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　５ 年　 ６ 月 １６ 日現在

機関番号：１２１０２
研究種目：基盤研究(C)（一般）
研究期間：2019～2022
課題番号：１９Ｋ０５３８１

研究課題名（和文）機械学習を用いた溶媒和モデルの精度向上に関する研究

研究課題名（英文）A theoretical study for the improvement of solvation model by machine learning

研究代表者
松井　亨（Matsui, Toru）

筑波大学・数理物質系・准教授

研究者番号：７０７１６０７６

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：本研究で、我々は量子化学計算と溶媒和モデルとの組み合わせ用いて、次の3点を改善した。(1) 機械学習を使用して溶媒和エネルギーの補正を行い、より正確な分配係数(logP)の算出を試みた。(2) 酸化電位を解析し、Lasso回帰を使用して誤差の重要な因子を特定した。114種の有機化合物の実験値と計算値を比較し、機械学習による解析を行った。(3) 酸解離定数の計算においては、従来は官能基が同じ化合物で計算し線形近似を行っていたが、その理由や化合物の選択に不自然な点があったため、多重回帰を用いて酸解離定数の導出を提案し、より正確な結果を得ることを試みた。

研究成果の学術的意義や社会的意義
本研究を通して、量子化学と情報科学・データサイエンスとが融合する形態が第3ステージの量子化学になると考えて「量子化学3.0の時代」という造語を提唱するに至る段階になったと考えている。　「機械学習」や「人工知能」「自動化」など情報科学分野の進展は目覚ましい。それに付随して、機械学習・深層学習が多くの分野で普及が進んでいる昨今ではデータベース化がより進行している。したがって、今後の量子化学ではコンピュータによる自動的なデータ収集などが主流になると予測できるが、アウトプットとゴール(大抵の場合は実験値)との「差」をどう解釈するかは今後も課題であり続けるだろう。

研究成果の概要（英文）：In this study, we improved the following three aspects by combining quantum chemical calculations with a solvation model:
(1) We obtained more accurate partition coefficients (logP) by using machine learning to correct solvation energy. (2) We analyzed oxidation potentials and identified significant error factors using Lasso regression. We compared experimental and calculated values for 114 organic compounds and performed an analysis using machine learning. (3) In the calculation of acid dissociation constants, conventional methods involved linear approximations based on calculations for compounds with the same functional groups. However, there were unnatural aspects in the reasoning and compound selection. Therefore, we proposed the derivation of acid dissociation constants using multiple regression to obtain more accurate results.

研究分野：計算化学

キーワード：溶媒和モデル　機械学習　酸解離定数　酸化還元電位

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

近年、大量の計算データを処理して量子化学計算の結果を情報の一部として解釈するケモインフォマティクスの手段とする研究が増えている。「機械学習」や「人工知能」「自動化」など情報科学分野の進展は目覚ましい。それに付随して、機械学習・深層学習が多くの分野で普及が進んでいる昨今ではデータベース化がより進行している。元来、計算は情報科学との親和性が高いはずであり、実際に自動化により化合物ライブラリを用いた計算結果のライブラリなどが公開されているなど大量計算・大量データ化の波は量子化学の分野にも到達している。したがって，今後の量子化学ではコンピュータによる自動的なデータ収集などが主流になると予測できるが、アウトプットとゴール(大抵の場合は実験値)との「差」をどう解釈するかは今後も課題であり続けるだろう。その「差」についての新たな解釈・指針を「機械学習」が与えてくれる、と考えた.

その中で，溶媒和モデルの発展は、単純な手続きで溶媒に関係する多くの物理量を算出可能となったことによる。これまで我々は酸解離定数や酸化還元電位など多くの物理量を溶媒和モデルで記述できるための計算スキームを提唱してきたが、根元にあるのはモデルが不正確であるが故のパラメータを含んだ補正項導入であった。恣意的とも取れるパラメータで問題を解決している現状を打開すべく「機械的に」これらの補正を入れて理論や計算制度を向上させたいという考えがあって本研究課題を進めた。

２．研究の目的

溶媒和モデルによる計算から生じる誤差の要因を系統的に理解するために、対象とする分子の特徴量と誤差の相関を「機械学習」により把握して、化合物の溶媒内における物理量 (今回は pK_a, 酸化還元電位, logP)の算出を改善することを目的とした。

３．研究の方法

具体的な手法として、溶媒和モデルを用いた量子化学計算の結果に対してガウス過程モデルや重回帰分析 (主成分解析)を行って、誤差を系統的に見積もり、回帰式を導出することによって、より正確な物理量を算出できるシステムを構築した。その詳細と結果は次項にて述べる。

