまうので、そこでしかお金が来ない構造になっていて、マルチになるとなかなか評価も難しくなってしまう。実際は、マルチディシプリンの組織運用を「やれよやれよ」と言われても、具体的にそれを本当にやれる組織は限られているのではないかと思います。
- ・普通の学部では、学部教育を行っているため、基本的なところはやらないといけない。なので、いろいろなところを混ぜるのはなかなか難しいと思う。うちは学科が3つの専攻に分かれてしまっていて、分け過ぎだろうと思っており、いろいろな先生の背中を見たいというのがなかなかできない。結局、専攻の中に教授の先生も数人しか実質的なことになってしまったりもしている。
- ・人材育成に関して、トップサイエンス的な研究のための人材的なところと、メタデータのとか、標準スキルの持つておかなければいけない人たちを育てる人材というのは大きく分けるべきではないかと考える。また、国が横串的に支援をしたら、メタデータの人材を増やしていく点では、非常に重要だろうと感じた。その場合、メタデータ系の人には、当然根っこがないと駄目で、結局それはマテリアルのことをきちんと勉

強して、根っこがついている人がそのスキルを身につけることが重要だと。そのために必要なものとしては例えば何なのか、期間的にはどれくらい必要なのか、そういう議論があると嬉しい。

- ・メリーランド大学では、学部生全員に対して機械学習が必修になっている。これは、全学部対象で、例えば経済学の人たちはそれを経済指標の解析に使うし、文学系の人々はテキストデータラーニングで、例えばアメリカの作家の文章のスタイルとか、日本で言えば夏目漱石風の文章、偽文章と本物があつたときに判断できるかとか、法学部の人たちは判例の分布を見るとか、そういったところに使っているらしい。そういう基礎的な機械学習のスキルは今後誰でも必要で、日本でも、東工大が、全学部必修か何かにしたと聞く。ああいう試みはボトムラインをつくるという意味では重要。
- ・人材に関して考えておきたいのは、今、言われたようなメタデータの的な人に関するキャリアパスを何とかしなければいけない。研究者のキャリアパスはあるのだが、研究者以外のキャリアパスが日本には極めて脆弱であることを違う分野でもすごく感じていた。
- ・装置的な視点からの意見はないか
- ・今日の話は、加工等にも展開できるかと思って聞いていた。今日のお話を伺っていて自分のところでもやってみようかなと本当に思ったぐらいで、楽しく聞いていた。ただ、自分の経験で言うと、装置の経年変化等もあって、それにどうやって対処するのかというの聞いていて感じた。あと、同じMBEでもA社とB社で同じ条件でやっても違うものができるが、それもちょっとケアしたほうがいいかなと感じた。

【オーガナイザによるクロージング】

今日は第1回目ということであったが、我々としては、プロセスインフォマティクスは日本がこれから材料の分野で勝ち続けるために非常に重要なテーマだと強く思っており、この後、2回目も行う予定である。また、今日は俯瞰が目的だったが、何とかこれを国のプログラムという形で結実できたら、そのときにはどういう視点でつくっていくべきなのかということも考えていきたい。

今日は本当にたくさんいろいろなキーワードが得られたし、この分野の状況が良く理解できて非常に勉強になった。これからこれをまとめて、その次のステップへとまだ道は長い、我々としても頑張りたいと思う。今後も、御協力をお願いしたい。最後になるが、本日の総合討論の場で、中心となって議論頂いたワークショップの話題提供者の皆様方、また貴重な意見を多数頂いたコメンテータの先生方に深く感謝したい。

3. 第2回ワークショップ（触媒・有機材料分野）

3.1 触媒インフォマティクス：材料組織とプロセス条件を考慮した触媒設計

高橋啓介（北海道大学）

触媒設計においてどのようにインフォマティクスを使っているか紹介する。

最初に従来の研究とインフォマティクスの違いについて述べる。従来の研究では、最初に人が物をつくる。実験ではサンプルを作り、計算ではモデルを作る。次に、実験では様々なキャラクターゼーションを行う。電子顕微鏡で観察したり、XRDで構造を解析する。計算化学であれば計算結果から物性や反応が分かる。以上の過程を試行錯誤的に繰り返し、いいものが見つかったらそれが科学的な発見になる。これが従来の研究。一方、インフォマティクスは進め方がこれと全く逆になる。我々人間が欲しい物性、反応を提案し、機械もしくはデータ科学の手法が「こういう材料がいい」、「こういう実験条件がいい」、「こういう反応だ」と、答えを教えてくれる。我々が欲しいものを要求して答えが返ってくるという、今までとは全く逆のプロセスがインフォマティクスでは期待されている。

触媒反応や化学反応もしくは材料設計に関して、我々が実験で観察しているものを「森」に例えると、理論化学、計算化学では森の中の「木」一本を見ることに例えられる。計算で一本の木を完璧に理解できたとしても、森がどのような形をしていて、どのようなエコシステムが形成されているかという全体像はなかなか見ることができない。それに対してデータ科学はそういう詳細を見るのではなく、森全体を見ることに主眼が置かれる。木の一本一本は分からないが、関係する温度や圧力といったいわゆるプロセス条件、それらを全部考慮した上で見るので、全体の何らかの傾向が見えて、それが設計につながる。それがデータ科学の正しい使われ方だと思う。

2010年頃に材料と触媒の両方でインフォマティクスという動きが始まった。マテリアルプロジェクトではクリスティーナ・ピアソン、触媒インフォマティクスではノロスコフ、この2人が中心となって推進した。彼らは計算化学の専門家だったので、ハイスループット計算でデータを用意し、データ科学を適用して触媒や材料を見出そうというのが最初の動きだったと思う。実際に革新的な成果があがっており、機械学習を使うことによって化学反応経路探索の計算が非常に高速化できた。

似たようなことが材料科学でも起きている。米国の Materials Genome Initiative は2008年、日本の MGI が2015年に発足したが、両者とも、最初には実験データがないため、計算データをたくさん集めて何かを生み出そうとした。去年あたりまでの成果をまとめてみると、データ科学と計算化学は非常に融合がうまくいくことが分かってきた。しかし、良い材料の候補や大量の情報はたくさん出てくるが、どれも仮想的な情報なので、実際にそれができるかどうか分からないという問題に、今、我々は直面している。

その大きな理由は、プロセス条件の情報が欠如していること。いかに理想的な反応、理想的な材料、もしくは触媒が見つかったとしても、それを実現するプロセス条件がわからないと全く合成できない、全く反応が起きないということが起こる。実際の材料設計、化

学反応設計はプロセス条件あってこそだということを痛感している。

こうした問題意識は米国の研究者も持ち始め、材料や触媒をどんどん合成できるインフォマティクスの研究に米国はシフトしている。ハイスループット実験、ハイスループット計測に主眼を置いた研究にシフトしている。

我々が進めているCRESTプロジェクトでは、触媒インフォマティクスでメタンを直接的にエタンやメタノールに変換させる研究に取り組んでいる。自分たちでプロセス条件も含めたデータを用意し、データ科学や機械学習その他の方法を使いながら予測して、自分たちで実験実証している。この後の谷池先生の発表で、実験についてより詳細に説明いただけたと思う。

最初に取り組んだのがメタン酸化のカップリング反応。メタンと酸素を衝突させてエチレンと水を作る反応だが、関連する過去30年分の文献や特許情報を約1,800件集めて、解析を行った。当初我々は触媒組成だけに注目していたが、エチレンの収率を上げるために重要な記述子が何かを評価したところ、触媒組成だけでなく、実験条件がそれと同じくらい大事だということが見えてきた。

例えば触媒調整方法が非常に大事だという解析結果が得られた。いい触媒が見つかっていても、触媒の調整方法1つで全く駄目になったり、もしくは全く違う方向に行ったり、逆に計算で駄目だと思われていても調製方法次第でかなりよくなるということが起こる。さらに、反応の温度、メタンと酸素の圧力、触媒とガスの接触時間というようなプロセス条件も非常に大事であることが分かってきた。これは触媒インフォマティクスに限ったことではなく、材料でも化学反応でも、原料とプロセス条件が同時に最適化されないと目的の物質は生まれないことがデータ科学からもここ3年ではっきり見えてきた。

1,800件の過去の文献データだけでは、一貫性やプロセス条件が欠如している部分が多かったのも、次に発表される北陸先端大の谷池先生に独自のハイスループット触媒実験装置を作製いただいた。この装置は20個の触媒を入れて同時にいろいろな条件で実験でき、1日に数千件の実験ができる。このハイスループット装置で得られたデータを使うと、機械学習が非常にうまくいく。過去30年の文献データ約1,800件に対して機械学習の手法を使ってもうまく予測できない。一方、ハイスループット実験で取得した12,000点のデータでは、回帰が非常にうまくいく。ハイスループット実験ではプロセス条件も含めた均質で一貫性のあるデータが得られるので、ものづくりが大幅に加速することが分かった。ハイスループット実験とデータ科学は相性がよく、実際の材料設計に役立つという感触を持っている。

12,000点ほどのデータは1週間程度で取得できる。今は5万件程度集まっている。この均質なデータに対して機械学習による予測を行うと、C2収率がどれくらい実験条件によって変わるのかを機械学習で予測できる。例えば、ある特定の条件のときに温度を変化させたときとメタンと酸素の比率を変化させたときとで、C2の収率がどれだけ変わるのかがわかる。機械学習の結果、温度800℃程度、メタン／酸素の比率が3程度で、最適点があることが分かった。従来の実験方法で検証した結果で、機械学習の予測と一致することが確認できた。

触媒インフォマティクスの敷居を下げるため、誰でも簡単にできるプラットフォームも我々は作成している。ウェブ上でデータの公開、管理、利用ができる。今後はよりオープンソースの時代になると思われるので、自分が持っている材料や触媒のデータを公開し共有できるようなものを作っている。クリック1つで対象データを可視化することもできる。散布図をつくると、同時にエクセル形式でデータを出力することもできる。

機械学習にはプログラミングのスキルが必要だが、使いたい機械学習の方法をセットして記述子を入力するだけで実行できる環境も提供している。このようなプラットフォームは、今後この分野を発展させるには必要ではないかと思う。

今後この分野が広がっていく上で一番の障壁は、研究者人口の少なさにあると思われる。インフォマティクスは自然科学と情報科学の両方にまたがる。自分でプロセス条件を含めたデータをつくり、データを扱う能力が必要。最先端のデータ科学技術を使いこなし、さらにデータから変換した知識を科学的に理解できる人が今は圧倒的に少ない。人材育成環境も整備していかないと、この分野はなかなか発展しないと危惧している。

計算化学先導型は米国が既に断念しているので、よく考えながら進めないといけない。合成ができるインフォマティクス、プロセスデータを含めたデータ取得環境も整備していかないと、プロセスインフォマティクス分野の発展は難しいかなという印象がある。

【質疑応答】

Q: 機械学習を通じて、結果を予測する理論に到達することはできるか。

A: 予測された結果を統計的に解析してみると、例えばこの組成では高温で絶対出てくるといった傾向が見えてきたということはある。

Q: マテリアルズインフォマティクスでは第一原理計算が中心に据えられていたが、他にも動力学計算とかいろいろなシミュレーション手法を使ってもプロセスインフォマティクスにはあまり効果がないということか。

A: 触媒の調整方法が非常に重要で、同じ組成のはずなのに調整法次第で全く違うパフォーマンスが出てきたりする。そのあたりの予測は、計算化学では難しいのではないかという印象を抱いている。

Q: プラットフォームでデータを公開するという話をされていたが、このデータにはプロセスデータも入れるのか。

A: 上げる人が決めることだが、ここに示した例では、温度や収率など、全部含まれたデータをたくさん上げている。

Q: ハイスループット実験で予測できたとしても、過去の文献にあるようないろいろな条件、ほかの条件の何が効いているのか、どこが重要なのか、そういう再考察ができにくくなるのかなという印象を持ったが、その点はいかがか。

A: ハイスループット実験のデータは、ほかの実験装置とも整合性があるデータなので、

誰がやっても同じような結果になると思う。文献データのほうが予測できなかった1番の理由として、時代による実験装置の違いが挙げられる。また、特に特許のデータに言えることだが、触媒やプロセスの条件だけでなく、実験装置自体をかなり特殊に改造しているものが多い。実験装置に関する条件が異なるものがたくさん存在していることが原因で、うまく行かなかったのではないかと解析して思っている。

Q: インフォマティクスの場合、因果関係を導き出すのが非常に難しいのでは。

A: 因果関係を導き出すのは難しい。私たちの今回の目的は物質を見つけることなので、見つかった後で、キャラクターゼーションの専門家が因果関係を説明してくれば良いという考えで取り組んでいる。

Q: 触媒をつくる段階もかなり効くという話もあったが、ハイスループット実験では触媒のつくり方の要素は考慮されているか。

A: 今回の一連のデータでは、合成方法は一定にしていたはず。調整方法がかなり効くという解析結果が出ているので、今後は触媒の合成条件も入れていく必要があると思う。

3.2 ハイスループット実験を基盤としたデータ駆動型材料科学研究

谷池俊明（北陸先端科学技術大学院大学）

高橋先生から総括的な話をしていただいた。私は実際の材料化学の実験という立場に立って、ハイスループット実験を基盤としたインフォマティクス研究の現在の状況について、我々の研究室の成果を中心に述べる。

これまでの研究スタイルでは、各研究者が独自のプロセスやプロトコルによって研究者独自の対象を中心に成果を蓄積してきた。こういう状況で生じるデータは、データが局所的に集中した島と、データが欠乏した海から成るような、非常に一貫性のないまばらなデータとなる。これを、プロセスを気にせずにマテリアルズデータとしてだけ見たときには、このプロセスの変数が隠れた軸になる。

一方でマテリアルを固定したとき、プロセスを幅広く試したケースは少ないので、インフォマティクスに足るデータ量が集まらない。この理由から収集データを限定せざるをえず、データ不足や局所探索に陥るといった問題があった。そこで、ハイスループット実験により自分で収集したデータを前提としたマテリアルズインフォマティクスを展開してきた。

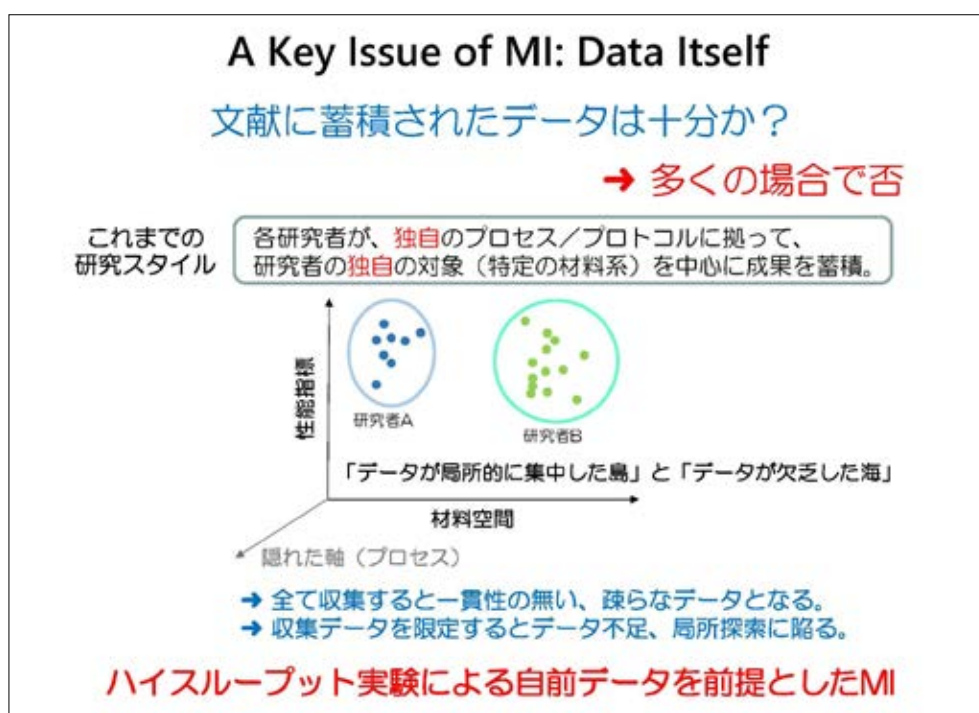


図3.2.1 これまでの研究スタイルの問題点と改善策

本日は、2つの研究について話をする。1つ目は、ハイスループット実験による高分子材料の安定化、2つ目は、高橋先生との共同で実施したハイスループット実験による触媒インフォマティクスである。最後に、私のプロセスインフォマティクスに対する思いを簡単に述べる。

近年、プラスチックの分野でサステナブルプラスチックインダストリーという構想が進んでいる。この分野で求められることは、プラスチックの劣化を抑制するための安定化剤の量を増やさずに耐久性を向上することである。複数存在する安定化剤の配合は一般則がなく、安定化剤配合の性能は高分子の寿命でのみ評価可能であるため、1つの組み合わせの評価におおよそ1か月の時間が必要となる。22種類の市販されている安定化剤から10種類を選ぶ場合、100万通りを超えるような組合せになる。

この問題に対し、私の研究室ではハイスループット化学発光イメージング装置という、100検体の高分子の寿命を同時決定できる装置と、遺伝的アルゴリズムを使った研究を実施した。化学発光法は、高分子が劣化する際に発する化学発光を検出する方法であり、通常の機械的特性の劣化による寿命計測の約20倍のスループットがある。100検体同時に検出可能な装置を開発したので、2,000倍のスループットとなる。この装置を使って機械学習が未知の配合を導きだせるかを研究した。

研究プロセスは、最初に、乱数で10種類を選んで配合を決定し、ポリプロピレンに含浸・溶融混練してフィルム化することで検体を得る。それぞれの検体の寿命を化学発光イメージング装置で計測し、検体寿命を指標として遺伝的アルゴリズムを通じて配合を進化させる。なお、実験温度は150℃である。

概念実証実験の結果、たった6世代の検討によって、安定化剤配合を0.1 wt.%加えるだけで、ポリプロピレンの劣化を150℃で200時間抑制できるような新しい配合を見つけることに成功した。この評価時間は合計すると5.5年に相当するが、先ほどの装置を使用して50日間で測定が完了した。

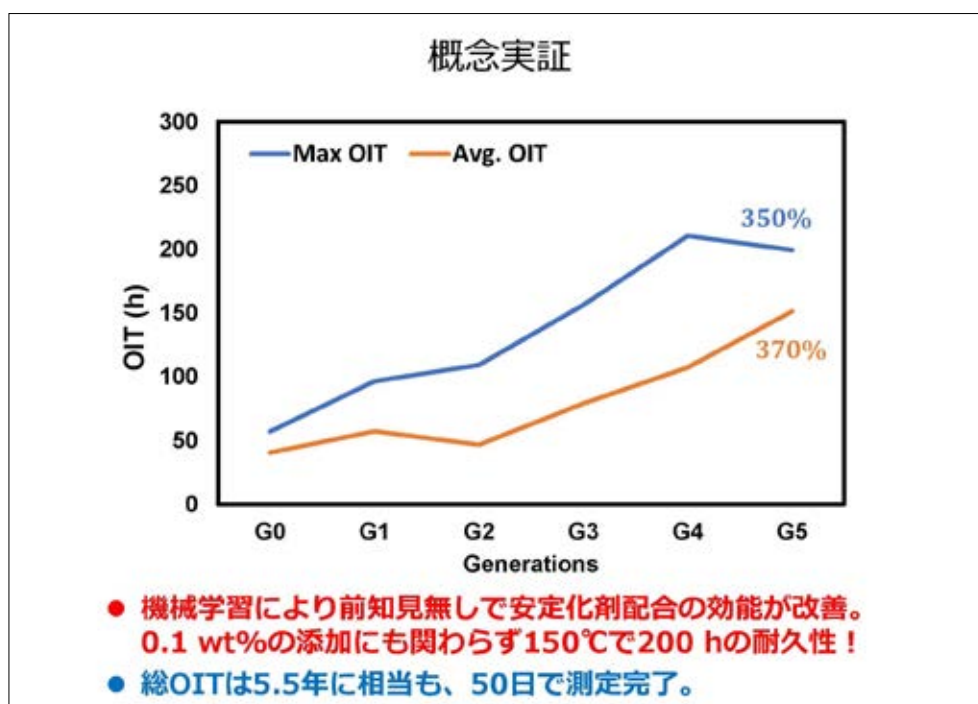


図3.2.2 遺伝的アルゴリズムを使用した概念実証結果

また、多数のデータが集まったことで、配合の中にどういう相乗効果が働いているのかを分析できるようになった。高性能な配合は互いに相乗関係を持つ安定化剤を多く含むも

のであることが分かった。例えば、単独で加えたときに高性能な安定化剤であっても、ほかの安定化剤と拮抗する場合は、配合物としてよい効果が得られない。

最初の研究をまとめると、ハイスループット化学発光イメージングの開発によって高分子の寿命決定のスループットを 100 倍向上し、さらに機械学習を併用することで 10^7 組の組み合わせ問題を取り扱えるようになった。この方法で、50 日間という短期間でポリプロピレンの高温酸化耐性を劇的に改善できる配合を見いだしたということである。

