



入門講座

インフォマティクス入門ー8

第一原理計算データを活用した材料探索

Materials Exploration Using First Principles Datasets

田中 功

Isao Tanaka

京都大学
大学院工学研究科
教授

1 第一原理計算の材料研究への適用

材料研究に第一原理計算が広く利用されるようになった。第一原理計算とは、量子力学の基本原理解から出発し、実験により既知である物理量や経験的法則を使わない計算のことであり、獲得される情報は、電子状態だけでなく、結晶構造、物性値、エネルギーなどの熱力学的性質など極めて多岐にわたっている。21世紀に入った頃から計算機と計算技術が進歩し、高精度の第一原理計算を現実的な時間で多数実行できるようになった。それにより格子振動状態を第一原理計算し、これまで実験によって評価されていた熱力学量、たとえば自由エネルギーの温度依存性を求めることが可能になった。また実験では評価が難しかった材料系や、粒界・表面などの機能コア、拡散や化学反応などのダイナミクス等についても理解が進んだ。さらに特筆すべき進歩は、このような第一原理計算結果を収録するデータベースが、誰もがアクセスできる形で公開されたことである。たとえば米国の Materials Project¹⁾には、既知の無機結晶を中心として現在13万件あまりの化合物についての計算結果がデータベース化されており、誰もが手軽に利用できる。また格子振動の第一原理計算をもとに各種の熱的性質の温度依存性を算出した結果は、東後が構築した phonondb²⁾ に収録されており、自由にアクセスできる。

図1に示すのは Materials Project 収録の遷移金属炭化物についての計算結果を筆者が抽出したものである。Cr-C 二元系について、Cr 金属と黒鉛を基準にした絶対零度での生成エネルギーを、組成別に、計算の結果エネルギーが最低であった結晶構造についてプロットしたものが図1 (a) である。この図で凸包上に存在する化合物とその構造が、絶対零度での安定構造に対応することになる。それは Cr-C 二元系では3つの炭化物になっており、ASM Alloy Phase Diagram Database³⁾ に収録の化合物と過不足なく一致している。図1 (b) から (d) は、

この生成エネルギーの大きさを組成に応じて色変化で示したものである。遷移金属の左の元素のほうが、生成エネルギーの絶対値が大きく安定性が高いことや、炭素組成が50at%を超える化合物が存在しないことが一目で理解できる。その傾向は遷移金属元素の周期によらず同じである。これらは常識と思われるかもしれないが、そのような知識は、第一原理計算が出現する前までは、丁寧な熱測定や状態図作成の実験を通して獲得されたものであることを思いだして頂きたい。実験結果に先んじて常識は無かったのである。

2 第一原理計算に基づくハイスループットスクリーニング (HTS)

第一原理計算と一括りにしても、対象によって計算コストは大きく異なる。10原子程度の単位胞の結晶の電子状態や結晶構造、エネルギー、体積、弾性定数などは、手元のPCでも比較的短時間に計算できる。一方で、周期性が損なわれている欠陥、とくに転位や界面・粒界について妥当と考えられるモデルを作製して計算するには、大規模計算機クラスターやスパコンが必要になる。これらを易計算物性と難計算物性と分けて呼ぶことにする。計算機の能力と時間の積としての計算資源を固定した場合、難計算物性を1つ計算するあいだに、易計算物性は何桁も多い種類の計算が可能である。先に述べた計算データベースとして公開されているデータの多くは、易計算物性についてである。

易計算物性について最適な物質をスクリーニングすることは、探索する範囲(探索空間)を網羅する結果を予めデータベースとして持っていれば単純工程である。スクリーニングの対象が易計算物性でなくても、何らかの経験知で、対象となる特性と強い相関を持つ易計算物性が存在する場合も同様である。このような探索方法は、第一原理計算に基づくハイスループットスクリーニング (HTS) と呼ばれている。た