４．研究成果

(1) 分配係数(logP)の機械学習的な導出

物質の溶解度・輸送に関係する物性値である分配係数（logP）は薬品の体内への輸送や、化学物質の環境評価のために必要な物理化学的性質である。logP を求めるには、実験を行うよりも計算によるスクリーニングを行なう方がより効率的であることが多い。

logP を計算化学的に求めるためには、オクタノールと水中での溶媒和エネルギーΔG(oct), ΔG(wat)の差が必要となり、量子化学計算による算出も可能と考えられる。一方で、分子動力学の場合はインプットの作成に時間を要してしまい、自由エネルギー計算に多くの時間を要する。溶媒和エネルギーを量子化学計算で予測するには、簡易な計算にするため分子の周辺に空洞を作り、その周辺がε (溶媒の誘電率)の誘電体で埋める溶媒和モデルを用いる。代表的な溶媒和モデルである conductor-like polarizable continuum model (C-PCM)は、この計算法から、大きな分子や不揃いな形をした分子にも適用できるといったメリットがある。しかし、C-PCM では、cavity が仮想的で特定分子間の相互作用を考慮することができないことから、溶媒和エネルギーがεの値にのみ依存 (電子の溶媒和エネルギーが 1/ε と線形に相関)してしまう。そのため、どの分子

を計算しても logP がほぼ一定に算出され精度・コストの両方で悪い結果をもたらす。そこで、まずは C-PCM の精度向上のため、溶媒和エネルギーについての誤差をそれぞれの化合物の性質と溶媒の種類から求める必要がある。そこで、これらの誤差を機械学習の手法により見積もることで、溶媒和エネルギーを補正することでより正確な溶媒和エネルギーの算出を試みることを本研究の目的とした。

実験値は、水溶媒での自由エネルギーの実験値（263 種類の化合物）n-octanol 溶媒での自由エネルギーの実験値（205 種類の化合物）のデータベースから取り出した。MP2/6-31++G(d,p)のレベルに焦点を絞る。全ての計算は Gaussian16 を用いて実行した。また、実験値の自由エネルギーと計算値の自由エネルギーの差をとって計算値の誤差とした。

本研究で用いた溶媒和モデル C-PCM と他の溶媒和モデルである SMD の水溶媒での溶媒和エネルギーの実験値と計算値の MP2 と B3LYP でそれぞれもとめ、その誤差の相関をとったものである。それぞれのグラフから C-PCM が他のモデルに比べて計算手法に依存しないことが分かる。これは、「C-PCM の場合では、計算手法に依存しない何かが溶媒和エネルギーの誤差を与えている」と解釈できる。SMD の場合は、元々M05 などの密度汎関数で良いパフォーマンスを出すために開発されてパラメータを駆使した手法であることから、計算手法を変更すると予期せぬエラーが生じる可能性があるとも言える。誤差の原因がなるべく化合物の部分構造から機械学習で解明できるように今回は計算手法に依存しにくい C-PCM を補正するに至った。

フィンガープリントを用いた回帰の結果、水溶媒では主に(A)NH2 基を持つ (B) C=O の結合を持つ、 n-octanol 溶媒では主に(C) ベンゼン環を持つ (D) 主鎖が 4 以上の炭素を持つ のような特徴を持つ分子において大きな溶媒和エネルギーの誤差が生じることが分かった。これは、具体的には C-PCM で欠落している溶媒と溶質の間に生じている相互作用に起因するものと考えられる。この誤差を補正した自由エネルギーをそれぞれ代入すると補正前より実験値に近い logP が得られた。決定係数についても 0.86 となり傾きも 1 となりほぼ実験値を再現できた。この補正式を用いると、これらのデータセットにない薬品の logP について平均絶対誤差が 0.5 に収まるような精度で算出することができた。また, これに関連して, logP を多重解析によって求める方法も考案された.