続いて、ハイスループット実験による触媒インフォマティクスの例として、メタンの酸化的二量化の成果を説明する。過去データを調べると、第 1 に、僅か 1,870 点のデータしかない、第 2 に、各研究者が特定触媒系に研究を集中させた結果として、材料空間における触媒材料の分散が非常に悪い、という 2 つの問題点がある。触媒はプロセス依存性が非常に強く、反応器の設計の違いや反応条件の不統一により、同一触媒でも全く違う性能が現れることがある。その結果、機械学習による予測や知識抽出に大きな問題が生じる。

私たちはハイスループット触媒評価システムをつくった。この装置は、20 個の触媒の性能を、あらかじめ登録した 216 条件で、1 日に約 4,300 点のデータを自動取得できる装置である。同レベルのスループットを保有する装置は報告例が無い。

この装置を用いて、3 日間の運用で 59 の触媒に対して 216 条件、欠損データもあったが、合計 12,708 点のデータを取得した。これは触媒の良し悪しにかかわらず同一プロセスを用いて広範な条件で体系的に取得したデータ、という点に特徴があり、このようなデータは私の知る限り報告例がない。最高収率は 21% で、低収率まで含めた一様なデータを使うと回帰がうまくいくが、収率が 10% を切るデータを敢えて除いたデータセットでは回帰がうまくいかないことから、ネガティブデータの存在が機械学習には非常に重要であることを示している。

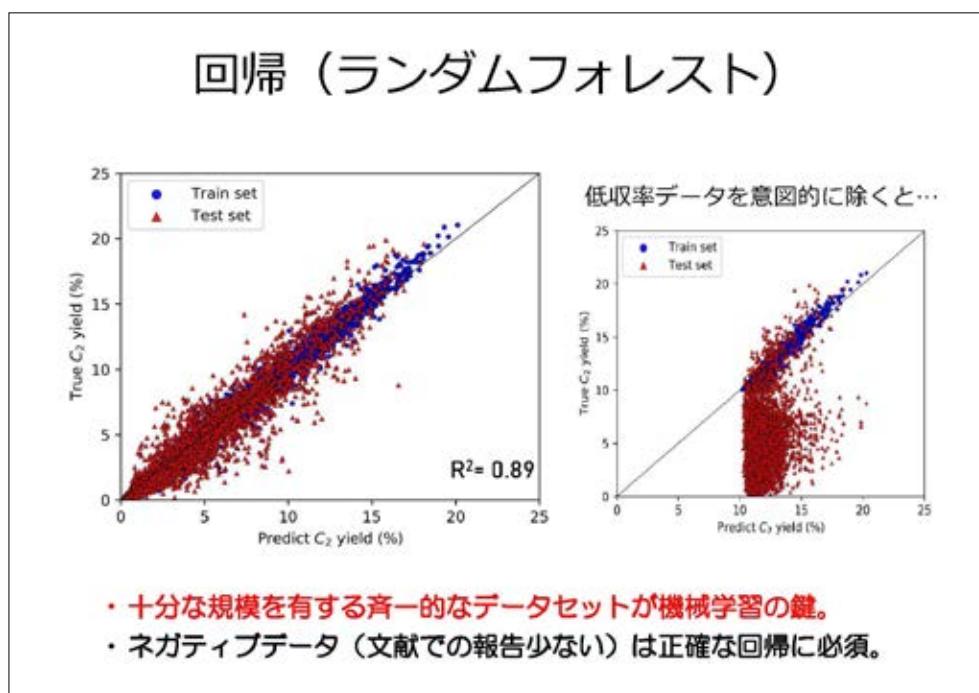


図3.2.3 メタン酸化的二量化収率の実験結果。予測 (横軸) と計測結果 (縦軸)

次に触媒開発における知識抽出、一般化の一例を示す。こちらの散布図は主生成物である C_2 化合物の収率を色彩軸として、横軸、縦軸に副生成物である一酸化炭素や二酸化炭素の選択性を示す。収率の高い黄色や赤の点を見ると、二酸化炭素の生成を 10% 以下に抑える触媒や反応条件が存在しないことが分かる。二酸化炭素の生成は、転換率と選択性の双方に悪影響を与える。二酸化炭素の生成を抑える触媒、反応条件が存在しない原因として、メタンよりも不安定なエチレンやエタンなどの生成物が、気相中でフリーラジカル反応により酸化していることが考えられる。そこで、触媒層の温度と触媒以外の部分の温度を別々に制御することで、最高収率を 21% から 26% まで改善することに成功した。本研究例は、触媒性能がプロセス条件によって大きく異なり、触媒探索をするときにプロセスの探索も必要であることを示している。したがって、プロセス探索を行わないと、本当によい触媒を見つけているのではなく、検討に使用したプロセス条件に合った触媒を見つけているだけになってしまう、ということの意味する。

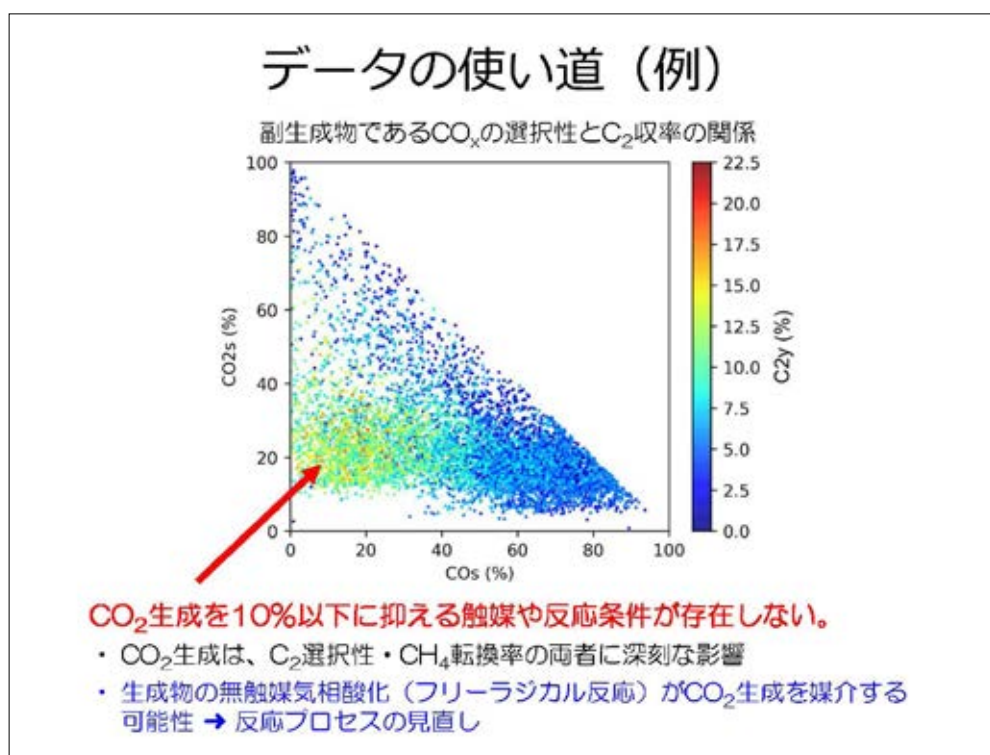


図3.2.4 メタン酸化的二量化収率の実験結果。
CO 選択性（横軸）と CO_2 選択性（縦軸）に対するプロット

2 つ目の研究をまとめる。ハイスループット触媒評価システムを開発し、触媒性能の 4,320 点/日の自動取得を可能にした。同一プロセスを用いて斉一的に取得した触媒ビッグデータは、機械学習などによる予測や知識抽出を可能にした。最後に、気相での過剰酸化をプロセス的に抑えつつ、高性能な触媒を用いることが収率の改善には必須であることが分かった。

マテリアルズインフォマティクスとしての結論としては、データセットの欠如は致命的問題であり、短期間に斉一的なデータセットを与え得るハイスループット実験は、データ科学的な意味でのマテリアルズインフォマティクスの果実を最大限に生かす手段だと認識

した。そして、ハイスループット実験とデータ科学により、従来設定し得なかったような広大な材料空間を対象とした研究が短期間で実践できることを示した。

プロセスインフォマティクスについて最も進んだ分野は、予測された有機化合物を合成する経路を探す逆合成解析の分野だと思う。有機化合物の場合は、目的生成物が決まれば、どの前駆体を使ってどのように合成すべきか、ディープラーニングなどを使って予測できる状況である。具体的には、分子構造や反応の類似性に関する記述子をつくり、これに基づいて可能な経路や前駆体を提案できる。

同様な方法を無機材料に展開するためには、無機材料そのものの構造やそれをつくるための反応のフィンガープリンティングが必要である。類似性の高い文献を見つけるのが出発点と考える。ある種の材料性能は合成だけでなく、これを成形や運用するプロセスに強く依存する。例えば高分子の成形機や触媒を用いる反応器などによって、これらの材料の性能は大きく異なる。ポリオレフィンプラスチック材料の生産の半分以上を占めているが、ポリオレフィンがここまで使われるようになった理由は、ポリオレフィンが合成できるようになったことだけでなく、ポリオレフィンを合成する効率的なプロセスと、これを成形して物性を制御するための効率的な成形機が開発されたことが非常に大きいと考える。これまでマテリアルズインフォマティクスでは、プロセスを統一した上で材料を最適化するということを主にやってきたが、材料を絞った上でインフォマティクスによって最適プロセスを設計するという研究も考えられるし、究極的には、マテリアルとプロセスを同時にインフォマティクスの対象とすることが必要だと考える。

【質疑応答】

Q: ハイスループット実験でデータを取るときに、計測条件を単に変えるだけか、それとも、パラメータ間の相互の関係、多重共線性を排除できるように、なるべく独立な変数に関して振るように、計測条件を選ぶのか。

A: 条件は文献などを見て、よく変えている条件を選んで、できるだけ幅広く取るようにする。触媒の候補材料は種類が多く、ある触媒は 700℃ で性能が最大になって、別の触媒は 900℃ で最大と、最適条件が大きく変化するので、網を大きく広げる、網の目はあまり小さくしないように、意識している。網の目を小さくするときは実験計画法のような方法を使うと良い。

Q: 化学発光イメージングというのは、劣化の何を計測しているのか。

A: 劣化のときにラジカル的な酸化をするが、酸化したときに生成物として生じるカルボニル化合物が発光するので、その発光を計測する。

Q: 最適制御を使って目的の機能を自動的に最適化することが可能だと思うが、研究例は無いのか。

A: 可能だと考えるが、現状、広い材料空間から新規の触媒を発見することに主眼が置かれている。今ある触媒では、最高収率が 26% ほどなので、プロセス条件を最適化して

も劇的に収率が向上するとは考えにくい。したがって、我々は材料開発に注目している。

Q: プロセスインフォマティクスの1つの形として、目的生成物から経路とか前駆体を提案する、という話が有った。これとデータ科学とは、どのようにつながるのか。

A: 逆合成解析では、文献から反応を収集して、分子構造との構造の類似性、反応の類似性、反応性の類似性などを記述子に変換する。そこから成功率の高い方法を提案すると認識している。

Q: 最初の話で、複数の安定化剤で相乗効果がある、という話だった。これは化学的に相乗効果がありそうだと予測できるものが良かった、ということか。それとも、機械学習を使ったら予想以上に相乗効果がある組合せが見つかったということか。

A: 相乗効果はポリマーの劣化において、例えば、ある安定化剤はラジカルをトラップし、別の安定化剤は酸化によってできた不安定化合物を分解するなどの場合、役割的に相乗である。しかし、分子構造によって本来 相乗的であると予想できても、拮抗的である場合もあり、予測は難しい。これが、本格的な配合開発がなされてこなかった1つの原因だと考える。

Q: ハイスループット実験を行うとなると、例えば一個一個丁寧に実験するのに比べ、何かを諦めているのではないか。どういうものが諦め得るか、方針を持っているか。

A: 今回のハイスループット実験では、触媒の性能評価は従来の性能評価方法とほぼ同一であるが、触媒調製では、この並列の速度に追いつくためには、ある程度ラフな方法を使わざるを得なかった。その調製方法のために、同じ触媒でも性能が上下することはあり得る。ハイスループット実験の役割の一つは、シーズを幅広い空間から見つけてくることである。触媒調製が最適でないため除かれてしまうものは、諦めざるを得ないと思う。

3.3 反応経路のデータベースと機械学習を用いる触媒・発光材料の理解・分子設計

畑中美穂（慶應義塾大学）

反応経路のデータベースと機械学習を用いる触媒や発光材料の理解・分子設計について紹介する。まだプロセスインフォマティクスといった状況には至っておらず、ほとんどマテリアルインフォマティクスのような状態ではあるが、今まで私がやっていた発光材料におけるマテリアルインフォマティクスのイメージをつかんでもらい、それを触媒反応に展開するとどのような未来が待っているか紹介したい。

最初に、図 3.3.1 に示す私の基本思想を説明しておく。マテリアルインフォマティクスやプロセスインフォマティクスといったインフォマティクスにおいては、第一原理計算や量子化学計算の結果を使っている、ターゲットとする量に対する新しい知見はそれほど出てこない。なぜなら、量子化学者は様々なものを計算して、計算する前の段階でいろいろなことが予想できるためである。そのため、目的変数を実験データにしてこそ新しい知見が出てくると考えられる。しかし、実際に実験データを多量に持っている実験グループは非常に少ないのが現実である。文献から集めてくることもできるが、年代によって装置が歴史的に変わっていきなりするので、データが均質ではない。

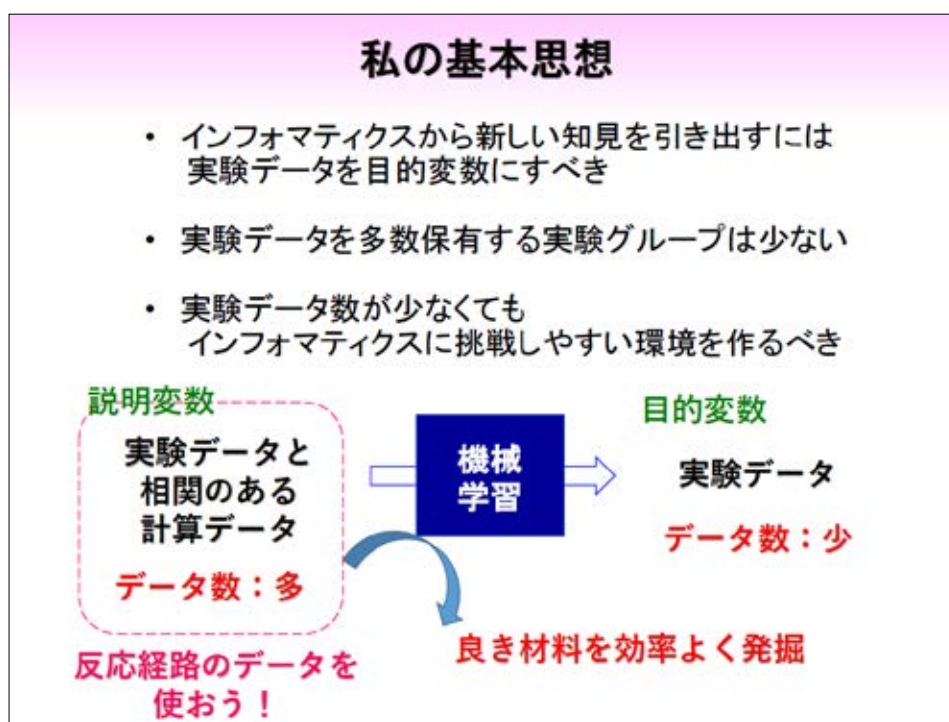


図 3.3.1 私の基本思想

そこで、基本的に実験データは多くないという現状を受け入れて、実験データ数が少なくても、多くの実験グループがインフォマティクスに挑戦しやすいような環境をつくるのが大事な仕事だと思っている。これを達成するために私が考えているのが、図 3.3.1 の下の図である。機械学習モデルに入れて出てくる目的変数は、実験データとしての材料の特性である。これを少ないデータで済ませるためには以下のようにする。説明変数として、

あらかじめ実験データと相関がある（因果関係がある）計算データを大量にためておく。そうすれば、実験データが少なくとも何とか機械学習モデルをつくることができる。ここに相関関係があれば線形に近いモデルでつくれるはずである。これをつくっておき、まだ実験データのない未検討の化合物に対する入力を行うことで、良い材料をこの中から見つけることが重要だと思っている。この実験データと相関のある計算データとして私が着目しているのが、反応経路のデータである。

題名にある発光材料や触媒反応については、両方とも反応経路が定義できる。反応経路とは、現象（化学反応、発光 [消光]）の進行に伴うエネルギーと分子の構造の変化である。

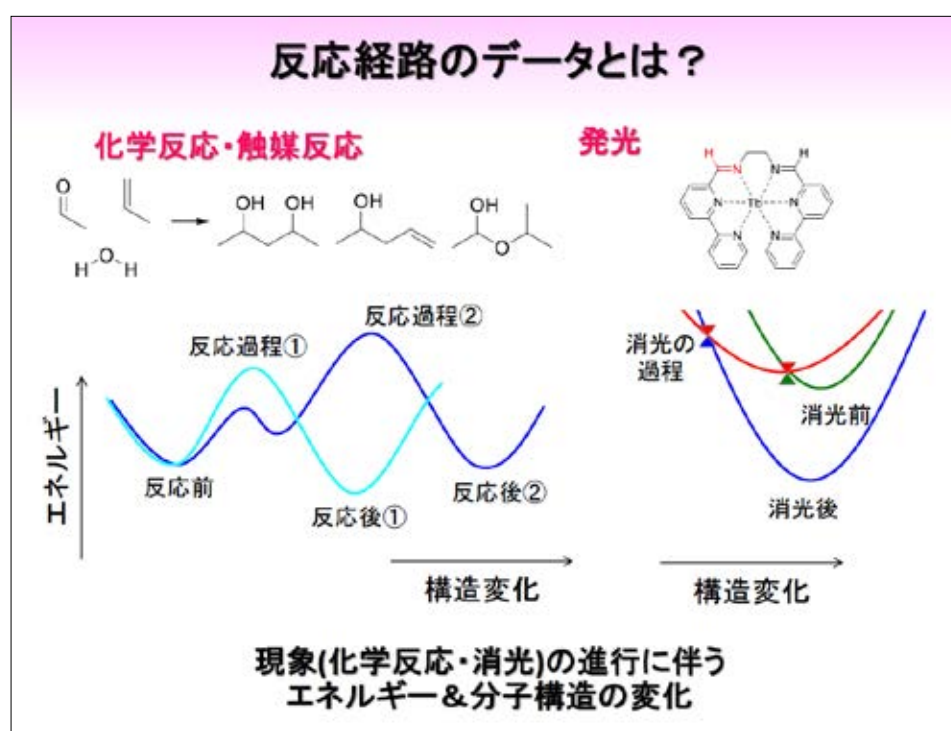


図 3.3.2 反応経路のデータとは？

左側の化学反応の場合、横軸に構造変化、縦軸にエネルギーを軸として反応過程を図示している。例えば、水色で示すような構造変化に沿っていくと、一つの山を越えればいいだけだが、濃い青色のような経路に沿っていくと、水色の経路よりも高い山を越えなければならないので、こちらの反応が進みにくくなる。このため、反応物よりも生成物の方が圧倒的に安定になる場合（つまり逆向きの反応が起こらない場合）は、水色の経路に沿った生成物が得られると予測できる。これと同じように、発光材料も反応経路を定義できる。発光の効率を見積もる量として、発光量子収率という量がある。これは、吸収された光子数に対する放出された光子の数の割合を示す量である。発光量子収率を予測するためには、発光の速度だけではなく、消光の速度も同時に考慮しなければならない。消光の速度は、光を吸収して励起状態に至った後の反応経路、つまり右側の図の赤色や緑色で示す励起状態が、青色で示す基底状態と交差するまでの経路を見ることで議論できる。

