

jh240059

## 統合機械学習分子動力学システムの構築

奥村雅彦（日本原子力研究開発機構）

### 概要

人工ニューラルネットワークを用いて低計算コストで高精度な原子分子スケールシミュレーション手法である「機械学習分子動力学法」を誰でも簡易に実施することができる「統合機械学習分子動力学システム」の構築を目的として研究・開発を行った。この手法は、大量の第一原理計算結果を生成してそれらを教師データとし、それらの一部を用いて人工ニューラルネットワークを訓練する。そのため、効率的な教師データの生成、データの貯蔵と抽出、人工ニューラルネットワークの訓練が必要になる。今年度は、そのために必要なソフトウェアの開発を継続した。また、一部の計算資源を用いて、機械学習分子動力学法による材料物性評価シミュレーションを実施し、蛍石物質の比熱異常が局所的対称性の破れで説明されることを示した。

### 1. 共同研究に関する情報

#### (1) 共同利用・共同研究を実施している拠点名

東京大学 情報基盤センター

mdx

#### (2) 課題分野

データ科学・データ利活用課題分野

#### (3) 参加研究者一覧と役割分担

奥村雅彦（代表・原子力機構）：研究統括・システム開発、永井佑紀（副代表・東大）：コード作成、華井雅俊（副代表・東大）：データベース (DB) 開発、鈴木豊太郎（東大）、芝隼人（兵庫県立大学）、町田昌彦（原子力機構）：統合システム開発、横井達矢（名大）、小林亮（名工大）、島村孝平（熊大）、中村博樹、小林恵太（原子力機構）：材料科学 DB 作成、安藤康伸（東工大）、：固体物性 DB 作成、河野秀俊、石田恒（量研機構）：生命科学 DB 作成、板倉充洋（原子力機構）、森英喜（産技短大）：金属材料 DB 作成、山口瑛子（原子力機構）：地球化学 DB 作成、渡邊聡（東大）、清水康司（産総

研）：表面界面物性 DB 作成、沖田泰良（東大）：原子力工学 DB 作成、加藤幸一郎（九大）：ナノ構造 DB 作成、富谷昭夫（大阪国際工科専門職大学）：素粒子物理 DB 作成

### 2. 研究の目的と意義

機械学習分子動力学法 (machine learning molecular dynamics, MLMD) は、密度汎関数法等の量子力学計算の結果を人工ニューラルネットワーク (artificial neural network, ANN) 等で学習して機械学習ポテンシャル (machine learning potential, MLP) を作成し、系全体のエネルギーや原子にかかる力を高精度かつ低計算コストで評価するミクロスケール現象のシミュレーション手法である。MLP は、入力が原子の粒子配置、出力がエネルギーの非線形関数 (ポテンシャルエネルギー曲面) を ANN で実現するものであり、量子力学計算の結果を基にしているため、経験的なパラメータを含まず、また、共有結合の生成/切断を伴うような化学反応のシミュレーションも可能である。最近では、さまざまな手法が提案され、パッケージも多数リ

リースされている。また、グラフニューラルネットワークを用いた手法の中には、マテリアルズプロジェクトなどのデータベースを教師データとして学習し、さまざまな原子種に対応した「ユニバーサル機械学習ポテンシャル (universal machine learning potential, UMLP)」も存在する。しかし、ある特定の物質の物性を調べるためには、その物質に対応するためのファインチューニングが必要である。UMLP 以外の MLP では必ずユーザー自ら教師データを準備して訓練が必要であり、結局、どの MLP アーキテクチャーを選んでも、訓練は避けて通れない。

MLP の訓練には、【問題 1】大量の学習データ (量子力学計算の結果) 及びそれらを作成・学習するための計算機資源が必要、【問題 2】計算対象ごとに MLP を作り直すことが必要、【問題 3】適切な教師データセット作成の決まった手順がなくデータセット作成の経験が必要、という 3 つの問題が存在する。これらを効率的に実施するためには、経験が必要のため、MLMD は今でも、誰でも手軽に使える手法ではないと言える。

そこで、本研究では、これらの問題を解決して誰でも MLMD 計算を実施できる環境を提供するために、a. 大量の量子力学計算を実施して適切な教師データセットを作成するシステム、b. 教師データのデータベース、c. データベースを利用して MLP を作成する機械学習システム、d. 作成した MLP のデータベース、e. MLP を用いて大規模 MLMD を実施する計算環境、を備えた統合 MLMD システムを mdx に構築し、運用することを目的とする。

