東京科学大学 TSUBAME 共同利用
令和６年度利用終了課題 利用成果報告書集

東京科学大学　情報基盤センター
共同利用支援室
https://www.t4.cii.isct.ac.jp/tsubame-kyodo

本報告書集および個別の報告書の PDF ファイルは、以下の URL にあります。
令和６年度採択実績および利用終了課題報告書
https://www.t4.cii.isct.ac.jp/sites/default/files/kyodo/R06_seika.pdf

■令和６年度産業利用 利用成果報告書 一覧

■令和６年度産業利用 利用成果報告書（未提出）

■令和６年度学術利用 利用成果報告書別紙の提出免除課題 一覧
（利用成果を論文/学会等にて発表した要旨等の提出により提出免除）

申請課題名

所属機関／利用課題責任者

論文/学会等における発表済利用成果の情報

植物ポリフェノールの化学構造および反応機構に関する研究

長崎大学生命医科学域（薬学部）／松尾 洋介

[1] Ryo Ohtsuka, Yosuke Matsuo, Yoshinori Saito, Koji Yamada, Takashi Tanaka, Fumika Yakushiji, Structural Revisions of Macrocyclic Ellagitannin Dimers Bridged by Two Hexahydroxydiphenoyl Groups, ChemRxiv (preprint)
DOI: 10.26434/chemrxiv-2025-k7dtx

深層生成モデルによる人工タンパク質・人工ペプチド設計（※ 提出免除）

東京医科歯科大学難治疾患研究所／島村 徹平

分子動力学シミュレーションによる新興再興感染症に関する研究

国立感染症研究所病原体ゲノム解析研究センター／横山 勝

[1] 小谷 治、横山 勝、佐藤裕徳, 分子動力学計算を用いた HIV-1 ゲノム RNA LTR 領域の構造機能予測系の開発. 第 38 回日本エイズ学会学術集会・総会 The Journal of AIDS Research Vol.26 No.4 2024
https://jaids.jp/pdf/2024/20242604/20242604197353.pdf

界面・局所構造制御による熱物性変調メカニズムの解明

産業技術総合研究所 計量標準総合センター／志賀 拓麿

[1] Takuma Shiga, Emi Minamitani, Yuichiro Yamashita, Takashi Yagi, Naoyuki Taketoshi, Yuzo Shigesato, Makoto Kashiwagi, Thermal transport and topological analyses of the heat-carrying modes and their relevant local structures in variously dense amorphous alumina, Applied Physics Letters. 125, 012201 (2024)
DOI: 10.1063/5.0200729

[2] Takuma Shiga, Yukihiko Terada, Takashi Kodama, and Shohei Chiashi ,Influence of perturbative intertube interactions on ballistic and quasi-ballistic phonon transports in double-walled carbon nanotubes, International Journal of Heat and Mass Transfer, 233, 126030, (2024).
DOI: 10.1016/j.ijheatmasstransfer.2024.126030

[3] Kaiyao Zhou, Kan Ueji, Takuma Shiga, Hiroyuki Nishidome, Shigeki Saito, Sota Nakamura, Takahiko Endo, Yohei Yomogida, Yasumitsu Miyata, Junei Kobayashi, Takahiro Yamamoto, Takahi Yagi, and Kazuhiro Yanagi,Correlation between the cross-plane thermal and electrical conductance in randomly stacked MoS_2, ACS Applied Electronic Materials, 6, 5934-5941 (2024).
DOI: 10.1021/acsaelm.4c00911

[4] Takuma Shiga, Takashi Yagi, and Hiroshi Fujihisa, Inorganic Chemistry,Structural phase transition of Mg_2NiH_4 induced by NiH_4 molecular ion distortion and metal lattice deformation under high pressures, ACS Inorganic Chemistry 63, 21121-21129 (2024).
DOI: 10.1021/acs.inorgchem.4c03336

申請課題名

所属機関／利用課題責任者

論文/学会等における発表済利用成果の情報

機能性メタレンズのマルチスケール解析

東京農工大学／岩見　健太郎

[1] Masakazu Yamaguchi, Mitsutoshi Hada, and Kentaro Iwami
Polarization-independent full-color holographic movie with a single metasurface free from crosstalk
DOI: 10.1364/OE.550169

金属組織形成および物性の大規模フェーズフィールド・分子動力学解析

大阪公立大学／三好　英輔

[1] Tomoo Fujiwara, Eisuke Miyoshi, Akinori Yamanaka, Evaluating inclination-dependent anisotropic grain boundary energies: Bayesian data assimilation approach using molecular dynamics and phase-field simulations, Computational Materials Science, February 2025, 113605, Volume 248
DOI: 10.1016/j.commatsci.2024.113605

継続的ベイズ推論の改善

理化学研究所革新知能統合研究センター／モハマド エムティヤズ カーン

[1] Sin-Han Yang, Zhedong Liu, Gian Maria Marconi, Mohammad Emtiyaz Khan, Variational Learning Induces Adaptive Label Smoothing, arXiv:2502.07273v2 [cs.LG] 4 Mar 2025 (Under review)
DOI: 10.48550/arXiv.2502.07273

南海トラフ域の地震波速度構造モデルに基づく走時計算ニューラルネットワークの構築・公開

海洋研究開発機構／縣　亮一郎

[1] Ryoichiro Agata, Satoru Baba, Ayako Nakanishi, and Yasuyuki Nakamura, HypoNet Nankai: Rapid hypocenter determination tool for the Nankai Trough subduction zone using physics-informed neural networks, Seismological Research Letters (2026) 97 (1): 524-536
DOI: 10.1785/0220240377

超大規模焼結シミュレーションによる緻密化挙動の定量的解析

国立研究開発法人 物質・材料研究機構／石井　秋光

[1] 石井秋光, 大規模固相焼結フェーズフィールドシミュレーションによる緻密化挙動の定量評価, 第 30 回計算工学講演会 計算工学講演会論文集 Vol.30 (2025 年 6 月)
https://confit-sfs.atlas.jp/customer/jsces30/abstract/E-03-05.pdf