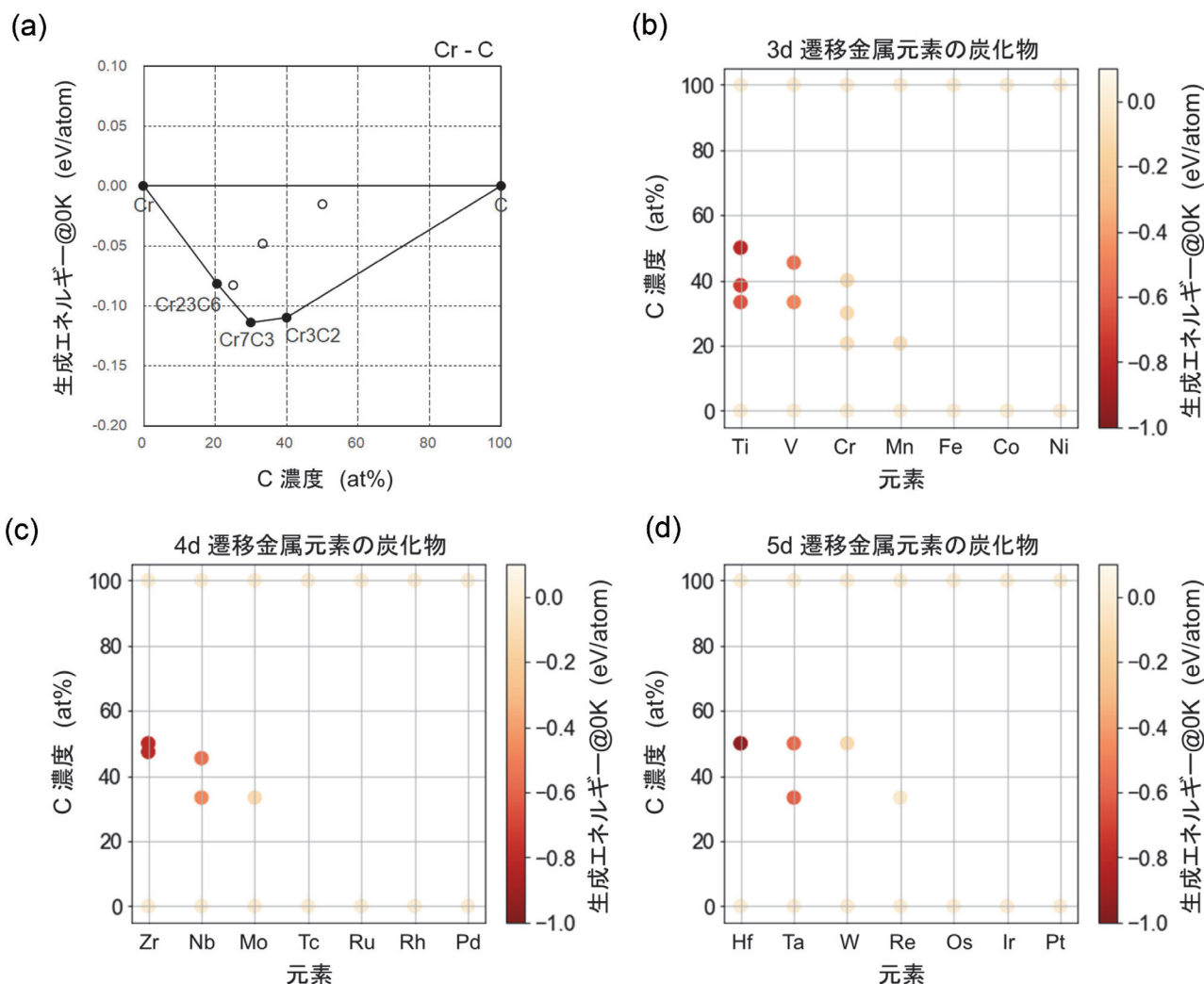


図1 (a) Crの炭化物の生成エネルギーの第一原理計算結果と二元系での凸包。(b) (c) (d) 周期表の第4から第6周期の遷移金属元素と炭素の二元系での生成エネルギーの凸包上にある炭化物。いずれも単体金属と黒鉛を基準とし、温度効果を取り入れない絶対零度での生成エネルギーである。なお、たとえばセメンタイトFe₃Cは、BCC-Feと黒鉛を基準とした凸包よりも生成エネルギーが0.055eV/atom高い準安定物質であり、図には示されていない

例えば、低変態ヒステリシスの形状記憶合金を探索する目的で、Leeら⁴⁾は48元素の範囲で、二元系XY (B2), X₃Y (D0₃)と三元系X₂YZ (L2₁)化合物の第一原理計算を実施し、相転移前後でのエネルギー変化や構造適合性、そして両相の力学的安定性、生成エネルギーなどの条件を与えてHTSを実施した。機能性物質では、LiFePO₄/FePO₄の体積変化に対する置換固溶元素の効果について網羅的な第一原理計算を実施し、この体積変化がリチウム電池のサイクル寿命に関係すると言う経験知(仮説)のもとでHTSを実施し、実際に長サイクル寿命のリチウム電池の創出に成功した西島らの例⁵⁾や、第一原理計算に基づき、生成エネルギー、動的安定性、バンドギャップや有効質量等を用いたHTSを実施し、新しいCa-Zn系窒化物半導体の合成に成功した日沼らの例⁶⁾などが挙げられる。

