ハイスループット計算と機械学習を活用した光機能性有機材料の新規構造探索

Searching for Novel Photo-Functional Organic Materials using High-Throughput Calculation and Machine Learning

大越孝洋 Takahiro Ohgoe
大越昌樹 Masaki Okoshi
長尾宣明 Nobuaki Nagao
四橋聡史 Satoshi Yotsuhashi

要　旨

筆者らはデジタライゼーションによる新材料開発の高速化を目的として，量子化学計算と機械学習の融合による光機能性有機分子の探索スキームを構築した. 本探索スキームは，(1) 多数の候補分子に対して網羅的に量子化学計算を行うハイスループット計算と，(2) モンテカルロ木探索とリカレント・ニューラルネットワークを用いて広大な材料空間のなかから分子探索を行う人工知能からなる. 特に後者では, ハイスループット計算で得られたデータを元に機械学習による特性予測モデルを構築し, これを利用することで分子生成ループを高速化した. さらに, 本スキームを非線形光学材料探索において実践し、計算機上での分子の生成・評価に基づく設計プロセスを約100倍高速化することに成功した．本汎用スキームを活用することでさまざまな機能性有機材料の探索を加速することが期待できる.

Abstract

To accelerate new material development, we propose a scheme for searching for photo-functional organic molecules by combining quantum chemistry calculations and machine learning methods. This scheme consists of two parts: (1) high-throughput calculation for a large number of molecules and (2) de novo molecular generation based on the Monte Carlo tree search and recurrent neural networks. In the latter, we employed a machine-learning model for predicting properties, which was constructed from the numerical data obtained in the high-throughput calculation. We applied the scheme to a problem on searching for non-linear optical materials and accelerated the computational molecular design about 100 times faster than before. We expect that the proposed scheme can accelerate the development of functional materials.

1. はじめに

創薬, 材料分野において, 所望の特性を示す構造設計は，従来科学者の知識，経験と勘に基づいて行われてきた．このようなアプローチは通常，実験のトライアルアンドエラーを通して構造を最適化することが必要であり，実験には時間，コストを要するため，限られた実験回数のなかで最適な構造を発見することが求められる.

このようなフィジカル空間（現実空間）での構造設計に対し，近年の計算機と計算科学，機械学習分野の発展に伴い，サイバー空間（計算機上の仮想空間）あるいはフィジカル空間と融合したサイバー・フィジカル空間において高速かつ低コストに構造探索を行うアプローチが急速に進展してきている．特に分子構造の探索を行う創薬や機能性有機材料といった分野では，無機材料における結晶構造探索と比べて構造生成が容易であることから，応用が広がりつつある.

計算科学を用いたサイバー空間での構造探索スキームとしては，ハイスループット計算による材料スクリーニングが知られている．本スキームではまず，分子構造の候補を大量に発生させた後，量子化学計算によって網羅計算を行うことで所望の特性の分子構造を探し出す．本アプローチの課題は，材料空間があらかじめ生成した範囲に限られる点である．有機分子全体の数，すなわち材料空間は少なくとも10^{60}個とも言われており，所望の特性の分子が設定した材料空間内にあるとは限らない.

一方で，近年深層学習を活用し，広大な材料空間の中から新規分子構造を探索する人工知能（artificial intelligence: AI）手法がいくつか提案されている．そのなかでも東京大学の津田教授らが開発したモンテカルロ木探索（Monte Carlo tree search: MCTS）とリカレント・ニューラルネットワーク（recurrent neural network: RNN）を組み合わせた手法であるChemTS [1][2]は，変分オートエンコーダ（variational autoencoder: VAE）[3]やgraphical VAE（GVAE）といった他の探索手法に比べ, 単位時間当たりの分子生成数が1桁以上大きいことが報告されている．一方で，実応用で求められる特性の計算のためには多くの場合，量子化学計算を行う必要があり, 膨大な数の分子を生成しながら探索を行うには, 大きな計算コストを要することが問題となる.