申請課題名

所属機関／利用課題責任者

論文/学会等における発表済利用成果の情報

フェイクメディア検知の基礎研究と応用実験

国立情報学研究所コンテンツ科学研究系／山岸　順一

[1] Xuechen Liu, Xin Wang, Junichi Yamagishi, A Preliminary Case Study on Long-Form In-the-Wild Audio Spoofing Detection, 2024 International Conference of the Biometrics Special Interest Group (BIOSIG) (Oral), Darmstadt, Germany, 2024 年 9 月
DOI: 10.1109/BIOSIG61931.2024.10786724

[2] Yoshihiko Furuhashi, Junichi Yamagishi, Xin Wang, Huy H. Nguyen, Isao Echizen, Exploring Active Data Selection Strategies for Continuous Training in Deepfake Detection, 2024 International Conference of the Biometrics Special Interest Group (BIOSIG), Darmstadt, Germany, 2024 年 9 月
DOI: 10.1109/BIOSIG61931.2024.10786748

[3] Xin Wang, Héctor Delgado, Hemlata Tak et al, ASVspoof 5: crowdsourced speech data, deepfakes, and adversarial attacks at scale, Proc. The Automatic Speaker Verification Spoofing Countermeasures Workshop (ASVspoof 2024), 2024 年 8 月
DOI: 10.21437/ASVspoof.2024-1

[4] Xuechen Liu, Junichi Yamagishi, Md Sahidullah, Tomi Kinnunen, Explaining Speaker and Spoof Embeddings via Probing, 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, 2025 年 4 月
DOI: 10.1109/ICASSP49660.2025.10887897

生体脂質膜環境下でのシグナルタンパク質の分子動力学シミュレーション

東北大学大学院薬学研究科／井上　飛鳥

[1] Yuya Sugiura, Tatsuya Ikuta, Yuji Sumii, Hirokazu Tsujimoto, Kohei Suzuki, Ryoji Suno, Putri Nur Arina Binti Mohd Ariff, So Iwata, Norio Shibata, Asuka Inoue, Takuya Kobayashi, Hideki Kandori and Kota Katayama, Discovering Key Activation Hotspots in the M2 Muscarinic Receptor, Journal of the American Chemical Society, Soc. 2025, 147, 14, 11754-11765
DOI: 10.1021/jacs.4c14385

マルチモーダル言語モデルの開発

東北大学大学院情報科学研究科／斉藤　いつみ

[1] Itsumi Saito, Haruto Yoshida, Keisuke Sakaguchi, Sketch2Diagram: Generating Vector Diagrams from Hand-Drawn Sketches, The Thirteenth International Conference on Learning Representations (ICLR 2025)
https://openreview.net/pdf?id=KvaDHPhhir

計算機の文章読解・生成能力向上に関する研究

早稲田大学基幹理工学部／河原　大輔

[1] 織田 宥楽, 小川 隼斗, 河原 大輔, 教師なし学習によるラップの形式の学習, 人工知能学会全国大会(第 39 回)
https://www.t4.cii.isct.ac.jp/sites/default/files/2025-05/R06_gaku_24IDH.pdf

申請課題名

所属機関／利用課題責任者

論文/学会等における発表済利用成果の情報

次世代医療創出のための疾患・創薬研究（※ 提出免除）

東京医科歯科大学 M&D データ科学センター／清水　秀幸

現実世界の逐次的環境変化に適応するマルチモーダル自然言語理解モデル

東北大学大学院情報科学研究科／坂口　慶祐

[1] Itsumi Saito, Haruto Yoshida, Keisuke Sakaguchi, Sketch2Diagram: Generating Vector Diagrams from Hand-Drawn Sketches, The Thirteenth International Conference on Learning Representations
https://openreview.net/forum?id=KvaDHPhhir

オープンソースかつ日本語に強い大規模言語モデルの構築

産業技術総合研究所人工知能研究センター知識情報研究チーム／高村　大也

[1] Youmi Ma, 水木 栄, 藤井 一喜, 中村 泰士, 大井 聖也, 島田 比奈理, 塩谷 泰平, 齋藤 幸史郎, 前田 航希, 服部 翔, 岡本 拓己, 石田 茂樹, 横田 理央, 高村 大也, 岡崎 直観, 模倣学習による大規模言語モデルの指示チューニング, 言語処理学会 第 31 回年次大会 発表論文集（2025 年 3 月）
https://www.anlp.jp/proceedings/annual_meeting/2025/pdf_dir/Q8-21.pdf
DOI: 10.48550/arXiv.2503.23714

深層学習を用いた文書画像解析手法の評価

東北大学大学院工学研究科／宮崎　智

[1] Yaohou Fan, Zhengmi Tang, Tomo Miyazaki, Yongsong Huang, Shinichiro Omachi, A real-driven image synthesis framework for adapting pre-trained scene text detectors, Image and Vision Computing, Volume 172, August 2026, 106027
DOI: 10.1016/j.imavis.2026.106027

※令和 6 年 10 月 1 日より東京工業大学と東京医科歯科大学が統合し東京科学大学となったため、提出免除となります。

■令和６年度学術利用 利用成果報告書 一覧

■令和５年度学術利用 利用成果報告書（公開延期分）

■令和６年度学術利用 利用成果報告書（公開延期）

申請課題名 所属機関／利用課題責任者	頁
GPU クラスタを用いたミリ波帯大規模広帯域電波伝搬シミュレーション 国立研究法人情報通信研究機構／チャカロタイ・ジェドヴィスノプ	公開延期

