jh220057

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。そのために必要なソフトウエアを作成し、計算環境を整備した。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、鉄の劈開、酸化アルミニウムの粒界の原子スケール構造と電視状態、アルミニウムの粒界の熱物性を明らかにした。

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（該当するものを残す）

東京大学 情報基盤センター

mdx

(2) 課題分野（該当するものを残す）

データ科学・データ利活用課題分野

(3) 共同研究分野（HPCI 資源を利用している研究課題のみ，該当するものを残す）

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、鈴村豊太郎、芝隼人(東大)、町田昌彦(原子力機構):統合システム開発、浦田新吾、今村穣、吉田拓未(AGC 株)、 横井達矢、浜島明宏、松浦茉耶(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、渡邉聡、清水康司、飛田倫太朗、大塚竜慈、竇穎(東大):表面界面物性 DB 作成、沖田泰良、網谷駿、田川秀明、松田那由多、山本耀二郎(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。また、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ると期待され、実社会インパクトが期待される。

3. 当拠点の公募型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

5. 今年度の研究成果の詳細

2 の研究目的に向けて今年度は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 Oakbridge-CX、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 Oakbridge-CX、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あり、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、今年度は数行の 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[1]。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。

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[2]。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。

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[3]。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。

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開始: 100%（作成を開始）

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開始: 100%（作成を開始）

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）（ Oakbridge-CX 、 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: 20%

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。一方で、データを蓄積していくソフトウエアの開発が遅れており、データベースの作成まで至っていない。また、適切な教師データ作成手法の開発については、共同研究者が各々進めている研究において、有用な知見を集約する必要があるが、まだ集約できておらず、次年度の課題としたい。独自コードの開発についてはまだ完成に至っていないが、一定のペースで開発が進んでいる。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、目標を達成できた項目も多いが、予定していたよりも開発のペースが遅れている部分もあるが、どのような開発に時間がかかるのかを把握できたので、この経験を今後の研究開発に活かしていきたい。

今後は、今年度進まなかった「データベースの構築」と「適切な教師データ作成手法の開発」を実施し、mdx とスーパーコンピューターの連携部分の実装を進めたい。

7. 研究業績

(1) 学術論文 （査読あり）

[1] T. Suzudo, K. Ebihara, T. Tsuru, H. Mori, “Cleavages along {110} in bcc iron emit dislocations from the curved crack fronts,” Scientific Reports **12**, 19701 (2022).
DOI: 10.1038/s41598-022-24357-5

[2] T. Yokoi, A. Hamajima, J. Wei, B. Feng, Y. Oshima, K. Matsunaga, N. Shibata, Y. Ikuhara, “Atomic and electronic structure of grain boundaries in a-Al2O3: A combination of machine learning, first-principles calculation and electron microscopy,” Scripta Materialia **229**,

115368 (2023).
DOI: 10.1016/j.scriptamat.2023.115368
[3] T. Yokoi, M. Matsuura, Y. Oshima, K. Matsunaga, “Grain-boundary thermodynamics with artificial-neural-network potential: Its ability to predict the atomic structures, energetics, and lattice vibrational properties for Al,” Physical Review Materials **7**, 053803 (2023).
DOI: 10.1103/PhysRevMaterials.7.053803

(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）

[1] 奥村雅彦（原子力機構）「粘土鉱物分子動力学シミュレーションの新展開：機械学習分子動力学法」第 65 回粘土科学討論会, 2022 年 9 月 7 日, 島根大学.【招待公演】

[2] 奥村雅彦（原子力機構）「粘土鉱物の機械学習分子動力学法シミュレーション」, 原子層鉱物の機能開拓に向けた計算・計測連携研究会, 2022 年 5 月 17 日, 筑波大学. 【招待公演】

[3] 森英喜（産技短大）「材料強度の解析および予測的評価に向けた高精度ニューラルネットワーク型原子間ポテンシャルの開発と適用」, 第二回マルチスケールマテリアルモデリングシンポジウム, 2022 年 5 月 29 日, 大阪科学技術センター.

(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）