要旨

結晶構造解析の初期段階における結晶系と空間群の決定は、材料科学ワークフローのボトルネックとなっており、多くの場合、手作業による調整が必要である。ここでは、シミュレーションパターンを用いた概念実証として、粉末X線回折（XRD）パターンに基づく結晶系と空間群の分類のための機械学習（ML）ベースのアプローチを提案する。我々のツリーアンサンブルベースのMLモデルは、三斜晶の場合を除き、結晶系分類においてほぼ90%以上の精度で動作し、5つの候補を用いた空間群分類においても88%の精度で動作した。また、専門家の間で曖昧に共有されている経験的知識を定量化することにも成功し、解釈可能なMLアプローチを用いることで、実験データに埋め込まれた認識されていない特徴をデータ駆動で発見できる可能性を示した。

はじめに

結晶構造は材料特性を決定するため、結晶構造解析は材料開発における最も重要なタスクの一つである1,2。結晶構造は、格子対称性、格子定数、原子の種類と位置、サイト占有率で定義される。粉末X線回折（XRD）と粉末中性子回折は、結晶構造を解明するための主要な実験手法である。これらの手法を用いて得られたデータは、例えば無機材料やタンパク質など、特定のクラスの材料について様々なデータベースに保存されている3。粉末回折パターンを結晶構造情報にデコードするには、ピークのインデックス付け、空間群の決定、結晶構造の初期パラメータ推定、構造精密化など、いくつかのステップが必要である4,5,6,7,8,9。最も骨の折れるステップはリートベルト法10を用いた構造精密化であり、通常、数十のパラメータを手作業で最適化する必要があるが、構造解析の初期段階における空間群の決定も手作業による試行錯誤が頻繁に必要である。世界中の放射光施設で毎日大量の粉末XRDパターンが生成されていることを考えると、人間の専門家が手作業で行うこれらの時間のかかるプロセスは、材料研究において明らかにボトルネックとなっている11,12,13。このようなプロセスにおける人間の関与を可能な限り排除することは、状況を改善し、ハイスループット（HiTp）実験の実現に役立つ。そこで我々は、経験豊富な研究者が与えられた回折パターンから結晶系を推測できるという事実にヒントを得て、機械学習（ML）アプローチを用いた結晶系と空間群の分類に焦点を当てている。

回折データ解析へのMLや関連技術の応用は、最近のホットな研究トピックである13,14。パターン分解や位相同定15,16,17,18、クラスター解析や位相マッピング19,20,21,22,23、回折データ比較のための類似度メトリクス24,25,26、結晶対称性の分類27,28,29,30,31,32,33など様々なサブトピックの中で、Parkらの論文34が本研究に関連する。Parkらは、模擬粉末XRDパターンに畳み込みニューラルネットワーク（CNN）を適用して、結晶系と空間群を分類した。彼らは、ポアソン・ノイズや装置分解能によるデータの劣化にもかかわらず、高い分類性能を達成した。しかし、CNN（ディープ・ニューラル・ネットワーク）は複雑であるため、その内部プロセスを解釈して意味のある洞察を引き出すことは困難である。多くのマテリアルズ・インフォマティクス（MI）研究は、高い分類精度や予測精度を目指しているが、人間が解釈可能なモデルを使用することで、データ駆動型の知識発見の可能性を維持することも重要であると考えている。MLモデルが回折パターンに基づいて結晶クラスと空間群を分類できるのであれば、それは分類規則を持っていなければならない。モデルを分析することで、経験豊富な研究者が持つ経験則を定量的に特定することができる。

本論文では、シンプルで高速なML手法（図1）により、粉末XRDパターンに基づいて結晶系（7クラス）と空間群（230クラス）を高い精度で分類できること、また解釈可能なMLモデルを用いて経験的な専門知識をデータ駆動で定量化できることを示す。本研究の目的は、従来の手法に取って代わることでも、他の新しいMLベースの手法の中で最先端の精度を達成することでもなく、知識発見や実世界の実験に適したシンプルなML手法の可能性を示すことであることを強調する。本研究は、理想的な模擬回折データ（ノイズなし、ピーク幅拡大なし、ピーク重ね合わせなし、不純物ピークなし）を用いて学習させたMLモデルを用いた概念実証（POC）の段階であるが、以下の節で述べるいくつかの興味深い知見が得られた。