■令和６年度学術利用 利用成果報告書（未提出）

申請課題名 所属機関／利用課題責任者	頁
モデル崩壊の観察と対処方法の開発 理化学研究所革新知能統合研究センター／幡谷　龍一郎	未提出

TSUBAME 共同利用　令和 6 年度　産業利用　成果報告書

利用課題名　マルチモーダル基盤モデルとモデル間連携技術の研究開発
英文：Research and development of inter-model adaptation technology and multimodal foundation models

シーン　誠

Sakana AI
https://sakana.ai/

邦文抄録（300 字程度）

大規模言語モデルとマルチモーダル基盤モデルは高い性能を示す一方で、そのサイズが実用展開の障壁となっている。本研究では、異なるモデル間の効率的な知識連携を実現する「Temporally Adaptive Interpolated Distillation（TAID）」を提案した。TAID は生徒モデルの初期分布から教師モデルの分布へと徐々に移行する時間依存的な中間分布を導入することで、容量ギャップ問題とモード平均化/崩壊問題を同時に解決する。実験の結果、TAID は言語モデルとマルチモーダルモデルの両方で既存手法を上回る性能を示し、その有効性が国際会議 ICLR での Spotlight 採択によっても認められた。

英文抄録（100 words 程度）

While large language models and multimodal foundation models demonstrate remarkable capabilities, their size poses significant deployment challenges. We introduce Temporally Adaptive Interpolated Distillation (TAID), a novel inter-model adaptation technique that dynamically bridges teacher and student models through time-dependent intermediate distributions. TAID gradually transitions from the student's initial distribution to the teacher's distribution, effectively addressing capacity gap and balancing mode-averaging/collapse issues. Our experiments demonstrate TAID's superior performance in developing both language and multimodal models, showcased through our state-of-the-art TAID-LLM-1.5B and TAID-VLM-2B, advancing efficient AI deployment.

Keywords: モデル間連携技術、マルチモーダル基盤モデル、知識蒸留、時間適応型学習、効率的 AI

背景と目的

大規模言語モデル（LLM）は様々な分野で革新的な能力を示しているが、そのサイズが増大するにつれて、リソース制約のある環境への展開が大きな課題となっている。例えば、エッジデバイスでの実行や、リアルタイムアプリケーションでの低遅延応答、省エネルギー運用などが困難になっている。
このような課題に対応するためには、異なるモデル間の効率的な連携技術が重要となる。知識蒸留（Knowledge Distillation）はその代表的なモデル間連携技術であり、大型の教師モデルから小型の生徒モデルへと知識を移転することで、小型でありながら高性能なモデルを作成する有望なアプローチである。しかし、従来のモデル間連携手法には二つの根本的な課題がある。一つは教師モデルと生徒モデル間の大きな容量ギャップで、もう一つはモード平均化（生徒モデルが教師の全モードを過度に平滑化）とモード崩壊（生徒モデルが特定のモードのみに集中）のバランスの問題である。
本プロジェクトでは、これらの問題を解決するための新しいモデル間連携技術「TAID: Temporally Adaptive Interpolated Distillation」の技術実験を実施し、大規模言語モデルおよびマルチモーダル基盤モデルの効率的な開発を実現することを目的とした。これにより、小型で高性能な言語モデルおよびマルチモーダルモデルの開発を可能にし、リソース制約のある環境での AI 技術の普及に貢献する。

概要

TAID の核心は、従来の知識蒸留法と比較して根本的にアプローチが異なる点にある。下図に示すように、従来の知識蒸留（左図）では生徒モデルを固定された教師分布に直接最適化するのに対し、TAID（右図）では時間に依存した中間分布を通じて段階的な最適化を行う。

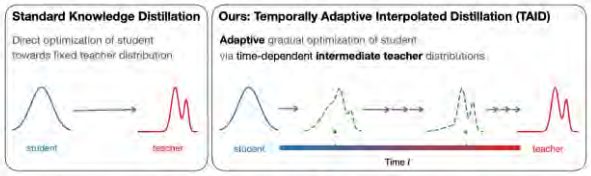


Figure 1: **Comparison of standard KD and TAID. (Left)** Standard KD methods typically employ direct optimization towards a fixed teacher distribution. **(Right)** TAID creates a dynamic bridge through adaptive, time-dependent *intermediate teacher* distributions (green dashed lines), enabling gradual optimization of the student. This approach facilitates a flexible transition from the student's initial distribution towards the teacher's distribution over time, effectively addressing the capacity gap and balancing knowledge transfer across varying model sizes.

この中間分布は、生徒モデルの初期分布（図 1 右側の左端）から教師モデルの分布（図 1 右側の右端）へと時間の経過とともに徐々に変化する。緑の破線で示される中間分布は、補間パラメータ t によって制御され、次式で定義される：

Definition 3.1 (TAID Interpolated Distribution)**.** *For any input sequence $y^{<s} \in \mathcal{Y}^{s-1}$ and any output token $y_s \in \mathcal{Y}$, the TAID interpolated distribution p_t is defined as:*

$$p_t(y_s|y^{<s}) := \text{softmax}\left((1-t)\cdot \text{logit}_{q_\phi'}(y_s|y^{<s}) + t \cdot \text{logit}_p(y_s|y^{<s})\right) \quad (1)$$

where $t \in [0,1]$ is a time-dependent interpolation parameter, $\text{logit}_{q_\phi'}$ represents a detached version of the student logits (i.e., treated as a constant without being backpropagated), and logit_p represents the teacher logits.

