様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　６ 年　 ６ 月 １３ 日現在

機関番号：８２５０２
研究種目：挑戦的研究（萌芽）
研究期間：2020～2023
課題番号：２０Ｋ２０９０７

研究課題名（和文）分布適合原理と機械学習に基づくシミュレーション技法の開拓

研究課題名（英文）Distribution Matching Principle for Machine Learning Based Molecular Simulation

研究代表者
櫻庭　俊（Sakuraba, Shun）

国立研究開発法人量子科学技術研究開発機構・量子生命科学研究所・主幹研究員

研究者番号：９０６４７３８０

交付決定額（研究期間全体）：（直接経費）　　4,800,000円

研究成果の概要（和文）：古典分子動力学(MD)シミュレーションは、分子構造を計算機上に再現し計算することで分子の種々の物理化学的特性を解析する手法である。古典MDシミュレーションでは原子間相互作用や原子グループ間の相互作用を記述する力場関数、並びに力場関数のパラメータが分子の計算機中での振る舞いを規定するため、適切な力場関数およびパラメータ決定は極めて重要である。本研究では計算コストの低い関数系を計算機に自動で探索させるため、記号回帰のアルゴリズムを分子シミュレーションに合わせて実装し、探索を行った。実際に既存の量子化学計算に基づくデータセットから、古典の力場モデルとしての関数系を作成することに成功した。

研究成果の学術的意義や社会的意義
これまで、分子シミュレーションの力場の提案は人の直感に基づく関数系の提案と、長い時間を掛けた人手によるパラメータ改善の試行錯誤により実現されてきた。本研究では実際のデータからnon-trivialな関数系の「発見」を行っており、分子シミュレーションの力場の提案をデータ中心に行う一助となることが期待される。これにより、現在は高コストな計算（量子化学計算、全原子シミュレーション）がより低コストな計算（古典、陰溶媒、粗視化）で近似できるシステマティックな手法が整備され、階層的なシミュレーションがより容易に、低コストで、大規模に実現されていくことが期待できる。

研究成果の概要（英文）：Classical molecular dynamics (MD) simulations enable us to analyze various physicochemical properties of molecules by reproducing and simulating molecular structures on computers. In classical MD simulations, the function of the force field that describes the interactions between atoms (or groups of atoms), as well as the parameters of the force field function, determines the behavior of the molecules in the computer. Finding appropriate functions and their parameters is thus vital in the simulation. In this research, I aimed to find force field functions with low calculation costs by computer. I implemented a symbolic regression algorithm fitted to the molecular simulation and searched the functions. In an existing dataset of quantum chemical calculation, functions that fit well with the dataset were successfully found.

研究分野：計算科学

キーワード：記号回帰　分子動力学シミュレーション　パラメータサーチ

※科研費による研究は、研究者の自覚と責任において実施するものです。そのため、研究の実施や研究成果の公表等については、国の要請等に基づくものではなく、その研究成果に関する見解や責任は、研究者個人に帰属します。

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

古典分子動力学(classical molecular dynamics; 以後 MD)シミュレーションは、分子構造を計算機上に再現し、その原子間や原子グループ間の相互作用をニュートン力学的に表し、時間発展を計算することで種々の物理化学的特性を解析する手法である。MD シミュレーションはこれまで材料設計、溶液構造解析、生体分子解析など広範な現象に適用され成功を収めてきた。MD シミュレーションに於いては原子間相互作用や原子グループ間の相互作用を記述する力場とそのパラメータが分子の計算機中での振る舞いを規定する。このため、高精度な力場パラメータの作成は MD において根幹を為すステップとなっている。