一方で、望みの特性に対応する物理量が不明な場合や、難

計算物性である場合には、上記のような直接的なHTSは困難である。しかし探索空間のなかで少数だけでもデータがあれば、機械学習法を適用することでHTSを実施できる場合がある。これを仮想スクリーニングと呼ぶ。図2には、直接スクリーニングと仮想スクリーニングを比較して示す。前者では、探索範囲の全物質の目的特性を物理法則あるいは経験知を通じて予め獲得したうえでHTSするのに対し、仮想スクリーニングでは、探索範囲の一部の目的特性を機械学習したモデルを使ってHTSするため、その結果の妥当性に注意を払う必要がある。図2 (b) に示すように、仮想スクリーニングの結果を検証し、結果が妥当でない場合には、検証結果を初期データに加えてモデルを更新するベイズ最適化過程と組み合わせることで、予測能を逐次的に向上させることが可能である。仮想スクリーニングは、目的特性を決める物理法則や経験知がなく、広い探索範囲を対象にHTSする場合に有効である。次

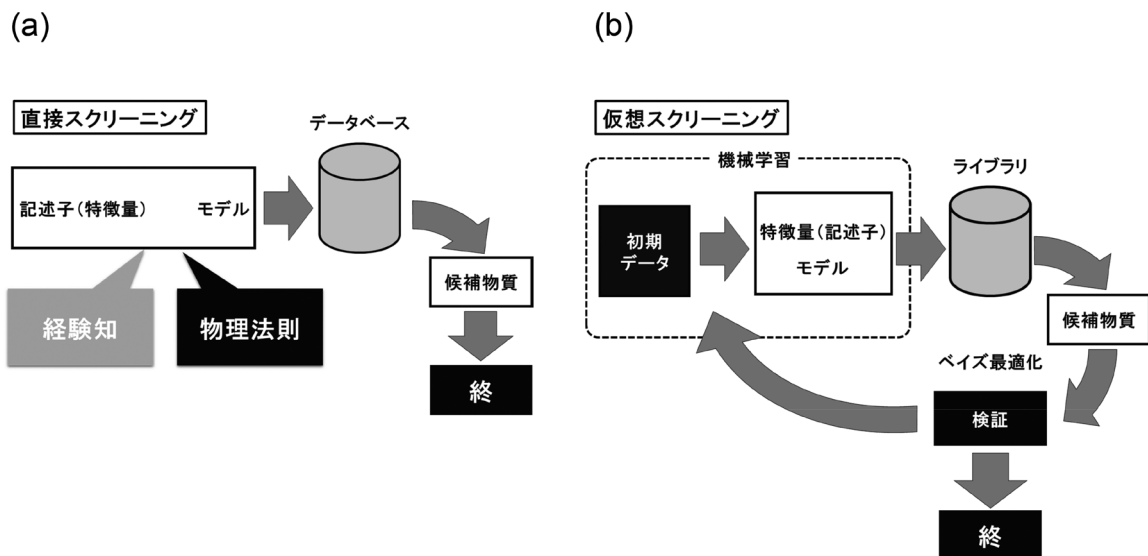


図2 (a) 直接スクリーニングと、(b) 仮想スクリーニングの流れ

に低格子熱伝導物質の探索に適用した例を紹介する。

3 低格子熱伝導度物質の仮想スクリーニング

熱伝導度は材料の基本的な特性値の一つである。電子機器での熱制御や、熱電材料、熱遮蔽材料など、工学的に重要なテーマも数多い。物質の熱伝導は、伝導電子と格子振動（フォノン）による寄与に分けることができる。絶縁体では前者の寄与が無視でき、フォノンによる格子熱伝導が支配的となる。熱伝導度の測定は簡単に思うかも知れないが、格子欠陥や粒界、不純物など外的要因を取り除いた格子熱伝導度が実験的に既知である無機結晶の種類は100に満たない。東後らは、第一原理計算により格子熱伝導度を求めるプログラム phonopy および phono3py を作成・公開した⁷⁾。図3 (a) に、それを用いて計算した300Kでの格子熱伝導度の計算結果と文献に掲載されている実験結果とを比較して示す。第一原理計算により、実験と比肩できる精度で格子熱伝導度が求められることがわかる。格子熱伝導度の物理モデルとしては、原子の質量数の平均値 $\bar{M}$ 、原子当たり結晶体積 V 、デバイ温度 θ_D 、グリュナイゼン定数 γ を用いて $\frac{\bar{M}\theta_D V^{1/3}}{\gamma^2}$ に比例するという Slack が定式化したもの⁸⁾ が知られている。第一原理フォノン計算により θ_D と γ を求めて Slack モデルに当てはめ、それを格子熱伝導度の計算結果と比べた結果を図3 (b) に示す。多くの化合物で、格子熱伝導度が Slack モデルで説明できているが、例外的に低い格子熱伝導度を示す化合物が幾つか存在することもわかる。これらの化合物は、Slack モデルが前提としている調和振動子モデルが近似として悪いためである。