この消光の経路を求めることで、発光材料の発光強度についてどんな議論ができるか紹

介する。着目するのはランタノイドと呼ばれる化合物であり、発光の色が変わらないという特徴がある。しかし、発光強度は周りの影響を大きく受ける。例えば、ランタノイドの一つであるテルビウム (Tb^{3+}) はほぼ全て緑色に光るが、周りが変わると発光強度が変わる。 Tb に配位する分子が違ってても、 Tb 錯体が含まれている溶液内の添加剤の濃度を変えても、温度が変わっても発光強度が変わる。この理由を明らかにするために見るべきところが、消光の経路である。

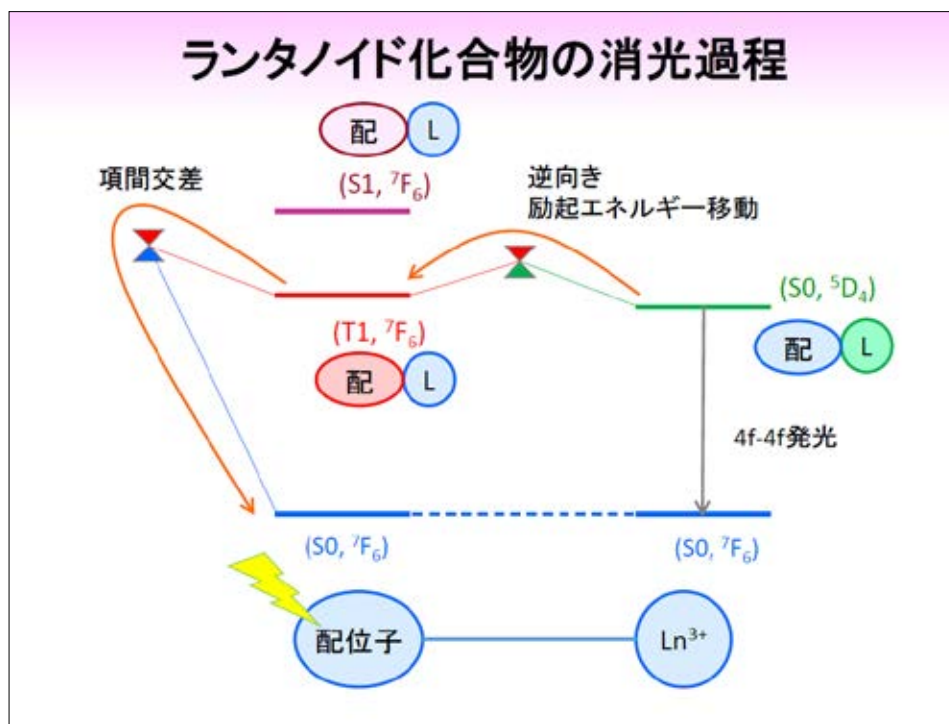


図 3.3.3 ランタノイド化合物の消光過程

図 3.3.3 にランタノイド化合物の消光過程を示す。光を吸収して励起された電子が、発光する状態に移動しても発光せずに元に戻り、連続的なエネルギー変化を伴う構造変化を経て一番下の基底状態に戻ってくることによって消光する。左上の山のところは、ポテンシャルエネルギー面が交差する点なので、簡単に交差点と呼ぶことにする。この交差点を計算するためには工夫が要るが、今回は私のオリジナルの計算手法を使ってこの交差点を求め、実際に消光の起こりやすさや、発光強度が予測できるかを示す。

分子の構造が似ているにも関わらず、発光強度が大きく異なる 2 つの分子に着目する。図 3.3.4 に示すように、よく光る分子 $\text{Tb}(\text{bpy})(\text{NO}_3)_2$ に着目すると、消光が起こる山の高さ（障壁）は高い。しかも、この山における分子の構造を可視化することができる。青色で描いてあるところは光を吸収する前であり、消光が起こるポイントが黄色で描いてあり、どのように構造が動いてこの山を越えていくかを可視化できる。この分子の場合には、高い山を越えないといけなないので、消光はほとんど起こらない。つまり、よく光ることになる。

一方、図 3.3.5 に示す光らない分子 $\text{Tb}(\text{L1})(\text{NO}_3)_2$ に着目すると、この消光のために越えるべき山がほとんどない。この越えるべき山における分子構造を見ると、山を越える直前

とほとんど構造が変わらない。しかし、CとNの二重結合のところが少しだけひずんでいる。面が変角振動しているような構造変化が消光を誘起していることが分かる。

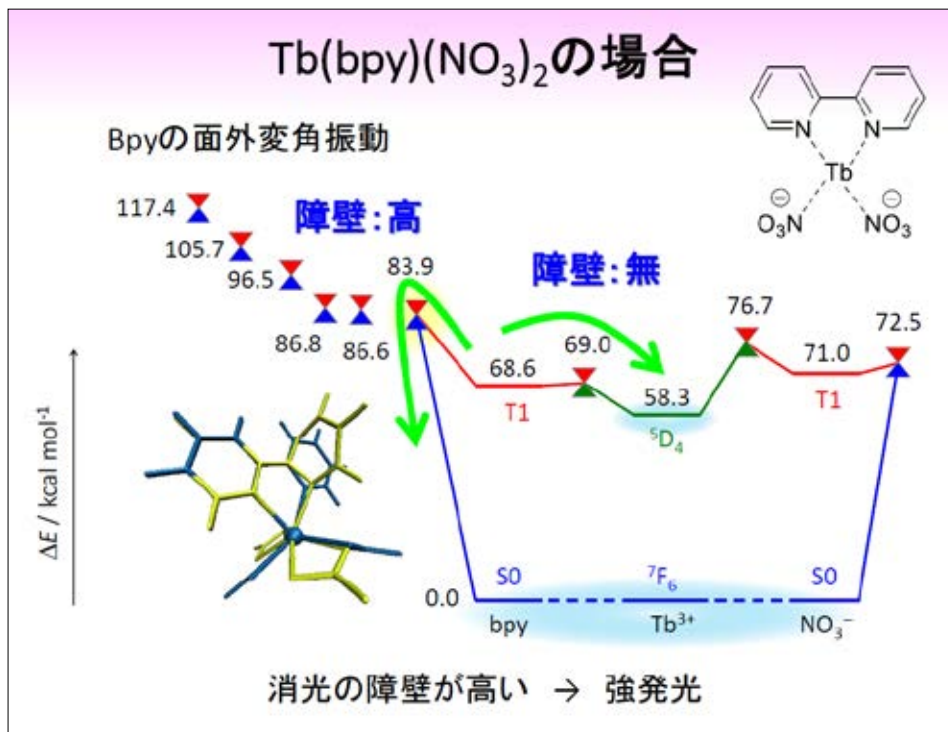


図 3.3.4 Tb(bpy)(NO₃)₂の場合

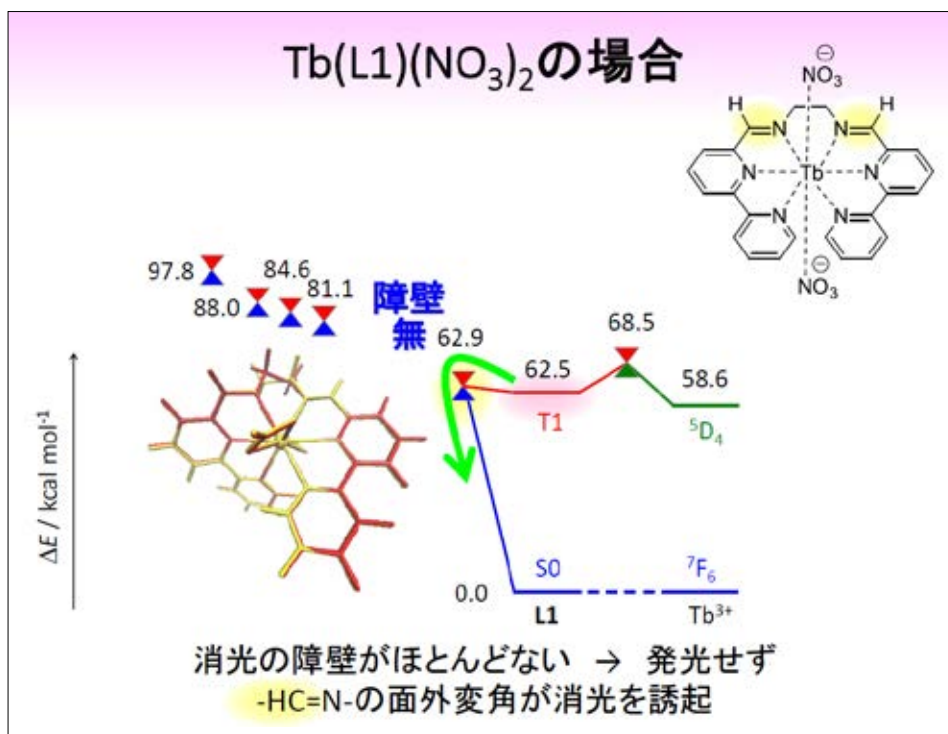


図 3.3.5 Tb(L1)(NO₃)₂の場合

この交差点のエネルギーや構造変化を量子化学計算で求めるのは大変で、1つの計算に3週間程度かかる。このため、もう少し簡単に計算できるようにならないかと考え、様々な分子の交差点のエネルギーと構造のデータを集めてデータベースをつくった。はじめは機械学習モデルをつくろうと思っていたが、あまりにもルールが簡単過ぎて機械学習の出番がなかった。消光を最も誘起しやすいのは、分子振動がどこか一箇所だけでも曲がるという構造変化だった。一箇所しか曲がらないのであれば、光らないものに対してその一箇所を取り除いてしまえばよいということになる。図 3.3.5 の消光の原因になっていた箇所(二箇所)を取り除き、C-N や C-O 単結合に置き換えた分子をコンピュータの中で作り、消光の起こりやすさを計算すると、確かに消光の障壁が高くなることがわかった。また、同じ分子を実験グループが合成して測定したところ、計算上消光が起こりにくくよく光りそうな分子が、実際によく光ることが証明できた。このように、消光の経路におけるエネルギー変化を追うことで発光の強度が予測できる。

これを使って様々なランタノイド材料の説明ができたので、ランタノイドのことは何でも分かるようになったと思って鼻高々だったのだが、鼻を折られる出来事が起こった。ランタノイドは様々な種類があるが、その中でも緑色に光るテルビウム (Tb) と赤色に光るユウロピウム (Eu) の2つが突出して論文の数が多い。テルビウムについては実験と理論の結果にあまり矛盾がなく何でも説明できるが、ユウロピウムは実験と私の計算の結果が矛盾していた。私のオリジナルの近似を使って計算しているので、近似のやり方に問題があったのではないかと考えた。近似に大事な効果が入っていないのではないかと考えた。近似方法がよくないのなら、その近似の精度を上げるのが一般的だと思うが、何をどうしたらいいか分からなかった。

そこで、機械学習の出番である。もともと、発光の強さの実験結果と私のオリジナル近似方法の計算結果の間には高い相関関係があると信じていた。実際に、緑色に光るテルビウムの化合物については、この相関関係が成り立っている。しかし、赤色に光るユウロピウムではこの相関関係が成り立っていないので、説明変数の数を増やした。図 3.3.6 に示すように、考えられるパラメータを集めて入れた。例えば、量子化学者が好きな軌道のエネルギーや、実験の条件、溶液の粘度や分極率など、化学の事典に載っているデータから計算データまで拾ってみた。これらの説明変数と実験データの間の相関モデルを機械学習を用いて構築すれば、抜けていた重要なパラメータが何だったのか分かるのではないかと考えた。

このモデルをつくるために、実験データを文献値から人海戦術で集め、計算データも人海戦術でと思ったが、計算に時間がかかる。そこで、周りの配位子が同じようなものは消して、例えばユウロピウムは計算が大変なのでナトリウムにして、消光の経路のデータだけでなく軌道のエネルギーや実験データを集めて実際に機械学習モデルをつくった。

このモデルの中で重要度の高いパラメータを並べてみると、あらかじめ予期していた、私のオリジナル近似で出てくるパラメータが重要度の高いパラメータとして出てくるが、予期していなかった様々なパラメータも重要なパラメータとして上がってきた。

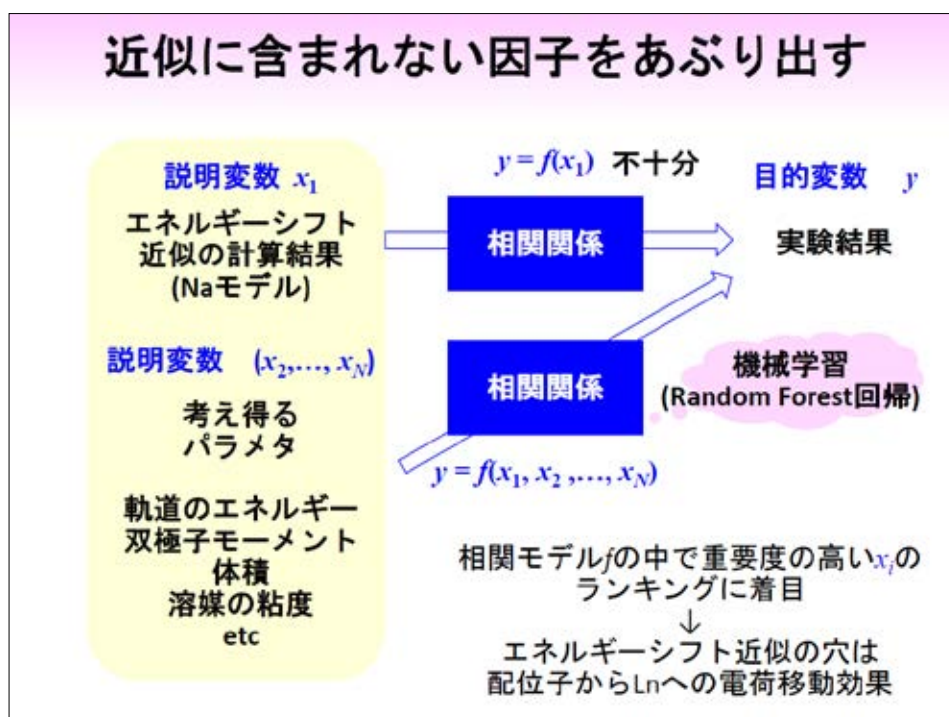


図 3.3.6 近似に含まれない因子をあぶり出す

これらに対して科学的な意味を考察すると、きちんとと解釈を与えることができる。この解釈から、近似に抜けていたのは配位子（錯体の中の金属の周りについている分子）からランタノイドへの電荷移動の効果であり、これが実験と計算データの違いの原因だったことを明らかにすることができた。この経験から、メカニズムに一番効いてくるパラメータがきちんと入っていれば、200 個程度のデータしかなくてもよい機械学習モデルをつくることができることが分かった。つまり、データが少ない場合は、メカニズムに直結した説明変数が選んでいるかどうか、成功の鍵を握るわけである。

メカニズムに直結するデータである“消光経路のエネルギー”の情報がよい説明変数になっていたということは、触媒反応も反応のエネルギーの情報を使うのはよい戦略だと考えられる。化学反応の経路上で最も不安定な点、つまり遷移状態のエネルギーを下げるのが触媒の仕事である。このため、発光材料における交差点に着目したのと同じように、触媒反応では遷移状態に着目し、このエネルギーを目的変数にすれば、少ないデータでもよい機械学習モデルがつくれそうである。

ここでは均一系触媒（溶液に溶かして利用する触媒）を対象として考えよう。実験室でフラスコの中の均一系に使われている反応は、カラムの中で触媒をシリカゲルに担持する形で中に詰め込み、反応物を流すという形で工業的にも使われている。実際にはフラスコの中での触媒反応と、カラムを通すフローシステムの触媒反応の経路は全く違うが、ミクロな視点で見ると一つ一つの経路のパターンは似ているので、計算データは工場やプラントで使われるフローシステムの考察にも十分に流用可能である。

均一系触媒としてよく用いられるのが、金属のまわりに有機分子を持つ“金属錯体触媒”である。金属錯体触媒は、触媒構造や、ターゲットとする反応を少し変えただけで、触媒の性能は大きく変わる。しかし、一つの触媒を作るのに 2～3 週間かかることも珍しくな

いため、販売ルートに乗っていない触媒をわざわざ実験室で検証することはほとんどない。このため、多くの触媒が埋もれている。日本の中には有機化学の研究室が多くあり、多くの触媒がつくられて論文に載っているが、他のグループで検証されることはほとんどない。

そのため、高い性能を示す触媒を機械学習を用いて推薦するシステムを作ることは急務である。さて、先程、触媒反応の場合、反応経路を調べて遷移状態のエネルギーを説明変数にすれば少ないデータでもよい機械学習モデルを作れるはずと述べたが、遷移状態計算はそんなに簡単ではない。発光材料の場合は、一つの分子に着目すれば良かったが、複雑な触媒反応の場合は、多数の分子が触媒と相互作用して反応が進んでいくため、分子の接近する順番、接近する方向などが異なる様々な反応が同時に進行する可能性がある。つまり、一種類の触媒反応の中に、多数の遷移状態がある。しかし、どの遷移状態に着目すれば良いかは、計算するまで分からないことが多い。この場合、考えられる反応経路を網羅的に調べるしかないため、計算に膨大な時間とコストがかかってしまう。

ならば、遷移状態の情報の収集を諦めるのがよいかというところも言えない。図 3.3.7 に示すのは、金属に結合して金属錯体を形成する分子（配位子）であるが、配位子の中の一つの原子が金属と結合を作るものもあれば、二つの原子が結合を作るものもある。結合を作る原子の種類も多様だ。このように金属との結合の仕方が全く異なるものが混在しているため、全ての金属錯体触媒に統一的に利用できるパラメータ／記述子は非常に少ない。

素反応のTSの情報を記述子に

Cc1ccc(cc1)P(c2ccccc2)C(c3ccccc3)C4=CC=CC=C4 (XPhos)
 Cc1ccc(cc1)P(c2ccccc2)C3=CC=CC=C3 (JohnPhos)
 Cc1ccc(cc1)P(c2ccccc2)C3=CC=CC=C3 (Xantphos)
 Cc1ccc(cc1)P(c2ccccc2)C3=CC=CC=C3 (IPr)
 c1ccc2c(c1)nc3ccccc3n2 (phen)