TAID の最適化過程では、生徒モデルは常にこの中間分布を目標として学習を進める。学習初期には中間分布が生徒モデル自身に近いため学習が容易であり、徐々に教師モデルの特性を取り入れることができる。この段階的アプローチにより、容量ギャップによる学習の困難さを軽減し、同時にモード平均化とモード崩壊のバランスも取ることができる。

さらに、TAID では補間パラメータ t を生徒モデルの学習進度に応じて適応的に更新する機構を導入している。これにより、学習過程を通じて一貫した難易度の学習タスクを維持し、効率的かつ安定した知識移転を実現している。

このような高度なモデル間連携技術により、単一モダリティの言語モデルだけでなく、複数のモダリティを扱うマルチモーダル基盤モデルの効率的な開発も可能となる。後述するように、TAID を活用してテキストのみを扱う言語モデル（TAID-LLM-1.5B）と、画像とテキストを統合的に処理するマルチモーダル基盤モデル（TAID-VLM-2B）の両方を開発することに成功した。

結果および考察

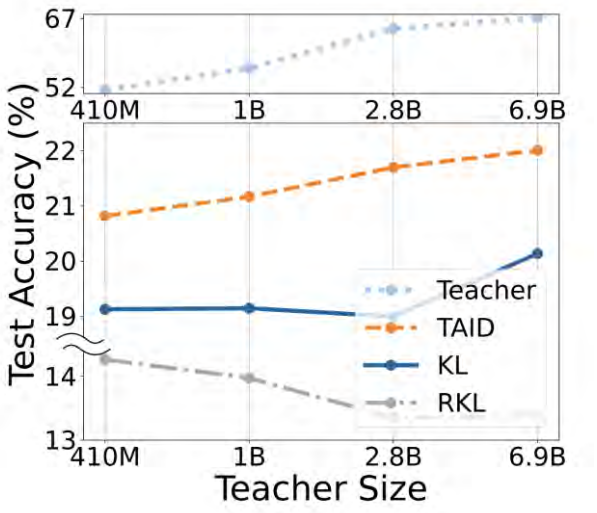
TAID の有効性を検証するため、まず UltraChat 200k データセットを用いた指示チューニング実験を行った。下表に示すように、様々な教師-生徒モデルペアでの MT-Bench スコアによる評価において、TAID は既存の知識蒸留手法を一貫して上回るパフォーマンスを示した。

Table 1: **Evaluating distillation methods for LLM instruction tuning.** The MT-Bench scores after training are listed, where higher scores indicate better conversational performance. For each of the three teacher-student pairs, different distillation algorithms, including the proposed TAID method, are compared. The highest score in each column is highlighted in **bold**.

Method \ Teacher / Student	Phi-3-mini (3.8B) / TinyLlama (1.1B)	Llama-2 (6.7B) / TinyLlama (1.1B)	StableLM Zephyr (2.8B) / Pythia (0.4B)
SFT	2.00	3.94	2.57
KL (Hinton et al., 2015)	2.71	3.99	2.74
RKL (Wen et al., 2023; Gu et al., 2024)	3.48	3.92	2.53
TVD (Wen et al., 2023)	3.27	3.64	2.57
Adaptive KL (Wu et al., 2024)	3.27	3.77	2.64
GKD (Agarwal et al., 2024)	2.24	3.82	2.59
DistiLLM (Ko et al., 2024)	3.23	3.97	2.97
CTKD (Li et al., 2023b)	1.78	2.84	1.39
DKD (Zhao et al., 2022)	2.70	4.14	2.90
(Ours) TAID w/o adaptive update	3.44	4.18	2.88
(Ours) TAID	**4.05**	**4.27**	**3.05**

特筆すべきは、TAID が学生生成出力（SGO）を必要とせずにこの性能を達成している点である。これにより、SGO ベースの手法と比較して訓練時間が約 2〜10 倍短縮された。また、適応型更新機構を導入することで、適応型更新なしの TAID と比較して 2.2%から 17.7%の性能向上が確認された。

また、TAID の容量ギャップへの耐性を検証するため、固定サイズの生徒モデル（70M）に対して異なるサイズの教師モデル（410M から 6.9B）を用いた実験を行った。下図は、教師モデルのサイズと生徒モデルの性能の関係を示している。



TAID は教師モデルのサイズが増加するにつれて単調に性能が向上しており、容量ギャップの問題を効果的に解決していることが示されている。一方、KL や RKL といった従来手法では教師モデルのサイズ増加に伴う一貫した性能向上が見られず、容量ギャップの問題を抱えていることが確認された。

研究の実用的インパクトを示すため、TAID を用いて二つの最先端モデルを開発した。TAID-LLM-1.5B は 2B 未満のパラメータを持つモデルで最高のスコアを達成し、TAID-VLM-2B は 4B までの視覚言語モデルでトップパフォーマンスを示した。特に TAID-VLM-2B の開発成功は、本研究の主目的であるマルチモーダル基盤モデ

ルの効率的な開発が達成されたことを示している。これらの結果は、TAID というモデル間連携技術が単一モダリティだけでなくマルチモーダルモデルの開発においても有効であることを実証している。

まとめ、今後の課題

本研究では、マルチモーダル基盤モデルとモデル間連携技術の研究開発を目指し、効率的なモデル間知識移転手法「TAID」を提案した。TAID は時間依存的な中間分布を導入することで、異なるモデル間の容量ギャップを橋渡しし、モード平均化とモード崩壊のバランスを取る。広範な実験により、TAID は指示チューニングと事前学習の両シナリオにおいて既存のモデル間連携手法よりも優れたパフォーマンスを示し、その学術的重要性は ICLR 2025 での Spotlight 採択によっても認められた。
実用面では、TAID-LLM-1.5B と TAID-VLM-2B の開発により、リソース制約のある環境での高性能言語モデルおよびマルチモーダル基盤モデルの実装可能性が示された。特に視覚とテキストを統合処理する TAID-VLM-2B の成功は、本研究の主目的であるマルチモーダル基盤モデル開発の達成を示す重要な成果である。
今後の課題としては、以下の点が挙げられる：
1. 他の距離メトリクスへの TAID の拡張：現在の KL ダイバージェンス以外のメトリクスでの有効性検証
2. 非線形補間の探索：より複雑な補間スキームによる性能向上の可能性
3. マルチモーダル知識蒸留への適用：視覚-言語タスクなど複数モダリティにまたがる知識移転
4. マルチテーチャー蒸留への拡張：複数の教師モデルからの知識統合手法の開発
これらの課題に取り組むことで、TAID はより広範な AI アプリケーションにおいて効率的なモデル開発を可能にし、AI 技術のアクセシビリティと実用性をさらに向上させることが期待される。