世古らは、このように Slack モデルで説明できないような低い格子熱伝導度を示す化合物を探索することを目的とした研究を行った⁹⁾。探索範囲を、当時の Materials Project¹⁾ に収録されていた既知の無機化合物のデータベース ICSD¹⁰⁾ 掲載の全化合物54779件とし、101件の化合物についての格子熱伝導度の計算結果を機械学習することで仮想スクリーニングを実施した。対象とする54779件すべての格子熱伝導度を第一原理計算で評価できれば、単純に直接スクリーニングが可能である。しかし格子熱伝導度の第一原理計算は、通常の電子状態を評価するだけの第一原理計算に比べて約10万倍の計算時間を要する難計算物性であり、現在の計算機性能でもすべての化合物について計算することは現実的ではない。全データが無くても、しっかりした物理モデルがあれば直接スクリーニング可能であるが、いま探したいのは、図3 (b) で例外となるような化合物である。これらの例外に当てはまる物理モデルは不明なのである。

機械学習を行うためには、予測モデルと特徴量が必要である。本研究では、物理的考察よりも、簡便に求められることを優先した単純な特徴量を用いることにした。化合物の特徴量としては、結晶構造についての情報（構造情報）と構成元素についての情報（元素情報）が考えられる。ここでは構造情報として1原子あたりの結晶体積と結晶の密度という2つの特徴量だけを用いた。また元素情報として、101種類の学習データに含まれる化合物の34種の構成元素について、存在すれば1、存在しなければ0というワンホット表現を採用した。たとえば化合物 BN については、元素特徴量 B と N が1となり、それ以外の元素特徴量が0となる。これら合計36種類の特徴量から成る特徴量ベクトルを用いて、学習データをガウス過程回帰した。ガ