研究成果

データの準備と特徴抽出

199,391個の粉末XRDパターンを、Pymatgenミドルウェアを用いて、Inorganic Crystal Structure Database (ICSD)のエントリーからトレーニングデータセットとして計算した（詳細は "Methods "のセクションを参照）。このパターンをそのまま使用しなかった理由は以下の通りである。典型的な粉末 XRD パターンは数千のデータポイントを持ち、そのまま使用すると ML モデルでは数千次元のベクトルとして扱われる。このような極めて高次元のデータを用いてMLモデルをトレーニングすると、必然的に「次元の呪い」35に悩まされることになる。つまり、トレーニングに必要なデータ量は、データの次元が大きくなるにつれて指数関数的に増大するという事実である。この問題を回避するため、我々は数千のデータポイントを、各XRDパターンを特徴付ける一握りの数値に削減した。これらの数値は通常「特徴量」または「記述子」13 と呼ばれ、本稿では前者の用語を使用する。画像認識に用いられる約1,400万枚のラベル付き画像を含むImageNetデータセットのように、品質管理された大量のデータが利用可能な分野では、膨大なデータセットを用いてディープニューラルネットワークを学習させることにより、分類モデルと同時に特徴抽出器を構築することが主流となっている。しかし、このようなアプローチは、利用可能なデータ量が限られていることと、データの品質管理が困難であるという2つの理由から、MIにおいては不適切であることが多い。我々の場合、低品質なデータは可能な限り排除しているが、データサイズはImageNetよりもおよそ2桁小さい。さらに、自動的に選択された特徴は、必ずしも人間が理解できるような単純な形にはなっていない。そこで、人間の専門的知識を用いて手動で特徴を選択することにした。XRDパターン（図1左）には、結晶構造のフィンガープリントとして機能する多くのピークがあります。ピークの特徴のうち、強度（高さ）はほとんどが原子位置によって決定され、一部の例外を除き、本研究の対象である結晶の対称性には関係しません。そこで、本研究では以下の11の特徴を選択した：

(1) 低角度領域における最初の10ピークの位置。

(2) 0°～90°までの2θのピークの総数。

実際の実験データでは、高角度域のピークは重なりが多く、特に対称性が低い場合、個々のピーク位置を特定することは不可能であるため、低角度域のピークのみを考慮し、残りを無視することは正当である。特徴空間を調べるために、多次元データを2次元平面上に可視化する一般的な手法であるt分布付き確率的隣接埋め込み法（Stochastic Neighbor Embedding with t-distribution: t-SNE）を用い、元の多次元空間の局所構造を保存した。図2は、t-SNEを用いて11次元データを2次元空間に次元削減した結果を示している。XRDパターンは特徴空間内に分布し、結晶系ごとに緩やかなクラスターを形成していることから、11個の特徴が少なくとも結晶系の観点からXRDパターンの特徴を捉えていることがわかる。

ここでは、ピーク位置の単位の曖昧さについて言及する。ほとんどの回折実験では、回折パターンの横軸は2θを表します。

これは入射ビームと散乱ビームの角度である。ブラッグの法則 2dsinθ=λと散乱ベクトル長Q=4πsinθ/λの定義によりここでdとλははそれぞれ散乱面の間隔と入射ビームの波長を表し、2θはd、1/d、1/d2、Qに変換されることがある。

これらのピーク位置の表記が分類精度に与える影響を調べたところ、その差はごくわずかであることがわかった（表1）。したがって、本書では2θを使用する。これは最も一般的に測定される変数であり、したがって多くの読者にとって最も馴染み深いものである。提案した方法は、上記の波長に依存しない式を用いて、任意の波長のXRDパターンに適用できることに留意すべきである。