分子シミュレーションには長らく主に古典分子動力学法と第一原理分子動力学法が用いられてきたが、前者は計算コストが低いとその性能は経験的パラメータの調整に強く依存し、後者は特別な調整の必要はないが計算コストが高いため、適用対象が限られていた。MLMD はこれらの手法のデメリットを

解消する手法であるが、普及しているとはいえない。その理由は、上記 1、2 に示した通りであり、本申請により統合 MLMD システムが完成すれば、MLP 作成の経験がないことや教師データ作成・学習の計算機資源がないことを理由に MLMD の利用を躊躇・断念していた研究者に研究の機会を提供することが可能になる。MLMD は化学反応等の既存のシミュレーション手法で扱うことが難しい物質・現象の研究を可能にする手法であり、多くの研究者が利用可能になれば、様々な分野における原子・分子スケールでのシミュレーション研究が大きく進展することが期待される。また、将来的に多くの研究者にシステムを利用してもらうことによって教師データが増え、システムの利便性が向上することが予想され、システムとユーザー双方にメリットをもたらす循環的な発展が期待される。

これまで、研究が終わった後の計算データは削除されるか研究者個人によって保管され、再び使用される事は稀であった。本研究により、使い終わった計算データが他の研究者によって教師データとして利用される可能性が生まれ、「データのリサイクル」と呼べるデータ利活用推進が期待される。さらに、近年では、科学研究費助成事業において、研究過程において、どのような種類の「研究データ」をどのように管理・利活用するか等について整理した計画書であるデータマネジメントプランの作成が要求されており、研究不正防止等の観点からもデータの保存と検索が重要視されるようになってきている。

本研究によって、迅速な材料開発 (材料科学)、エネルギー技術開発の加速 (エネルギー科学)、環境問題の解決 (地球化学)、素粒子物性の解明 (素粒子物理)、生命機能の解明 (生命科学) 等、社会問題の解決につながる科学的知見が得られれば、実社会へ強いインパクトが与えられると期待される。

### 3. 当拠点の公募型共同研究として実施した意義

本研究課題で扱う機械学習分子動力学法は、教師データ作成のための量子力学計算を大量に実行する必要があるため、大規模並列計算が必要である。その一方で、その大量の計算結果を貯蔵し、その一部を選択して教師データとして利用する必要があるため、大規模データ処理に適した mdx を利用する。このように、本研究は異なる目的に適した計算機を有機的に連携させることにより、これまでにないシステムを構築することを目的としており、本研究は JHPCN の公募研究でなければ実現が難しい研究開発課題である。

### 4. 前年度までに得られた研究成果の概要

上記2の研究目的に向けて前年度までに下記のシステム開発を行なった。

1. mdx から東大情報基盤センターの Wisteria/BDEC-01 にジョブを投入し、結果を mdx に回収するスクリプトを作成した。これは、データベースに自動で第一原理計算の計算結果、すなわち教師データを集積するために必要な機能であり、現段階では試作段階であるが、教師データのデータベースを自動で作成するために必要な機能の一つの実装を完了した。
2. 東大情報基盤センターの Wisteria/BDEC-01 において実行、コンパイルの手間を省くため、よく用いられる第一原理計算パッケージの VASP と Quantum Espresso について、Singularity コンテナを作成した。これにより、ユーザーを、計算機環境に強く依存する第一原理計算コードのインストールから解放し、インストールに要していた時間を有効利用することができるようになった。
3. よく用いられているオープンソースの機械学習分子動力学法パッケージの一つに aenet があり、共同研究者の森英喜博士

(産技短大) が古典分子動力学オープンソースコード LAMMPS で aenet を動作させるプラグインを開発している。今年度はプラグインの改良を行い、実行速度を2倍に高めた。