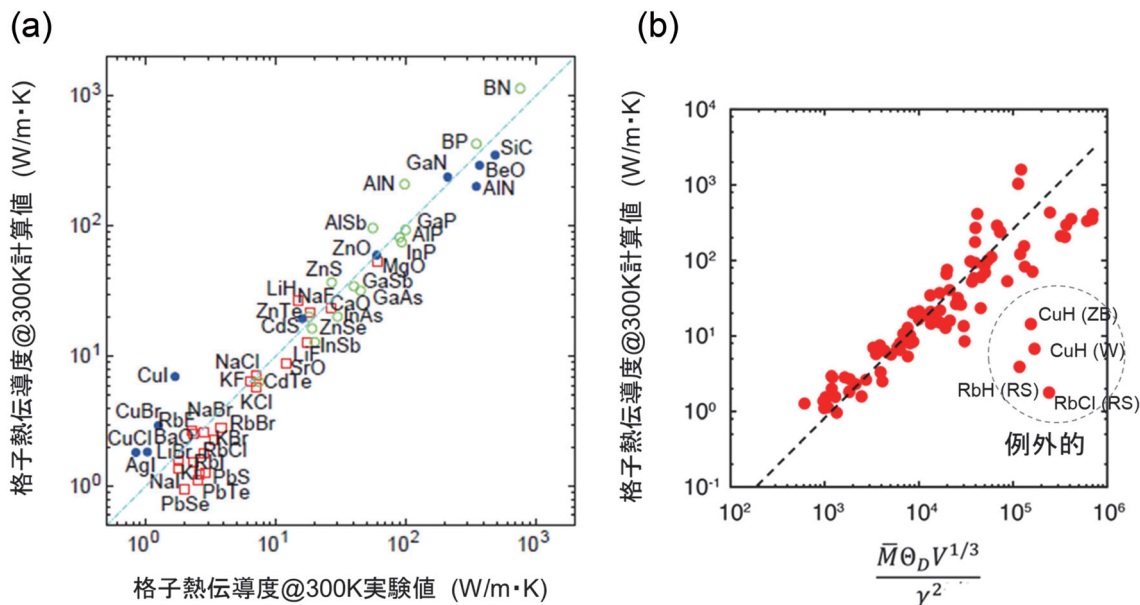


図3 (a) 300Kでの格子熱伝導度の第一原理計算結果と実験値の比較⁹⁾。(b) 300Kでの格子熱伝導度の第一原理計算結果とSlackモデルとの比較

ウス過程回帰の概要を図4 (a) に示す。このモデルを使うと、与えられた特徴量ベクトル $\mathbf{x}^*$ に対し、その予測値 $f(\mathbf{x}^*)$ とともに、ガウス分布 (正規分布) に基づく予測誤差 $\sigma(\mathbf{x}^*)$ が与えられる。これらを用いて $z(\mathbf{x}^*) = \frac{f(\mathbf{x}^*) - f_0}{\sigma(\mathbf{x}^*)}$ という z スコアを定義する。ここで f_0 は基準値で任意に与えてよい。この z スコアが高いほど、目的変数が f_0 を上回る確率が高いことになる。図4 (b) は、目的変数として格子熱伝導度 κ_L の対数に負符号をつけたもの $f = -\log \kappa_L$ をとり、既知の無機化合物のデータベースに収録されている全化合物に対し、 z スコアでの順位が上位の化合物を示す。 f_0 の基準としては、学習データ内で最小の κ_L である PbSe の値とした。これらの化合物について z スコアが正の値であることは、 κ_L が基準値である PbSe を下回る確率が高いことを意味する。これら7種の化合物について、実際に第一原理計算により κ_L を算出すると、高順位のもの κ_L が $0.1 \text{ W/m} \cdot \text{K}$ 程度と極めて低いことが検証された。

上記の研究では、単純で容易に求められる特徴量を用いて予測モデルを作った。構造情報と元素情報ともに特徴量の数を増やしてモデルの表現力を向上させると、予測誤差を対数値で半分まで低減できることが分かっている¹¹⁾。しかし、検証を行うことを前提とした仮想スクリーニングの場合、粗いモデルでも目的が達成できていることを強調したい。

4 五元系までの安定化合物組成 (CRC) の推薦システム

推薦システムとは、Amazon や Netflix などの電子商取引において、蓄積されている膨大な購買データをもとに、各々の

客の嗜好を予測し、推薦する商品を決めているシステムである。客AとBの嗜好が類似していると判断されると、同様のものを購入する可能性が高いと判断し、購買を推薦する。一般に電子商店は、図5左のような購入履歴リストを持っている。この履歴に基づき、顧客と購入商品の類似度を計算し、購買傾向が類似した顧客に未購入の商品を推薦するのである。図5に示すように、購入履歴リストを行列として取り扱い、それがランクの小さい行列に分解できるという近似のもとで行と列を入れ替えて行列分解し、図中*で示す未購入の商品を推定し、購買を推薦している。

世古ら¹²⁾ は、この推薦システムの方法論を、既知化合物のデータベースをもとに、未知の安定化合物組成 (chemically relevant composition : CRC) を発見することに応用した。CRCとは安定化合物が出現する化学組成のことを指し、図6に示すように、その一部が既知化合物として ICSD¹⁰⁾ や ICDD¹³⁾ などのデータベースに収録されている。第一原理計算では、CRCは擬二元での終端物質を基準とした生成エネルギーの凸包上に化合物が存在する組成に対応する。

世古らの研究では、66種の陽イオンと10種の陰イオンから構成される五元系までの化合物組成を対象に、CRCの推薦システムを構築した。図7に示すように、五元系組成 $A_aB_bC_cD_dX_x$ で整数 a から x が20以下となるものの総数は231億とおりあるが、ICSDなどに収録されている既知化合物は、僅か1000通り程度と7桁の違いがある。それでも、ICSD収録化合物組成を学習セットとし推薦システムによりCRCを予測した。正答率を評価するために、他の2つのデータベース ICDD¹³⁾ と Springer Materials¹⁴⁾ に収録されている化合物