そこで今回，ハイスループット計算と自動分子設計を行うAIであるChemTSを活用し，サイバー空間における光機能性有機分子の高速な構造探索スキームを構築した．本スキームでは, ChemTSのボトルネックとなる特性評価部において，ハイスループット計算で得られたデータを用いて構築した機械学習による予測モデルを活用することにより高速化を実現した．また，生成分子のうち，一部の新規分子

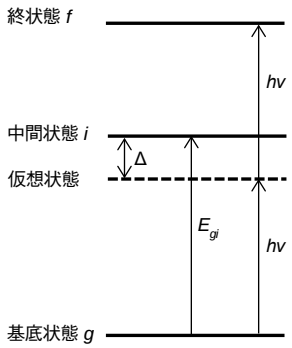
については量子化学計算を行い，データを増やすことで，予測モデルの精度を担保する仕組みを取り入れた．以下では，構築した本スキームの詳細と，非線形光学材料である二光子吸収材料の探索において実践した結果について報告する．

2. ハイスループット計算

非線形吸収材料である二光子吸収材料をターゲットとし，約4700件の候補分子に対して網羅的に量子化学計算を実施した．計算対象とした候補分子は，フラグメントと呼ばれる分子の構成要素を複数用意し，それらの組み合わせによって生成した．

2.1 二光子吸収の原理と計算方法

二光子吸収とは，光吸収の際に2個の光子を同時に吸収する現象である [4]．その特徴として，吸収量が光強度の2乗に比例するため，レンズで集光すると焦点付近でのみ吸収を起こすことができ，この空間選択性を生かして生体イメージングなどさまざまな用途に応用されている．**第1図**に二光子吸収の原理の概念図を示す．ここでエネルギー$h\nu$をもつ励起光を吸収する準位は実在しないが，集光したレーザの強い光が照射されると電場によって中間状態iが摂動を受け，ごく短時間ではあるがエネルギー準位が下がる（仮想状態）ことによって励起光の光子が吸収可能となり，2つの光子を同時に吸収しそのエネルギーの和に相当する高エネルギーの最終準位fへ遷移する．



第1図　二光子吸収に関わる状態のエネルギー準位図

Fig. 1　States and energy levels for two-photon absorption

二光子吸収の大きさを表す指標としては，一般に二光子吸収断面積σ_fが用いられ，次の（1）式で与えられる．

$$\sigma_f = \frac{2\pi h\nu^2 L^4}{\varepsilon_0^2 n^2 c^2 \Gamma}\left[\sum_i \frac{\langle \boldsymbol{\mu}_{gi}\cdot\boldsymbol{\mu}_{if}\rangle}{(E_{gi}-h\nu)}\right]^2 \quad \cdots\cdots\cdots\cdots (1)$$

ここで，$\boldsymbol{\mu}_{gi}$は基底準位gから中間準位iへの遷移双極子モーメント，$\boldsymbol{\mu}_{if}$は中間準位iから最終準位fへの遷移双極子モーメント，E_{gi}は基底準位gと中間準位i間のエネルギー差，Γはローレンツ型の線幅関数である．σ_fは通常GM単位（1 GM=10^{-50} cm^{4}s photon^{-1}molecule^{-1}）で示される．

具体的に，二光子吸収断面積の計算は以下のようにして行った．まず，遷移双極子モーメントおよび各準位のエネルギーレベルは量子化学計算によって計算した．量子化学計算ソフトにはGaussian16, 計算手法としては時間依存密度汎関数法を用い，交換相関汎関数と基底関数にはそれぞれtuned-CAM-B3LYP [5]，def2-SVP [6]を用いた．次に，これらの計算結果とGeneralized Few-State Model（GFSM）法 [7]と呼ばれる計算方法に基づき，σ_fの値を算出した．最後に，ターゲットのエネルギー領域に存在する終状態fに関する和をとった．ここでは，窓関数w_fを導入して$\sigma=\Sigma f\ w_f\,\sigma_f$を計算した．今回$w_f$としては，終状態$f$への励起エネルギー$E_{gf}$ [eV]が区間 [5.2, 5.6]で0から1へ滑らかに変化し，区間 [5.6, 6.4]で1，区間 [6.4, 6.8]で1から0へ滑らかに変化し，それ以外の区間では0となる関数を使用した．