4. 教師データとなる量子力学計算の結果は原子配置だけでなく、エネルギーカットオフや k 点数などの計算条件を含んでいるため、複雑な情報構造を持っている。そのため、リレーショナルデータベースを用いて情報を整理するのは容易ではない。そこで、json 形式のデータを MongoDB で扱うこととし、VASP の出力 XML 形式ファイルを json 形式に変換する Python スクリプトを作成した。
5. オープンソースの機械学習分子動力学法パッケージ n2p2 を簡易に実行できる Python パッケージの作成に着手した。今年度は、下記の操作を自動で実施するスクリプトを作成した。
  - (ア) n2p2 のインプットファイルを生成する
  - (イ) 大量の VASP の実行結果を収集し、n2p2 の教師データフォーマットに変換する
  - (ウ) 教師データの規格化を実施する
  - (エ) 原子分布の特徴量である記述子を規格化する
  - (オ) 学習を実施する
 これらは下記のようにそれぞれ1行のコマンドで書くことができ、それぞれ簡易に実行可能である。
  - (ア) `n2p2.write_nn()`
  - (イ) `n2p2.write_data()`
  - (ウ) `n2p2.normalize()`
  - (エ) `n2p2.scaling()`
  - (オ) `n2p2.train()`
 上記「n2p2」は作成したスクリプトで定義される N2p2 クラスのインスタンスである。上記、(ウ)、(エ)、(オ)は n2p2 の

実行形式のファイルを実行するコマンドであるが、直接実行する形式と PBS 等にジョブを投入しする形式を選べるようになっている。

最終的に訓練が終わった機械学習ポテンシャルを用いて、分子動力学法パッケージ LAMMPS による機械学習分子動力学計算が実施可能であることを確かめた。

これまで、様々なパラメーターをユーザーが決める必要があったため、初心者はどのパラメーターをどの程度の値にすべきか迷うことがあったが、開発中の Python スクリプトはデフォルト値が設定しており、どのような物質に対しても大きく失敗しないようなパラメーターを使ったインプットファイルが作成可能であるため、初心者でも気軽に機械学習分子動力学法の研究を始めることができる。また、規格化や訓練等の実行形式ファイルについても、コンテナを用意しておくことで、ユーザーはコンパイルの手間なく、確実に実行が可能になる。

上記の一連の開発によって、数行の Python スクリプトを書くだけで機械学習分子動力学法が実行できる環境が整い、mdx からのジョブ投入及び結果の回収、データベース作成の基礎部分が完成した。

記の開発に加え、計算時間の一部を用いて下記の機械学習分子動力学計算を実施し、下記の成果を得た。

- A) 体心立方晶遷移金属 $\alpha$ -Fe について、機械学習ポテンシャルを作成し、劈開破壊についての分子動力学シミュレーションを実施した。その結果、従来の embedded-atom-method と呼ばれる古典力場を用いるとアーティファクトが出てしまう場合においても、正しいシミュレーションが可能であることが示された。さらに、作為的な理想的な劈開モデルではなく、より現実的な歪みがあ

るような亀裂についても機械学習分子動力学シミュレーションを実施し、今後のより現実的な金属劈開シミュレーションの可能性を拓いた。(T. Suzudo *et al.*, Sci. Rep. **12**, 19701 (2022).)

- B) 機械学習ポテンシャル、密度汎関数法、透過型電子顕微鏡を用いて、酸化アルミニウムの粒界の原子スケール構造とその電子状態を調べた。機械学習ポテンシャルはエネルギーが最低状態のスクリーニングに使用された。機械学習ポテンシャルは、密度汎関数法と同程度の精度かつ計算がより高速であるため、効率的なスクリーニングが可能であることが示された。そして、シミュレーションによって得られた原子スケール構造が透過型電子顕微鏡による観測結果と一致することが示された。得られた原子スケール構造を基に、粒界の電子状態を調べた結果、粒界のバンドギャップはバルクに比べて小さくなっていることがわかり、これは先行研究の実験結果と一致することが確かめられた。(T. Yokoi *et al.*, Scr. Mater. **229**, 115368 (2023).)
- C) アルミニウムの機械学習ポテンシャルを構築し、アルミニウムの熱物性を評価した。その結果、学習データに含まれない結晶粒界の原子スケール構造についてもフォノン状態密度と振動エントロピーを正確に予測できることが示された。また、古典力場は結晶粒界とバルクにおけるフォノン周波数と原子間力に大きな誤差が見られることもわかり、粒界の熱物性の正しい評価のためには機械学習ポテンシャルを用いるのが有効であることが示された。(T. Yokoi *et al.*, Phys. Rev. Mater. **7**, 053803 (2023).)

