

準結晶及び近似結晶の物性の機械学習に向けた データセットの構築

Establishment of dataset for machine learning of physical properties of quasicrystals and their approximants

藤田 絵梨奈¹, 桂 ゆかり^{1,3,4}, 劉 暢², 吉田 亮², 木村 薫¹

Graduate School of Frontier Sciences, University of Tokyo¹, The Institute of Statical Mathematics²,
MaDIS, National Institute for Materials Science³, RIKEN⁴

E-mail: fujita@phys.mm.t.u-tokyo.ac.jp

近年、新規材料の発見に向けた機械学習の研究の増加に伴って訓練データとしてのデータセットの需要が高まっている。特に実験に基づくデータセットは試料の製造工程や測定条件を内包した情報であり、未知材料の予測に大きく寄与するものと考えられる。

しかしながら、テキストマイニングや図表の自動処理などによる実験データの収集は発展途上であり、精度の高いデータの収集には人間の知能を以ての収集が必要とされている。

今回は、我々が取り組んでいるディープラーニングを用いた準結晶・近似結晶の組成予測の研究で得られた予測結果の評価、及び訓練データセット拡充プロセスについて発表する。複数の媒体を対象とする場合、フォーマットに応じたツールを採用することで収集効率を高めることができる。補助的に小回りの効く自動化ツールを自作することによって人間の判断を容易にし、正確なデータの収集・評価に活用している。予測相図と実験相図の比較においては、198 報の論文から収集した 266 の相図の画像加工ツール作成することで対象領域内の視覚認識を容易にした。また、実験相図からは、63 の相図から

WebPlotDigitizer [1]を用い相の位置を座標化し、変換ツールを作成することで 577 の領域の組成の抽出を可能とした。テーブルからの組成収集では OCR (Optical Character Recognition) とマクロを使用し、データ収集を行なっている。現在は準

結晶及び近似結晶に特化したデー

タセットの構築も行っている。データの抽出・管理にオープンデータベースである Starrydata web system [2]を使用し、現在までに 217 の準結晶・近似結晶の温度依存性の電気抵抗率、比熱、磁化率などを論文の図中からカーブとして収集した。

今後はこれらの実験に基づいたデータセットを用い機械学習の予測精度の向上を図る。

[1] WebPlotDigitizer <https://automeris.io/WebPlotDigitizer/>

[2] Y. Katsura, *et. al.*, Sci. Tech. Adv. Mater. **20**, 511 (2019).

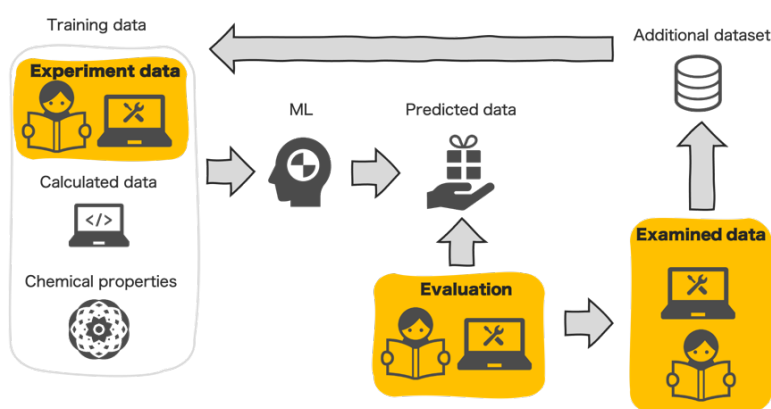


図 1.マテリアルズ・インフォマティクスにおけるデータの収集・評価のプロセス