

中谷アドバンスプログラム参加報告書

中島 春樹 (東京科学大学理工学系物質理工学院材料系学部 4 年)

留学先研究所: Deutsches Forschungszentrum für Künstliche Intelligenz

(ドイツ人工知能研究センター, DFKI)

Organizational unit: Smart Data & Knowledge Services

受け入れ: Prof. Dr. Prof. h.c. Andreas Dengel

アドバイザー: Dr.-Ing. Muhammad Nabeel Asim

留学期間: 2025 年 2 月 12 日 – 2025 年 5 月 4 日

1 インターン先での研究内容

1.1 要旨

化合物の組成のみに基づいて安定な結晶構造を予測する問題である結晶構造予測 (CSP) において、複数のモデル同士をより対等な状況で比較するためのフレームワークを構築し、そのフレームワーク上で複数のデータセットや評価指標を用いてモデルの評価を行った。

1.2 背景

材料開発において対象としている物質の構造を知ることが、その物質の特性を知るために重要である。しかし、化合物の組成から結晶構造を正確に予測することは困難である。これは結晶構造が周期的とはいえ多様であり、取り得る構造が膨大であることに起因する。1950 年代からこの問題は考えられており、密度汎関数理論 (DFT) や機械学習ポテンシャルと最適化アルゴリズムを組み合わせ、エネルギー的に安定な構造を探索する手法などが提案されてきた [1]。近年、深層学習や生成モデルを使い、安定な構造のデータセットからエネルギーを介さず、直接安定な構造の分布を学習する手法が様々な提案されて注目されている [2]。従来の手法よりもよりはやくより正確にエネルギー的に安定な構造を予測できるとして様々な研究されている。また、機械学習で結晶構造を扱うために使われている手法は結晶構造予測の問題だけではなく、結晶構造から物性値を予測する問題や物性値から結晶構造を予測する問題など他の問題に対しても応用可能であり、結晶構造予測に留まらないテーマとして研究されている。結晶構造予測だけでも様々なモデルが提案されている一方で、データセットがモデルによって異なる場合や評価指標の計算方法が異なる場合、実装コードが公開されていない場合など、モデル同士の直接的な比較が困難であるものも多い。

本研究ではこの問題を解決するために、同じデータセット、同じ評価指標を使い、異なるモデルを正確に比較するためのフレームワークを構築した。このフレームワーク上で複数のモデルを学習、評価することで正確にモデル同士の比較を行えるだけでなく、評価指標やデータセットの共通化による新たなモデルを提案するときの簡易化や新たな評価指標の実装とその指標での既存モデルの評価の簡易化も達成した。

1.3 方法

1.3.1 論文の分析

研究の初期段階として、機械学習を用いた化合物の構造予測を行っている論文、約 100 本をデータセットや問題設定、解決手法の観点から分析した。また、データセットとモデルの学習と評価の両方が実装されたコードが公開されているものについてはそれらのコードもレビューした。それらのレビューは、拡張性や拡充性のあるフレームワークを構築するときに活用した。特に実装のコードのレビューからは、既存の実装からフレームワークに移すときに変更が少なくなるようにフレームワークのアーキテクチャを考えるために活用した。

以下に、論文レビュー時に確認した項目例を示す。

- Type of Compound
- Tasks Name
- Datasets
- Dataset Statistics
- Database/s
- Experimental Setting
- Evaluation Metrics
- Inputs/Outputs

1.3.2 フレームワークについて

フレームワークのアーキテクチャは、既存の多くの結晶構造予測のモデルの実装で共通して使われているアルゴリズムの実装を変更せず使えるように、共通部分をモジュール化し、必要に応じて組み合わせるアーキテクチャとした。また、データセット、モデル、評価指標の実装がそれぞれ大きなモジュールで構成されるようにし、データセットの変更、モデルの変更、評価指標の変更が容易に可能な設計にした。

1.3.3 モデル比較

評価には、ペロブスカイト型 ABX_3 構造 18928 件から成る公開ベンチマーク Perov-5 データセットを使用し、訓練 60%・検証 20%・テスト 20% に分割した。

■対象モデルの概要

- DiffCSP — 拡散確率モデルにより格子と原子座標を同時に生成し、回転・並進・周期対称性をネットワーク内部で保証する [2]。
- CDVAE — 結晶グラフを潜在変数に圧縮し、デコーダで逐次デノイズして構造を復元する変分オートエンコーダ型モデル [3]。
- EquiCSP — DiffCSP を拡張し、格子の置換を含む対称操作に対しても厳密な同変性を満たすよう設計された手法 [4]。