- ・ 触媒の性能： 分子の微小な違いで大きく変化
- ・ 多用な触媒に共通して利用可能な記述子は限定的
- ・ 複雑な触媒反応の経路：
 - 十～数十種類の「素反応」の集まり
- ・ 触媒の性能を決めるのは、そのうち数個の素反応
- ・ **素反応の反応経路の情報(TS・中間体)を説明変数にすれば**
 目的変数(触媒性能)の実験データ数が少なくても
 そこそこ良い機械学習モデルを作れるのでは？

図 3.3.7 素反応のTSの情報を記述子に

しかし、どんなに複雑な触媒反応でも、一つ一つのステップに分解してみると、教科書に載っている素反応の組合せでしかないことに気付いた。そこで、素反応の反応経路を計算して、遷移状態の情報をデータベース化する。素反応の計算は北大の反応経路自動探索 (GRRM) を使うとほぼ自動的に計算データを集めることができる。遷移状態の情報は、メカニズムに直接的に関わるデータであるため、これを説明変数、実験データを目的変数にすることで、少ない実験データから効率的な機械学習モデルをつくれると期待できる。一度モデルができれば、未検討の触媒の中からよりよいものを推薦できるはずであり、触媒の分子を少しずつ変えていきチューニングするという苦行から解放されると思っている。

一番のアピールポイントは、競合する実験グループの全てが私の提供する反応経路のデータベースを使えることである。図 3.3.8 にそのイメージを示す。実験データは各実験グループから提供してもらう。私のデータを説明変数、実験グループのデータを目的変数として機械学習モデルを作るが、機械学習モデルは実験グループ側に帰属する。様々なグループがこれを使っても、実験データが違えば出てくるモデルが違うので、競合する実験グループ間でのデータの開示は必要ない。

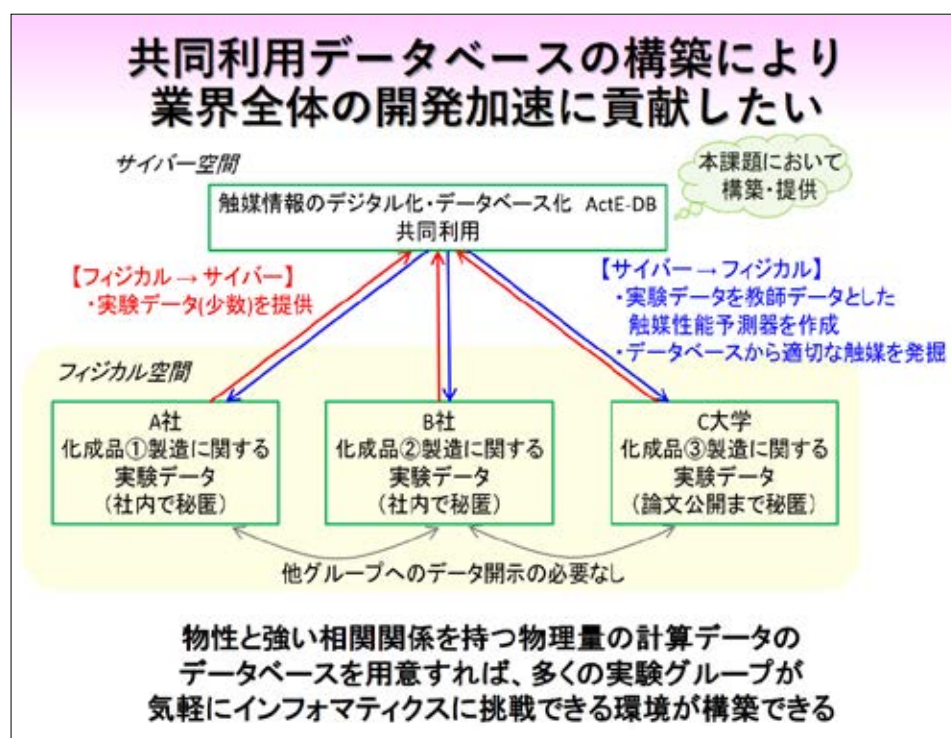


図 3.3.8 共同利用データベースの構築により業界全体の開発加速に貢献したい

このように、材料の性能（発光強度や触媒能）と強い相関関係や因果関係を持つ量を量子化学計算で求め、このデータベースを用意しておけば、様々なグループが気軽にアクセスして機械学習モデルをつくることができる。つまり、実験データさえあれば、だれでもインフォマティクスに挑戦できる環境が構築できるので、その中からよい事例が多く出てくると考えている。

【質疑応答】

Q: この素反応のデータベースで、実際の反応は素反応が幾つかがあるので、その幾つかの素反応と実験データの間を関係づけると思うが、その素反応の順番も考慮するのか。

A: 素反応の順番が違ったとしても、結局は全体の活性化障壁を決めるのは一番安定な構造のエネルギーと一番不安定な遷移状態のエネルギーであり、これは、幾つかの素反応の和で簡単に説明でき、順番がばらばらだったとしても全体の触媒の性能に効いてこないことが分かっているので、順番はあまり気にしない。

Q: 素反応がパラレルに起こればよいが、シークエンシャルに起こる反応もあり、素反応がシークエンシャルに起こるときにはどこかが律速になっているときにはその順番が効くわけなので、順番も考慮しなければいけないのかという質問である。

A: 順番を考慮しないとうまくいかないときもあると思うが、基本的にターゲットとする反応と全く同じ素反応のデータベースを集めるわけではない。実際の反応に関係なく素反応の様々なデータをたくさん集めて大きな入力データにして、そことの関係の機械学習モデルをつくっている。

Q: 計算精度のレベルがある程度担保されないと、幾ら AI をうまく使っても合わないということになるのか。

A: 今後はそれが気になってくると思うが、まずは金属錯体を触媒とする系に限って考えている。金属の種類が同じなら、計算レベルを変えても、エネルギーの絶対値は変わっても、エネルギー差はあまり変わらない。反応経路のエネルギーは、反応経路上のエネルギー変化のことなので、ほとんど問題ないと考えている。また、計算は一番簡単なパラジウム錯体からスタートしているが、計算レベルはそんなに心配しなくてもよいことが分かっている。今後、コバルト、マンガン等にも手を出すと、計算方法と結果に大きな依存性があるので、考慮する必要があると考えている。

Q: 経路探索で遷移状態を押さえるという話をされたが、それは反応領域を全部サーベイしているのと結局は一緒なのではないかと思うが、それは違うのか。

A: それとは違う。反応領域を全部サーベイすると、触媒と反応したときに様々なものができる可能性がある。ここでは、幾つか素反応の種類を決めて反応経路を探す。例えば、水素からメタンができる経路と限ってしまって、その遷移状態に着目するのである。網羅探索はせず、どういう種類の反応が起こるかを指定することで計算量を削減している。

3.4 マイクロフロー合成とプロセスインフォマティクスとの融合と展開

永木愛一郎（京都大学）

インフォマティクスは専門ではないが、プロセス開発という観点から研究を進める中で、最近、インフォマティクスを指向したことを考えるようになった。そのあたりの内容を含めて紹介する。

まず、我々はプロセスとして、フロー型のマイクロリアクターを活用してものづくり等を進めている。フロー型のマイクロリアクターというのは名前の通り、マイクロメートルサイズの微細な流路を持つ反応器のことであり、通常はフロー型として流しながら反応をおこなう。

マイクロリアクターの研究分野は合成化学において非常に活発化しており、例えば Sci finder で検索すると、論文数が軒並み増加傾向にある。また、合成化学のコミュニティにおいて代表的な出版物である「有機合成化学」でも、2015 年にフローケミストリーにフォーカスした特集が組まれた。最近では ACS の OPR&D というジャーナルで日本プロセス化学会の特集号が組まれたが、その背景にはフロー型の合成が物をつくる上で非常に重要視されているという状況がある。

では、なぜフロー化学・マイクロリアクターを、物をつくるプロセスとして使うのか。まず研究・開発、学術面においては、通常のバッチ型反応器に代表されるフラスコ化学では困難、または不可能な反応やものづくりが実現できる。また、産業界では研究・開発面に加えて、最終的には物をたくさんつくらなければならないという観点がある。フロー化学を利用するとスケールによらない安定的な連続生産が可能であり、今後、21 世紀の生産技術の概念を革新するのではないかとということで、注目を集めている。さらに、インフォマティクスにも関係するが、フローを用いると通常のバッチ型反応器にはない非常に信頼性の高いデータが取得できるという大きな特徴がある。例えば、通常のバッチにおいて人に依存していた部分がフローの場合にはプロセス依存になるため、こういった装置を使うかによって極めて再現性の高いものづくりが実現できる。また、真の反応速度論に従った反応が実現可能になるため、反応・生産プロセスの構築において大きなアドバンテージになると期待されている。

では、バッチではできない反応やものづくりを実現する、フローマイクロリアクターならではの特徴は何か。1 つ目に、精密な温度制御ができることが挙げられる。バッチのスケールにもよるが、おおむね 100 倍、200 倍という高い温度制御性が実現できる。2 つ目は混合の制御で、複数の溶液を混ぜるという点においても、通常の攪拌混合では数秒程度かかるのに対して、数ミリ秒で瞬間的な混合が達成できる。3 つ目の特徴は反応時間の制御。非常に短い時間を、正確にしっかりとコントロールできる。

そこで、我々がこれまで進めてきたのが、フラスコでは不可能な非常に速い反応を使った合成化学の体系化と理論構築である（図 3.4.1）。通常、バッチ型反応器では、非常に速い反応のエリアでは残念ながら反応速度論に従うことなく、反応が物質や熱の移動による影響を受けてしまう。一方で、マイクロリアクターを用いると、物質移動が良い、混合が速い、熱移動が効率化できるといった特徴から反応のより速い側まで反応速度論で扱える

領域としてカバーできる。非常に速い反応において反応速度論に従った反応が実現できるので、フラスコでは不可能な合成化学を進めることができる。



図3.4.1 フラスコでは不可能な超高速合成化学

この考え方に基づいてこれまでに取り組んだ例を2つ紹介する。図3.4.2は官能基選択的な反応の制御、非常に速い反応において反応速度論に従った制御をおこなった例である。この反応をバッチ型反応器でおこなうと3つのプロダクトの混合物になるが、マイクロリアクターを用いると、反応性の高いアルデヒドと反応した生成物が非常に高い収率で選択的に得られる。真の反応速度論に従った反応選択性を、マイクロリアクターで実現できたということであり、バッチ型反応器では困難な選択性の制御を実現した一例と言える。

また、図3.4.3は時間軸をコントロールすると何ができるかという例である。これまでの常識では、ケトンはリチウムとすぐ反応するので、ケトンを持つ有機リチウム種という中間体は存在し得なかった。この中間体をマイクロリアクターを用いて発生させ、3ミリ秒という時間でR2の滞留時間をコントロールすることで、中間体が分解する前に、続く反応に利用することができる。これはまさにバッチでは実現できない分子変換の例になる。

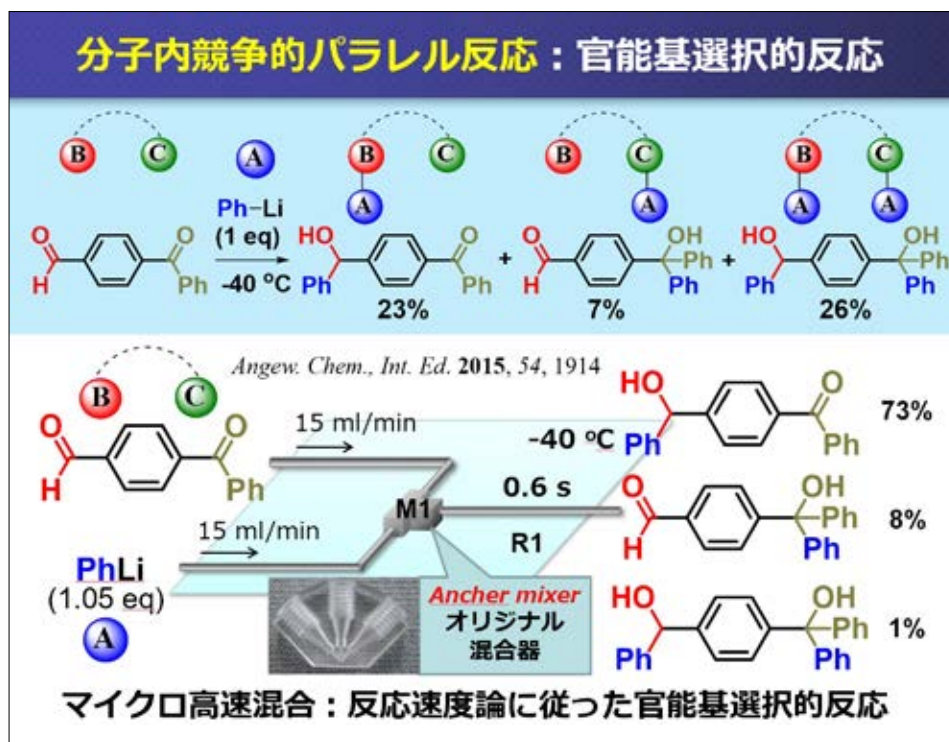


図3.4.2 分子内競争的パラレル反応

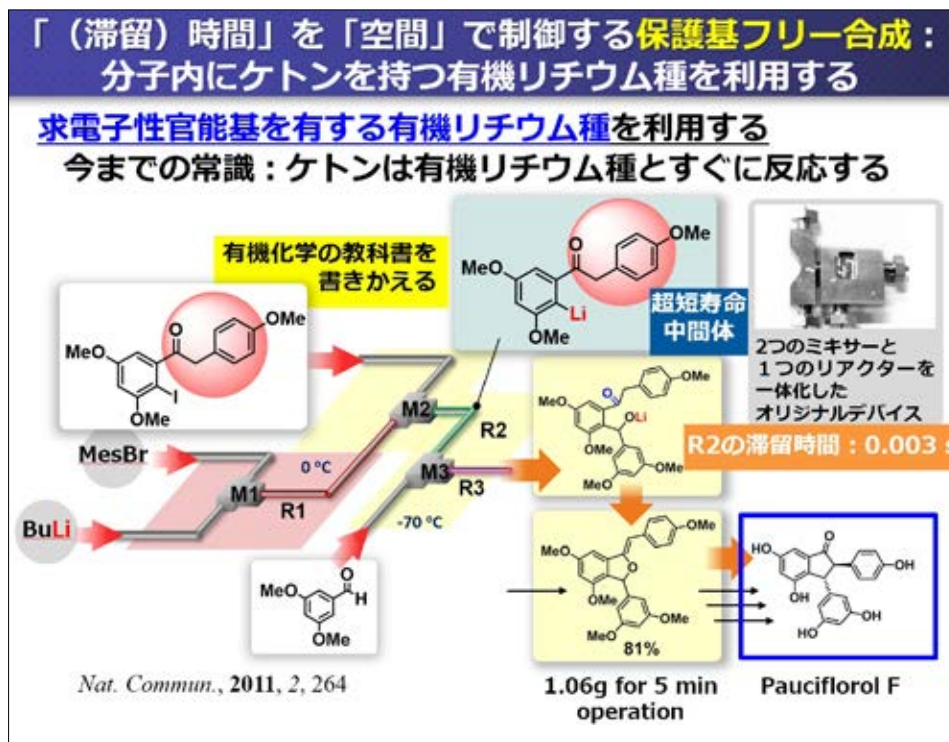


図3.4.3 「（滞留）時間」を「空間」で制御する保護基フリー合成

このように我々はこれまでプロセス化学として、フラスコでは不可能な反応や物質合成が実現できるという点からフローを進めてきた。一方で、フローには非常に信頼性の高いデータを取得できるという特徴があり、インフォマティクスに展開できないかということで、1年程前から始めた結果について紹介する。

まず低分子を考えたときに、実在の分子とその機能との間をインフォマティクスやAIで解読していく（図3.4.4）。その中で、さらに高機能な仮想分子を提案し、これを何らかのプロセスを使って合成することが合成化学者は求められる。このときに、分子構造を基点に考えることが多いのが低分子の一つの特徴である。ここでプロセスという側面で考えると、フラスコ化学というのは曖昧であり再現性、信頼性が低い。一方、フロー化学を用いるとプロセスとして非常に正確であり、かつ再現性等が極めて高いという特徴がある。したがってフロー化学を利用することによって、プロセス情報から実際にその機能をもつ分子につなげることできないか、インプットとしてプロセス情報を上手く使い、分子構造との相関を取りながら、すなわちサイエンスを理解しながら高機能な分子を生み出すことをワンパスで行うことができないかと考えている。

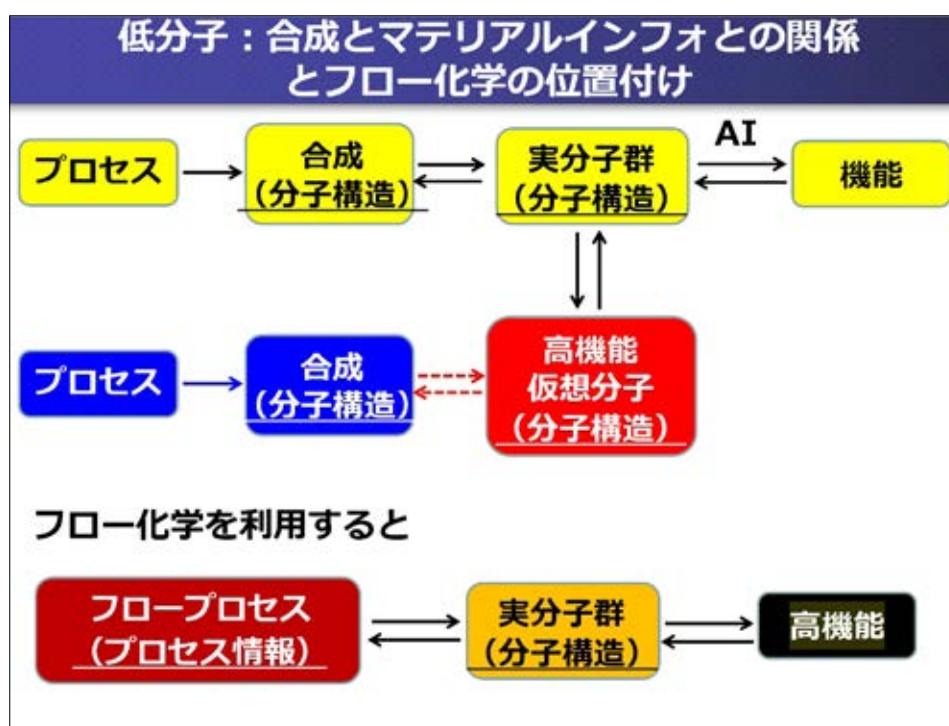


図3.4.4 低分子：合成とマテリアルインフォとの関係とフロー化学の位置付け

今、一番力を入れているのは実は高分子材料である（図3.4.5）。インフォマティクスへ利用するときの課題として、高分子は分子構造が曖昧なケースが多いことが挙げられる。配列、分子量、分子量分布などは精密重合を使えばある程度可視化できることもあるが、基本的には非常にアバウトに捉えられている。さらに、それをつくるプロセス自体も経験や曖昧さに基づく場合が多い。したがって高機能な仮想高分子を提案することが容易ではなく、高分子化学の機能探索においては、たくさんの組合せをハイスループットで生み出してポリマーの機能を評価していくことが主たるやり方になってくると考えられる。