(2) 機械学習的手法を用いた酸化電位の算出値補正

量子化学計算において溶媒効果を考慮する際に用いられる代表的な溶媒和モデルの一つに conductor-like polarizable continuum model (C-PCM) が存在する。C-PCM は非常に低コストなモデルだが、特に荷電溶質を扱う際に誤差が大きくなることが知られている。そこで我々はその計算誤差の特性を解析することで、誤差に対してどのような因子が重要であるのか考察を行った。ここでは、荷電溶質が関与する現象として酸化電位に注目した。解析手法として機械学習の回帰手法の一つであり、重要な説明変数の絞り込みを可能にする Lasso 回帰を採用した。

アセトニトリル溶液下での有機化合物 114 種の酸化電位の実験値を文献[2]より取得し、酸化ギブズエネルギー$\Delta G_{\mathrm{ox}}^{\mathrm{expt}}$を計算した。対応する計算値$\Delta G_{\mathrm{ox}}^{\mathrm{calc}}$を Gaussian16 rev. A.03 を用いて算出した。基底関数や汎関数等による誤差を抑えるために、G3B3/C-PCM レベルによる自由エネルギー計算を実行した。ここから、計算誤差$\delta = \Delta G_{\mathrm{ox}}^{\mathrm{expt}} - \Delta G_{\mathrm{ox}}^{\mathrm{calc}}$を計算した。

以上で得られたデータを基に機械学習による解析を行った。まず、溶質の構造を数学的に扱えるようにベクトルに変換した。即ち、溶質に含まれる 44 種の部分構造の数を数え上げ、その並びをベクトルとした (counting substructure fingerprint, CSFP)。このベクトルを説明変数とし、誤差δを目的変数として Lasso 回帰を行った。また、5 分割交差検証によって得られた回帰式の評価を行った：4/5 のデータで回帰し、残りの 1/5 のデータを用いて予測・評価をすることを 5 回繰り返した。

比較のために UB3LYP/6-31+G(d, p)/C-PCM レベルの構造最適化及び調和振動子計算と CBS-QB3/C-PCM レベルの自由エネルギー計算による結果も併せて記載した。2 つの近似高精度手法はどちらも大きく過大評価しているのに対して、B3LYP は僅かに過小評価しているのが読み取れる。これは C-PCM による荷電溶質の溶媒和エネルギーの過大評価と、B3LYP の自己相互作用誤差に起因するイオン化エネルギーの過小評価による結果と考えられる。

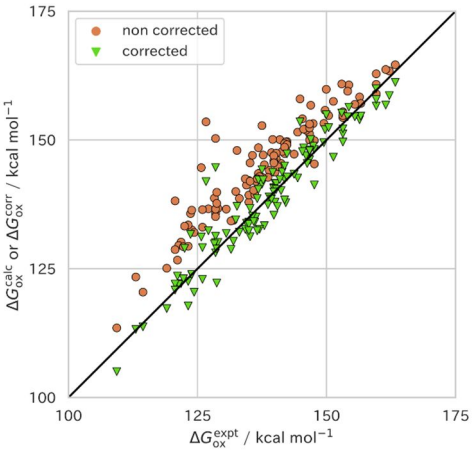

図 1. Plot for $\boldsymbol{\Delta G_{ox}}$: experimental value (horizontal axis) and calculated and correction value (vertical axis).

Lasso 回帰によって補正を行った結果を右図に示す。この補正によりプロットが$y = x$の直線上にフィットする傾向が読み取れる。元の誤差が MAE で 6.99 kcal/mol であったのに対し、補正後は 2.57 kcal/mol となったことから 37%と大幅に誤差を削減することに成功している。また、回帰係数について解釈を行ったところ、主に分子サイズに比例して G3B3/C-PCM の系統誤差が生じていたことが分かった。

(3) 酸解離定数における補正の提案

(a) 多重回帰を用いた酸解離定数の導出

量子化学計算と誘電体モデルを併用した酸解離定数は、これまで官能基が同じものでまとめて計算し線形近似を行うこと、またその線形近似に必要な化合物を選択することで 0.3 pK_a unit 程度の誤差に収まる精度を得ることができていた。しかしながら、その線形近似をする理由が不明であることに加えて，化合物の選択に不自然な点が残ること、特殊な官能基を持つ場合などに対応できない、など多くの問題がそのままにされている。そこでこの研究では、約 420 種類の化合物の酸解離定数から、より一般的な酸解離定数計算を求めるために線形近似だけではなく、より多くの変数を取り入れる新しい手法を構築することを目的とした。

結果と誘電体モデルの結果から、プロトン脱理反応前後のエネルギー・局所的な溶媒和エネルギー・官能基周辺の電荷・HOMO の軌道エネルギー・プロトンとドナーの結合距離などの観点から 10 種類の変数を導入して多重解析を行うことによって、より一般的な酸解離定数の算出が可能となることがわかった。また、この多重解析の結果を用いることで薬剤の酸解離定数を 0.5 pK_a unit 以内の精度で求めることも可能となった。一方で、アミンを含む系については 1.0 pK_a unit 程度の誤差が出るケースが残ってしまい、改善の余地を残す結果となった。