力場パラメータの決定はこれまで、MD シミュレーションの種別を問わず、実験から得られた結果を再現することを目標として作成されてきた。多くの場合、これは手動での職人芸的な改良の積み重ねによって達成されてきた。例えばリボ核酸(RNA)の全原子 MD の力場パラメータとして現在最も広く使われている AMBER99bsc0χOL15 力場は、(1)核酸主鎖・塩基の電荷決定 (2)電荷補正 (3)主鎖二面角の補正 (4)側鎖二面角の補正　と改良を積み重ねて作られている。これらのいずれの段階も、実験結果との乖離が確認されたことを発端として改良のためのアドホックな修正を加えたものである。このような「実験値との乖離が明らかとなり、充分に周知された段階になってはじめて力場が改良される」プロセスは、これまで力場決定において半ば当然のこととして受け入れられてきた。しかしながら、このような力場決定のプロセスは「実験値がないものをシミュレーションすることができない」問題を引き起こし、MD シミュレーションを予測手法として用いる際にその適用範囲を狭めて来た。

２．研究の目的

本研究はこのシミュレーションの「当たり前」の制約に対して疑問を投げかけるところを出発点としている。すなわち、このような力場パラメータの決定プロセスは「当たり前」で動かせないものなのだろうか？　実験値に対するアドホックな修正を重ねることなく高精度な力場パラメータを用意することは可能であろうか？　可能であれば、どのような手段を用いることで実現が可能であり、既存のパラメータを上回る高精度なパラメータを作成することができるだろうか？

本研究ではこうした力場パラメータとその決定法に対し、統一的な原理原則に基づき決定プロセスを再構成することを長期的な目標とした。本研究計画では「シミュレーションから得られる統計分布は、より高度なシミュレーション結果から得られる統計分布と一致する」という自然な原理原則（以下、分布適合原理と称する）に則り、新たな力場決定手法の構築とその応用を行うことを目的として研究を実施した。

分布適合原理の考え方は、シミュレーションどうしをつなぐ上で自然なものであり、全く奇特なものではない。例えば全ての原子を質点として扱う全原子 MD シミュレーションは、より高精度でより適切に実験を再現するシミュレーションである量子力学(QM)計算から推定される配座分布と結果が一致することが好ましい。また、いくつかの原子グループを集め、グループ間の相互作用で系を記述する粗視化 MD シミュレーションは、全原子 MD シミュレーションから得られる統計分布を再現することが好ましい。しかしながらこれまで、分布適合原理に基づく力場パラメータの決定が行われてこなかった背景には、技術上の困難がある。分子配座の分布関数を一致させるためには分布関数そのものを何らかの形で評価する必要があるが、多次元の分布関数を直接的に扱う数学的な道具がこれまで不足していた。代替としてこれまでは電荷のみを QM 計算から決定する、動径分布関数など 1 次元に射影した分布関数を用いる、少数の実験値を再現するようなパラメータを設定する、最小エネルギー構造のごく近傍についてフィッティングを行うなど、「パラメータのごく一部に着目するか、分子の取り得る構造のごく一部に集中する」方法が採られてきた。結果として着目した評価指標だけを再現する過適合状態が発生し、分布適合原理の適用を難しくしていた。

本研究では、いままで分布適合原理を用いる上で存在していた技術的な障害を
(1) 近年劇的に発展した多次元分布関数に対する機械学習アルゴリズム
(2) 古典・量子双方のシミュレーション技術の発展
(3) 計算機の並列性能向上と並列性能を生かすアルゴリズム
の組み合わせにより解決し、新たな力場決定手法を構築することを目指した。

３．研究の方法

研究申請当初の時点では、上記の力場改善に対して敵対的生成ネットワーク(Generative Adversarial Network; GAN)などを援用しつつ、力場改善の手段を模索することを考え、理論基盤の構築と立式を行っていた。しかしながら、申請から開始までの間の世界的な研究の進展、および研究開始後に得られた知見などから、サンプルに対する学習効率が低く実用性が低いことが予想され、toy example 以上への適用に難があることが想定された。また、研究計画段階に比して、世界的なニューラル力場手法の開拓が急速に進行したことで内挿を目的とする力場は（高コストではあるものの）ある程度カバーされるようになった。これら知見の蓄積と急速な状況の