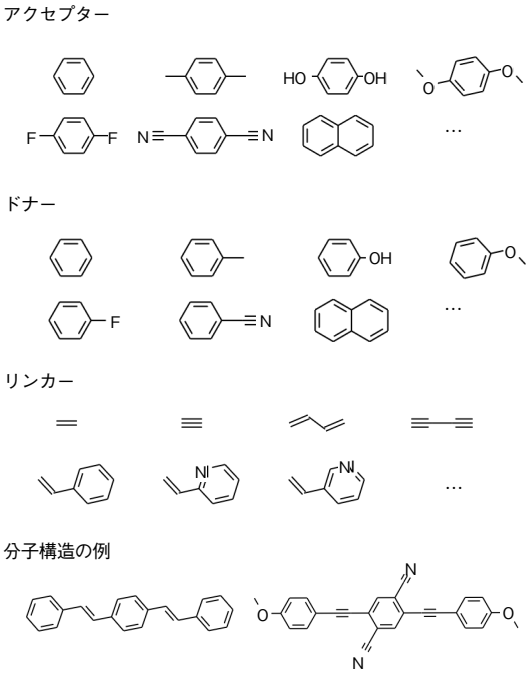
2.2　分子構造生成

有機分子全体（材料空間）の数は少なくとも10^{60}個とも言われており全てを計算し尽くすことは不可能である．そこで二光子吸収が起きる可能性のある分子構造を，フラグメントと呼ばれる複数の小さな構成要素を組み合わせることで大量に発生させ，候補分子とした．

実験事実や理論計算から以下の条件を満たすものが二光子吸収を起こしやすいことが知られている：1. 長いπ共役系を有する分子，2. π共役系の末端にドナー，アクセプターを有する分子，3. 中心対称な分子．具体的には，ドナー（D）とアクセプター（A）をつなぐ部分構造をリンカー（L）とすると，D-L-A-L-D型の分子構造，さらにはアクセプターにリンカー+ドナー部が3つあるいは4つ付いた3分岐構造や4分岐構造などが候補構造となる．

ドナー，アクセプター，リンカー部の分子構造を用意する際には，分子の文字列データ表現であるSMILESを使用した．用意したアクセプター，ドナー，リンカーの化学構造式の例を**第2図**に示す．また，生成される分子構造の例として，D-L-A-L-D型のものを示した．

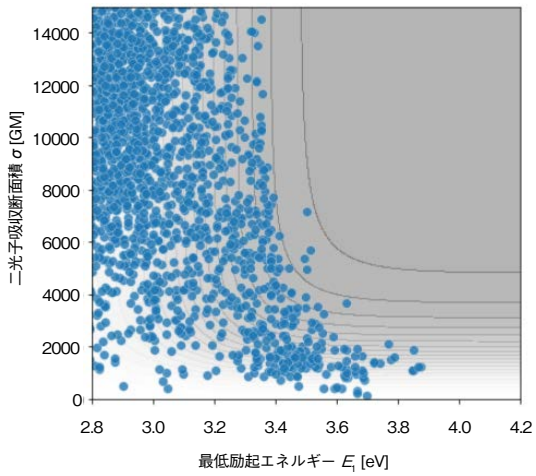
次に，ケモインフォマティクスのライブラリーであるRDKitを活用し，各フラグメントのSMILESを結合することでD-L-A-L-D型などの分子構造のSMILESを作成した．さらに，各SMILESからOpen Babel [8]のpythonラッパーであるpybelを用いて3次元の立体構造とし，その後，計算コストが低い半経験的量子化学計算法であるGFN2-xTB [9]を用いて分子の構造最適化を行った．



第2図　フラグメント構造と生成される分子の例
Fig. 2　Examples of fragments and generated molecules

2.3　ハイスループット計算結果

第3図にD-L-A-L-D型の分子構造約4700件に対する量子化学計算の結果を示す．ここでは，縦軸を二光子吸収断面積σ [GM]に，横軸には最低励起エネルギーE_1 [eV]をとり，得られた計算データの分布を示した．E_1が小さい分子のなかには，二光子吸収断面積が大きいものが多数見つかったが，E_1が大きい分子は二光子吸収断面積が著しく小さくなる様子が見られ，相反する関係にあることがわかる．これは（1）式において，分母にあるデチューニングエネルギー