TSUBAME 共同利用　令和6年度　産業利用　成果報告書

利用課題名　GPU の産業分野の流体解析（車両空力、ファン騒音、建築物風荷重）への適応検討
英文：Study on adapting GPU to industrial fluid analysis (vehicle aerodynamics, fan noise, building wind load)

利用課題責任者
Koichi Ijuin

所属
Altair Engineering
Altair（アルテアエンジニアリング株式会社）

邦文抄録

近年，計算機性能の向上に伴い，格子ボルツマン法(Lattice Boltzmann Method: LBM)による数値流体計算は複雑な形状への適用や並列計算の効率性など，実務的な利点が注目されている。一方で，特に建築物の風荷重評価において，従来の有限体積法(Finite Volume Method: FVM)と比較した場合の LBM の特徴や適用性については十分な知見が得られておらず，日本建築学会「CFD 適用ガイド」1)においても，FVM による検討方法が位置づけられているのみであり，LBM による荷重評価方法は適用外とされている。そこで本研究では，LBM を用いた CFD の建築分野における実務利用を目指し，LBM による建物の風圧力の予測精度検証を実施した。本研究の流体解析には商用流体解析ソフト(ultraFluidX)と TSUBAME4.0 を用いて実用性についても検討を行った。

英文抄録

In recent years, with the improvement of computer performance, computational fluid dynamics using the Lattice Boltzmann Method (LBM) has been attracting attention for its practical advantages, such as its applicability to complex geometries and its efficient parallel computing. However, there is not enough knowledge about the characteristics and applicability of LBM compared to the conventional Finite Volume Method (FVM), especially in wind load evaluation of buildings. In the "Guide to CFD Application" by the Architectural Institute of Japan 1), only the FVM method is positioned as a method for evaluation, and the LBM method for load evaluation is not applicable. Therefore, in this study, aiming at practical use of CFD using LBM in the field of architecture, we verified the prediction accuracy of wind pressure on buildings using LBM .For the fluid analysis in this study, we used a commercial fluid analysis software (ultraFluidX) and TSUBAME4.0 to examine the practicality of the method.

Keywords:

格子ボルツマン法(Lattice Boltzmann Method: LBM)
有限体積法(Finite Volume Method: FVM)
日本建築学会「CFD 適用ガイド」1)
風荷重評価
TSUBAME4.0

背景と目的

近年，計算機性能の向上に伴い，格子ボルツマン法(Lattice Boltzmann Method: LBM)による数値流体計算は複雑な形状への適用や並列計算の効率性など，実務的な利点が注目されている。一方で，特に建築物の風荷重評価において，従来の有限体積法(Finite Volume Method: FVM)と比較した場合の LBM の特徴や適用性については十分な知見が得られておらず，日本建築学会「CFD 適用ガイド」1)においても，FVM による検討方法が位置づけられているのみであり，LBM による荷重評価方法は適用外とされている。そこで本研究では，LBMを用いたCFDの建築分野における実務利用を目指し，LBM による建物の風圧力の予測精度検証を実施した。

概要

格子ボルツマン法による風圧力の予測精度検証として以下の5つの検証を行った

（1）辺長比 D/B の異なる二次元角柱の風力係数
（2） 二次元角柱周りの風圧分布の比較検討
（3）単体角柱に作用する風荷重の評価と格子解像度の影響
（4）市街地内の角柱建物に作用する風荷重の評価と格子解像度の影響
（5）気流条件と建物形状による影響

結果および考察

本研究で行った上記5つのテーマの内容は 2025 年 9 月の日本建築学会大会で発表が行われる予定です。
（詳細は 9 月以降に公開される投稿論文に記載）
結果としては、LBMの予測精度はFVMと同等の結果が得られました。

まとめ、今後の課題

本件研究では角柱などの基本形状に対してLBMの精度検証を行ったが、今後は LBM の実用化を目指し、実際の高層建築模型を使った精度検証を行う必要がある。

論文:格子ボルツマン法による外装ルーバーから発生する風切り音の解析
土屋 直也, 田中 英之, 今野 大輔, 伊集院 浩一, 溝口 晶弘, 髙舘 祐貴, 柴田 良一
日本建築学会学術講演梗概集 環境工学I 209-210 2025年9月

TSUBAME 共同利用　令和 7 年度　産業利用　成果報告書

利用課題名　深層強化学習を用いた超人間級 AI 運転技術の研究

英文：Development of super human-grade AI driving technology using deep reinforcement learning

利用課題責任者 山口祐
Yu Yamaguchi

所属 Turing 株式会社
Turing Inc.
URL https://tur.ing/

邦文抄録

本課題では、深層強化学習を用いて人間を超える自動運転 AI 技術の実現を目指し、画像の可変品質圧縮を可能にする「One-D-Piece」手法と、世界モデルにおける指示追従性能を測る「Act-Bench」ベンチマークを開発しました。これらにより、膨大な映像データを効率良く扱いつつ、人間の指示に沿った柔軟な行動決定が可能な自動運転 AI の基盤が整い、安全で高度な自動運転の実現に寄与します。これらの成果を通じて、人間の判断を超える高度で安全な自動運転 AI の実現へ一歩近づき、将来的に交通事故の減少や移動の効率化など社会的課題の解決に寄与することが期待されます。

英文抄録

In this project, we aimed to realize an autonomous driving AI that surpasses human driving capabilities using deep reinforcement learning. As key outcomes, we developed “One-D-Piece,” a novel method for image tokenization enabling variable-quality compression, and “Act-Bench,” a benchmark for evaluating instruction following in world models. These achievements allow efficient handling of massive visual data and provide a foundation for AI agents to follow human instructions, bringing us closer to safer and more advanced autonomous driving. We expect that this progress will contribute to reducing traffic accidents and improving transportation efficiency in the future.