変化を鑑み、研究目的を振り返った場合、真に求められるのはパラメータだけの決定ではなく、外挿に頑健で記述力のある低コストな関数系の構築であると考えられるようになった。このことから、当初の研究計画からピボットし、低コストで外挿能力を有する関数の学習に向けて軌道修正を行うこととした。ニューラル力場により内挿・高コストの部分が強くカバーされるようになったため、これらの手法と相補的な部分の研究を進めることで研究の目的である高精度な力場の自動決定を目指すことができると考えられた。このため、近年急速にその能力を増している記号回帰の手法に着目した。記号回帰は関数の入力値と出力値から、構文木として表記される関数を推定する手法である。シミュレーションに対する記号回帰の応用を目指した研究を行った。

４．研究成果

単純化のために、本研究では最も単純な loss として force matching error を用い、高レベル（量子化学計算等）の計算から得られた原子に掛かる力を再現する関数系を構築するような記号回帰アルゴリズムとソフトウェアのプロトタイプを作成した。ソフトウェアの概略を図 1 に示す。本アルゴリズムは最低限に単純化した系を用いており、原子に掛かる力から、2 体間ポテンシャル関数$U(r|\theta)$を導出するよう構成されている。ここでrは二体間距離、θは関数の持つ(関数系により、複数あることもある)パラメータである。ポテンシャル関数には種々の制約が実用的に必要であることが多い。たとえば、カットオフ半径で力が 0 となる制約は、高速な分子間相互作用の計算には必要不可欠である。このような制約の処理には augmented Lagrangian method が有効であった。そのほか、分子シミュレーションの回帰のために特化したプログラムを作成し、探索が可能かどうかをテストした。

実際に本アルゴリズムを用いて、H20-13 データセット[1]に対する回帰を行った例を図 2 に示す。このデータセットは水 2 分子からなり、水の座標分布に加え、量子化学計算により得られた高精度なエネルギーおよび力の情報を含んでいる。この系では分子内の相互作用を除いた、分子間相互作用の関数系を探索するように設定し、関数の単純さと loss のどちらかに優れる(パレート最適)関数を優先して探索する形式で回帰を行った。原子間相互作用関数にはカットオフ半径で力が 0 となる制約を入れた。探索の結果、1/r 型の最もシンプルな相互作用の他、loss の低いより複雑な相互作用関数を提案することができた。これらの相互作用関数は q-spcfw モデルなどの古典的非剛体水モデルより小さな loss となった（尚、元々古典的な水モデルでは水素に LJ 相互作用を割り付けないのが通例のため、この結果の解釈には注意する必要がある)。すべての原子間ポテンシャルは遠距離で 0 に向けて減衰しており、結果的に関数の外挿が実現されている。

さらに、本アルゴリズムの陰溶媒モデルへの応用を試すため、水中 alanine dipeptide が溶媒からの受ける力を真空中 alanine dipeptide で近似する問題設定での回帰を試みた。しかしながら、データセットのバリエーションが探索空間に比べて乏しいことから、制約を満たすだけの関数をアルゴリズムが出力してしまい、充分に意味のある陰溶媒力場の関数を出力するに至らなかった。