**結晶系と空間群の分類**

POCとはいえ、高速な学習速度、容易なハイパーパラメータ調整、研究者が実験条件（2θ

の範囲や波長などの実験条件に合わせてモデルを調整し、補足情報の助けを借りて最も妥当な候補を選択することができる。

分類タスクのための様々なMLアルゴリズムの中で、ランダムフォレスト（RF）とその関連アルゴリズムは、上記のすべての要件を満たしている。表1は、結晶系分類タスクにおけるRFと他の代表的なMLアルゴリズムExtremely randomised trees（ExRT）41、k-nearest neighbour（KNN）、ロジスティック回帰、決定木4の性能比較を示す。ExRTはRFベースのMLアルゴリズムであり、通常のRFにおける有意性ベースの選択基準とは対照的に、意思決定に使用される変数がランダムに選択される。ExRTにおけるこのランダム性は、訓練データへのオーバーフィッティングを減らし、未知のデータに対する分類性能を向上させる。さらに、ExRTやRFを含む多数決アルゴリズムの頑健性に注目する。実際の実験では、一部の弱いピークが認識しにくかったり、測定範囲外のピークを見逃してしまうことがある。このような欠陥は、個々の決定木の一部に誤った決定を引き起こす可能性があるが、木のアンサンブルによって取られる多数決は、多くの場合このような摂動に影響されない。比較の結果、ExRTが要件を満たし、最高の性能を示したので、我々はExRTを選択した。モデル学習に要した計算時間は数分であり、1つのXRDパターンに対する分類に要した時間は、一般的なワークステーション（3.3GHz 10コアCPU、96GB RAM）で数ミリ秒であった。

テストデータセットを用いた各水晶システムの分類性能を表2と図3に示す。図3の値は、与えられた結晶系データセット（実際のラベル）に対する、ExRTモデルによって分類された結晶系（予測ラベル）の比率を表す。対角線上の値が精度に相当する。我々のモデルは、三斜晶の場合を除き、我々の予想を上回る約90％の精度で結晶系を分類することに成功しました。

三斜晶の場合の精度が低いのは、ICSDでは三斜晶構造が稀で（わずか4%）分類の妨げになっているためと考えられる。

結晶系とともに、空間群分類の性能も評価した。モデルによって提案された最も可能性の高い候補の分類精度は80.46%でした（図4a）。各空間群の学習データ数が各結晶系の学習データ数よりかなり少ないことを考慮すると、結晶系分類の精度を大幅に下回るこの精度は妥当である。しかし、我々のモデルによって提案された複数の候補のリストを考慮すると、リストに正解が含まれる確率として定義される精度は、最も可能性の高い候補が5個と10個の場合、それぞれ92.42%と94.35%に上昇する。これらの値は、従来の空間群決定法でも複数候補のアプローチが適用されることが多いため、単一候補での精度に比べ、実際の空間群決定タスクに関連する値である。

図4aにおいて、正しい予測を示す対角線以外にも注目すべき特徴があり、それらの特徴には、我々のモデルの欠点を示す2つの大きな原因と、1つの些細な原因があると結論づけられる。まず、些細であまり重要でない方を指摘する。いくつかの孤立した非対角の太い青い画素［例えば（行、列）表記で（50, 12）、（97, 139）、（132, 116）］がある。これらは、少数のテスト構造を持つ空間群の偽予測であり、テストデータ数で正規化することで欺瞞的に強調されている。次に、2つの大きな原因のうちの1つ、つまり機械学習でよくある問題であるオーバーフィッティングについて説明する。オーバーフィッティングの問題は、いくつかの空間群（図4a、bの2、12、14、15、62、139）において、縦に並んだ太い青い画素として見えるが、これはMLモデルが特定の空間群を選んで予測する傾向があることを意味する。この問題は、補足情報のFig.S1に学習データのヒストグラムとして示した、学習に使用したICSDデータの空間群の分布が不均一であること、つまり、我々のモデルが頻繁に学習したものから答えを出すように偏っていることが原因です。2つ目の大きな原因は、ピークの数と位置のみから空間群を決定しようとする我々のアプローチ自体に内在する欠陥です。回折パターンから結晶構造が中心対称かどうかを判断するには、ピーク強度が不可欠です。したがって、ピーク強度を無視した我々のMLモデルは、本質的に回折パターンから中心対称性の有無を決定することができない。