第3図　ハイスループット計算結果
Fig. 3　High-throughput calculation results

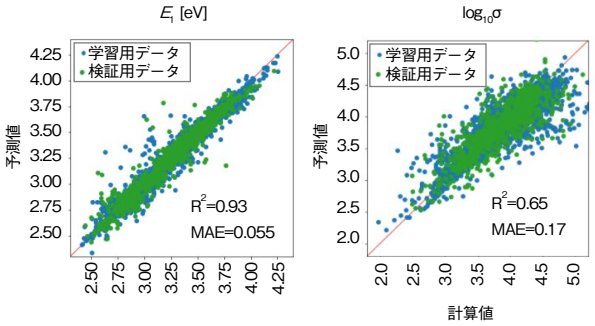
$\Delta = E_{gi} - h\nu$に着目することで理解できる．すなわち，E_1が小さい分子はΔが小さくなる中間準位が存在すれば（1）式よりσが大きくなるが，最低励起エネルギーであるE_1が大きいと**第1図**より全ての中間準位に対してΔが大きくなるため，デチューニングエネルギーによるσの上昇が起きなくなる．非線形吸収材料では，線形吸収に対して非線形吸収が支配的である必要があるため，一光子の励起エネルギーが高く線形吸収が抑制され，かつ二光子吸収が高いような，**第3図**における右上の領域に位置する分子材料が望ましい．**第3図**中の等高線は、4.3節で説明するChemTSによる探索用に使用した報酬値の等高線であり，本節の計算で使用されたものではないが，4.3節の**第7図**との比較のためにここにも示した．報酬値の定義については4.3節を参照のこと．

3. 機械学習による予測モデル構築

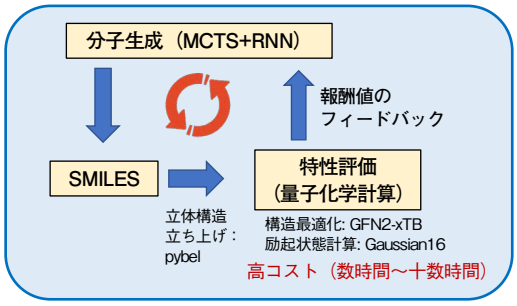
筆者らは，ハイスループット計算で得られたデータを活用し，機械学習による特性の予測モデルを構築した．これにより未計算の分子に対しても特性値の予測が可能になる．機械学習モデルはサポートベクターマシンを使用し，記述子の違いによって以下の3つのモデルを構築した: モデル1. RDKit記述子(1D, 2D)200個を使用，モデル2. GFN2-xTBの基底状態計算で得られた分子軌道エネルギー80個（HOMO-39～LUMO+39）を使用，モデル3. モデル1と2の両方の記述子を使用．モデル1の予測モデルは，数秒で予測値が得られる一方で，モデル2, 3の予測モデルでは，記述子を得るためのGFN2-xTB計算（基底状態計算）には，約1分程度の計算時間を要する（16 CPUコア使用時）．**第1表**に各モデルにおける精度（決定係数R^2と平均絶対誤差MAE）を示した．一般に記述子数が多いモデル3が最も高精度となるが，この例のようにモデル間で精度に大きな差異はない場合では，速度の観点からモデル1が有用である．**第4図**にはモデル1における計算値と予測値の相関プロットを示した．

第1表　各予測モデルの精度
Table 1　Accuracy of prediction models

予測モデル：記述子	R^2 (E_1)	MAE (E_1)	R^2 ($\log_{10}\sigma$)	MAE ($\log_{10}\sigma$)
モデル1：RDKit記述子 200個	0.93	0.055	0.65	0.17
モデル2：分子軌道エネルギー 80個	0.88	0.079	0.61	0.19
モデル3：RDKit記述子 200個 + 分子軌道エネルギー 80個	0.93	0.053	0.68	0.17



第4図　計算値と予測値の相関プロット（モデル1）
Fig. 4　Correlation plots for numerical and predicted values



第5図　MCTS+RNN (ChemTS)による分子生成ループ
Fig. 5　Loop for molecular generation based on MCTS+RNN

4. 計算科学・AI融合による新規分子構造探索

ハイスループット計算で得られたデータをもとに構築した特性予測モデルを用いることで，自動分子設計AIであるChemTSの高速化を行った．さらに，二光子吸収が大きく、かつ最低励起エネルギーが大きい材料を探索するという2特性の同時最適化問題で実践した．