## 5. 今年度の研究成果の詳細

昨年度に残された課題であった、第一原理計算の結果を蓄積するデータを制御するコードを作成した。データベースのソフトウェアには、柔軟な検索に対応するため、MongoDB を用いる。今年度作成したコードは、第一原理計算コード Vienna Ab initio Simulation Package (VASP) の計算過程及び計算結果が記されている XML ファイル（以下、結果 XML と呼ぶ）について

- ・結果 XML の情報のデータベースへの登録 (bson で格納)
- ・結果 XML と同等なデータ (json 形式) の検索、取得
- ・MLP のトレーニングに必要なデータ (原子種、原子位置、各原子にかかる力など) のデータベースからの検索、取得

が可能である。また、起点とするディレクトリから再帰的に VASP の計算結果を収集し、データベースに登録することが可能であるため、ユーザーの既存の VASP の計算結果及び今後の計算結果を全てデータベースに格納し、検索して MLP 作成等に利用することが可能になる。

上記の開発に加え、計算時間の一部を用いて下記の機械学習分子動力学計算を実施し、下記の成果を得た。

- 代表的な核燃料物質であるウラン、プルトニウム、トリウムの二酸化物は蛍石構造を持つ。したがって、蛍石構造の熱物性の研究は原子炉の安全性を高めるために、非常に重要である。蛍石構造が持つ特徴的な熱物性に、比熱異常がある。この比熱異常は、超イオン転移と呼ばれ、格子を構成する 2 種類の原子のうち、一種類が格子構造を保ったままもう一種類がフレンケル対のような欠陥を形成することに起因すると考えられている。

この研究では、MLP を用いて  $\text{ThO}_2$  など

の局所対称性の破れの観点から比熱異常を調べた。具体的には、八面体構造の乱れを表す局所対称性を調べた結果、過冷却水や液体シリカに似た振る舞いをすることがわかった。超イオン転移前には、格子欠陥のような局所構造と液体のような局所構図が共存するが、転移後は液体のような振る舞いが支配的になることがわかった。(J. Chem. Phys. に採択済)

これらのシステム開発の進捗や研究成果について国際会議の招待講演や日本原子力学会の講演で報告した。

## 6. 進捗状況の自己評価と今後の展望

申請時の研究計画とその進捗状況をまとめる。

### 研究テーマ 1：システム開発 (mdx)

#### 6.1.1 教師データのデータベースの構築

- フォーマットの決定：100%
- 既存データを用いたデータベース作成：0%

#### 6.1.2 教師データ作成環境の構築・開発

- 密度汎関数法実施環境の整備 (Quantum Espresso 等を予定)：100%
- 計算結果の自動データベース化コードの仕様決定：100%
- 計算結果の自動データベース化コード作成：50%

#### 6.1.3 学習環境の構築・開発

- 学習実施環境の整備 (オープンソースコード n2p2 等を予定)：100%
- データベースとの連携コードの仕様決定：100%
- データベースとの連携コード作成：50%

#### 6.1.4 MLMD 実施環境の構築・開発

- 分子動力学法コード LAMMPS 及び MLP インターフェイスのインストール (n2p2 等を予定)：100%

#### 6.1.5. 適切な教師データ作成手法の開発

- a. 教師データ作成手順に関するノウハウの整理: 10%
- b. 教師データ作成手順: 10%

#### 6.1.6. 独自コードの開発

- a. 統合システムに最適化された独自の MLMD コードの開発: 10%

### **研究テーマ 2：本計算（教師データ作成、学習、MLMD 実行）（Wisteria/BDEC-01）**

#### 6.2.1. 材料科学分野、エネルギー科学分野、地球化学分野、素粒子物理分野、生命科学分野: 100%（継続して成果を創出）

上記をまとめると、機械学習ポテンシャル作成や機械学習分子動力学法実行のためのコード実行環境についてはコンテナを作成することで目標を達成しており、データベースのフォーマットについても、データに含まれる情報に対応するため、json 形式で非リレーショナルデータベースを構築する基礎部分が完成した。一方で、データを蓄積していくソフトウェアの開発が遅れている。研究テーマ 2 の本計算については、本来であれば、完成した統合機械学習分子動力学システムを用いて研究を実施すべきであるが、まだ実用に至っていないため、共同研究者がそれぞれのノウハウを活かして研究を実施した結果が得られている。

今年度の全体的な進捗状況としては、開発が予定していたよりも大幅に遅れているが、ソフトウェアの基盤部分が出来つつあり、遅いながらも開発は前進している。