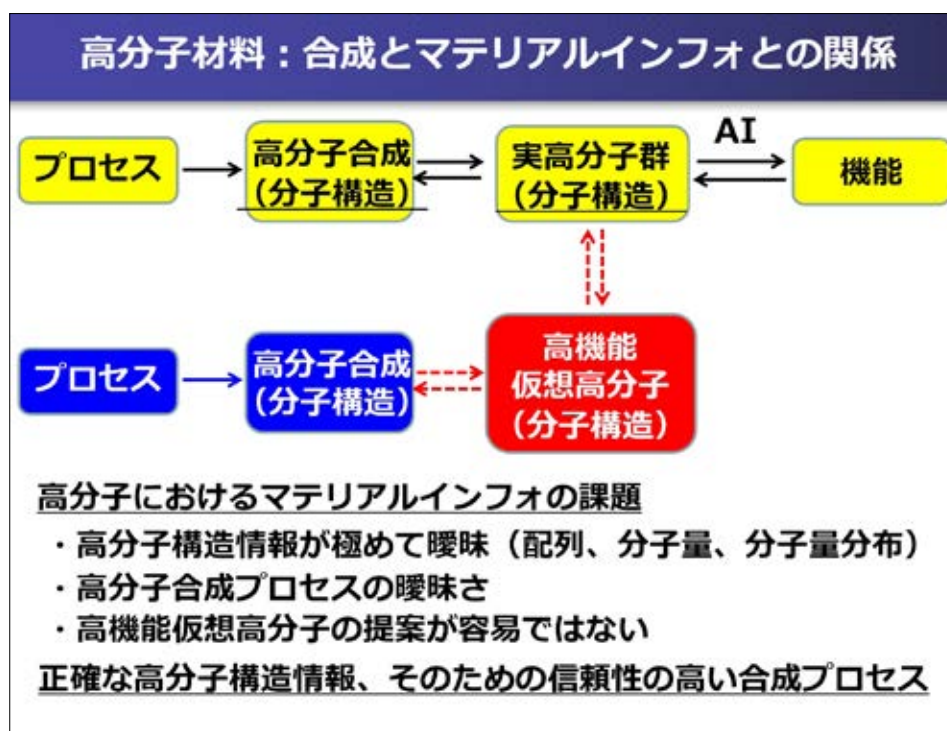


図3.4.5 高分子材料：合成とマテリアルインフォとの関係

そこで、高分子の正確な構造情報、またそれをつくるための信頼性の高い合成プロセスを情報として確立できれば、機能に対してダイレクトにつなげることができ、またそこから理論構築ができるのではないかと考えて研究をスタートした。これまでに我々はアニオン重合を使って、構造が精密に制御されたポリマー群を報告してきた。通常のスターポリマーまでの代表的な構造制御ポリマー、リビング重合を使った構造制御ポリマーの合成、連続生産の実証等についても進めてきた。ただし、機能性の高分子を考えたときには、共重合によるものが主になると考えられる。しかしながら、共重合でできるポリマーはモノマーの反応性によって時にはグラジエントになり、また、時にはランダムになるといったように、モノマーの反応性に依存した重合法を取ることがほとんどである。また、多くの場合、構造情報があまり明確ではない。特にフリーラジカル重合を使ってつくるような場合、シーケンスレベルではほとんど構造情報が明確になっていない。

そこで、リビングのアニオン共重合において、合理的に設計してシーケンスをしっかりと可視化する方法というのを課題として設定した。そこでの大きな問題は、アニオン重合というのは非常に速いということ。重合が速いために一次シーケンスの可視化が非常に難しく、共重合の合理的な設計のための Fineman-Ross プロットや Mayo-Lewis 式の導出をおこなうことが極めて困難である。そこで、マイクロリアクターを用いて共重合の反応速度論を実際に求めた（図 3.4.6）。具体的には R1 の反応時間、図 3.4.6 のグラフに示した 0 秒～25 秒ぐらいまでの時間軸を順番に切っていく。すると、最初の段階では赤のモノマーがより高く消費されるが、どんどん青色の消費量が上がっていく。すなわち、マイクロリアクターで反応速度をしっかりと見ていくことによって、一次シーケンスの配列にグラジエントがかかっていることを可視化できる。また、どうしてもグラジエントがかかってしまうものに対して Mayo-Lewis 式等を活用することで、いろいろなモード設計が

可能になる。そうすると、速い時間軸においても逐次添加で補うことによって、2つのモノマーの反応性に差があるにもかかわらず、ランダムさを担保したポリマーの合成ができる。A・Bモノマーの共重合体から得られる3つのパターンについて機能評価をしてみると、同じ組成比でシーケンスが変わることによって、溶解度等の変換ることが分かった(図 3.4.7)。

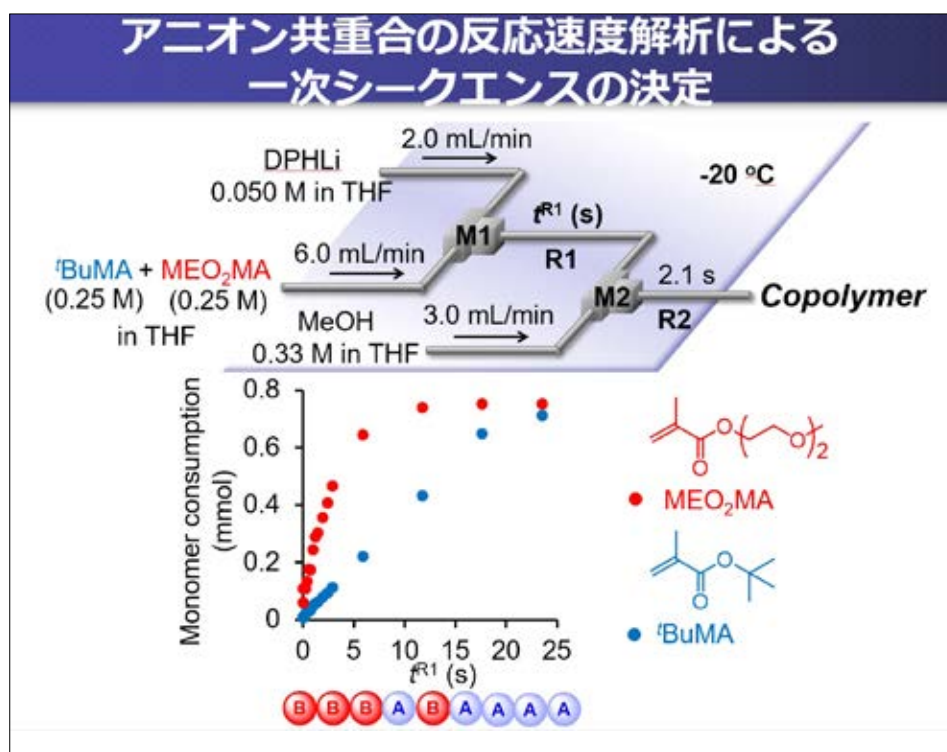


図3.4.6 アニオン共重合の反応速度解析による一次シーケンスの決定

このように共重合をフロープロセスでおこなうことによって、高分子合成プロセスの曖昧さをなくし、インプットの情報に変えてやることができる。また、高分子の構造情報の曖昧さという問題に対しても、速度論解析によりシーケンスの可視化ができる。そうすると、高分子構造やプロセス情報によってより高機能の仮想高分子の提案につながり、さらにそこから理論構築が可能ではないかと考えている。正確な構造情報と信頼性の高いプロセスを利用することで、情報と構造と機能をつなげていく、そこにインフォマティクスをうまく使っていければと考えている。

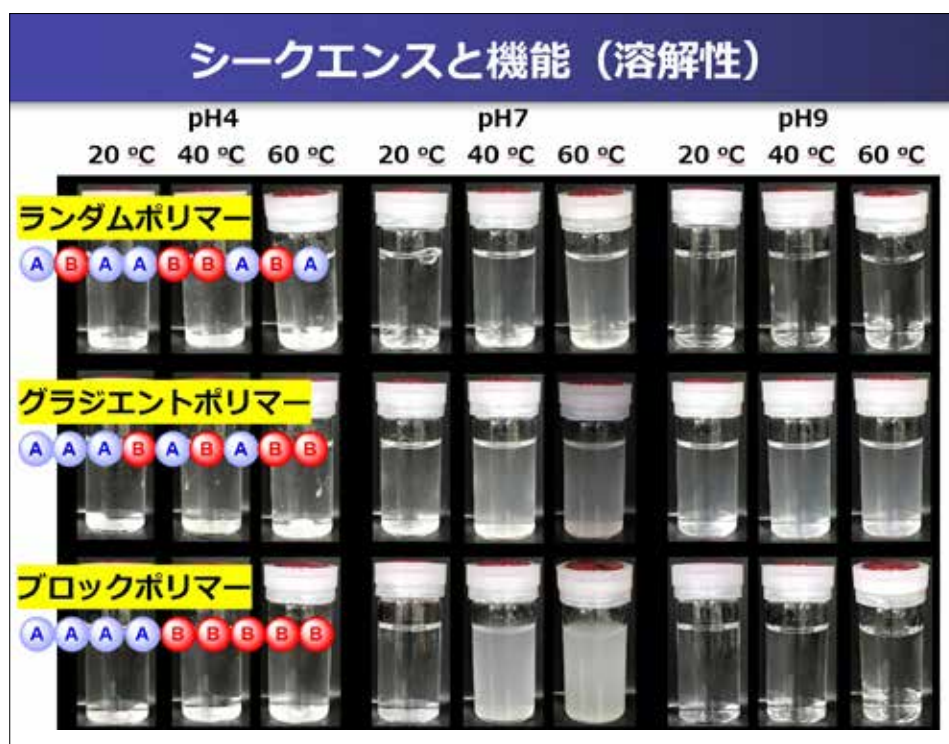


図3.4.7 シークエンスと機能（溶解性）

最後に、フローにおいて機械学習を使って取り組んでいる例を2つ簡単に紹介する。まずフローの課題として、操作変数が正確である一方、非常に多いということがある。したがって、変数をいろいろ変えてそれら全部の掛け合わせをやらなければならない。例えば、2つの変数に対して通常はグリッドサーチ的な検討をおこなうところ、ベイズ最適化を使うことによって4回の試行回数で一番高い収率の得られる条件が求められる。このようにベイズ最適化を活用することで、多変数に対しても少ない試行回数で良い条件を求めるという展開を今後進めていく。また、フローというのはインラインの分析手法を活用できることが大きな特徴ではあるが、インラインの分析手法は利用法が限定的である。例えば原料から生成物になる過程でIRを見たときに、波形としてほとんどが吸収であると、どこを見ればよいか分からない。こういった場合は通常であればIRでは測定できないという位置づけになるが、ここでニューラルネットワークを使い、収率情報とデータとの間でソフトセンサを構築することによって、IRの波形をそのまま読み込めば収率情報に置き換えられるようなネットワーク構築を進めている。

まとめると、合成とインフォマティクスについては、プロセス情報を使って機能につなぐ。それによって従来の試行錯誤的な手法ではなく最終的に合理的な機能の予測等につなげていく。また、迅速なプロセス構築については、ベイズ最適化の利用やインライン分析をうまく活用する。インラインの情報を最終的にはそのサイエンスにフィードバックすることによって、実際に高機能な材料を瞬時に、自動的に求めていくようなプロセス構築を進めていきたいと考えている。

【質疑応答】

Q：インラインでの分析は高分子という分野的に難しいということか。構造情報については現状ではフローであってもバッチで取ってから分析をしているという理解で良いか。

A：現状はそうである。今、NMRを導入してインラインでソフトセンサ解析をしようとしており、ニューラルネットワークでNMRの情報から良い機能や分子量、何か新しい相関関係等を出せば良いと考えているが、現状はまだ具体的ではない。

Q：インライン分析とマイクロリアクターをうまく組み合わせると、相当ハイスループットの実験が可能になり、条件を振った結果いろいろなものができるというのをかなりのスピードでやれるように思うが、そのようなアプローチは取られているか。

A：現状ではまだ無いが、まさにそれをやりたいと思っている。インラインというのは、例えば複数の分光学的な手法を並列化して使うことができるので、上流の変数にフィードバックして大量の情報を短時間で得るというスキームを動かしていきたい。その結果、機能のある材料が最終的には勝手に見つかるというのが理想である。

Q：マイクロフローを使ったアニオン重合で、モノマー配列の制御・可視化という話があったが、さらには高分子の特徴である分子量あるいは分子量分布、立体規則性の制御に向けて、将来的にマイクロフローリアクターを使うことは可能か。

A：リビング重合に関しては1.1～2くらいの分子量分布で、分子量も時には10万くらいをカバーするようなポリマーをつくることが可能。それを上手くアイテムとして載せて、共重合に展開しようとしている。

Q：ハイスループットへの展開において、特に並列化するような状況では何がボトルネックになると考えられるか。

A：機能と相関が取れるような、インラインでの直接的な分析をいかに上手く立ち上げていけるかが一つのポイントと考えている。ハイスループットに使える装置というのは、フローリアクターではシステムベースでいろいろ考えられる。やはり出口の機能評価というところが重要なポイントになる。

3.5 NEDO 超超プロジェクト概要と高分子 MI に対する取り組み

青柳岳司（産業技術総合研究所）

現在 5 年目となる新エネルギー・産業技術総合開発機構（NEDO）「超先端材料超高速開発基盤技術プロジェクト」（「超超プロジェクト」と呼称）の取組について紹介し、超超プロジェクト内外から高分子のマテリアルズ・インフォマティクス、プロセスインフォマティクスに関する幾つかの事例を紹介する。

超超プロジェクトは 2016 年度から始まった 6 年間のプロジェクトで、主に有機・高分子材料を対象とし、シミュレーション・人工知能を活用した材料設計、実際のプロセス開発、計測技術開発を行っている。本プロジェクトの立ち上げ時には、計算により出力した大量のデータについて機械学習を行うことを計画していたが、プロジェクトを進行する中でハイスループットの反応装置を使って出力したデータを機械学習するというアプローチも多く行われるようになった（図 3.5.1）。



図3.5.1 超超PJのコンセプト

基本的に産業技術総合研究所と 18 社の企業の共同体（先端素材高速開発技術研究組合（ADMAT））で実施し、各企業が様々な課題を持ち込んで取り組んでいる。

プロジェクトの期間内に成果を上げた上で、プロジェクト終了後も産業界で継続的に使えるような基盤を提供するというミッションがある。プロジェクトにおいて開発した基盤技術、個別テーマに対してある程度蓄積した技術、データ、設備等を材料設計プラットフォーム（MDPF）としてワンセットで提供し産業界のニーズに答えを出していく基盤構築の提供をプロジェクト終了後の姿として謳っている（図 3.5.2）。

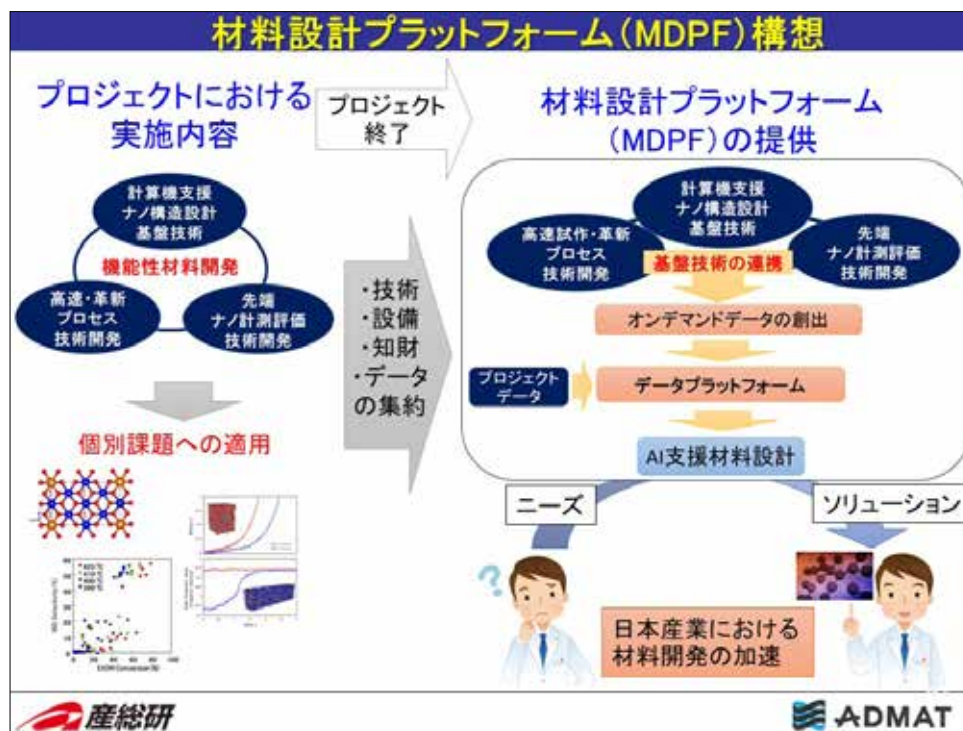


図3.5.2 材料設計プラットフォーム構想

有機・高分子材料に限らず、金属微粒子を含んだフィルム、無機フィラーを入れたコンポジット、カーボンナノチューブなどを含み、ターゲットは広い（図 3.5.3）。

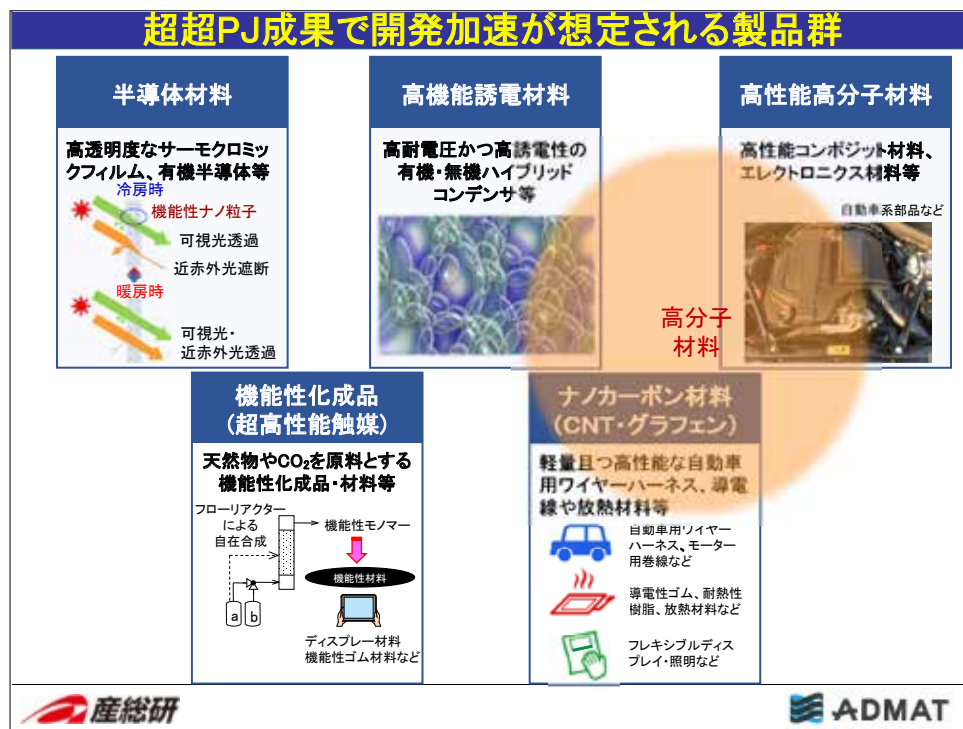


図3.5.3 超超PJ成果で開発加速が想定される製品群

高分子のマテリアルズ・インフォマティクス (MI) に関しては、経験的な物性予測は昔から行われている。モノマー構造と諸物性を相関づける構造物性相関 (Structure-property relationship) はおそらく 1970 年頃から取り組まれており、有名な方法としては原子団寄与法がある。昨今、高度な機械学習の手法を駆使してこういったアプローチが継続されているが、モノマー構造から物性機能を予測する範囲では、基本的な考え方は変わっていないと認識している。

