

jh230069

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・開発を行なった。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、体心立方格子の鉄における線欠陥（転移）の大規模ダイナミクス及びシリカガラスの中距離秩序の起源を明らかにした。

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) 共同研究分野 (HPCI 資源利用課題のみ)

超大規模数値計算系応用分野

(4) 参加研究者の役割分担

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・原子力機構）：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、浦田新吾、今村穰、吉田拓未（AGC 株）、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（産総研）、：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、清水康司、飛田倫太郎、大塚竜慈、寶穎、劉玉璽、田村達也（東大）：表面界面物性 DB 作成、沖田泰良、網谷駿（東大）：原子力工学 DB 作成、加藤幸一郎、松本大夢（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ーを含まず、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、最もユーザー数が多い密度汎関数法コードの一つ Vienna ab initio simulation package (VASP)にも機械学習ポテンシャルが導入され、また、Preferred Networks 社が ENEOS 社と機械学習分子動力学法を提供する合弁会社を設立するなど、機械学習分子動力学法が大きな注目を集めている。しかし、この手法には現時点で、【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ため、誰でも手軽に使える手法ではない。

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がその性能は経験的パラメーター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、普及しているとは言い難い。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。

これまでは、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、初心者ほ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。

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。

- A) 体心立方晶遷移金属 α -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

シヨ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

今年度は、前年度に引き続き、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 パッケージの開発を継続した。前年度までに、計算実施に関する一連のメソッドが完成したので、今年度は、各計算のアウトプットをデータベースに格納するためのメソッドを実装した。例えば、教師データを作成した VASP の xml フォーマットのアウトプットファイル (vasprun.xml) から、json ファイルに自動で変換するメソッドを実装した。このメソッドを用いると、VASP の xml アウトプットファイルが含む MLMD に不必要な情報 (電子状態計算における収束中の計算情報など) を取り除いて json ファイルに変換することが可能になる。さらに、研究データの保存という観点からは、必要な情報だけを取り出してデータベースに格納し、いつでも計算結果が取り出せるので、研究不正防止や、新たに関連研究を始める際に有用な機能と言える。この実装により、MLMD に必要な教師データのデータベースの自動作成に近づいたが、計算実行機能とデータベース作成機能の連携部分の実装が遅れている。

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。

- A) 金属は原子が規則的に並んだ結晶構造をとるが、実際の金属においては一部の原子が抜けてしまっている欠陥が存在し、特に、原子の並びが線状に乱れた領域における線欠陥を転移と呼ぶ。転移のダイナミクスは金属の機械特性に大きく影響するため、金属の機械特性向上のためには転移の物理的性質を明らかにし、制御する必要がある。転移は原子スケールの構造変化であるが、その構造は 100 万個の原子についてダイナミクスを評価する必要があるため、これまで計算が高速な古典分子動力学法が用いられてきた。しかし、より精度の高い第一原理計算の結果を再現でき

ておらず、転移のダイナミクスの理解は進んでいなかった。本研究では、体心立方格子の鉄を対象として機械学習分子動力学法シミュレーションを実施し、古典分子動力学法では実験結果と矛盾していた転移ダイナミクスについて、実験と整合する結果を得た[1]。この結果は、機械学習分子動力学法が転移ダイナミクス評価に非常に効果的であることを示しており、今後、金属の転移ダイナミクス評価手法として、機械学習分子動力学法が標準的な手法となることを示唆している。

ガラスは、液体の原子配列が凍結した無秩序な原子配列を持つと考えられてきた。しかし、シリカガラスの X 線・中性子回折実験データには低散乱ベクトル領域に鋭い第一ピーク (First Sharp Diffraction Peak: FSDP) が現れ、無秩序なガラスの原子配列の中には一定の原子配列の繰り返しパターンが存在すること、すなわち中距離秩序構造が存在することが示唆されていた。この FSDP の構造的起源がどのような構造なのか、長い間議論されてきた。特に、FSDP は圧縮により大きく変化することが報告されており、低温で圧縮したシリカガラスと高温で圧縮したシリカガラスの FSDP はそれぞれ減少、増大することが実験で確かめられている。われわれは、シリカの液体状態、固体状態、ガラス状態の全てを記述する機械学習ポテンシャルを作成し、液体状態からの急冷シミュレーションを実施し、実験で得られた X 線・中性子回折スペクトルを高精度で再現することに成功した。さらに、低温圧縮、高温圧縮の際の FSDP の振る舞いも再現することに成功した。これらの結果を基に、FSDP に直接関係する構造を調べた結果、圧縮時