■評価指標

1. 一致率 (Match Rate) テスト用の各組成について予測構造を生成し、pymatgen の `StructureMatcher` (`ltol=0.3`, `stol=0.5`, `angle_tol=10`) で正解構造と一致するかを判定した [5]。一致した組成を「成功」とみなし、その割合を計算した。
2. 原子位置 RMSD 一致が確認できた組について、予測構造と正解構造の原子間 RMSD を計算し、 $\sqrt[3]{V/N}$ (単位胞体積 V を原子数 N で割った立方根) で正規化して平均した値を用いた。

以上の設定により、3 手法を 同一データセット・同一指標 で簡潔かつ公平に比較した。

1.4 結果・考察

1.4.1 モデル比較

構築したフレームワーク上で、DiffCSP と CDVAE と EquiCSP の 3 つのモデルを Perov-5 データセットで学習させ、テストデータセットで一致率と原子位置 RMSD を評価した。値は表 1 に示す。

表 1: Perov-5 データセットにおける 3 手法の評価結果

手法	一致率 ↑	原子位置 RMSD ↓ (Å)
DiffCSP	0.50436	0.072998
CDVAE	0.42959	0.11483
EquiCSP	0.50753	0.067158

これらの値は論文 [3, 4] で述べられている値とおおよそ一致している。このことからこれらのモデルの学習と評価において、フレームワーク上での学習でそれぞれの論文を再現できていると評価できる。厳密に一致していない理由は、訓練・検証・テストに分割されて入っているデータが乱数設定が異なることで違うものになっている点やモデルを学習させたコンピュータの違い等が考えられる。

また、論文 [3, 4] ではデータセットの分割によるそれぞれの学習・検証・テストに入っているデータが同じであるかや学習環境が同じであるかなど統制できない要因を含む。それらの統制できない要因の影響は既存の論文 [3, 4] の結果の値だけからでは判断できなかった。しかし、今回のフレームワーク上の学習においては同じデータセット分割であることや評価指標の実装方法が同じことなどが保証されている。そのため、上で述べたような統制できない要因の影響を無視した直接的なモデルごとの比較が可能となった。

1.5 結論

結晶構造予測の3つのモデルをフレームワーク上で学習・評価した。同じフレームワーク上で学習・評価することで、既存の論文の結果だけでは保証することができなかった厳密な比較が可能となった。今後はこのフレームワーク上で他のモデルの学習・評価を行うことや異なるデータセットでの学習・評価を行うこと、また新たな評価指標を扱うことで、多岐にわたる結晶構造予測のモデルを正確に比較し、その性能の差異について厳密に議論することが可能となると予想される。

参考文献

- [1] Gautam R. Desiraju. Cryptic crystallography. *Nature Materials*, Vol. 1, No. 2, pp. 77–79, Oct 2002.
- [2] Rui Jiao, Wenbing Huang, Peijia Lin, Jiaqi Han, Pin Chen, Yutong Lu, and Yang Liu. Crystal structure prediction by joint equivariant diffusion, 2024.
- [3] Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi Jaakkola. Crystal diffusion variational autoencoder for periodic material generation, 2022.
- [4] Peijia Lin, Pin Chen, Rui Jiao, Qing Mo, Cen Jianhuan, Wenbing Huang, Yang Liu, Dan Huang, and Yutong Lu. Equivariant diffusion for crystal structure prediction. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, Vol. 235 of *Proceedings of Machine Learning Research*, pp. 29890–29913. PMLR, 21–27 Jul 2024.
- [5] Shyue Ping Ong, William Davidson Richards, Anubhav Jain, Geoffroy Hautier, Michael Kocher, Shreyas Cholia, Dan Gunter, Vincent L. Chevrier, Kristin A. Persson, and Gerbrand Ceder. Python materials genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science*, Vol. 68, pp. 314–319, 2013.

2 アドバンスプログラムで習得したもの

2.1 研究活動における日米欧の違い, RIES と ARIP の違い

RIES では米ジョージア工科大学に、ARIP では独ドイツ人工知能研究センター (DFKI) で研究した。私は RIES では留学した研究室のテーマを、ARIP では私の日本での専攻と CS に関連したテーマをと、行ったテーマとの関係が異なり、単純な米独の比較はできないため、ここからは RIES と ARIP でどのように研究が違ったか同じだったかについて書いていく。