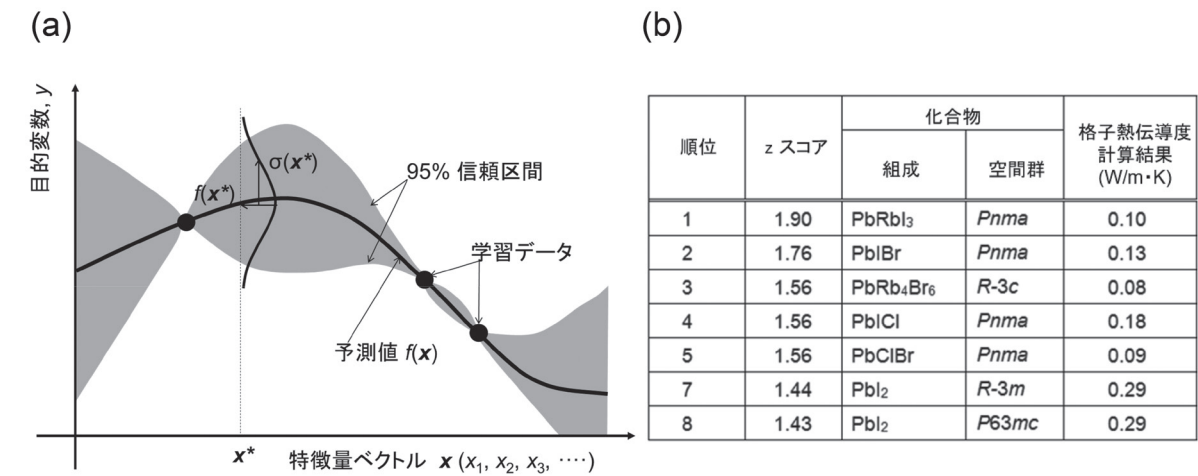


図4 (a) ガウス過程回帰の模式図。(b) 300Kでの格子熱伝導度の予測値についてのzスコア上位の化合物と、検証のために行った300Kでの格子熱伝導度の第一原理計算結果⁹⁾。なお順位6位の化合物は単位胞に含まれる原子数が多すぎて格子熱伝導の検証計算ができなかった

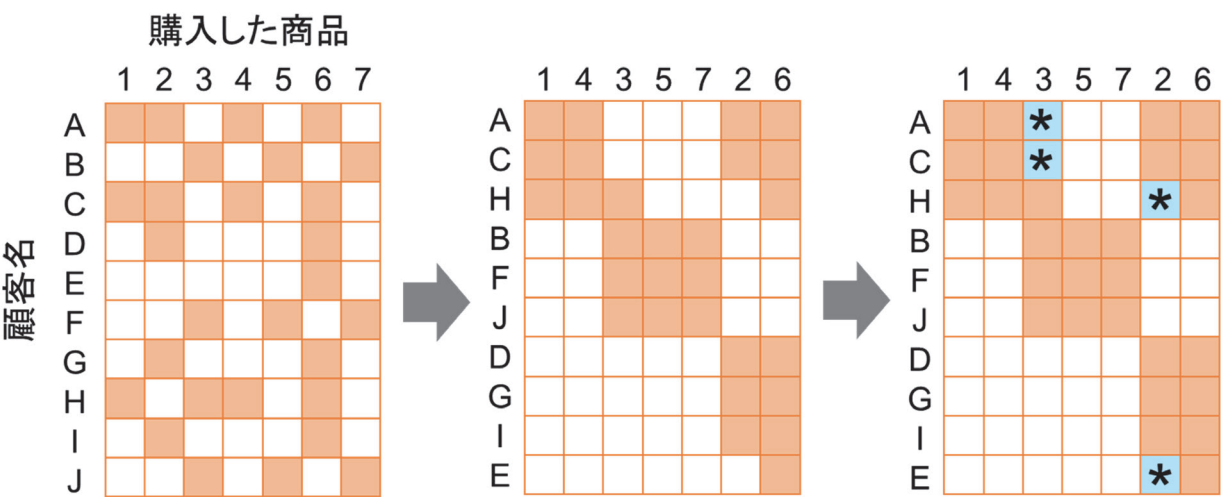


図5 行列分解を利用した推薦システムの模式図

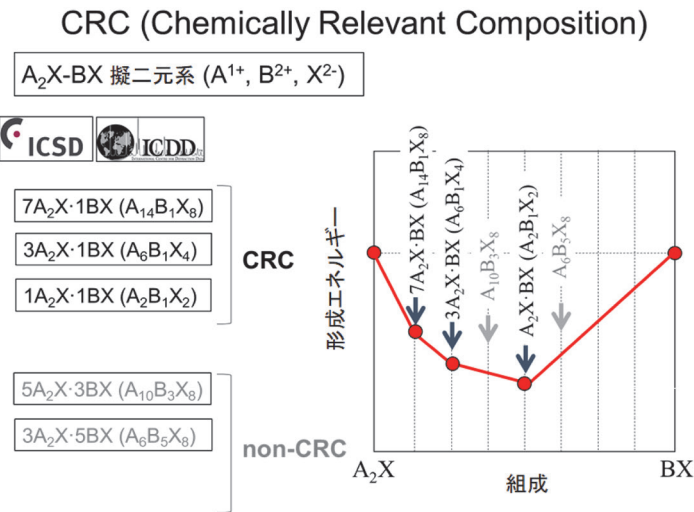


図6 A₂B-BX擬二元系における安定化合物組成 (CRC)