Keywords

- Autonomous Driving
- Image Compression
- Image Tokenizer
- World Models

背景と目的

自動運転技術は、交通事故の削減や移動の効率化に大きな期待が寄せられています。しかし現在の自動運転システムは、天候の変化や予測不能な歩行者の挙動など、人間ならではの柔軟な判断を要する状況で限界を露呈することがあります。人間ドライバーは経験と直感でこれらに対応しますが、注意力の低下や判断ミスによる事故も後を絶ちません。そこで、人間の能力を上回る高度な判断力と安全性を備えた「超人間級」の AI ドライバーを実現することが求められています。

近年、ディープラーニング（深層学習）の発展により、囲碁やビデオゲームにおいて人間のトッププロを凌駕する AI が誕生しました。特に、試行錯誤を通じて学習する深層強化学習は、人間が直面し得ないほど膨大な経験をシミュレーション上で積むことで、人間以上の判断能力を獲得できる可能性を示しています。この手法を自動運転に応用すれば、現実では危険すぎるシナリオも仮想環境で繰り返し学習させることで、安全かつ効果的に AI ドライバーの能力を向上させられると期待されます。

しかし、自動運転への深層強化学習の適用にはいくつかの課題があります。その一つが、カメラ映像など膨大なセンサーデータの扱いです。高精細な画像を逐次 AI に処理させるには計算資源を圧迫するため、効率的なデータ圧縮・表現方法が必要となります。またもう一つの課題は、AI に人間の与える指示や交通ルールを理解させ、それに従って行動させることです。現行の強化学習エージェントは与えられた報酬に基づき自律的に行動を学習しますが、「指示に従う」能力を評価・訓練する仕組みは十分ではありません。たとえば、「次の角を左折せよ」「このエリアには進入するな」といった指示を AI が正確に解釈し実行できることは、安全な自動運転に不可欠です。

そこで本研究では、深層強化学習を用いた超人間級 AI 運転技術の実現を目指し、上記課題の解決に取り組みました。具体的には、(1) センサーデータとしての画像情報を効率よく圧縮・利用するための新しい手法の開発、(2) AI が人間の与える指示に従う能力を評価・向上させるための基盤整備、この二点を主な目的としました。前者により限られた計算資源下でも高品質な視覚情報処理を可能にし、後者により AI エージェントがルールや指示を遵守できるようにすることを狙います。なお、本取り組みは外部の計算資源と TSUBAME4.0 を活用して進めました。

概要

本プロジェクトでは、上記の目的を達成するために東京科学大学のスーパーコンピュータ TSUBAME4.0 を活用しました。

センサーデータ処理の効率化に向けて、画像を効率よくデジタル化する新手法「One-D-Piece」[i]を開発しました。これは、画像を小さな断片（ピース）の一列に変換し、必要に応じてその一部だけを使って画像を再構成できる技術です。重要な情報を持つピースを優先的に使用し、それ以外の細部は省略するといった柔軟な圧縮が可能なため、通信帯域や計算資源に応じて画像品質を調整できます。One-D-Piece の実現には、大規模な画像データセットを使った深層学習モデルの訓練が必要でしたが、TSUBAME4.0 上で多数の画像を並列処理することで、開発期間内にモデルの構築とチューニングを完了できました。

図 1 「One-D-Piece 各トークン長で再構成イメージ」

また、自動運転開発に利用する世界モデル（弊社が開発した Terra 含む）における指示追従能力を評価・向上させるための基盤として「Act-Bench」[ii]というベンチマークを策定しました。Act-Bench の設計と評価の多くは外部の計算資源を活用しましたが、本研究においてはその理論的枠組みや応用可能性についての検討を実施しました。

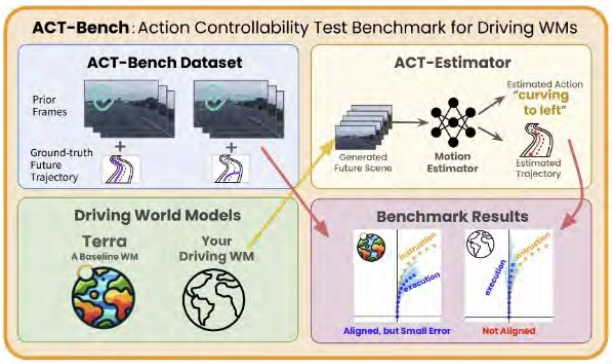


図 2 「Act-Bench 構成イメージ」

まとめ・今後の課題

本報告書で述べた成果により、深層強化学習を用いた超人間級 AI 運転技術の実現に向けて大きな前進が得られました。しかし、目標とする「人間を超える自動運転」の実現には、まだ乗り越えるべき課題が残されています。

まず、One-D-Piece については、現在は静止画ベースでの圧縮手法ですが、実際の自動運転では連続する動画データをリアルタイムに処理する必要があります。今後は動画への適用や、車載システム上での実装・最適化に取り組むことでより低コストでの推論処理を可能にしていきます。また、圧縮率と画質のトレードオフをさらに改善し、安全に関わる重要な物体（歩行者や標識など）は劣化させずに圧縮する工夫も課題です。

次に、Act-Bench に関しては、現実世界のあらゆる状況を網羅しているわけではありません。今後は、より多様で複雑な指示やシナリオをベンチマークに追加し、AI エージェントの包括的な理解力を評価できるように拡張していきます。例えば、自然言語で曖昧さを含む指示への対応や、複数の目標が同時に与えられる状況など、より現実に近い課題設定が考えられます。その上で、Act-Bench を活用して、指示に忠実かつ柔軟に対応できる新たなアルゴリズムの開発にも取り組んでいきます。