ただ、高分子の場合、モノマー構造により屈折率や誘電率などの電氣的・光学的な物性はかなり予測できるが、機械的な物性については高分子鎖構造、高次構造を考慮せねばならない。場合によっては相分離構造、結晶構造、分子の複雑な絡み合いの構造などを考慮しないと機械的な物性が予測できないため、単純な構造物性相関ではなく、階層的な構造と物性の相関を取ることが高分子の MI には求められている。

加えて、この各々の階層構造形成あるいは物性発現に対して、全てプロセス条件が関与してくる。重合条件や成形条件、機能、物性など、高分子の場合測定条件によって変わってくるものが多々ある。例えば、ガラス転移温度は測定する手法によって異なり、また、例えば同じ示差走査熱量測定器を用いて測定しても昇温速度や種々の測定条件によって値が異なってくる。様々なプロセスの条件をこの一連の相関を取るために加味する必要がある、最近、Process-Structure-Property-Relationship (PSPR) という言葉が高分子に限らず頻出するようになった。製品の性能まで含め Process-Structure-Property(-Performance) Relationship と表現する場合もある。プロセスと構造は切っても切れない関係であるというのが高分子材料の特徴である。

現在、我々は、モノマー構造、高分子鎖の構造、高次構造、成形条件等と物性との相関について研究を進めている。文献データや計算データ、実験データなど、様々なアプローチを試みているところである。以下に具体的な事例を幾つか紹介する。

まず比較的 conventional な方法として、モノマーユニットの構造から記述子としてフィンガープリントを抽出し回帰するものがある。低分子のフィンガープリントの作製には様々な手法があるが、ポリマーの場合は組成比等を考慮する必要がある。ある程度重みづけをし、その結果、例えば、密度汎関数理論 (DFT) 計算で導出した屈折率を教師データとして機械学習を行い、計算値を精度よく再現できることを確認できている。

このように、光学物性や電子物性を含め、屈折率、バンドギャップあたりはモノマーユニットの化学構造を与えればそれなりに相関をとることができる。ガラス転移点などはローカルなダイナミクスに由来するため、これも化学構造と比較的相性よく回帰できる。しかし機械的物性には高次構造が起因し予測が難しい。

次に、もう少し高分子鎖の構造を考慮し、物性との相関を取るという試みで、液晶エラストマーを題材にした事例を紹介する。

液晶エラストマーというのは、側鎖にメソゲンがついていて、メソゲンが液晶相として相転移したり、あるいは配向したりしていくものだ。主鎖は架橋性の高分子になっており、液晶とゴム両方の性質を示すようなポリマーで、単純なゴムの弾性挙動ではなく液晶の配向や相転移によってソフトエラストシシティと呼ばれるような複雑な弾性挙動、いわゆる応力-ひずみ曲線、SS カーブという挙動を示す。主鎖の中にメソゲンがついているタイプ

もあり、鎖の構造として多様なバリエーションがある。同じメソゲンの化学構造、同じソフトセグメントの化学構造でも、鎖の構造としては非常にバリエーションが多い。その分子鎖構造をシミュレーションで様々に変え、SSカーブを計算した例もある。

粗視化分子動力学という手法で応力-ひずみ曲線を計算し、構造パラメーターとSSカーブを回帰し、どういった構造因子、特徴量が重要であるかランダムフォレストで検討し、重要度の高いものを基に、最終的に設計の方向を示すという事例もある。

ブロックポリマーの場合、異種の単量体からなるブロックは混ざらず相分離する。相分離構造を変えた時にSSカーブがどう変わるか、シミュレーションを行いデータを集めている。本来は化学構造、高分子鎖の構造などから一連を検証すべきであるが、ここでは途中段階として相分離構造を読み取るディープラーニングを用いている。マイクロ相分離構造を与え、相分離構造とSSカーブで回帰した結果、未学習の相分離構造からSSカーブを予測し、シミュレーションで得られたSSカーブと深層学習で予測したSSカーブがよく一致するという事例である（図 3.5.4）。

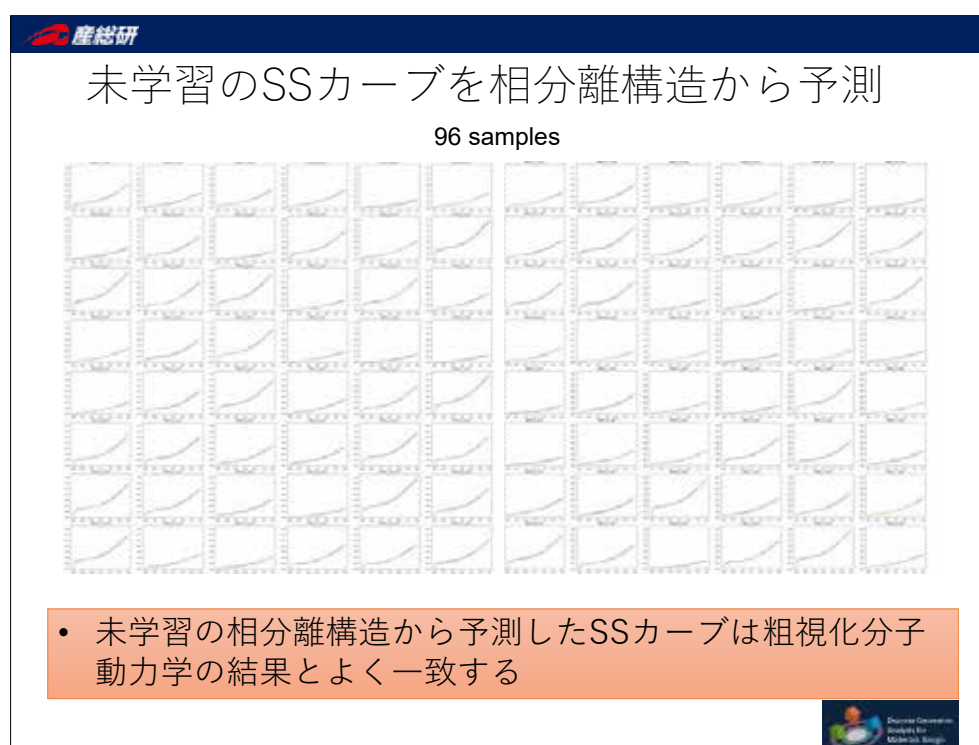


図3.5.4 未学習の相分離構造からの予測

最後に実験データを用いた例を紹介する。例えば高い耐薬品性を持ち、高温剛性と耐衝撃性を併せ持つポリマー材料の作製を目標とし、まずは高温剛性と衝撃強度を上げるためにポリマーブレンドを考える。ポリマーブレンドの場合、同じ物質材料、同じ素材を混ぜても、混ぜ方によって出来る物や物性が違ってくるといふ実験事実がある。組成やプロセス条件、実際に測定された押出機上の温度までも機械学習のデータとして用い、物性と回帰する。ランダムフォレストにより主要な因子が温度であることを見出し、温度を制御することによって物性を向上させた事例である。

プロセスインフォマティクスでは、データをどう取得するか、どう扱うか、特定の装置で取得したデータに汎用性があるか、時系列で異なる装置・条件で取得したデータを一緒に使えるかなどが課題として挙げられる。ハイスループット実験により今後同じ条件でデータベースを揃えていくことも一つの流れとしてあるが、特に産業界などでは、今現在存在するデータを何とか使えないかという希望も当然持っている。今後議論していきたい（図3.5.5）。

産総研

材料開発におけるプロセスインフォマティクス
～材料合成プロセスへのデータ科学適用の現状と展望～

- **どういう原料を使って、どのように作れば目的の材料ができるか？**
 - プロセスは必須の記述子
- **プロセスインフォマティクスの課題**
 - データをどう取得するか？どう扱うか？
 - 実験者のノウハウとして、そもそも記録されてこなかったデータ（記述子）が多々あるはず
 - 特定の装置で所得したデータにどこまで汎用性があるか
 - 時系列で異なる装置、条件で取得したデータはいっしょに使えるか

➤ 結局、大規模シミュレーションによる材料DB構築(ex. Materials Projects <https://materialsproject.org/>)のように、マスター装置を用いたハイスループット実験によるDB構築が必要なのか？

➤ 現状（手持ちのデータ、装置等）で何ができるのか？

図3.5.5 プロセスインフォマティクスの課題

【質疑応答】

Q：現在超超プロジェクトに参加している18社間では、研究データは全て共有・使用する仕組みになっているのか。

A：実際に使っている事例はさほどないが、プロジェクトの中でデータを共有することについては知財も含め立ち上げ時に合意書等で規定されており、データとして集積している。国のプロジェクトであり成果は公開が原則だが、自社の事業に大きな影響を与えるデータである等の場合には、その都度、知財運営委員会で議論し判断する運用としている。

Q：ネガティブデータ、うまくいっていないデータはどの程度集約しているのか。全く集約していないのか。

A：課題によって異なる。例えばハイスループットの合成装置等を入れ、その分析もハイス

ループットで自動的に集めるようなシステムを構築しているような場合は、ネガティブデータであっても蓄積している。全てが自動ではないため、ある程度人の判断が入る場合はある。

Q：集約対象のデータのフォーマットは共通化しているか。

A：ポリマーの押し出しから触媒の反応まで様々なデータが集まってくる中、共通フォーマットをつくることは半ば諦めている。その都度、必要に応じて構造化したデータをつくれるように、中間状態として蓄積するようなイメージである。

Q：途中段階からシミュレーションを実施した例があったが、最初から最後まで、パラメータをたくさん入力し、全てをシミュレーションすることは難しいのか。

A：各段階のデータを同じような量と質で揃える必要があり、現状困難である。

3.6 機械学習から見た関心と課題

瀧川一学（理化学研究所、北海道大学）

機械学習側から材料プロセスの課題がどう見えているのかについて簡単な紹介をしたい。例えばバイオインフォマティクスの場合は、元々ヒトゲノムプロジェクトが国際プロジェクトだったこともあり、また、材料・プロセスと比べると、ゲノムの配列を公開したところでそれ自体がそのまま価値を生むものではなく、データが非常にオープンである。一方、材料・プロセスに関わる内容は、研究所や企業の競争力そのものであるために、共有できるデータが少ない。機械学習の専門家が入りにくい一番の障壁になっているのではないだろうか。バイオインフォマティクスでは、そもそもプラットフォームがたくさんあり、“オレオレ・データベース”や“オレオレ・ウェブサービス”みたいなものが乱立したが、それを統合する仕組みや標準化機構・コンソーシアムなど、集約を担う組織が欧米にできた結果、風通しがよくなった。

私からのメッセージは3つあり、1つはデータ科学的なやり方を取り入れるときには様々なレベルがあり、どのレベルの話をしているかに食い違いがあると、訳が分からなくなってしまうということ。この点を心に留めて置いていただきたい。2つ目は、データ科学を使おうと思ったら、データ取得の戦略や方法、データに偏りがいいかなどは、どの手法を使うのか以上に非常に重要であるということ。材料プロセスではより難しい部分になると感じている。3つ目は、機械学習に丸投げするとあまり幸せにならないということ。基本的には、機械学習にどのパートを任せるかをよく考えたほうがいい。

本日の3つのメッセージ

1. 「データ科学」的なやり方を取り入れると言っても**色々なレベルがあり、各々のニーズに応じて必要な範囲で取り入れてみるのが重要です。**
2. 「データ科学」的なやり方の実効性を左右する「**心臓部**」は「**データそのもの**」で**データ取得の戦略・手法・計画や品質評価は(アルゴリズム以上に)重要です。**データが取れない対象や問題には使えないので不向きな問題もたくさんあります。
3. 立場として「**この不毛な作業を誰かが代わりにやってくれないかなあ**」を解消し「**私たちが私たちができないこと(やりたいこと)に集中するため**」の**委任技術**であって、当然あらゆることに適用できるものではありません(“予知”や人的介入の完全排除は原理上不可能)。「何をどの程度任せるか」が重要です。

図3.6.1 3つのメッセージ

様々なレベルがあるということについて、私自身は有機合成の研究と関わることが多いが、実験をおこなう際にこれまでのデータに基づいて判定しようというレベルのインフォマティクスは、問答無用に全員が取り組んだほうがよいと思っている。それより上のレベルでは、必ずしも全員がこの方向に行く必要はないと考えている。機械学習としては、それぞれが考えているレベルによって向いている手法も基本的に異なる。例えば私は機械学習自体の研究をしているが、より発展的で挑戦的な技術研究を行うケースが多く、ほとんどの場合はそれがすぐ身近な実問題に使えるわけではない。有機合成で有名なインフォマティクス活用の例に Sigman グループの一連の研究があり、新しい化学反応をみつけるためのデータの活用戦略として極めて示唆に富んでいるが、一方で機械学習技術として新しいことはこういう研究にはあまりない。技術的には線形回帰を用いた伝統的な実験計画法の応用で、おそらく機械学習側の人にはそのような研究結果は届いていないのではないかな。

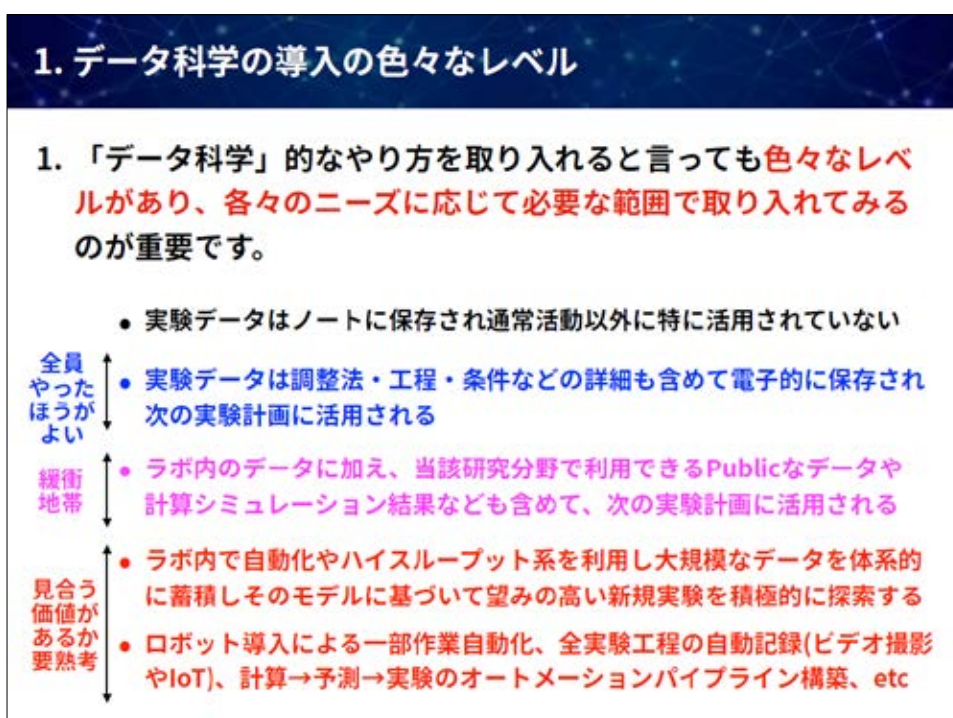


図3.6.2 データ科学導入の色々なレベル

2つ目としては、材料プロセスの最適化では、実験計画法は長い歴史を持っており、まずはそういう確立された方法論を取り入れることから始めることが非常に大事である。つまり、どういうデータを取るかを、実験をおこなう前に計画するという。似たような話は触媒開発などのところでも、最近色々と適用されている。

機械学習側の関心として最近注目を浴びているのは、第一原理計算とハイブリッドさせるやり方。どこかに大規模なデータがある場合、非常に大きなモデルを学習させておけば、それを細かい個別のタスクに転移することは可能かもしれない。

2. 「データ科学」では「データ」が全てです

2. 「データ科学」的なやり方の実効性を左右する「心臓部」は「データそのもの」で**データ取得の戦略・手法・計画や品質評価は(アルゴリズム以上に)重要**です。データが取れない対象や問題には使えないので不向きな問題もたくさんあります。

例. 手元に30件の事例があるのでこれと機械学習とやらで何かできないかな？

- 最もありえる事態だとは思いますが「失敗リスクの高い」サインです。この場合「**この30件がどう取られたデータか**」に激しく依存するので機械学習に投げるよりは通常の探索的データ解析やデータ可視化などで次の実験計画や対象理解につなげる方が個人的にはオススメです。
- そもそも論として本当に30件が結構大変な実験系であればあるほど事前によく「実験計画」をしたほうが良いと思います。また件数が3000件とか30万件とかあればこの点を気にしなくて良い訳でもありません。

図3.6.3 「データ科学」では「データ」が全て

もう一つの関心としては、人間がデータをつくると何らかの偏りによって、どうしてもバイアスが反映されるという問題がある。材料だけではなく、AI の技術全般として非常に問題になっている。

最後のポイントとして、基本的には全自動化できるものではないので、インフォマティクスにそもそも不可能な期待をしている事例がある。これは機械学習外の問題であり、やはり何をどの程度任せるのか、事前にしっかり考えておく必要がある。明らかに人間がやらなくても何とかできるのではないかというタスクが発生している場合は、機械学習が有用な可能性があるが、そもそも2つ目に言ったデータが取れるか次第。最近、機械学習の分野で注目されているのは、ロボットを用いた自動合成系を使うケース。創薬では10年20年の歴史があるが、ハイスループット系では後発の中国企業等が、やはり全自動にして追いつこうとしている。

機械学習の関心としても探索というのは、材料だけではなく非常にホットな分野である。アルファ碁が有名だが、ゲームだけでなく機械学習自体にユーザーがチューニングしなければならないパラメータがたくさんある。それ自体も機械学習でチューニングしようという“AutoML”というジャンルがあり、技術的にも商業的にも非常に研究が盛んな分野になっている。これらの技術は、もしデータや問題がきちんと設定されたら、材料・プロセスにも使えるのではないかと考えている。

3. 何をどの程度任せるか

3. 立場として「この不毛な作業を誰かが代わりにやってくれないかなあ」を解消し「私たちが私たちができないこと(やりたいこと)に集中するため」の委任技術であって、当然あらゆることに適用できるものではありません("予知"や人的介入の完全排除は原理上不可能)。「何をどの程度任せるか」が重要です。

- ご神託のように有望な結果や示唆が得られるわけではありません。いくつかの仮定のもとで確率を改善していく技術です。(特に探索の文脈においてセレンディピティが不要になるわけではありません)
- 自然科学は「人間の存在とは独立した自然法則を研究している」ように見えて、「自然法則を見つけたい・理解したい」というのは人間の勝手なニーズなので、どこかで必ず人間の特性の考慮が必要です。
- 材料合成プロセスの最適化・合理化・効率化のような話では現実的制約の中の最善策の模索になりどこでどういう「人的タスク」を入れるか熟考する必要があるように思います。

図3.6.4 何をどの程度任せるか

合理論(計算化学)と経験論(機械学習)の融合に向けて

Theory-driven 【合理論】

- 対象現象の複雑化
- シミュレーション技法も複雑化
- "経験的に決める"パラメータや初期値
- 汎関数、交換相関項の設計

(人工知能分野)

- 知識ベースと論理推論(記号AI)の限界
- 厳密推論や探索の計算爆発(NP困難性)
- 大量データの知識化の問題
- 制約プログラミングや組合せ最適化

データ同化、模倣学習、論理合成、etc

モデルベース最適化・強化学習、メタ学習、ドメイン適応、生成モデル、etc

新たな方法論へ？

Data-driven 【経験論】

- 小サンプル・低カウントの問題
- 外挿の不可能性の問題
- 帰納バイアスのモデルエンコード
- Blackbox性・解釈性の問題

(人工知能分野)