RIES では電池の材料の研究を行った。私は材料科学専攻だが、RIES で行った研究テーマに関しては全く馴染みがなかったため、留学直前に身につけた知識やメンターに教わった知識、実験をして考えたことなどをもとに研究を進めた。そのため、私がわかっていないことに対してメンターや PI に教わるということも多く、自分の力で研究のすべてを進めているという感覚は少なかった。ARIP では私が材料科学専攻であり、機械学習を使った研究を行ったことがあるというところから、機械学習を使った化合物の構造予測のテーマをやることになった。そして、その研究プロジェクトでは私以外はすべてコンピュータサイエンス専攻の研究者や学生であった。そのため、この研究プロジェクトにおいては私が材料の知識を持っていることを前提に研究は進んだ。もちろんプロジェクトの中で CS のテーマを進める上での大まかな流れの提示などはあったが、プロジェクトの中で材料科学のドメイン知識を持った独立した一人の研究者として扱われたように感じた。これは RIES ではなかった経験であり、日本でもなかった経験であった。こうした扱いは研究をより大いに刺激的であった一方で、責任を負っている感覚があり、自分の材料科学の知識が十分なもののなか不安になった。

また、このプロジェクトでは研究が進んで、次にやることをアドバイザーの先生に相談したときに毎回”Up to you”と言われ、私がどのように研究を進めたいかを非常に重視していただいた。これによって、研究テーマについてどうすればより深く考えるきっかけとなった。

私の日本での所属大学と米独での留学先の一番の違いは、他の国から来ている人の数の割合であった。アメリカでもドイツでも大学や研究所では毎日英語以外の言語が聞こえてくるのが日常であった。世界中からここで学びたいと思われている環境は設備の面では優れており、人材の面でもバックグラウンドが多様で面白い人たちが集まっていた。

アメリカでもドイツでも PhD の学生が本当にハードワーカーであることを目の当たりにした。研究室の人たちは、朝早くから夜遅くまで研究をしており、研究に対する情熱を感じた。私も研究をやると決めたからには情熱を持ってやっていこうと思った。

2.2 現地の文化・生活面での発見・苦労等

ドイツでは、日曜日にほとんどの店が閉まっていることに驚いた。土曜日に買い物を済ませておかないと、日曜日は何も買えないというのは、私にとって大きな文化的な違いであった。一方で、ドイツ以外のヨーロッパの国に行くと、日曜日に開いている店も多く、ヨーロッパと言っても国に

よって大きく文化が異なることを実感した。

また、ドイツの鉄道は時間通りに来ないことが多く、特に長距離列車では遅延や運休が頻繁に発生していた。乗り換えの時間を 10 分以上の乗り換え余裕を確保するのが常識で、日本の時間の正確さとは大きな違いを感じた。

ドイツの食文化も日本とは大きく異なり、ジャガイモが主食的立ち位置にある点に驚いた。ドイツの学食であるメンザでは、フライドポテトと肉の揚げ物が昼食として提供されることが多く、最初はなにが主食なのか分からなかった。

また、私はドイツ語がほぼ話せなく、その国の第一言語が全く話せない土地で生活するというのは私にとっては初めての経験であった。私の留学したカイザースラウテルンには外国人がそこまで多くないため、ほとんどの人が外国人であってもドイツ語で話しかけてくる。そのため、私は相手の第一声は必ず理解できないという問題があった。今回、このような経験を初めてしたことで、その国の言語ができないことの大変さを思い知らされた。

3 本プログラムに参加の成果・意義

このプログラムを通して、多くのことを学ぶことができた。

ひとつ目は、研究の面白さを再認識することができたことである。今回は RIES よりも増して、私の今までの経験や知識が研究の成果に直結する形の研究であった。そのため、研究に「挑戦」しているという実感が RIES よりも増して感じることもできた。また、私にとってはその挑戦は非常に面白く、もっとその挑戦をしたいと思うものであった。加えて、今回の留学中には研究とは直接関わらないが関連する論文や本も多く読むことを心がけていた。この習慣が他の論文を理解するときの助けになることもあり、自分の研究テーマの範囲だけでなく、広く知識を持つことの重要性を感じた。今後も第一原理計算やマテリアルズインフォマティクスに関連する研究を行っていくつもりだが、より広い問題に挑戦できるように、特定の領域に留まらない学習をしていきたい。

ふたつ目は、単身での研究についてである。RIES では自分以外の JP Fellows も同じ場所に滞在して研究を行っていた。一方の ARIP では同じ場所には自分一人しかおらず、ただ研究をしていればいい RIES とは大きく違った。なにか問題があった時に簡単に相談できる相手がいないという不安がある一方で、生活においても研究においても自分のやりたいことを自分で決めて進めることができるという自由度があった。この自由度を活かして、研究を進めることができた。自分のやりたいことを自分で決めて進めることができたという経験は、今後の研究をする上でも大きな自信となると感じた。