さらに、今回得られた技術を統合し、総合的な自動運転 AI システムへ発展させることが今後の目標です。具体的には、One-D-Piece によって効率化された視覚情報処理と、Act-Bench で鍛えられた指示遵守型の意思決定を組み合わせ、実車両や高度なシミュレーション環境で試験を検討しています。シミュレーションと実世界のギャップを埋める研究も重要であり、必要に応じて実証実験や他分野との協調も視野に入れています。最終的には、本研究の成果を産業応用につなげ、安全で快適な自動運転社会の実現に貢献することを目指します。

なお、本研究の遂行にあたり、東京科学大学のスーパーコンピュータ TSUBAME4.0 の計算資源提供を活用しました。外部の計算資源と合わせた利用ですが、計算リソースの支援なくして、これらの成果を短期間で得ることは困難でした。この場を借りて深謝いたします。

---

[i] Keita Miwa, Kento Sasaki, Hidehisa Arai, Tsubasa Takahashi, Yu Yamaguchi, "*One-D-Piece: Image Tokenizer Meets Quality-Controllable Compression*," arXiv:2501.10064 [cs.CV], 2025. https://arxiv.org/abs/2501.10064

[ii] Hidehisa Arai, Keishi Ishihara, Tsubasa Takahashi, Yu Yamaguchi, "ACT-Bench: Towards Action Controllable World Models for Autonomous Driving," arXiv:2412.05337 [cs.CV], 2024. https://arxiv.org/abs/2412.05337

TSUBAME 共同利用　令和6年度　産業利用　成果報告書

利用課題名　プラントによる気象変化究明
英文：Meteorology change prediction due to Plant Construction and Operation

利用課題責任者　胡 希東
Xidong Hu

所属　日揮グローバル(株)
JGC Corporation
URL: https://www.jgc.com/jp/

The micro-meteorology change due to the energy plant construction and anthropogenic heat release was studied using Multi-Scale Simulator for the Geoenvironment (MSSG). Three cases were compared to reveal the change in the wind direction, wind speed, ambient temperature. It is found that although the wind direction does not change significantly, the horizontal wind speed and ambient temperature increases significantly after the plant construction and operation.

Keywords: Micrometeorology, Anthropogenic Heat, Urban Heat Island, Energy Plant

1. Introduction

Micro-meteorology, which focuses on the lower atmosphere closer to the surface and at finer scales, significantly impacts energy plants regarding site selection, design, and operation. For a typical LNG plant with two trains, the air-cooled heat exchangers (ACHEs) located in an area measuring 500 m x 500 m can discharge approximately 40,000 to 60,000 m^3/s of air, carrying between 200 and 1000 MW of heat, which corresponds to a heat intensity of 4 to 10 kWm^2. Additionally, tree removal and land surface alteration are commonly conducted over an area of 2 km x 2 km for construction.

This study aims to elucidate the mechanisms and extent of the impact resulting from the construction and operation of energy plants on micro-meteorology at the construction site. The simulation compared micro-meteorological data before and after plant construction and operation, considering land surface changes and artificial heat release under the climatic backgrounds of tropical region.

In this study, the weather simulation was performed using TSUBAME and the impact of the plant construction and artificial heat release due to the plant operation on the performance of ACHEs was revealed.

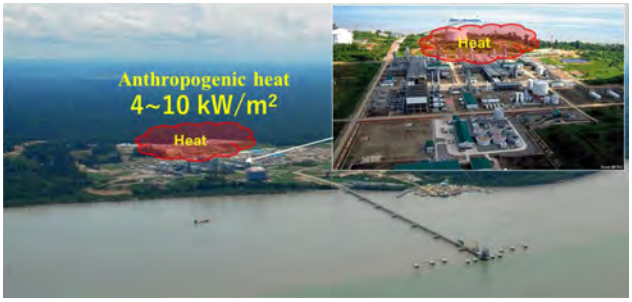


Fig. 1 LNG plant construction and artificial heat release

2. Simulation Model and Conditions

2.1 Simulation Tool

Multi-Scale Simulator for the Geoenvironment (MSSG) was used in the simulation. The simulation tool was developed by Japan Agency for Marine-Earth Science and Technology (JAMSTEC), Institute of Science Tokyo and Waseda University. As shown in Fig.2, the MSSG can simulate the spatial scales from global scale to street scale. The features of MSSG can be

found in the literature [1].

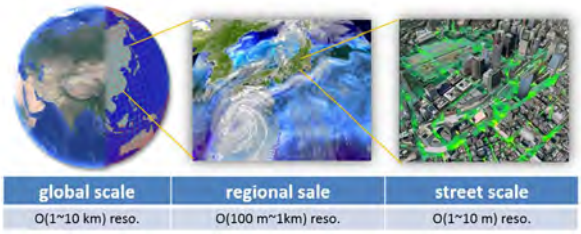


Fig. 2 Spatial scales and resolution of MSSG

2.2 Simulation Model

The simulation adopted four two-way-coupled nested systems, with horizontal resolution of 6 km (nest-1), 1.5 km (nest-2), 400m (nest-3) and 100m (nest-4) centered at the energy plant area as shown in Fig. 3. For all four nest systems, 86 vertical layers (lowest-layer height: 30 m) were used for the 40-km height domain. The boundary and initial conditions for the simulations were taken from the Japan Meteorological Agency (JMA) mesoscale analysis data (MANAL).

The elevation change was considered in the model as shown in Fig. 4. The elevation in the plant area was lowered because of the tree cutting. In addition, the land index change was considered as shown in Fig. 5. Furthermore, the anthropogenic heat of 5 kW/m^2 released from an area of 300m x 400m and 200m x 200m was considered in Case-3. The anthropogenic heat is released in the first cell above the ground and the height of the first cell is 30 m.