- Data-Driven手法(機械学習)と人間の論理的思考との大きなギャップ
- Dataがない領域の探索や「ひらめき」
- モデル適用範囲と信頼性・安全性

図3.6.5 合理論(計算化学)と経験論(機械学習)の融合に向けて

3.7 第2回 総合討論

ファシリテーター 伊藤聡（計算科学振興財団、JST-CRDS）

プロセスに関しては、データの共有ができないことがかなり問題であるが、一方で、データは産業界がかなり持っているというのが現状だろう。今回のワークショップでも、産業界の方に講演いただきたかったが、やはりプロセスデータを社外に出すことはできないので講演することが難しいとのことであったので、コメンテータとして参加していただいている。まずは企業からの2名のコメンテータに本日の各講演、あるいは各社での知見を踏まえて、コメントを頂きたい。

【富谷茂隆（ソニー株式会社）】

我々は材料よりも、特にそれを使ったデバイスを作製するということに注力している。もちろん、新材料探索は行っており材料探索は非常に重要だが、その材料を使って実際にデバイスをつくるのが我々にとってはキーとなる。材料を実際に使いこなしてデバイスに組み込むというところで、プロセスが非常に重要になる。

材料と個々のプロセス装置との相互関係が非常に強いので、プロセス装置が変われば実際に特性も変わってしまうという課題がある。同じ装置を導入しても、実際にできたものの特性が変わってしまうことがある。例えば、結晶成長装置においては、僅かな装置形状変更で特性が変化してしまい、思うようにいかない場合がある。また、有機半導体では結晶構造をきちんと理解することが必要なのだが、成膜条件により変化する場合があり、その制御が難しい。

もう一つは、材料メーカーから材料を入手してデバイス化した際に不具合が生じた場合、その原因究明が難しいことがある。材料そのものが原因なのか、デバイス化したときのプロセスが原因なのかの判断に困ることがある。この究明には、材料メーカーとのコラボレーションが非常に重要だが、材料メーカーも決められた範囲での特性は情報共有をしていただけるが、それ以外はなかなか共有化しにくいからである。例えば、デバイスを作って特性がおかしい場合、よくよく調べてみると、実は材料メーカーでプロセスのロットが変わっており、それによって、特性が微妙に違っていたということもあった。

このような情報共有は、課題の一つだと思う。なにか仕組みができないかと考えているが、なかなか前に進まない状況である。

【池端久貴（旭化成株式会社）】

旭化成では「MI」と称して、社内の様々な開発課題に対してインフォマティクスを活用している。「MI」と称しているが、実際のところ、製造条件などを含めて機械学習で特性を予測する、まさにプロセスインフォマティクスをやっている。

多くの製品の開発テーマで活用しているが、大体プロセス条件を入れて、目標特性を2ないしは3項目、多いときは7～8項目を目標の範囲に入れるような検討を、機械学習モデルをつくってベイズ最適化で探すというやり方をスタンダードにしている。

ただし、全てのケースでうまくいくわけではなく、今のところ成功率は充分とは言えない。うまくいくときはデータの数がそれなりにそろっている場合であり、感覚的には数百ぐらいのデータがあるテーマだとうまくいく。また、実験の誤差が少ないデータが得られる場合にはうまくいく。

また、過去のデータを活用しようとしても、装置が違っていたり、用途や顧客が異なり要求特性が異なるので、測定項目が違っている場合があり、既存の機械学習がそのまま使えない問題がある。

それに対して、一つは、ハイスループット実験（HTS）などでデータを増やす方法がある。もう一つは、過去の実験、類似しているテーマのデータを活用できるような機械学習モデルをつくる方法がある。

企業でそのようなプロセスインフォマティクスを進めようとするときに、どちらも一企業ではかなり難しいと考えている。

例えばHTSは実施できるテーマが限られてしまう。稼働し始めるとデータはたくさんとれるが、立ち上げまでに非常に時間がかかり、また予算がそれなりにかかる。企業では基本的に開発期間が決まっている。HTS装置を検討するのに1年ぐらいかかってしまうと、それだけで時間が経ってしまって難しい。

したがって、例えば、国、どこかの共同研究利用機関にHTS装置、分析機能、データベースを用意しておき、共同利用できるとありがたい。データは共有して蓄積されて、予測モデルに落とし込むことができる仕組みがあるとよい。データベース内のデータ自体は見ることにはできないが、予測モデルを活用できるようにする。ただし、モデルからデータが分からないようにする仕組みは必要ではあるが。これができると、企業でも気軽に使いやすと思う。この仕組みが作れば、データも揃ったフォーマットで蓄積していくので、とても重要な資産になる。

また、新組成、プロセス条件などをインフォマティクスで探す際になかなかうまくいわずに、気づいたら人間が新しいものを見つけていたという例は結構ある。これまでに使っていない方法を適用した場合に、こうなることがある。例えばこれまで単層膜でやっていたものを複合膜にするとか。方法を変えると、そもそもデータ形式が違うので既存の機械学習やベイズ最適化では到底たどり着かない。この状況で、機械学習ができるようにするには、何か新しい手法をつくらないといけないと思っている。

また、例えば、ある樹脂の組成物を開発しているデータはある開発部所があり、別の開発部所でそれを使って膜を開発しているケースがあるとする。このような場合、両方の開発部所のデータを活用したいのだが、それを実現するための機械学習モデルがない。この場合、例えば、組成物の物性を予測するモデルや、プロセス条件から膜の性能を予測するモデルなどのニューラルネットモデルを複数用意しておいて、それを直列につないで活用することが考えられるのではないかと。複合膜の場合には、組成物の物性を予測するモデルを並列につないで、さらにそのアウトプットを膜のプロセス条件に繋げるとか、フレキシブルに使えるモデルがあるといいと思う。それを一企業で作る人的リソースもないので、そういう材料の世界に特化したような機械学習モデルを一緒に作り上げていく必要があるのではないかと思う。

- ・もちろん実験データが必要だが、シミュレーションを活用するという方法もある。少ない実験データと何らかの相関を取ったような大きなシミュレーションデータを使って、それをサロゲートモデルに使うという議論もあった。
- ・ハイスループット実験で必要なデータを均質に集めると機械学習の精度が上がるという例もあった。また、ネガティブデータも重要であり、それを除外してしまうとあまりうまくいかないという報告もあった。
- ・プロセスデータは、産業界に非常に多く存在するが提供されないという問題があり、これをどうするのかは考えなければいけないことだろう。
- ・同じ装置でも、例えば装置をオーバーホールするとデータが変わるとか、そういう装置依存の問題、あるいは、装置も劣化するので導入初期としばらく経った後では同じパラメータを使っても同じものではないという問題もある。
- ・やはり拠点として何か考えなければいけない。ハイスループット実験のデータ収集ができるような拠点があるといい。デバイス関係ではベルギーにIMECがあり、産業界が必要に応じてデータをとれる。
- ・人材育成という面では、データサイエンティストにこういう場に参加してもらうことも必要。
- ・データをその場で取ろうとする際、何でもとれるというのはあり得ないので、分野ごとの拠点化とかプラットフォーム化も要るのではないか。

【現状 何ができて、何が難しいのか。将来はどこまでできるか、それに対する課題】

【データについて】

- ・ネガティブデータの重要性について。実際にうまく行かない点のデータがないと、その部分が実験データの外挿領域になってしまうので、正しい予測ができず、楽観的な予測になってしまう。プロセスインフォマティクスの場合は、この合成法だとはここは合成できないというデータも必要になる。
- ・マテリアルズインフォマティクスは理論からのデータだけでは不十分で、そこが限界になっている。実験データの重要性、特に予想しない結果を生み出すという意味でハイスループット実験データは重要。ネガティブデータも似た意味で重要であり、そこを考慮してプロセスデータを集める必要がある。
- ・“マイデータ問題”については、全く同じ条件で取ったつもりなのに、結果が違っていると実験の皆さんがぼやいているのをよく聞く。実験している人が同じ条件に思えるのなら、入力条件はそれ以上制御できないので、装置内を測定するしかない。その測定データを説明変数にし、プロセスでの生成物の物性などを目的変数にすることになる。こういう議論をする際に、センサーや測定技術に詳しい方が少なくて話が止まってしまうことが多い。
- ・センサーであれば、メーカーに聞いた方が話が早いかもしれないが、センサーメーカーの

開発者と実験でデータを集める研究者が直接話をする機会が少ない。測定機器メーカーと実験のデータを集めている現場の人たちが話しできるような機会がどこかにあると、話が進んでいくように思う。そういう場を誰かが作ってあげるといいと思う。

- ・プロセスインフォマティクスと計測は非常に密接に関係している。全てのプロセスパラメータが計測できているのであれば、それを基に解析できるはずだが、なかなかそれができていないのが現状。そこを改善できるような計測技術を入れていくのも確かに重要。マテリアルズインフォマティクスとプロセスインフォマティクスに加え、計測も含めて一体でやらないといけない。
- ・全く同じ条件で取ったつもりなのに結果が違うというのは、実際にある話。例えば、学生が月曜日から金曜日に暖房を入れて実験をやっている、金曜日に学生がセットして暖房を切って実験を続け、月曜日にその結果を見ていると、同じ実験で同じ材料を使っているのに全く答えが違うということはよくある。よくよく調べてみると暖房を切ったことによって配管が冷めていたから活性が落ちたのだった。また、夏と冬では同じことをやっているのに全く違うデータが出るということはよくある。実験条件以外のそういう環境因子（温度や気圧など）も全部パラメータ化してやると、よりよくなるのではないかと感じていた。

[実験データとシミュレーション]

- ・シミュレーションデータは確かに有効だが、ものをつくろうとする際には、見えない因子があまりにも多過ぎて、本来想定しないことが多いので、そういうものを全て想定してシミュレーションするのは難しい。
- ・マテリアルズインフォマティクスでは、シミュレーションデータをベースにして（不純物の効果は入れられないが）それで物性を予測して、実際に作って検証する方向もあったが、プロセスインフォマティクスに関しては、最初から実験データを中心にやっていかないと難しいのではないかな。
- ・自分自身でもアメリカの Materials Project の材料や触媒材料を、自分で計算した結果を基に実験チームに試してもらったことがあるが、うまく作れた例が少ない。理由は、プロセス中で想定していない挙動が起きていたことだった。プロセスの実験データは想定しにくい挙動も含んでいるので、ものづくりに関しては本当に重要である。
- ・ハイスループット実験装置であっても、環境因子は慎重に考慮する必要がある。再現性が取れないほど精密な温度制御を狙わないこと。熱容量を大きく取ること、雰囲気制御をすることなどをちゃんとやらないと、確かに環境依存というのは起きる。特に水に弱い材料を扱っているときは、しっかりやっているつもりでもその日の湿度の影響を受けることはある。ハイスループット実験装置をつくる前に、まず、制御可能な実験をするための条件を全部洗い出しておかないと、装置の運用は結構厳しくなる。
- ・触媒を作る際は速度論的要素がすごく強いが、そのプロセスの中で起きている化学反応の様子が誰にも正確には分からないので、シミュレーションをするのはかなり難しい。マルチスケールな反応シミュレーションで環境因子まで入れて、時間単位のプロセスを再

現するのは非常に難しい。そこがもし反応工学モデルと併せて本当にできるようになるのであれば、すごくいいのだが。

- ・超超プロジェクトでは、そのスケールに応じて、あるいは対象に応じて、計算がうまくいくところは計算を使いインフォマティクスを実行する。もう少しマクロな状況になって計算が難しくなる部分は実験でデータを取る。というように場合分けをしている。
- ・計算でできることと、できないことは明確に分けなければいけない。今までのマテリアルズインフォマティクスの一つのアプローチは、計算で入力も出力も出して、機械学習を使って予測できたと言っているような形だったと思う。しかし、実験データを予測するのが最終的な目標であるので、計算だけでは難しい。一方で、実験だけでも難しい。高分子の例にもあるように、モノマーユニット構造から高次構造まで全部を実験で収集するのは難しい。特に高次構造、相分離構造などは、ハイスループットで電頭で集めるようなことは現実的には不可能である。上手に記述子として計算や実験を併せて取り入れる、現実的には両者を組み合わせていって何とか精度を上げようとしている。これはやはりケース・バイ・ケースで、いろいろ試行錯誤をしながら、とにかく組み合わせていい方向を目指しているというのが現状だ。

[拠点]

- ・対象によってアプローチを変える必要があるとなると、ますます戦略を立てるのが難しくなる。戦略を立てていく上で重要なことは、将来どういう形でこれをやっていく体制を持つべきか、そういうことだと思う。
- ・みんなが使えるような拠点、プロセスのセンターを用意して、そこでいろいろなことを試す。そうすると、そこにデータも集まる。そのような仕組みが要るのではないか。あらゆるプロセスをその拠点に置くわけにいかないの、反応タイプごとに標準プロセスを選ぶことになる。
- ・液相合成ではマイクロフローリアクターのような技術が標準になるのではないか。均一系の反応では、データの信頼性、再現性の高さという意味で、全てフローでデータ化するのがいいと思う。不均一系においても、例えば触媒をカートリッジに詰めての実験例ではフロー合成での再現性の高さが一つの売りになっているので有効だと思う。また反応後の精製や分析が問題となることもある。フローリアクターでは、分離抽出や生成物分析など全てをインラインで実施でき、非常に再現性の高いデータがとれると思う。
- ・フロー化学では、MITのJensen先生はかなりそういうことを意識して研究しているのではないか。日本でフロー化学をやっているコミュニティはあまり大きくない。
- ・産業界の研究テーマでも、非常に小さいものからビッグインダストリーにつながるようなものまで様々なテーマがある。小さなテーマで始めるときには、拠点を利用できると非常に助かる。例えば、CVDで、コンビナトリアルで試したいようなときに、小さなチームでコンビナトリアル装置をいちいち導入するとそれだけで半年から一年たってしまうので、拠点が設置されていれば活用できる。その場合のデータあるいはテーマについてはコンフィデンシャルな部分もあるかと思うが、そこはルールをうまく作る必要がある。

- ・ 拠点に標準的な装置を置く場合でも、シミュレーションもうまく活用して、気象分野でよく使うデータ同化のような技術も加えいくと、非常にバランスよくプロセス開発、材料開発が進むのではないかと思う。そのときに、やはり計測が非常に重要になるので、図 3.7.1 は、まさにそこをうまく表している。
- ・ 拠点は複数持つ必要がある。その拠点ごとに、例えば、無機に強い拠点、高分子に強い拠点みたいな形で、要は素材に特徴があるような拠点を立てていくといったことを考えるのがよい。それにより、適したプロセスの技術や計測の技術、マテリアルズインフォマティクスのスペシャリストも集まってくる。そうなってくると、その後は、自動的にその材料分野でのデータが集まってくるだろう。この材料のこの領域は日本の中でここがオーセンティックなところなので一緒にやろうということになり、極めて分かりやすくなっていく。まさに IMEC などがそういうイメージ。
- ・ 材料のデータフォーマットが揃わないという問題も、素材の種類ごとにカテゴライズすることで最大公約数的にまとめていけるのではないか。
- ・ 企業側からデータが出てこない問題もあるが、これはやはりなかなか難しい。企業の競争原理、競争の資本になるところであるので。ただし、将来的なやり方として、例えばデータ取引みたいな、新たなビジネスモデルみたいなことがうまく立てられるのであれば可能性はあるかと思う。データの部分に関しては、もう少し別の視点での枠組みの議論が必要かと思う。
- ・ 材料ごとの拠点もその拠点間のネットワークも必要。データ共有はできなくても情報は共有する必要がある。そのためには様々なものの ID 化、トラッキングできる仕組みをつくる必要がある。NIMS がやっているような、材料データをプラットフォームとして集めていくアプローチがある。
- ・ プロセスは非常に膨大な領域をカバーしているので、領域ごとに特有の求められるものがある。ある程度のまとまりを持ってプロセスを考える必要があるため分野ごとに小規模なワークショップを何回か積み重ねている。拠点の考え方もワークショップの設計の考え方と類似しており、分野ごとに拠点を設置するのがよいと思う。また、データは、将来的には、特許と同じように、売買できるアセットになってくるだろう。そういう価値として、産業界や公的機関で認知され、アセットとして評価する仕組みができると、より動きやすくなるのではないか。

【人材】

- ・ 拠点を幾つか置く場合、データサイエンティストの数が不足しているという問題がある。データサイエンティストが実験者の近くにこそ、一緒になって議論し、データ科学的なアプローチができるのではないかと思う。それぞれの拠点にそれぞれデータサイエンティストを張りつけなければいけないのか、あるいはそうでなくてもいいのか。
- ・ 情報側としては、物理的に近くにいない必要は感じないが、コミュニケーションをすごく取りやすい状況でないとうまくいかない。お互いに「自分の水」で研究すると結局全然融合が進まないというようなことはいたるところにある。一方で、データサイエンティストにとって、例えばプロセスインフォマティクスの拠点に入る魅力が少ないことを危

惧している。あまりよく分からないところに飛び込むより、他にもいろいろデータサイエンティストにとって魅力的な出口があるのでそちらを優先する。今までと異なるところに行くのにはそれなりにインセンティブが必要。

- ・プロセスインフォマティクスでは、まずデータフォーマットや集め方などがまだ定まっていないので、データサイエンスの手法研究がなかなかできない。データ収集・整備に当たることに對して、データサイエンスを学んできた人が魅力を感じるかは疑問。多分、手法に関心がある人が多い。少なくともデータが数百件とかでは手法側で面白いことはなかなか難しい。ただし、データサイエンスにもいろいろなタイプがあるので、応用統計科学寄りであれば、こういう問題に関心が深いかもしれない。
- ・先端アルゴリズムをやっている人間はなかなかこちらに来てくれないだろう。当社では基礎的な機械学習の手法に関しては基本知識を持っている人がやればいいという認識で、底上げの活動をしている。
- ・拠点をつくる場合においても、データ科学者／応用統計科学者などのカテゴライズをうまくして、コミットしてもらう必要がある。
- ・大学で全専攻の学生に対して、データサイエンスを知らないマテリアル系の学生のデータサイエンス知識を底上げするカリキュラムがある。情報系の学生にプロセスや材料のことをアピールするより、マテリアル専攻の学生のデータサイエンスの知識を底上げする方がやりやすい。情報系の学生とマテリアル系の学生でプロジェクトベースラーニングのように何か研究テーマを立ち上げて、ちょっとした機械学習モデルをつくって発表するという演習が割と円滑に進む。どちらかというと、マテリアルやバイオの学生たちのデータサイエンスの知識を底上げする方が話は早い。

【オーガナイザによるクロージング】

マテリアルズインフォマティクスがここまで来て、限界も見えてきている。その中で、プロセスインフォマティクスがどういう方向へ進んでいくべきかを検討するのに必要な情報を非常に多く頂いた。

特に、実験的な手法がとても重要だということが、非常に強く印象に残っている。その意味で、ハイスループット実験は非常に重要である。また、インラインでいろいろ計測する、要するにそのプロセスの過程で何が起きているのかを見る、そういう計測、オペランド計測が進んできており、プロセス中でどういう反応経路をたどっているのかを知るとはとても重要だということを認識した。

マテリアルズインフォマティクスでは、アメリカのマテリアルゲノムイニシアティブが先行した形であったが、プロセスは日本のお家芸、日本の産業競争力との視点でも強くしていかなければいけない領域であるので、プロセスインフォマティクスでは、ぜひ日本発のコンセプトを出していきたい。今回のワークショップを通じてプロセスインフォマティクスが必要であり、何とか具現化できそうだという確信をかなり深めさせていただいた。最後になるが、本日の総合討論の場で、中心になって議論頂いたワークショップの話題提供者の皆様方、また貴重な意見を多数頂いたコメンテータの先生方に深く感謝したい。