4.1 自動分子設計 AI と実応用における課題

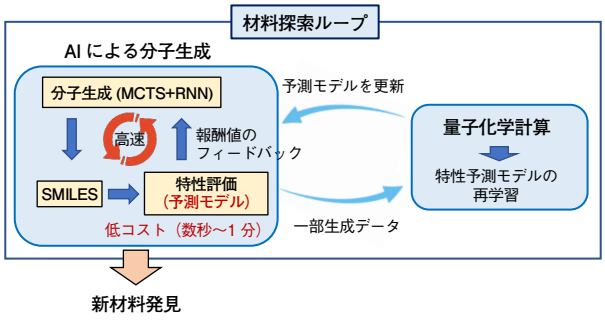
第5図にChemTSで行われる一連の流れの概略を示す．ChemTSにおける分子生成では，MCTSとRNNによってSMILES文字列を生成する．MCTSは文字列の先頭から一定の文字数（深さ）まで文字列パターンの探索を行い，有望な文字列パターンを選択する．不足した文字列は，RNNによって残りの文字列を予測することで補完し，1つの分子に対応した完全なSMILES文字列が作られる．RNNモデルの構築にはSMILESの学習データが必要であり，フラグメントの組み合わせで設計した約2万件の分子を活用した．生成分子に対しては，量子化学計算を用いて特性評価を行う．その後、特性値に応じた報酬値をMCTS部に返すことで結果のフィードバックを行う．このサイクルを繰り返すことによって、MCTS部がより高い報酬値を期待できる文字列パターンを探索していく仕組みになっている．MCTS部が探索する文字列長（深さ）は，サイクルを繰り返すことで長くなり，所望の特性の分子を発見するためには一般に十分にサイクルを繰り返すことが必要である．

しかしながら，ChemTSに限らず自動設計AIを実応用で用いる際の課題として，生成分子の特性評価のために行う量子化学計算が高コストになる点が挙げられる．特に，二光子吸収断面積のような高次の励起状態計算が必要な場合には，並列化を行っても数時間〜十数時間（16 CPUコア使用時）を要し，この部分がボトルネックとなるために十分に探索を進めることが困難となる．

4.2　高速な材料探索システムの構築

4.1節で述べた課題を克服するため，筆者らは**第6図**のような材料探索システムを構築した．**第5図**との違いは，ChemTSにおける特性評価部において高コストであった量子化学計算による励起状態計算の代わりに，機械学習による予測モデルを用いた点である．予測モデルを活用することで，数時間〜十数時間を要した部分が，数秒〜1分程度で二光子吸収断面積を予測値として得ることが可能になった．

ChemTSは，学習データ中の分子との類似性が低いような新規分子も生成する．そのような分子に対しては特性の予測精度は一般に低くなる．この問題に対応するために，一部生成分子に対しては量子化学計算を行い，学習データを増やしたうえで予測モデルの更新を行った．分子の選定には，分子間の類似度の指標であるタニモト係数を用いた．



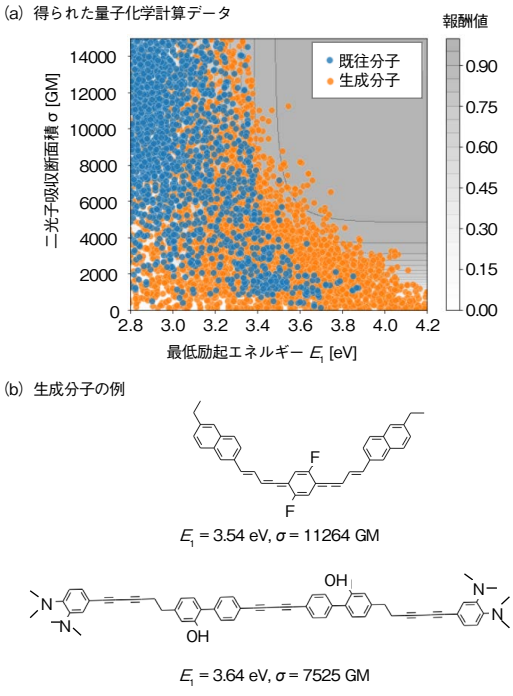
第6図　構築した材料探索ループ
Fig. 6　Proposed loop for material discovery

4.3　構造探索結果

第7図（a）にD-L-A-L-D型の分子構造の探索結果，および得られた量子化学計算データを示す．報酬値については，図中の等高線に示すように二光子吸収断面積と最低励起エネルギーが共に大きい領域で高くなるように設定した．具体的に報酬値rは，$x=E_1$，$y=\log_{10}\sigma$のそれぞれに対してシグモイド関数型の報酬関数 $r_1 = 1/\{1+e^{-a_1(x-b_1)}\}$，$r_2 =$

$1/\{1+e^{-a_2(x-b_2)}\}$を用意し，それらの相乗平均$r=\sqrt{r_1 r_2}$として定義した．また，関数中のパラメータについては所望のエリアで報酬値が最大値1に近くなるように設定した．材料探索ループとしては，**第6図**の全体ループを19回実施した．各回のAIによる分子生成では約20万件を生成し，そのなかの数百件を選定したうえで，量子化学計算を行った後，予測モデルの再学習を行った．その結果，探索前の既往分子（ハイスループット計算時に計算済みの分子）よりも高い報酬値領域に多数の分子が生成された．**第7図　(b)**には報酬値上位の生成分子の例を示した．