今後の展望として、アルゴリズ

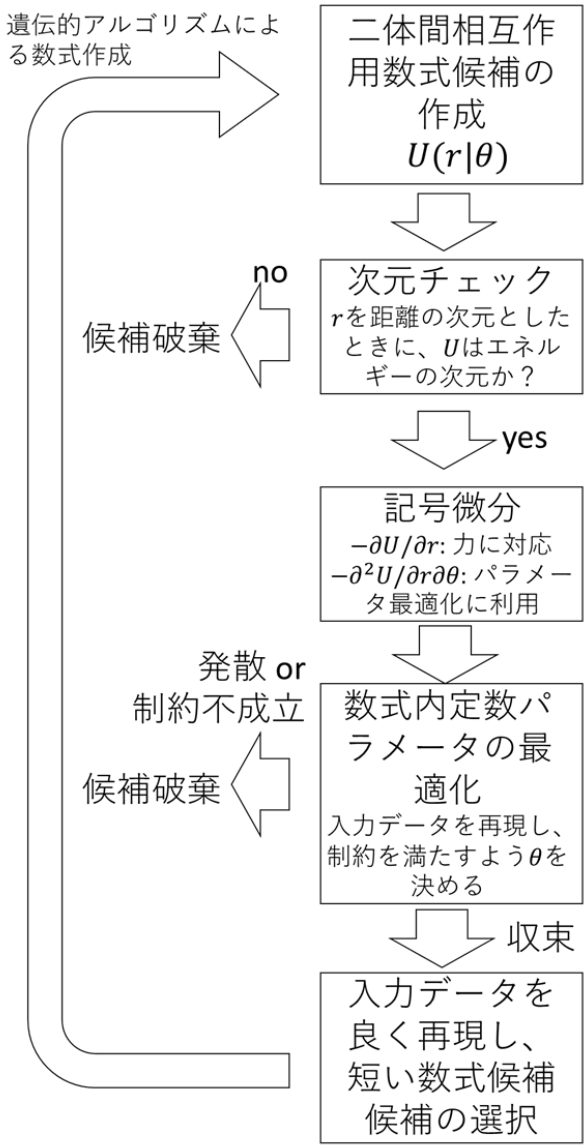

図 1: 作成したプロトタイプの単純化した実行フロー。入力されたデータセットを最もよく表す関数系を探索する。

ムの改良およびデータセットの拡充が望まれる。特に、構文木の生成アルゴリズムは改良の余地が大きく、近年急速に発展した Monte Carlo tree search を基調としたアルゴリズム群や言語モデルベースの手法の検討を行いたい。データセットは古典・量子ともにバリエーションと質が重要である。既存のデータセットだけでなく目的に即したデータセットの拡充を行っていく必要がある。

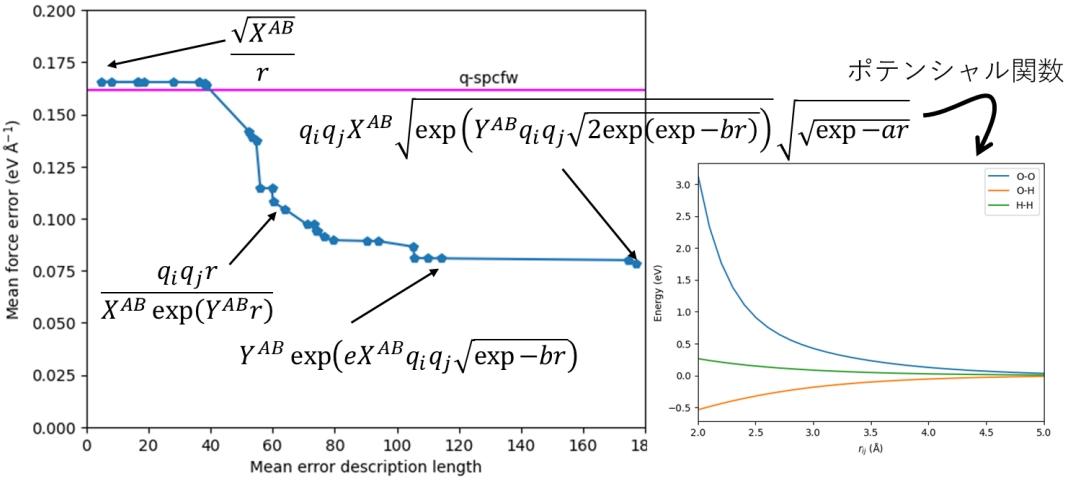

図 2: H2O-13 データセットで発見されたポテンシャル関数の例。左図の横軸は関数の複雑さを表し、縦軸は関数のデータセットからのずれを表す。マゼンタの横線は q-spcfw モデルでの分子間力から計算した loss である。X^{AB}やY^{AB}は$X^{AB} = X^A X^B$などと積の形で表せるような原子タイプに依存した定数である($A, B \in \{\mathrm{H}, \mathrm{O}\}$)。右図は最も loss が低かった関数候補で得られたポテンシャル関数を、原子タイプ別にプロットしたものである。