4. 総合討論（まとめ）

第1回（無機結晶・多結晶分野）、第2回（触媒・有機材料分野）の総合討論の論点を以下に、まとめる。（各文末の<1>、<2>はそれぞれ、第1回、第2回での総合討論でた論点であることを示す。）

【データについて】

- ・ 産業界の持つデータは開示されないと考えた方がよい。<1、2>
- ・ 他のデータ駆動型研究以上に産学連携が重要である。事業の機微にあまり関わらないところで産学連携を推進する方法を探すべき。<1>
- ・ マテリアルズインフォマティクスとプロセスインフォマティクスに加え、計測も含めて一体でやらないといけない。<2>
- ・ プロセス内で起きていることのその場計測を実現し、シミュレーションとの連携へ繋げることが必要。<1>
- ・ 装置の劣化（経年変化）による状態変化も大きな影響を与えることがある。直接観測できない現象に対してソフトセンサー的アプローチを活用することも有効。<1>
- ・ 実験条件以外の環境因子（温度や気圧など）も全部パラメータ化してやると、よりよくなるのではないか。<2>
- ・ できるだけ多くのデータを取得することは非常に重要。しかし、すべての条件データを取得しておくことは実質的に不可能。過去の経験・知識から重要だと認識されている項目を活用することが重要。<1>
- ・ メタデータの標準化は最低限必要。<1>
- ・ 特に予想しない結果を生み出すという意味で、ハイスループット実験データは重要。ネガティブデータも重要な意味を持ち、そこを考慮してプロセスデータを集める必要がある。<2>

【実験データとシミュレーション】

- ・ 実験データを取るのには膨大な時間がかかる場合が多い。シミュレーション技術を上手に活用すべきである。シミュレーションと実験データの連携も大切である。<1>
- ・ 欧米の市販ソフトの活用だけでなく、自ら作成すること、過去のプロジェクトで作成したソフトを有効活用することも重要である。<1>
- ・ シミュレーションデータは確かに重要であるが、ものをつくろうとする際には、見えない因子があまりにも多過ぎる。想定できていないことが多いので、そういうものを全て含めてシミュレーションするのは難しい。<2>
- ・ プロセスインフォマティクスに関しては、最初から実験データを中心にやっていかないと難しいのではないか。<2>

- ・ 計算や実験を上手に併せて取り入れることが有効。現実的には両者を組み合わせていつて何とか精度を上げようとしている。<2>

【拠点】

- ・ 実際に作るところから計測まで、一通りぐるっと回す「マテリアルファウンダリー」をそれぞれの分野で作るのが一つの将来像。たとえば、製膜装置と評価装置と、データを蓄積するサーバ、そこにデータサイエンティストがいるという形。<1>
- ・ 「メーカーズ」的な発想の、コンパクトな拠点を作るのがよい。それぞれの拠点がバーチャルにつながっている形。<1>
- ・ みんなが使える拠点、プロセス・センターを用意し、そこでいろいろなことを試す。そうすることで、そこにデータも集まる。そのような仕組みが要るのではないか。あらゆるプロセスをその拠点に置くのではなく、反応タイプごとに標準プロセスを選ぶ。<2>
- ・ 拠点到標準的な装置を置く場合でも、シミュレーションもうまく活用して、気象分野でよく使うデータ同化のような技術も加えていくと、非常にバランスよくプロセス開発、材料開発が進むのではないか。<2>

【プロセス分野におけるデータ科学】

- ・ 少ないデータで機械学習をやる場合に、間違いデータをいかに除外するかが重要。そのような研究も大切。<1>
- ・ 材料科学への機械学習の適用は、まだプリミティブな段階にあるのではないか。自然言語処理の分野では、非常にパラメータが多いが、パターン認識により上手にやっている。ステップ・バイ・ステップにやるのではなく、機械学習によりいきなり最適なプロセスを出してくれるようにするというのも、一つの方向になる。<1>

【人材】

- ・ 複数の専門性を持った人材が必要。マテリアルのことをきちんと学んだ上で、機械学習にも取り組む人材。こういうタイプの人間のキャリアパスが現在、日本にはない。<1>
- ・ データサイエンティストにとって、プロセスインフォマティクスの拠点に入る魅力が少ない。応用統計科学寄りであれば、こういう問題に関心が深いかもしれない。データ科学者／応用統計科学者などのカテゴライズをうまくして、コミットしてもらう必要がある。<2>
- ・ マテリアルやバイオの学生たちのデータサイエンスの知識を底上げする方が話は早い。<2>

【その他】

- ・ 将来的には、例えばデータ取引みたいな、新たなビジネスモデルがうまく立てられるのであればデータ流通の可能性はある。もう少し別の視点での枠組みの議論が必要。<2>

付録

付録 1. 開催趣旨・プログラム

開催趣旨

ナノテクノロジー・材料分野は物理学、化学、生物学を横断し、原子分子レベルでの観測や構造形成・機能発現などを通して、物質科学や材料技術、デバイス技術などを進展させ、さらには異分野の融合を促進しつつ進化する技術分野です。CRDS では 2019 年 3 月に、ナノテクノロジー・材料分野における世界各国の国家計画、投資戦略、研究ポテンシャル、技術進化、産業動向、重要な研究開発領域を含む俯瞰報告書（CRDS-FY2018-FR-03）を発行しました。この俯瞰報告書は、CRDS が政策立案コミュニティおよび研究開発コミュニティとの継続的な対話を通じて把握した研究開発の大きな流れを、研究開発戦略立案の基礎資料とすることを目的として、CRDS の視点からまとめたものになります。各分野における研究開発の方向性や主要な研究開発領域、さらに国際的なわが国のポジションを把握するのに役立つものとして、また、複数分野にまたがる新しい切り口からの研究開発戦略を立案することにも役立つことを期待しています。当該分野の動向を深く知りたいと考える政策決定者、行政官、企業人、大学・独法関係者、また、研究者にとっても、自身の専門分野を超えた範囲の状況を知る上で有益な資料となることを期待しています。

2019 年発行済のナノテクノロジー・材料分野の俯瞰報告書では、32 の重要領域について研究開発状況や国際比較を取り纏めています。さらなる調査・検討が必要な領域や、研究動向の変化・進展のきざしなどから新たに検討対象とすべき領域があると認識しています。そこで、昨年度から今年度にかけて、いくつかの領域に注目して調査しワークショップを計画・実施しています。

本ワークショップの主題である「プロセスインフォマティクス」は、必ずしも明確な定義があるタームではありませんが、広くは、材料合成プロセス開発にデータ科学の手法を導入し、これまでにできなかった材料の合成や、既存の材料の特性向上を可能にするサイエンス／テクノロジーとして理解されています。

材料探索にデータ科学を適用する試みとして、先行するケモインフォマティクス、マテリアルインフォマティクスにおいては、すでにいくつかの成果が得られており、新材料開発に一定の役割を果たし始めていますが、それらの成果のほとんどが、目的とする物性を持った物質の組成や結晶構造を見つけ出すものであり、新しい製造プロセスを見つけ出すことにデータ科学が使われる例はそれほど多くはありません。これは、材料の組成・構造を決めるパラメータ空間よりも、製造プロセスを特徴づけるパラメータ空間の方がはるかに広大であることが一つの要因であると思われます。

しかし、昨今の先端テクノロジー製品に用いられる材料特性には、従来以上に高度な性能や、相反する性能を併せ持つ（強磁性体の飽和磁化と抗磁力など）ことまでが求められるようになっており、これまで行われてきた材料合成手段を超えた合成手段、準安定状態や非平衡状態を積極的に利用するような、新たな物質合成スキームを検討する必要があります。このような、新合成プロセスの構築には従来よりもはるかに精緻なプロセスパラメータの制御が必要になると予想されるため、プロセス開発の手法自体を革新していくこ

とが求められています。プロセスインフォマティクスは、従来の試行錯誤的手法では手に負えないほどの、膨大な数のプロセス探索を実験とデータ科学の融合により加速し、求める最適プロセス構築を可能にする可能性があると期待されます。

以上の観点から、まずプロセスインフォマティクスの現状を理解し、将来の発展の可能性、解決すべき課題などを把握するためのワークショップを開催します。ワークショップ開催にあたり、材料系によってプロセス技術の違いが大きすぎ、共通の課題やタームがないため議論が深まらない可能性を考慮し、材料合成プロセス上共通点が多いと思われる材料の分野ごとに小規模なワークショップ行うこととします。第1回は「無機結晶・多結晶」分野のプロセスに、第2回は「触媒、有機材料」分野のプロセスに焦点を当てて実施します。

なお、本ワークショップの内容はワークショップ報告書として取りまとめるとともに、CRDS 俯瞰報告書 2021 年版にエッセンスを反映する予定です。また、CRDS における戦略立案に活用するとともに、各界における施策立案等に活用いただくことを念頭に置いています。

第1回 プログラム

(敬称略)

開催日時：2020年2月11日(火) 13:00～18:00

開催会場：JST 東京別館 2階 A会議室②

オーガナイザー

曾根 純一 (JST-CRDS)

13:00～13:05 開会挨拶

曾根 純一 (JST-CRDS)

13:05～13:20 ワークショップの趣旨説明

眞子 隆志 (JST-CRDS)

13:20～14:50 (話題提供1) (各発表20分+議論10分)

13:20～13:50 新材料開発のスマート化の潮流とプロセスインフォマティクスの重要性

知京 豊裕 (NIMS)

13:50～14:20 結晶成長モデルベース開発 (SiC 溶液法における大口径化を例に)

宇治原 徹 (名古屋大学)

14:20～14:50 結晶成長とインフォマティクス：温故知新 柿本 浩一 (九州大学)

15:00～16:00 (話題提供2) (各発表20分+議論10分)

13:20～13:50 多結晶インフォマティクスの現在地

宇佐美 徳隆 (名古屋大学)

13:50～14:20 地球科学におけるプロセスインフォマティクス

桑谷 立 (JAMSTEC)

16:10～17:50 総合討論

ファシリテーター

伊藤 聡 (JST-CRDS)

論点

1. プロセスインフォマティクスの現状 何ができて何が難しいのか
2. プロセスインフォマティクスの将来像 どこまでのことが出来るか
3. プロセスインフォマティクス研究開発の課題 技術的課題／制度体制的課題

17:50～18:00 閉会挨拶

曾根 純一 (JST-CRDS)

(敬称略)

開催日時：2020年7月29日（水）13:00～17:30

開催会場：オンライン会議（Zoom）にて開催

(JST 関係者+希望者) TKP 市ヶ谷カンファレンスセンター バンケットホール 8B
 オーガナイザー 曽根 純一 (JST-CBDS)

13:00 ~ 13:05 開会挨拶 曾根 純一 (JST-CRDS)

13:05 ~ 13:15 ワークショップの趣旨説明 福井 弘行 (JST-CRDS)

13:15 ~ 14:45 (話題提供 1) (各発表 20 分 + 議論 10 分)

13:15 ~ 13:45 触媒インフォマティクス:材料組成とプロセス条件を考慮した触媒設計
高橋 啓介 (北海道大学)

13:45～14:15 ハイスループット実験を基盤としたデータ駆動型材料科学研究
谷池 俊明（北陸先端大）

14:15 ~ 14:45 反応経路のデータベースと機械学習を用いる触媒・発光材料の
理解・分子設計 畑中 美穂（慶應大）

15:00 ~ 16:00 (話題提供 2) (各発表 20 分 + 議論 10 分)

15:00 ~ 15:30 マイクロフロー合成とプロセスインフォとの融合と展開
永木 愛一郎 (京都大学)

15:30 ~ 16:00 NEDO 超超プロジェクト概要と高分子 MI に対する取り組み
青柳 岳司 (産総研)

16:15 ～ 17:25 総合討論

ファシリテーター	伊藤 聡 (JST-CRDS)
コメンテーター	瀧川 一学 (理研)
	富谷 茂隆 (ソニー)
	池端 久貴 (旭化成)

論点

1. プロセスインフォマティクスの現状 何ができて何が難しいのか
2. プロセスインフォマティクスの将来像 どこまでのことが出来るか
3. プロセスインフォマティクス研究開発の課題 技術的課題／制度体制的課題

17:25 ~ 17:30 閉会挨拶 曾根 純一 (JST-CRDS)

付録 2. 参加者一覧

(敬称略、所属・役職はワークショップ開催日時点のもの)

第 1 回： 2020 年 2 月 11 日

招聘者

(発表者)

- ・宇治原 徹 名古屋大学未来材料・システム研究所 教授
- ・宇佐美 徳隆 名古屋大学大学院工学研究科物質プロセス工学専攻 教授
- ・柿本 浩一 九州大学応用力学研究所 教授
- ・知京 豊裕 物質・材料研究機構 情報統合型物質・材料研究拠点 副拠点長
- ・桑谷 立 海洋研究開発機構 火山・地球内部研究センター 研究員

JST-CRDS ナノテクノロジー・材料ユニット

- ・曾根 純一 上席フェロー
- ・宮下 哲 フェロー・ユニットリーダー
- ・荒岡 礼 フェロー
- ・永野 智己 総括ユニットリーダー・JST 研究監
- ・馬場 寿夫 フェロー
- ・八巻 徹也 フェロー
- ・眞子 隆志 フェロー
- ・福井 弘行 特任フェロー、旭化成 研究・開発本部
- ・伊藤 聡 特任フェロー、物質・材料研究機構 MI2I 拠点長
- ・佐藤 勝昭 特任フェロー、東京農工大学 名誉教授
- ・岩本 敏 特任フェロー、東京大学生産技術研究所 教授
- ・本間 格 特任フェロー、東北大学多元物質科学研究所 教授

関係府省・機関

- ・高橋 功 文部科学省研究振興局参事官（ナノテクノロジー・物質・材料担当）付参事官補佐
- ・有沢 俊一 物質・材料研究機構 経営戦略室 運営主席

JST

- ・小関 徳昭 未来創造研究開発推進部 主任調査員
- ・落合 幸徳 イノベーション拠点推進部 推進 PO

第2回： 2020年7月29日

招聘者

(発表者)

- ・青柳 岳司 産業技術総合研究所 機能材料コンピュータショナルデザイン研究センター 総括研究主幹
- ・高橋 啓介 北海道大学 理学研究院化学部門 准教授
- ・谷池 俊明 北陸先端科学技術大学院大学 マテリアルサイエンス研究科 准教授
- ・永木 愛一郎 京都大学 大学院工学研究科合成・生物化学専攻 准教授
- ・畑中 美穂 慶應義塾大学 理工学部化学科 准教授

(コメンテーター)

- ・池端 久貴 旭化成株式会社 研究・開発本部 インフォマティクス推進センター リードエキスパート
- ・瀧川 一学 理化学研究所 革新知能統合研究センター 研究員／北海道大学 化学反応創成研究拠点 主任研究者
- ・富谷 茂隆 ソニー株式会社 R&D センター 主幹研究員

JST-CRDS ナノテクノロジー・材料ユニット

- ・曾根 純一 上席フェロー
- ・宮下 哲 フェロー・ユニットリーダー
- ・赤木 浩 フェロー
- ・荒岡 礼 フェロー
- ・大山 みづほ フェロー
- ・永野 智己 総括ユニットリーダー・JST 研究監
- ・馬場 寿夫 フェロー
- ・福井 弘行 フェロー
- ・眞子 隆志 フェロー
- ・渡邊 孝信 フェロー
- ・伊藤 聡 特任フェロー／計算科学振興財団 チーフコーディネーター
- ・川合 知二 特任フェロー／大阪大学産業科学研究所 招へい教授
- ・清水 敏美 特任フェロー／産業技術総合研究所 名誉リサーチャー
- ・本間 格 特任フェロー／東北大学多元物質科学研究所 教授
- ・八巻 徹也 特任フェロー／量子科学技術研究開発機構高崎量子応用研究所
上席研究員

関係府省・機関

- ・木全 修一 内閣府 政策統括官（科学技術・イノベーション担当）付 上席政策調査員
- ・片山 郁文 文部科学省 研究振興局参事官（ナノテクノロジー・物質・材料担当）付 学術調査官
- ・野瀬 雅文 文部科学省 研究振興局参事官（ナノテクノロジー・物質・材料担当）付 文部科学調査員
- ・萩原 開 文部科学省 科学技術・学術政策局 企画評価課 新興融合研究開発調査戦略室 行政調査員
- ・米田 諭 文部科学省 科学技術・学術政策局 企画評価課 新興融合研究開発調査戦略室 調査員
- ・濱本 孝一 経済産業省 産業技術環境局 研究開発課 研究開発専門職
- ・関口 貴子 経済産業省 産業技術環境局 研究開発課 研究開発専門職
- ・山下 良之 物質・材料研究機構 機能性材料研究拠点 主幹研究員
- ・山下 勝 新エネルギー・産業技術総合開発機構 技術戦略研究センター 主任研究員
- ・寒川 泰紀 新エネルギー・産業技術総合開発機構 技術戦略研究センター 研究員
- ・福田 浩章 新エネルギー・産業技術総合開発機構 技術戦略研究センター 研究員
- ・関野 雅史 新エネルギー・産業技術総合開発機構 材料・ナノテクノロジー部 主査
- ・永井 良憲 新エネルギー・産業技術総合開発機構 材料・ナノテクノロジー部 主査
- ・吉岡 宏人 新エネルギー・産業技術総合開発機構 材料・ナノテクノロジー部 主査

JST

- ・的場 正憲 CRDS システム・情報科学技術ユニット フェロー
- ・竹内 良昭 CRDS 環境・エネルギーユニット フェロー
- ・庄司 真理子 未来創造研究開発推進部 調査役
- ・小関 徳昭 未来創造研究開発推進部 主任調査員
- ・三田 雅昭 未来創造研究開発推進部 主任調査員
- ・長谷川 一途 未来創造研究開発推進部 係員
- ・嶋林 ゆう子 戦略研究推進部 調査役
- ・酒部 健一 戦略研究推進部 主任調査員
- ・釣田 寧 戦略研究推進部 主任調査員
- ・山下 勝久 戦略研究推進部 主任調査員
- ・加藤 秀之 研究プロジェクト推進部 主任調査員
- ・水本 邦彦 研究プロジェクト推進部 主任調査員
- ・松本 秀一 イノベーション拠点推進部 主任調査員
- ・甲村 長利 PM 育成・活躍推進プログラム 研修生

■ワークショップ企画・報告書編纂メンバー■

曽根 純一	上席フェロー
宮下 哲	フェロー／ユニットリーダー
赤木 浩	フェロー
荒岡 礼	フェロー
大山みづほ	フェロー
永野 智己	総括ユニットリーダー・JST 研究監
馬場 寿夫	フェロー
福井 弘行	フェロー
眞子 隆志	フェロー
渡邊 孝信	フェロー
伊藤 聡	特任フェロー
岩本 敏	特任フェロー
川合 知二	特任フェロー
佐藤 勝昭	特任フェロー
清水 敏美	特任フェロー
玉野井冬彦	特任フェロー
本間 格	特任フェロー
村井 眞二	特任フェロー
八巻 徹也	特任フェロー

※所属は令和2年7月時点

CRDS-FY2020-WR-06

俯瞰ワークショップ報告書

ナノテクノロジー・材料分野 区分別分科会 「プロセスインフォマティクスの俯瞰 ～材料合成プロセスへのデータ科学適用の現状と展望～」

令和3年2月 February 2021

ISBN 978-4-88890-700-2

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

〒102-0076 東京都千代田区五番町7 K's 五番町
電 話 03-5214-7481
E-mail crds@jst.go.jp
https://www.jst.go.jp/crds/
©2021 JST/CRDS

許可無く複写／複製をすることを禁じます。

引用を行う際は、必ず出典を記述願います。

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

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