これらの結果から，機械学習による分子生成ループと量子化学計算を予測モデルで置き換えた新たな探索システムによって，従来の網羅的な計算では探索できなかった100万件以上の分子構造を探索することが可能であることを示すことができた．



第7図　分子探索結果
Fig. 7　Results of the molecular search

5. まとめ

デジタライゼーションによって新材料開発を加速することを目指し，ハイスループット計算と機械学習の融合によって光機能性有機分子を探索するスキームを構築した．本スキームの構築以前は，ハイスループット計算のみによる分子の特性評価を行っていたが，励起状態計算が高コストなために検証が行える分子は高々数万件であった．一方，本スキームでは自動設計AIによる有望分子の提案と，ハイスループット計算データを学習データとして構築した特性予測モデルによる高速な検証システムを導入することにより，同程度の開発期間において数百万件の分子を評価することが可能になった．言い換えれば，サイバー空間上での分子の生成･評価に基づく設計プロセスを約100倍高速化したと言える．そして，本スキームを実践した結果，ハイスループット計算では見つけられなかった報酬値0.95を超える高報酬値の分子を29件見つけることにも成功した．

一方で今回構築した枠組みでは，合成が困難な分子もAIによって多数生成される．このため，生成された分子については，逆合成解析などによる合成可能性の検討が必要になる．最近ではAIによって合成経路を予測する研究も行われており，これらと今回の材料探索システムを組み合わせて使うことで，高特性かつ合成可能な分子を高速に提案できるシステムを構築できると期待している．

最後に，本研究成果は（一財）高度情報科学技術研究機構の河東田道夫主任研究員と青野信治研究員との共同研究によって得られた．

参 考 文 献

[1] X. Yang et al., “ChemTS: an efficient python library for de novo molecular generation,” Sci. Technol. Adv. Mater. vol. 18, no. 1, pp. 972-976, 2017.
[2] M. Sumita et al., “Hunting for Organic Molecules with Artificial Intelligence: Optimized for Desired Excitation Energies,” ACS Cent. Sci. vol. 4, no. 9, pp. 1126-1133, 2018.
[3] R. Gomez-Bombarelli et al., “Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules,” ACS Cent. Sci. vol. 4, no. 2, pp. 268-276, 2018.
[4] M. Pawlicki et al., “Two-photon absorption and the design of two-photon dyes," Angew. Chem. Int. Ed. vol. 48, no. 18, pp. 3244-3266, 2009.
[5] T. Yanai et al., “A new hybrid exchange-correlation functional using the Coulomb-attenuating method (CAM-B3LYP)," Chem. Phys. Lett. vol. 393, no. 1-3, pp. 51-57, 2004.
[6] F. Weigend et al., “Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: design and assessment of accuracy”, Phys. Chem. Chem. Phys. vol. 7, no. 18, pp. 3297-3305, 2005.
[7] M. T. P. Beerepoot et al., “Benchmarking the performance of exchange-correlation functionals for predicting two-photon absorption strengths," J. Chem. Theory. Comput. vol. 14, no. 7, pp. 3677-3685, 2018.
[8] N. M. O'Boyle et al., “Open Babel: an open chemical toolbox," J. Cheminform. vol. 3, article no. 33, Oct. 2011.
[9] C. Bannwarth et al., “GFN2-xTB—an accurate and broadly parametrized self-consistent tight-binding quantum chemical method

with multipole electrostatics and density-dependent dispersion contributions," J. Chem. Theory Comput. vol. 15, no. 3, pp. 1652-1671, 2019.

執筆者紹介

大越　孝洋　　Takahiro Ohgoe
テクノロジー本部　マテリアル応用技術センター
Applied Materials Technology Center, Technology Div.
博士（理学）

大越　昌樹　　Masaki Okoshi
テクノロジー本部　マテリアル応用技術センター
Applied Materials Technology Center, Technology Div.
博士（理学）

長尾　宣明　　Nobuaki Nagao
テクノロジー本部　マテリアル応用技術センター
Applied Materials Technology Center, Technology Div.
博士（工学）

四橋　聡史　　Satoshi Yotsuhashi
テクノロジー本部　マテリアル応用技術センター
Applied Materials Technology Center, Technology Div.
博士（理学）