プロトンのギブスエネルギーやスケーリングファクターを一切導入しない方法を試したところ、線形近似を導入する結果よりも相関係数が 0.97 から 0.89 程度と小さくなった。そのため、線形近似を導入する必要性があることが確認された。

(b) 励起状態への応用

nPUA はアニオンセンサーとしての働きを期待されているウレア誘導体分子の一つである。nPUA は二か所の N-H 結合箇所を介して酢酸イオンと水素結合を形成する性質を持つ。またこの会合体を光照射すると、励起状態においてウレア誘導体に結合しているプロトンが水素結合を介して酢酸イオン側に移動する分子間プロトン移動 (Excited State Intramolecular Proton Transfer, 以下 ESPT と略す)を起こすことが知られている。ESPT によって生じた会合体構造 T*(Tautomer に由来) は特異な長波長蛍光を発するため、これを検出することで nPUA による酢酸イオンの識別が可能になる。また、ESPT 反応をウレア誘導体のセンシングに応用することで、選択性という観点からより優れたアニオンの識別が可能になり、有用なアニオンセンサー分子の開発が期待されている。しかし ESPT のタイムスケールは非常に速く、この反応各段階における会合体の構造、電子状態などの詳細が明らかになっていない。また、ウレア誘導体の ESPT に関する報告例は非常に限られていることなどから、ESPT の反応性がどのような要因によって制御されているのか不明である点が多い。本研究ではこれを踏まえ、① T*状態の会合体構造および電子状態、②ESPT の反応速度定数 k_{PT} に影響を与える要因、以上二つの観点に注目した ESPT の考察を行った。本会合体には nPUA-酢酸イオン間の水素結合を形成しているプロトンが二つ存在する。これらは nPUA のウレア部分の①アントラセニル基側、②フェニル基側、これらの N-H 結合箇所に存在しており、本会合体の ESPT はプロトンの移動の仕方によって①アントラセニル基側のプロトンの ESPT (T_a*)、②フェニル基側プロトンの ESPT (T_p*)の二つの構造がありうる。そのため本研究では、各 nPUA 会合体 T*構造として T_a*および T_p*をそれぞれ構造最適化計算により求めた。構造最適化計算は DFT,TD-DFT/B3LYP/6-31+G (d,p)のレベルで行った。

また ESPT の反応速度定数 k_{PT} に影響を与える要因として、本計算では励起状態における nPUA の酸性度 pK_a*は pK_a と酸解離前後のギブズエネルギー変化ΔG の関係を用いた線形近似により求めた。この線形近似を作成するための見積もりに nPUA と同様の化学種の pK_a を用いた。

各 nPUA のより安定な T*構造は 1, 2PUA が T_a*、9PUA が T_p*であった。また、これらの T*構造の安定性をエネルギーから計算した結果、9PUA,1PUA,1PUA の順序で T*構造が安定である結果が得られた。また pK_a *計算では 9PUA>1PUA>2PUA の順序で酸性度が大きいこと、活性化エネルギーの見積もりにおいても先ほどの順序で低い活性化エネルギーを取っている結果が得られた。以上の結果は 9PUA,1PUA,2PUA の順序で ESPT の反応性が大きいことを示唆しており、実測値の ESPT 速度定数はこれらの結果と一致している。

以上のように、本課題のような機械学習的な手法と量子化学計算を組み合わせた手法が有効であることを示すことができ、様々な応用計算も可能となってきた. その一方で、溶媒和エネルギーの「電荷に対する依存性」の問題はまだ解決しておらず、本研究課題の最終目標であった金属錯体などのレドックス(還元電位を含む)については完成に至らなかったため、引き続き研究の課題として検討する。