三つ目は、人との繋がりである。DFKI にはコンピュータサイエンスを他の分野と組み合わせて研究するというプロジェクトがたくさんあり、私の研究テーマもその一つであった。よって DFKI にはコンピュータサイエンス以外にも様々な分野の研究者が集まっており、その中には研究者同士のつながりから始まった共同研究も多くあった。私も今回の留学で知り合った研究者の方々とのコネクションを大切にして、今後の研究に活かしていきたい。また、DFKI の研究者の方々は非常にフレンドリーで、研究の話だけでなく、生活の話や趣味の話なども積極的にしてくれた。このよう

な人とのつながりは、研究を進める上でも非常に重要であると感じた。特に、異なるバックグラウンドを持つ人々との交流は、新しい視点やアイデアを得るために非常に有意義であった。今後も積極的に人とのつながりを大切にしていきたい。

四つ目は、話すことの重要性である。私は英語が完璧ではなく、言いたいことがうまく言えないこともある。しかし、今回の留学では色々な人と積極的に話すことを心がけていた。その結果として、現地の人と話すことでドイツと日本の文化の違いを理解することができ、また新たな友人を作ることができた。言語の壁はあるものの、話すことで相手との距離を縮めることができると感じた。また、研究においても、他の研究者と積極的にコミュニケーションを取ることで新しいアイデアや視点を獲得することができた。言葉の壁を乗り越えるためには、積極的に話すことが重要であると再認識した。

最後は挑戦することの重要性である。今回の留学では、日本人の方とも多く交流する機会があった。日本からヨーロッパに来て、ヨーロッパで自分のやりたいことを究めている姿を見ると、私ももっと挑戦しなくてはいけないと感じた。今回の留学を通じて、快適な場所は居心地がいいが、挑戦することでしか得られないものもあると感じた。そのため、特に研究においては挑戦し続けていきたい。

4 その他

今回の DFKI での研究では、上でも述べたように私だけが材料科学で他のメンバーはコンピュータサイエンスのバックグラウンドを持っており、研究テーマは材料科学とコンピュータサイエンスの学際的な研究であった。この学際的な研究を通して、片方の分野を理解しているだけでは、研究できないということを強く感じた。私は今回の研究以前に機械学習を研究に使った経験があり、またソフトウェアのアーキテクチャの設計を何度か行ったことがあった。今回の研究においてはこのようなコンピュータサイエンスの知識を事前に持っていたことが本当に研究の役に立ったと感じている。また、プロジェクトのメンバーも以前分子に関するプロジェクトに参加していたメンバーが多く、材料科学についてデータセットの特徴やモデルの特徴など理解していることが多く研究をするときに大変助かった。加えて、DFKI は人工知能や機械学習を他の分野と組み合わせて研究するというを行っているプロジェクトがたくさんあり、それらの研究をやっている研究者と話すと、お互いがお互いの分野を理解していない最初の段階では研究が全く進まず、理解し始めてからちょっとずつ進むということを言っていた。今回の研究で改めて学際や異分野融合を行う際には、両方の分野を理解する必要があるということを再認識した。

5 謝辞

今回のプログラムでは、非常に多くの方に支えられ、有意義な留学をすることができました。

Dr.-Ing. Muhammad Nabeel Asim 氏には、アドバイザーとして多大なご指導をいただき、心より感謝申し上げます。また、生活面でも私が困っていないか気にかけてくださり大変助かりまし

た。ありがとうございました。Prof. Dr. Prof. h.c. Andreas Dengel 氏には、DFKI に私を受け入れてくださり、私にとって大変貴重な経験をさせていただきました。深く感謝いたします。さらに、Shan Munir 氏と Tayyaba Asif 氏には同じプロジェクトのメンバーとして研究において具体的なアドバイスや研究手法を教えてくださいました。厚く御礼申し上げます。また、同じ部屋でパキスタン料理を振る舞ってくれた Zain Ali 氏や研究所で昼ご飯やパーティーに誘ってくださった研究所の皆様にも感謝いたします。

渡航前に、留学に関する様々な情報を提供してくださったり、不安に感じていた部分に対してアドバイスをくださった清水太さん、辺見桃音さんをはじめとする Alumni の皆様にも感謝いたします。

また、留学を快く許可して下さり大学の手続きなどのサポートもして下さった東京科学大学大場史康教授にも感謝いたします。

加えて、中谷財団の小川様、藤川様には、渡航や現地で様々な面でのサポートをいただきました。心より感謝いたします。また、体調を崩した際にサポートして下さったジャパンスタディツアーの林様にも心より感謝いたします。

本研究のご支援を頂いた公益財団法人 中谷財団に感謝いたします。RIES につづき 2 度目の研究留学の機会をくださり本当にありがとうございました。



(a) 研究所の方々と



(b) DFKI の前で



(c) 研究所の方々と 2