研　究　報　告　書

「ベイズ事後分布を探索重みに活用した物質構造探査の予測性向上」

研究タイプ：通常型
研究期間：　2016年10月～2020年3月
研　究　者：　本郷　研太

1．　研究のねらい

「物性の第一原理計算」と「機械学習などのデータ科学」との融合展開「マテリアルズ・インフォマティクス（MI）」において、その究極の目的は、「所望の物性を実現する構造を予測する事」である。本研究は、新規 MI 手法を開発して、「膨大な化合物空間」の中から所望の物性を持つ新規化合物、すなわち、「埋蔵分子」を効率的に探索する。そのための手法として、本研究ではベイズ推定に基づくアプローチを採用する。当該手法は、ベイズの定理に基づき、尤度を物性予測の回帰モデル、化合物構造の事前分布を自然言語モデルに基づき構築することで、それらの積をベイズ反転させることで、所望の物性を持つ化合物構造を見出す事後分布を構築する。すなわち、回帰モデルを逆に辿り、「化合物らしさ」の事前情報を統計的重み付けすることで、所望物性を持つ化合物を効率的に探索することが可能となる。このベイズ構造探索法は、所望物性を持つ化合物を探索する仮想スクリーニングの枠組みを超えた新しい化合物探索手法であり、広大な化合物空間から効率的・選択的に、所望物性を持つ化合物候補を提案することができる。本研究は、特に、回帰モデルと事前分布の構築で、物質・材料研究としては大規模なビッグデータを利用して、探索精度の向上を実現する。

本研究では、ベイズ構造探索法の確立、その適用事例として、有機太陽電池用途の新規化合物探索を実施する。ドラッグデザインやケモ・バイオインフォマティクスなどの関連分野を含め、逆問題としての物質探索の先行研究は多くない。特に、任意の物性量レンジに存在する化合物を探索する実用段階には至っていない。これらまでの予備的研究から、順問題回帰モデルの物性予測性能が、最終的な物質構造探索の信頼性に大きな影響を与えることがわかっている。回帰モデルの性能向上に向けては、モデル選択のみならず、パラメータ学習のために大量の学習データが必要となる。本研究課題で、ベイズ物質構造探索の実働事例を確立すれば、マテリアルズ・インフォマティクス分野の様々な問題に対する有効なアプローチとして、広く普及するものと期待される。

2．　研究成果

（1）概要

本研究では、(1) 学習データ生成のためのハイスループット第一原理計算、(2) 埋蔵化合物探索を目的とするベイズ構造探索法の研究基盤確立、及び、(3) ベイズ構造探索法の実材料開発に取り組んだ。

(1) ハイスループット第一原理計算

各種化合物データベースから既知化合物情報を入手し、第一原理計算による物性算定を行い、データベース化を進めている。本研究課題で得られた化合物の部分セットに対して、実験と計算結果を比較して、計算精度の検証を行った[論文発表 2/4/5]。これらの部分データに対

して解析を行った結果、物質探索に資する記述子の解明[論文発表 2/4]や高熱伝導率を持つポリマー結晶を発見できた[論文発表 4]。

(2) ベイズ構造探索法の開発

本研究では、具体的な問題設定として、HOMO-LUMO ギャップと内部エネルギーを対象物性量とする化合物探索に取り組み、ベイズ構造探索法の基盤を確立することができた。本研究では、尤度関数として線形回帰モデルを採用し、PubChem データベースからランダムに選んだ 10,000/6,674 個の化合物物性を学習/テストデータとした。対象物性の計算には Gaussian09 を用いた密度汎関数法(DFT/B3LYP)により算定した。事前分布としては、自然言語モデルに基づく化合物構造生成法を開発し、PubChem データベース収録の 50,000 化合物構造データを用いて学習を行った。化合物類似度を指標として、得られた仮想化合物とデータベース登録既知化合物を照合した結果、データベースには登録されていない新規構造を持つ化合物候補がベイズ構造探索法によって提案されていることがわかった[論文発表 1]。

(3) ベイズ構造探索法による実材料開発

本研究では、ベイズ構造探索による実材料開発の適用事例として、有機太陽電池のドナー材料探索に取り組んだ。実際の材料作成まで至っていないが、仮想分子ライブラリーを構築している。本研究課題に関連して、ベイズ構造探索法を活用した高熱伝導率ポリマー材料開発で、当該法の有用性を確認できた[論文発表 3]。