のリング構造の変形挙動が重要であることを明らかにした。低温圧縮ではシリカのつながりを保ったまま圧縮されるために、全てのリング形状が大きく歪み、FSDP の減少を引き起こす。一方、高温圧縮では、シリカのつながりの破壊と再結合が起こり、大きなリングの形状のみが優先的に変形することが分かった。結果として、高温圧縮では小さなリングから大きなリングまで長さスケールが揃い、FSDP の発達を引き起こしていることが示された。これらの結果は、機械学習分子動力学を用いれば、実験結果の情報なしにシリカガラスの非晶質構造を計算機上で作り出せることを示しており、ガラスに関する理論的な研究が進むことに加え、工業的なガラスの機能改善等への応用が期待される。

これらの研究成果について、日本原子力研究開発機構、AGC 株式会社、J-PARC センター合同でプレスリリースを行った[ア]。さらに、論文[2]は、2023 年 11 月 16 日という年末に出版されたにも関わらず、Scientific Reports 誌の Materials Science 部門で 2023 年に最もダウンロードされた論文の Top 100 (16 位) に選出された[イ]。また、液体・非晶質研究会にて招待講演を行った[a]。

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 [A, B] 及び日本原子力学会の講演 [b] にて報告した。

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

- a. フォーマットの決定 : 100%

- b. 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

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

- a. 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%
- b. データベースとの連携コードの仕様決定 : 100%
- c. データベースとの連携コード作成 : 20%

6.1.4. MLMD 実施環境の構築・開発

- a. 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 : 50% (ソフトウェアが完成していないため使用できないが、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に json 形式で非リ

レーショナルデータベースを構築することが決まった。また、教師データである第一原理計算の結果を json 形式に変換するコードを実装した。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おり、次年度の開発を予知している。また、適切な教師データ作成手法と独自コードについても、開発が遅れており本格的な開発は次年度に持ち越される状況である。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と言える。また、MLMD の応用研究の成果創出にも成功している。

7. 研究業績

(1) 学術論文 (査読あり)

- [1] H. Mori, T. Tsuru, M. Okumura, D. Matsunaka, Y. Shiihara, and M. Itakura, “Dynamic interaction between dislocations and obstacles in bcc iron based on atomic potentials derived using neural networks,” *Physical Review Materials* **7**, 063605 (2023).

DOI:

10.1103/PhysRevMaterials.7.063605

- [2] K. Kobayashi, M. Okumura, H. Nakamura, M. Itakura, M. Machida, S. Urata, and K. Suzuya, “Machine Learning Molecular Dynamics Reveals the Structural Origin of the First Sharp Diffraction Peak in High-Density Silica Glasses,” *Scientific Reports*

13, 18721 (2023).

DOI: 10.1038/s41598-023-44732-0

(2) 国際会議プロシーディングス (査読あり)

(3) 国際会議発表 (査読なし)

[A] M. Okumura, “Machine learning molecular dynamics simulations of materials with complex structures and behaviors,” The 7th International Symposium on Frontiers in Materials Science, Hinchu, Taiwan, January 21st–24th, 2024. **【招待講演】**

[B] M. Okumura, “Machine Learning Molecular Dynamics: A Fast and Highly Accurate Simulation Method for Materials,” 4th International Conference on Data Driven Plasma Science, Kyoto, April 16th–21st, 2023.

【招待講演】

(4) 国内会議発表 (査読なし)

[a] 小林恵太, “機械学習分子動力学法による高密度シリカガラスにおける中距離秩序の解析,” 2023 年度 液体・非晶質研究会, 水戸市, 茨城県, 2024 年 3 月 9 日.

【招待講演】

[b] 奥村雅彦, “統合機械学習分子動力学システムの開発とその応用”, 日本原子力学会 2024 年春の年会, 近畿大学, 2024 年 3 月 18 日.

(5) 公開したライブラリなど

(6) その他 (特許, プレスリリース, 著書等)

プレスリリース

[ア] ガラスの複雑な原子構造を高速・高精度な原子シミュレーションで再現！—ガラスの一見無秩序な構造の中に潜む秩序を解明—, 国立研究開発法人日本原子

力研究開発機構, AGC 株式会社, J-PARC
センター, 令和 5 年 11 月 16 日.

(<https://www.jaea.go.jp/02/press2023/p23111601/>)

論文ダウンロード数

[イ] Scientific Reports 誌の Materials
Science 分野において、2023 年に最もダ
ウンロードされた論文 Top 100 に選出
(16 位) :

<https://www.nature.com/collections/ehjfeidbbc>