混同行列を注意深く調べると、この限界の痕跡を見つけることができる。例えば、図4bでは、テストデータの非中心対称空間群1、4、6、8、9が、それぞれ中心対称最小同型超群2、11、10、12、15として頻繁に予測されている。つまり、約30％の誤予測（テストデータの6％）が、中心対称と非中心対称のペアで発生している。

我々は、いくつかの誤予測を中心対称/非中心対称の誤分類の問題として特定することは、理論的に明白であり、例外的に簡単であることに注意する。中心対称性の問題とは関係ない他の誤分類結果については、上記のような複数の原因が重なっている可能性があり、また、このような解析の経験が少ないため、結晶学的な観点から原因を分析することは難しいと結論づけた。

空間群と結晶系の間には階層的な関係があるため、予測された結晶系を空間群予測の特徴として用いることで、空間群予測の性能が向上する可能性がある。このように複数のMLモデルを組み合わせる考え方はスタッキングと呼ばれる46。Aguiarら32,33は、スタッキングによって電子回折による空間群決定の予測精度が向上することを報告している。しかし、我々のケースでは、スタッキングによる予測精度の向上は見られなかった。その理由として、結晶系予測における誤認識がノイズとして空間群予測器の学習に悪影響を与えたことが考えられる。

新物質に対する分類性能と実際の実験データ

新素材の結晶構造分類タスクに対する我々のモデルの有効性を評価するため、2種類のテストを実施した。最初のものは、すでに表2に結果を示しているが、ICSDのサブセットから計算された粉末XRDパターンを使用したものである。このサブセットはトレーニングデータセットに含まれていない材料、つまりモデルにとって未知の結晶構造で構成されています（詳細は「方法」のセクションを参照）。このテストは新しい物質に対するモデルの性能を表す。表2から明らかなように、ICSDに三斜晶系の結晶構造データがないため、本質的に困難な三斜晶系を除いて、我々のモデルは十分な性能を示しています。

これまでのところ、統計的ノイズや実験的ノイズによって劣化していないシミュレーションパターンのみを使用している。

2つ目のテストは、より実用的な状況でモデルを評価するためのものである。計算されたピーク位置の代わりに、Ca1.5Ba0.5Si5N6O3とBaAlSi4O3N5: Eu2+を文献34 の補足表から引用した。この2つの物質を選んだのは、その構造が非常にユニークで、Ref.34に記載されているICSDの19,000以上のプロトタイプ構造のどれにも属さないからです。

自動ピーク探索は、著者の説明に従って従来の方法で行われた。つまり、その結果を使用することは、我々の手元にない生のXRDパターンから始めることと実質的に等しい。表中の2θの範囲が限られているため、我々のモデルに適した検出ピークの数を得ることができなかった。このような状況を踏まえ、ピークの総数なしでモデルを再トレーニングし、分類に使用したのはピーク位置が最も低い 10 個のピークのみであった。それにもかかわらず、再トレーニングしたモデルは、両化合物の結晶系だけでなく、BaAlSi4O3N5:Eu2+の空間群も正しく分類した。

. Ref.34のCNNモデルは、両化合物の結晶系予測には成功したが、空間群予測には失敗した。

Crysfire2020のいくつかの標準的なピークインデクシングプログラム（ITO47, FJZN48, TREOR49, KOHL50, DICVOL51, LZON48）でもこれらのベンチマーク物質を分析した。しかし、いずれのプログラムもデフォルトの設定では、両材料に対して適切なインデックス作成結果を与えることができなかった。比較結果を補足資料の表S1に示す。Ref.34では、2つの素材をTREORで分析したが、この論文の著者も、TREORは人間の介入なしには適切な結果を出せないと結論付けている。この事実は、少なくともこれらの特定のケースについては、人間の介入をより少なくすることが望ましいのであれば、我々の方法が既存のソフトウェアよりも結晶系と空間群の推定タスクに適していることを間接的に示している。しかしながら、この比較は我々の手法の優位性を主張するものではない。両者ともピーク位置を入力とするが、出力の種類は異なる。従来のピークインデキシングプログラムの第一の目的は格子定数を与えることであり、我々の方法は結晶系と空間群のみを推定することを目的としている。この違いは、我々のモデルが長寿命の既存のピークインデキシングプログラムの代替にはなり得ないことを意味する。