組成のうち、ICSDに収録されていないものをテストセットとして用いた。化合物組成はテンソル表現とし、Tucker分解によりCRCとなる可能性を評価した。

CRCとなる確率が高いと予測されたTop100組成について、正答率、つまりテストセットに含まれていた割合は、三元、四元、五元系でそれぞれ、59%、52%、15%となった。この推薦システムでは、化学組成をテンソル表現して分解しただけであり、何ら事前知識や記述子（特徴量）を使っていない。それにもかかわらず、高い正答率が得られているのは、極めて興味深い。新材料探索において、多元系への展開が必要になると予想されるが、構成元素数の増加に伴い、候補数が大きく増える。本稿で述べたようなフィルタリング手法が活躍する場面が増えると予想される。

5 並列合成実験データに基づく合成条件推薦システム

第一原理計算や推薦システムに基づいて新規化合物が予測できても、それを実験的に合成し、評価できなければ『絵に描いた餅』である。合成条件についての情報がない場合、従来は試行錯誤に行うか、あるいは熟練研究者の勘と経験に頼っていた。しかし今後はデータ科学の活用範囲が広がるものと想定される。文献データを収集・解析するのが一つの方法であるが、文献データの問題は、目的に対してポジティブな結果であったデータに偏っていて、ネガティブなデータが不十分なことである。これを避けるためには、自動化された並列実験を設計し、新しいデータを系統的に蓄積するのが望ましい。林ら¹⁵⁾は、並列合成実験を実施し、獲得したデータをもとに推薦システムを活用して無機化合物の合成実験プロセスを効率的に見出した。

まず合成実験の成否に関するデータセットを獲得するために、4種類の合成方法に対し、原料、焼成温度など様々な合成条件について、系統的に合成実験を行うシステムを開発した。そして合成の成否を、試料自動交換機を備えた粉末X線回折実験装置により連続的に評価し、スコア化した。対象としたのは2種類の陽イオンと酸化物イオンから構成される三元系酸化物で、23種類の原料を用い、23種類の原料の配合比に対して、5種類の焼成温度で系統的に合成実験を行い、合計約1600件の合成成否に関するデータを収集した。探索対象となる全実験条件は約24万とおりで、これを合成条件テンソルとした。ここに上記の約1600件の合成成否に関するデータを入力し、テンソル分解することで、未実験条件についての合成実験の成功可能性をスコアとして与える推薦システムを構築した。図8にテンソル分解に伴う合成可能性スコアの変化¹⁵⁾を示す。横軸は焼成温度で、縦軸は2種の原料と配合比および合成手法という4系統の実験条件を併せて示す。テンソル分解前の合成実験データベースでは、橙色が合成成功、青色が合成失敗、白色は未実験条件を示す。テンソル分解後の推薦システムでは、未実験条件に対しても合成可能性が標準化スコアとして予測されている。橙色が濃いほど合成可能性が高いと期待できる。得られた結果の妥当性を検証するために、交差検定を行うとともに、合成実験の成功可能性を示すスコアが高いものから順に、実際に合成実験を行った。その結果、スコアの高い場合には合成が成功し、スコアが低い場合には、合成が失敗するという推薦システムの予想どおりの結果が得られることが見出された。この方法で、新規物質の合成条件が、試行錯誤に探索することに比べて効率化できることが実証された。このような並列実験と解析は、すでに創薬分野で実施されているように、自動実験ロボットを活用して大きく展開することが予想され、すでに萌

データベース	収録化合物数		
	三元系	四元系	五元系
ICSD (学習データ)	9,313	7,742	1,321
ICDD	9,278	7,864	1,326
Springer Materials	10,461	8,141	1,893
テストデータ数	4,134	4,961	1,616
候補組成数	7,405,200	1,188,038,460	23,104,706,560

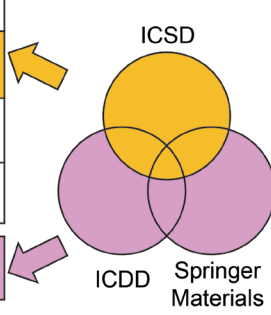
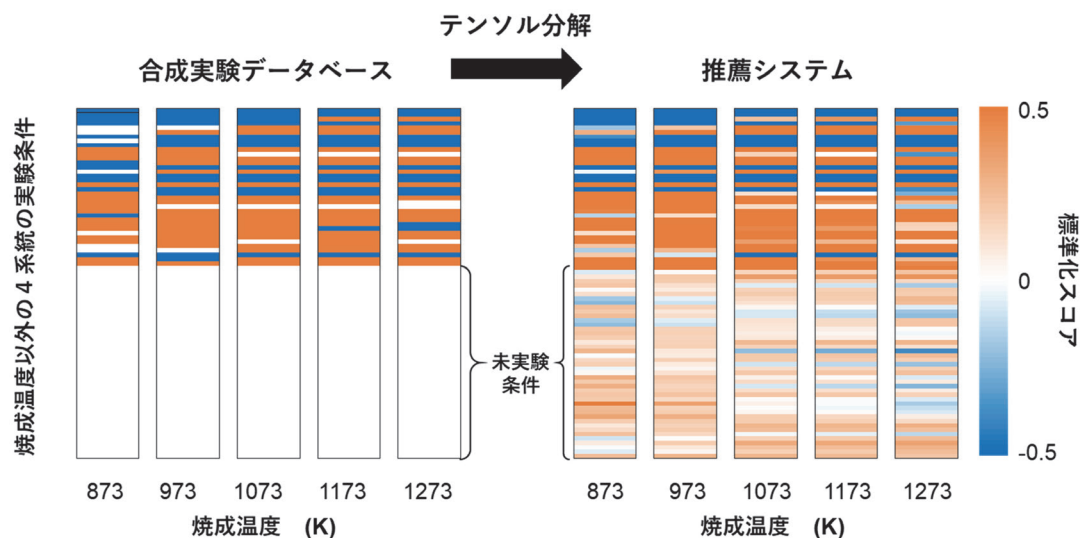