(2)詳細

研究課題(1) 学習データ生成のためのハイスループット第一原理計算

研究テーマ A: 既存化合物構造に対する第一原理計算の実施

化合物の公開データベースとして、PubChem、Polymer Genome、Materials Project などのデータベースから既知化合物の情報を入手し、第一原理計算を実施して、独自のデータベースを構築している。本研究課題は、後段の MI 研究展開への学習データ提供以外にも、それ自体に利用価値があり、次に示すスピンアウト的研究成果が得られた。発表論文 5:有機化合物の部分集合データを活用して、分子理論と組み合わせることで、液体プロセスなどの濡れ性制御に必要となるハマカー定数を評価し、実験との比較検証を行った。得られた計算結果は、十分な精度を持ち、ハマカー定数の計算科学的算定アプローチが確立できた。発表論文 2:層状化合物系の第一原理計算と呼応するマーデルング・ポテンシャルの相関関係を見出すことができた。当該ポテンシャルは、第一原理計算よりも計算コストの低い物理量であり、新規層状化合物の探索スクリーニングにおける良い記述子として活用できると期待される。論文発表 4: Polymer Genome Project に登録されているポリマー結晶構造を利用して、熱伝導率算定のベンチマーク計算を行った。その結果、計算科学の枠組みで、低温でポリエチレン結晶を超える熱伝導率を持つポリマー結晶を発見することができた。また、熱伝導率と相関する調和近似の範疇で算定可能な記述子を発見することができた。当該記述子を利用すれば、計算コストの非常に大きな熱伝導率計算を実施せずに、大量のポリマー結晶の仮想スクリーニングを実施することが可能となる。

研究課題(2) 埋蔵化合物探索を目的とするベイズ構造探索法の研究基盤確立

ベイズ構造探索法の構築では、ベイズの定理より、尤度関数 P(Y|S)と事前分布 P(S)の積(に比例する)として、所望の物性 Y を持つ化合物構造 S の事後分布 P(S|Y)を算定することができ、この事後分布に従ってサンプリングを行うことで、化合物候補を得ることができる。本課題では、次の小項目に分けて、研究を進めた。

研究テーマ B: 事前分布 P(S)の構築

化合物 S を SMILES 形式で記述することで、化合物を文字列として扱うことが可能となる。その結果、自然言語処理の言語生成モデルとして知られているn-gramモデルを適用することが可能となる。この n-gram モデルを化合物構造に対する文法制限を考慮して拡張することで、化合物生成器を構築できる。拡張 n-gram モデルの学習では、元素列の出現確率を既知化合物データベースから頻度として学習している。本研究では、具体的には、PubChem データベースからランダムサンプリングした数千～万程度の化合物を学習データとして活用した。

研究テーマ C: 順問題学習の回帰モデル P(Y|S)の構築

回帰モデルとしては、線形・非線形回帰の複数モデル(ガウス過程回帰、サポートベクトル回帰、ランダムフォレスト、勾配ブースティング等々)を検証した。また、記述子としては、各種フィンガープリント(PubChem、MACS 等々)を検証した。学習データとしては、数万程度の既知化合物の第一原理計算の結果を利用している。

研究テーマ D: ベイズ逆問題予測に基づく候補分子構造生成の実装

研究テーマ B と C の尤度関数 P(Y|S)と事前分布 P(S)を用いて、事後分布 P(S|Y)を構築し、この分布に従って、モンテカルロサンプリングを実施する。得られた候補化合物については、第一原理計算を行い、性能検証を行った。

研究テーマ E: 実装のパッケージ化

研究テーマ B/C/D のパッケージ化までは完了していないが、本研究課題に関連する一部の構造生成プログラムに関しては研究成果として独立に原著論文出版に向けて、Python パッケージ化を進めている(arXiv:1911.08071)。

研究課題(3) ベイズ構造探索法の実材料開発

研究テーマ F: 有機太陽電池のドナー材料探索/埋蔵分子の発掘

研究課題(2)で確立した研究基盤を利用して、有機太陽電池のドナー材料探索に取り組んだ。回帰モデルと事前分布の構築には、研究課題(1)の物性・構造データや HCEP(Harvard Clean Energy Project)データベースの物性・構造データを活用した。得られた候補化合物については、仮想化合物データベースに登録している。後述の今後の展開に示すように、実験研究者からの協力の下、得られた候補化合物から実際に有機太陽電池を作製し、エネルギー変換効率の測定に取り組む。なお、ベイズ構造探索法の実材料開発への適用事例として、関連研究として、高熱伝導率ポリマー開発を行い、実証研究を行った[論文発表 3]。

3. 今後の展開

本研究では、有機太陽電池ドナー材料の仮想化合物データベース構築まで到達することができた。候補化合物をドナー材料として実際に合成・作製することで、ベイズ構造探索法の実証研究事例として、本研究課題を完成させる。当該材料の実際の合成・作製では、学内実験研究者との共同体制を確立している。また、当該探索法の汎用性を実証するために、研究

課題(1)で得られた第一原理計算データベースを活用して、新規課題に取り組む。具体的には、高熱伝導率ポリマー結晶の探索を行う。現在、学習データを生成するための第一原理計算を継続している。また、候補ポリマーユニットからポリマー結晶構造を決定するためには、結晶構造探索手法の研究基盤確立が必要となるが、粒子群最適化法に基づく結晶構造探索法を利用する(既に別系統の化合物に適用し、新たな結晶相の発見につながっている)。ベイズ構造探索法と結晶構造探索法を組合せ、新規ポリマー結晶探索研究を展開する。なお、実際のポリマー合成・結晶作製についても、学内実験研究者との共同体制を確立している。