５．主な発表論文等

〔雑誌論文〕　計6件（うち査読付論文　6件／うち国際共著　0件／うちオープンアクセス　2件）

１．著者名 Takahashi Teruyuki、Matsui Toru、Hengphasatporn Kowit、Shigeta Yasuteru	４．巻 94
２．論文標題 A Practical Prediction of LogPo/w through Semiempirical Electronic Structure Calculations with Dielectric Continuum Model	５．発行年 2021年
３．雑誌名 Bulletin of the Chemical Society of Japan	６．最初と最後の頁 1807～1814
掲載論文のDOI（デジタルオブジェクト識別子） 10.1246/bcsj.20210035	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Fujiki Ryo、Matsui Toru、Shigeta Yasuteru、Nakano Haruyuki、Yoshida Norio	４．巻 4
２．論文標題 Recent Developments of Computational Methods for pKa Prediction Based on Electronic Structure Theory with Solvation Models	５．発行年 2021年
３．雑誌名 J	６．最初と最後の頁 849～864
掲載論文のDOI（デジタルオブジェクト識別子） 10.3390/j4040058	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Izawa Hironori、Yasufuku Fumika、Nokami Toshiki、Ifuku Shinsuke、Saimoto Hiroyuki、Matsui Toru、Morihashi Kenji、Sumita Masato	４．巻 6
２．論文標題 Unique Photophysical Properties of 1,8-Naphthalimide Derivatives: Generation of Semi-stable Radical Anion Species by Photo-Induced Electron Transfer from a Carboxy Group	５．発行年 2021年
３．雑誌名 ACS Omega	６．最初と最後の頁 13456～13465
掲載論文のDOI（デジタルオブジェクト識別子） 10.1021/acsomega.1c01685	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Hengphasatporn Kowit、Matsui Toru、Shigeta Yasuteru	４．巻 49
２．論文標題 Estimation of Acid Dissociation Constants (pKa) of N-Containing Heterocycles in DMSO and Transferability of Gibbs Free Energy in Different Solvent Conditions	５．発行年 2020年
３．雑誌名 Chemistry Letters	６．最初と最後の頁 307～310
掲載論文のDOI（デジタルオブジェクト識別子） 10.1246/cl.190946	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Kanamaru Yuki、Matsui Toru	4．巻 43
2．論文標題 Factor analysis of error in oxidation potential calculation: A machine learning study	5．発行年 2022年
3．雑誌名 Journal of Computational Chemistry	6．最初と最後の頁 1504～1512
掲載論文のDOI（デジタルオブジェクト識別子） 10.1002/jcc.26953	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Onozawa Shu、Nishimura Yoshinobu、Matsui Toru	4．巻 96
2．論文標題 A Theoretical Study on Rate Constants of Excited State Proton Transfer Reaction in Anthracene-Urea Derivatives	5．発行年 2023年
3．雑誌名 Bulletin of the Chemical Society of Japan	6．最初と最後の頁 215～222
掲載論文のDOI（デジタルオブジェクト識別子） 10.1246/bcsj.20220332	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計8件（うち招待講演　1件／うち国際学会　2件）

1．発表者名 金丸 雄基，松井 亨
2．発表標題 酸化還元電位算出における誤差の機械学習を用いた要因解析
3．学会等名 第15回分子科学討論会
4．発表年 2021年

1．発表者名 金丸 雄基，松井 亨
2．発表標題 酸化還元電位算出における誤差の機械学習を用いた要因解析
3．学会等名 日本化学会第102春季年会
4．発表年 2022年

１．発表者名
大﨑 象平，松井 亨

２．発表標題
FIVプロテアーゼとHIV-1プロテアーゼ阻害剤の相互作用解析

３．学会等名
日本化学会第102春季年会

４．発表年
2022年

１．発表者名
登坂　夏名、尾﨑　大和、松井　亨

２．発表標題
機械学習による溶媒和モデルの半経験的改善法

３．学会等名
日本化学会第100春季年会

４．発表年
2020年

１．発表者名
尾﨑　大和、藤田　健宏、松井　亨、寺山　慧、隅田　真人、守橋　健二

２．発表標題
長距離補正密度汎関数による領域分割パラメータの簡便な決定法

３．学会等名
第13回分子科学討論会

４．発表年
2019年

１．発表者名
Shohei Osaki, Toru Matsui

２．発表標題
Interaction analysis of FIV and HIV-1 protease inhibitor using the FMO method

３．学会等名
The 10th Asian Pacific Association of Theoretical and Computational Chemistry（国際学会）

４．発表年
2023年

１．発表者名 Yuki Kanamaru, Toru Matsui
２．発表標題 Machine Learning Assisted DFT Calculation Using Solvation Model
３．学会等名 The 10th Asian Pacific Association of Theoretical and Computational Chemistry（招待講演）（国際学会）
４．発表年 2023年

１．発表者名 原田　秦冴、松井　亨
２．発表標題 長距離補正密度汎関数理論を用いた有機薄膜太陽電池材料となる高分子の軌道準位の算出
３．学会等名 日本化学会第103春季年会
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