図7 安定化合物組成 (CRC) の推薦システムの学習データとテストデータに利用した化合物数および候補組成数

図8 テンソル分解による合成実験プロセス条件推薦システムの一断面¹⁵⁾

芽的な成果を上げている研究者が居られる。

6 おわりに

1978年の鉄と鋼に徳田昌則先生が「安定成長期における鉄鋼製錬技術と熱力学との係わり」と題した随想記事¹⁶⁾を執筆されている。その中で、昭和40年代の高度成長期には、熱力学などは現場では全く役に立たないと断言する方々がどこの会社にも居られたと書かれている。半世紀も前の話であり、いまの鉄鋼業における熱力学の位置づけを鑑みると隔世の感がある。本稿で述べた第一原理計算やデータ科学についても、直ちに鉄鋼プロセスや材質制御に結び付いているものではなく、鉄鋼業には役に立たないと思われる向きもあるだろう。しかし近年の計算機や情報科学を使った技術革新は著しく、これらを巧く活用する場面が出現するのに、長い年月は要しないと思われる。そのような展開にあたり、拙稿が少しでも参考になれば幸いである。

謝辞

本研究の内容の一部は、筆者が拠点長を務める京都大学構造材料元素戦略研究拠点ESISM(2012から2021年度)での活動の一環として行われた。拠点活動のなかで、多くの産官学の研究者から有用なコメントやアイデアを頂戴した。いちいちのお名前をあげることはできないが、深く御礼申し上げる。

参考文献

- 1) <https://materialsproject.org/>, (accessed 2020-11-27).
- 2) <http://phonondb.mtl.kyoto-u.ac.jp/>, (accessed 2020-11-27).
- 3) <https://matdata.asminternational.org/apd/>, (accessed

2020-11-27).

- 4) J.Lee, Y.Ikeda and I.Tanaka : NPJ Computational Materials, 3 (2017), 52.
- 5) M.Nishijima, T.Ootani, Y.Kamimura, T.Sueki, S.Esaki, S.Murai, K.Fujita, K.Tanaka, K.Ohira, Y.Koyama and I.Tanaka : Nature Communications, 5 (2014), 4553.
- 6) Y.Hinuma, T.Hatakeyama, Y.Kumagai, L.A.Burton, H.Sato, Y.Muraba, S.Iimura, H.Hiramatsu, I.Tanaka, H.Hosono and F.Oba : Nature Communications, 7 (2016), 11962.
- 7) A.Togo, L.Chaput and I.Tanaka : Phys. Rev. B, 91 (2015), 094306.
- 8) G.A.Slack : Solid State Physics, 34 (1979), 1.
- 9) A.Seko, A.Togo, H.Hayashi, K.Tsuda, L.Chaput and I.Tanaka : Phys. Rev. Lett., 115 (2015), 205901.
- 10) <https://icsd.products.fiz-karlsruhe.de/en>, (accessed 2020-11-27).
- 11) A.Seko, H.Hayashi, K.Nakayama, A.Takahashi and I.Tanaka : Phys. Rev. B, 95 (2017), 144110.
- 12) A.Seko, H.Hayashi, H.Kashima and I.Tanaka : Phys. Rev. Mater., 2 (2018), 013805.
- 13) <https://www.icdd.com/>, (accessed 2020-11-27).
- 14) <https://materials.springer.com/>, (accessed 2020-11-27).
- 15) H.Hayashi, K.Hayashi, K.Kouzai, A.Seko and I.Tanaka : Chem. Mater., 31 (2019), 9984.
- 16) 徳田昌則 : 鉄と鋼, 64 (1978), 1638.

(2020年11月27日受付)