4. 自己評価

本研究は、(1)ハイスループット第一原理計算、(2)ベイズ構造探索法の開発、(3)太陽電池のドナー材料開発へのベイズ構造探索法の適用を目的として、課題設定を行った。課題(1)と課題(2)については、当初の計画通り、目的を達成することができた。本研究提案手法は、既知化合物群からの仮想スクリーニングで発見できない新たな化合物を研究対象に据えることを可能とする新しいMI研究であり、当該研究分野における重要な研究成果であると考えている。また、その研究基盤を利用するための第一原理計算データを大量に確保することができた。この点では、研究費を利用して導入した計算サーバ(データ保存のストレージやバックアップシステムを含む)や、データ整理を目的に雇用した研究補助員、及び、学生が大いに貢献している。得られた計算データを素早く整理したことで、得られたデータセットの部分集合の解析を行う余裕が生まれ、その結果として当初の計画では想定していなかった研究成果も得られており、ベイズ構造探索手法に基づく新たな研究展開への途を拓くことが可能になったと考えている。課題(3)は、本研究提案手法の有効性に関する実証研究であるが、候補化合物の提案とそのデータベース化に留まり、研究期間内での新規材料の発見には至らなかった。特に、未知構造の化合物を対象とする場合、その合成可能性を考慮した研究課題設定が必要であった。合成可能性を検証する枠組みの整備と、それに基づく事前確率の再構築(すなわち、合成可能性を考慮した候補化合物生成)が今後の課題と考えている。課題(3)に関連して、来年度以降の研究費確保の見通しは立っており、当該課題を継続実施し、新規材料の発見を目指す。実際に新材料の発見に繋がれば、データ駆動に基づく新たな材料開発が大きく前進するものと期待される。MI 材料開発アプローチの導入は、これまで絨毯爆撃的かつ経験的な実験計画に基づき行われてきた材料開発を刷新するもので、材料開発に伴うコストと時間の削減に大きく寄与するものと期待される。

5. 主な研究成果リスト

(1)論文(原著論文)発表

1.	Hisaki Ikebata, Kenta Hongo, Tetsu Isomura, Ryo Maezono, Ryo Yoshida, "Bayesian molecular design with a chemical language model", Journal of Computer-Aided Molecular Design, 2017, 31, 379-391.
2.	Daichi Kato, Kenta Hongo, Ryo Maezono, Masanobu Higashi, Hironobu Kunioku, Masayoshi Yabuuchi, Hajime Suzuki, Hiroyuki Okajima, Chengchao Zhong, Kousuke Nakano, Ryu Abe, Hiroshi Kageyama, "Valence Band Engineering of Layered Bismuth

	Oxyhalides toward Stable Visible-Light Water Splitting: Madelung Site Potential Analysis", Journal of the American Chemical Society,2017, 139, 18725-18731.
3.	Stephen Wu, Yukiko Kondo, Masa-aki Kakimoto, Bin Yang, Hironao Yamada, Isao Kuwajima, Guillaume Lambard, Kenta Hongo, Yibin Xu, Junichiro Shiomi, Christoph Schick, Junko Morikawa, Ryo Yoshida, "Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm", npj Computational Materials, 2019, 5, 66:1-11.
4.	Keishu Utimula, Tom Ichibha, Ryo Maezono, Kenta Hongo, "Ab Initio Search of Polymer Crystals with High Thermal Conductivity", Chemistry of Materials, 2019, 31, 4649-4656, (selected in Virtual Issue on Machine-Learning Discoveries in Materials Science, Chemistry of Materials, 2019, 31, 8243-8247.)
5.	Hideyuki, Takagishi, Takashi Masuda, Tatsuya Shimoda, Ryo Maezono, Kenta Hongo, "Method for the Calculation of the Hamaker constants of Organic Materials by the Lifshitz Macroscopic Approach With DFT", Journal of Physical Chemistry A, 2019, 123, 8726-8733.

（2）特許出願

研究期間累積件数：0 件

（3）その他の成果（主要な学会発表、受賞、著作物、プレスリリース等）

1.	本郷研太、「第一原理計算とベイズ統計に基づく新物質探索」、マテリアルズ・インフォマティクスによる材料開発と活用集（技術情報協会） 第 3 章第 3 節、pp.62-69. （著書・執筆分担）
2.	Kenta Hongo, "Recent Advances in Materials Simulations and Informatics", International Congress on Pure & Applied Chemistry (ICPAC) 2018, 2018/03/10, Siem Reap, Cambodia (Invited).
3.	Kenta Hongo, "Computational materials design from ab initio simulations to ab initio materials informatics", The 10th International Conference of the Asian Consortium on Computational Materials Science (ACCMS-10), 2019/07/26, City University of Hong Kong, China (Invited).
4.	Kenta Hongo, "Data-driven approach to computational materials design", 20th International Union of Materials Research Societies International Conference in Asia (IUMRS-ICA), 2019/09/24, Perth Convention and Exhibition Centre, Perth, Australia (Invited).
5.	Kenta Hongo, "Data-driven approach to molecular design", The 5th International Conference on Molecular Simulation 2019 (ICMS 2019), 2019/11/04, Jeju, Republic of Korea (Invited).