テストサイズが非常に限られており（2つだけ）、我々のモデルは2つの空間群予測タスクのうち1つで失敗したが、それでも、たった10個の特徴しか持たない我々のモデルが、2つのテスト化合物で失敗した1万個のデータポイントを特徴として用いたCNNモデルよりも、空間群予測で良い結果を与えたことは注目に値する。これは、このタスクでは適切な特徴選択が重要であることを示している。さらに、不純物ピークやS/N比の悪さ、ピークの重なりなど、ピークの数を正しくカウントするのに適さないXRDパターンもあるため、ピークの総数を使わずに学習したExRTモデルの成功は、我々のアプローチのターゲット範囲を広げる可能性がある。

機械学習モデルからの知識抽出

予測性能に関しては、表1に示すようにKNNとExRTの間に有意差はなかった。しかし、ExRTや決定木のような木ベースのモデルは、KNNよりも多くの情報を提供してくれる。つまり、我々のMLモデルは、注意深い分析を通じて、関連するタスクに関する洞察を付加的な利点として提供するかもしれない。このような知識は、特定の目的のために、最小だが効果的な測定セットアップと実験計画を設計するのに役立つかもしれない。

(1)特徴量の重要度の評価、(2)特徴量（ピーク位置）の数に対する分類性能の依存性、(3)結晶系分類の決定ルールの定量化と可視化。

表 3 に結晶系分類タスクのモデルで用いた 11 個の特徴量の特徴重要度を示す。特徴量の重要度は、分割品質の向上に対する特徴量の寄与を定量化したものである。本研究では、ジニ不純物 を分割基準として用いる。

ピークの数と低い角度のピーク位置の重要性が高いことは、以下のように解釈できる。ICSDに保存されている結晶構造の格子定数は、ある範囲内で利用可能なものが多いため、2θ

ウィンドウ内のピークの数を決定する主な要因は、結晶構造の対称性である。同様に、同じ実験条件下では、10個の最低角ピークの位置は、主に構造の対称性によって決まる消光則に依存する。例えば、対称性が7つの結晶系の中で最も高い単純立方構造は、最低2θ

の範囲にある。構造が立方晶から斜方晶に変わると、元の立方晶構造の001反射と011反射がともに3つの反射に分かれ、場合によっては111反射の低2θ側に高次の反射まで現れる。

側に現れる。その結果、10個の低角ピークのほとんどがこのような分割ピークに支配され、その位置は立方晶の場合よりも低い2θの範囲にシフトする。同様の議論は、構造相転移を伴わない2つの結晶系の比較に関する一般的な場合にも有効である。したがって、多くのピークが低い2θ領域に多くのピークが現れる場合、対称性が低いことを意味し、逆もまた同様である。これらの推論は、専門家の第一印象と一致する。注目すべきは、これらの推論はMLモデルから定量的に抽出できることである。

次に、モデル学習で使用したピーク位置の数に対する分類精度の依存性をプロットすることで、実際の分類タスクにおける低いピークの重要性を評価する（図5）。ここでは、ピーク位置の影響を抽出するために、特徴量から回折ピークの総数を省略している。ほとんどの結晶系で、最も低いピーク位置が5つか6つで精度は飽和し、トレーニングでより高いピーク位置を追加しても、わずかな改善しか得られない。ここで、より複雑な結晶系ほど同定に多くの情報を必要とするという定性的な傾向が明らかになった。

はじめに」で述べたように、経験豊富な研究者であれば、結晶構造の対称性が高い場合、追加情報なしに与えられたXRDパターンから結晶系を推定することができる。一般に、経験によって得られるこのような専門的知識は曖昧であり、定量化することは困難である。ここでは、簡略化した結晶系分類タスク、すなわち立方晶系と非立方晶系の二値分類について、専門家の知識を定量化する試みを紹介する。