<引用文献>
[1] Albert P. Bartók, Michael J. Gillan, Frederick R. Manby, Gábor Csányi, Physical Review B 88(5): 054104, 2013.

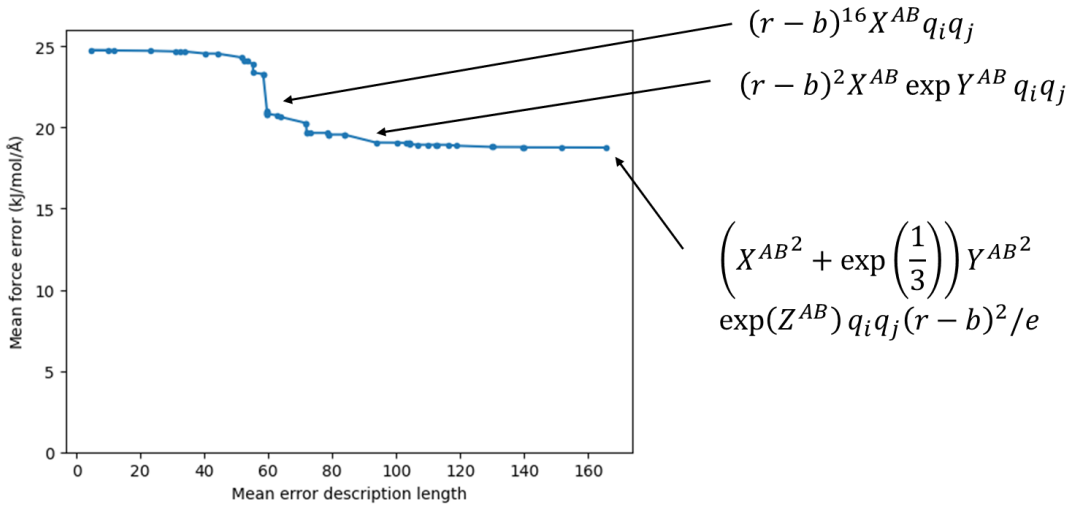

図 3: Alanine dipeptide データセットに対して回帰されたポテンシャル関数。表示されている 3 つの関数は代表例であるが、いずれも最適化後のbの値がほぼ 12Å となっており、カットオフ距離での力を 0 とする制約を満たすために overfit した関数系が出力されていることが推定される。

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計5件（うち招待講演　1件／うち国際学会　0件）

１．発表者名 櫻庭俊
２．発表標題 記号回帰による量子相互作用模倣
３．学会等名 量子生命科学会第4回大会
４．発表年 2022年

１．発表者名 櫻庭俊
２．発表標題 記号回帰による「都合の良い」ポテンシャル関数の探索
３．学会等名 第36回　分子シミュレーション討論会
４．発表年 2022年

１．発表者名 櫻庭俊
２．発表標題 AlphaFoldと分子シミュレーションの力学
３．学会等名 第22回 蛋白質科学会年会（招待講演）
４．発表年 2022年

１．発表者名 櫻庭俊
２．発表標題 記号回帰による軽量な相互作用関数の探索
３．学会等名 第37回 分子シミュレーション討論会
４．発表年 2023年

１．発表者名 櫻庭俊
２．発表標題 多腕banditによるタンパク質安定化提案の成功と失敗
３．学会等名 CPS研究会2023年勉強会
４．発表年 2023年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関