ExRTによる分類結果は、最大深度をハイパーパラメータとする数百の決定木の多数決に基づいているため、我々のExRTモデルから合理的な方法で分類ルールを抽出することは困難である。その代わりに、2値分類タスク（立方晶系か非立方晶系か）と結晶系分類タスク（7クラス）に対して、1つの決定木を異なるランダムシードで繰り返し訓練する。二値分類の代表的な決定木2本とその分割規則を図6に示す。

この2つの木の分割規則は、合計2つの特徴のみから構成されており、深さ1～2のこのような単純な木構造にもかかわらず、精度は驚くほど高い。一般に対称性の高い結晶構造は回折パターンのブラッグピークが少ないため、両木で採用されているピークの数は立方晶と非立方晶を区別するための妥当な基準である。逆に、もう1つの基準であるピーク3の位置が、2値分類タスクにどのように寄与するかを解釈するのは、ピークの数の場合ほど簡単ではありません。ヒントは図6bとdにあり、散布図で分割規則が視覚化されている。つまり、3番目の低角度ピーク（ピーク3）の位置を表す横軸に沿って、3次系のデータ点の分布が広がっている。既に述べたように、特徴量の重要度は、結晶系分類タスクにおいて、より低角度のピーク位置がより重要であることを示している。立方晶系と非立方晶系でのピーク3の分布の違いは、二値分類タスクに有効な閾値が存在することを意味し、これは図6aの決定木がとる戦略そのものである。この議論は、「立方晶か立方晶でないか」の分類という特定のタスクにおいて、より低いピークの重要性を示すケースであり、またMLモデルを用いた定量的な知識抽出の例とみなすことができる。

結晶系の分類においても同様の傾向を確認した。決定木は補足資料の図S3に可視化されている。

考察

本研究のハイライトは、訓練されたMLモデルの分析によって可能になる定量的な知識抽出である。材料科学における多くの実験は、依然として従来の経験則に基づくパラメータに依存しているため、このアイデアは実験セットアップの最適化に関連すると期待される。例えば、スキャン範囲、データポイント間隔、測定時間などの実験パラメータは、MLモデルがこれらのパラメータのガイドラインを示せば、特定の目的に合わせて客観的に調整することができる。効率的な実験デザインは、利用可能なビームタイムが限られているため測定時間の効率的な利用が不可欠である放射光X線や中性子ビームを使用する実験では特に重要である52,53,54。

結晶系と空間群の分類タスクのための2つのExRTモデルのトレーニングは、最新のラップトップコンピューターで数分しかかからない。これとは対照的に、文献34で提案されているCNNモデルは、現在最も高速なGPU（Nvidia TITAN RTX）を使ってもトレーニングに1日以上かかる。波長や2θの範囲などの測定パラメータは、回折実験や試料ごとに異なる可能性があるためです。

我々のモデルの顕著な欠点、すなわち非立方晶系、特に三斜晶の場合の精度の低さは、おそらくこれらの結晶系におけるトレーニングデータの不足に起因していると考えられます。この種の問題はクラス不均衡と呼ばれ、画像や音声認識タスクのようなMLのいくつかの分野では「データ増強」と呼ばれる手法によって解決される37,38,55,56。結晶系学習課題に対するデータ増強も、理論的には素直に可能である。実際の結晶構造を歪めたり拡大したりすることで、任意の数の人工的な低対称性結晶構造を生成することができる。しかし、人工構造の組み合わせは膨大な量になるため、データ補強による分類精度の向上は本稿の範囲を超えており、今後の課題とする。

本稿の範囲はMLを用いた新しい手法の提案に限定されており、現在のモデルが実データでうまく機能するとは現時点では考えていない。しかし、実データに対する我々のアプローチの現在の性能を確認することは興味深いので、いくつかのテストを行った。本文の「新素材と実際の実験データに対する分類性能」セクションと補足情報のセクションS4を参照してください。この手法を実データに適用するためには、MLモデルは実用的な測定条件下でのピークの重なりや不純物ピークのような不完全性に対して柔軟でなければならないが、現時点ではこれらの問題には対処していない。今後の研究では、モデルを更新し、実際の実験データで検証する予定である。