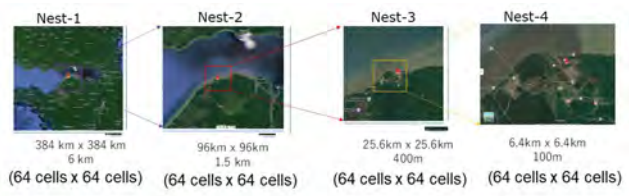


Fig. 3 Four nested domains in Indonesia

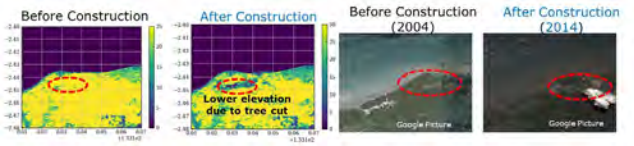


Fig. 4 Elevation change before and after plant cnstruction

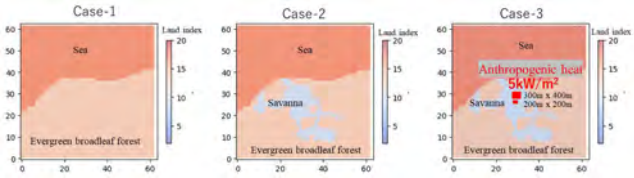


Fig. 5 Simulation model setting for Case-1, Case-2 and Case-3

2.3 Simulation Cases

Three cases were simulated to consider the impact of energy plant construction and operation. The main parameters are summarized in Table 1.

Table 1 Simulation cases.

Cases	Terrain (Elevation)	Land index	Anthropogenic heat
Case-1	No LNG plant (no tree cut)	Evergreen broadleaf forest	No
Case-2	With LNG plant (with tree cut)	Savanna in the plant construction area	No
Case-3	With LNG plant (with tree cut)	Savanna in the plant construction area	Yes 5 kW/m^2 for an area of 300m x 400m and 200m x 200m

3. Simulation Results

3.1 Wind direction change

The wind direction changes for Case-1 (before plant construction) and Case-2 (after plant construction but no anthropogenic heat) was compared in the wind rose of Fig. 6 and time series

data of Fig.7. It is found that no significant wind direction change is observed because of the energy plant construction.

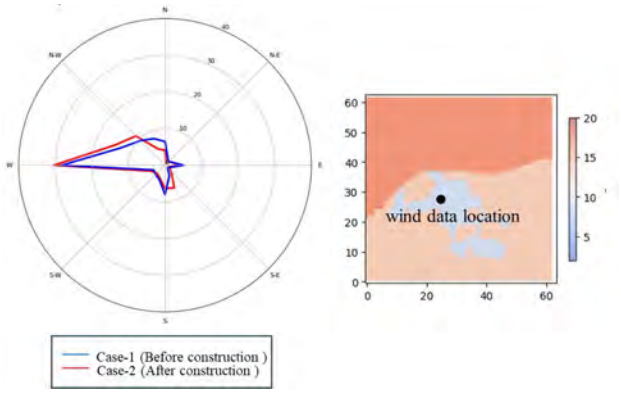


Fig. 6 Wind Rose of December 2012 inside the LNG plant for Case-1 and Case-2

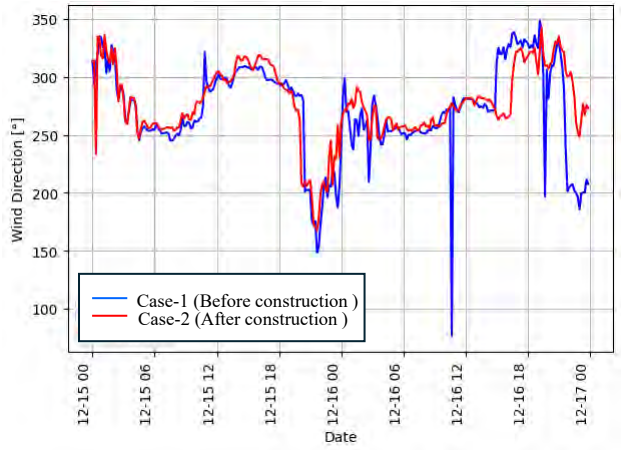


Fig. 7 Time series data of wind direction for Case-1 and Case-2 between 15th and 17th December

3.2 Wind speed change

The horizontal wind speed change at the location of anthropogenic heat release was compared in Fig.8 by plotting the wind speed data of three cases between 15th and 17th December. It is found that the wind speed in Case-2 is accelerated by about 0.5~1 m/s depending on the time because of the roughness of land surface is reduced due to the tree cutting. Also, it is surprising to observe the wind speed in Case-3 (case of anthropogenic heat release) is further accelerated by 1 m/s compared with that of Case-2. The reason is explored in Fig.9 by comparing the contour of ground surface temperature and vertical velocity for Case-2 and Case-3. The vertical velocity near the location of anthropogenic heat release is enhanced, and thus the horizontal velocity in Fig. 8 is accelerated. As shown in Fig. 8, the horizontal wind speed in Case-2 increases by approximately 0.5~1.0 m/s compared with that of Case-1.

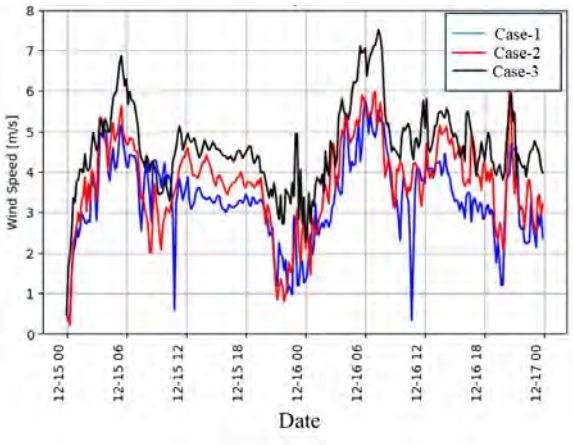


Fig. 8 Time series data of horizontal wind speed at the location of anthropogenic heat release for Case-1, Case-2 and Case-3 between 15th and 17th December

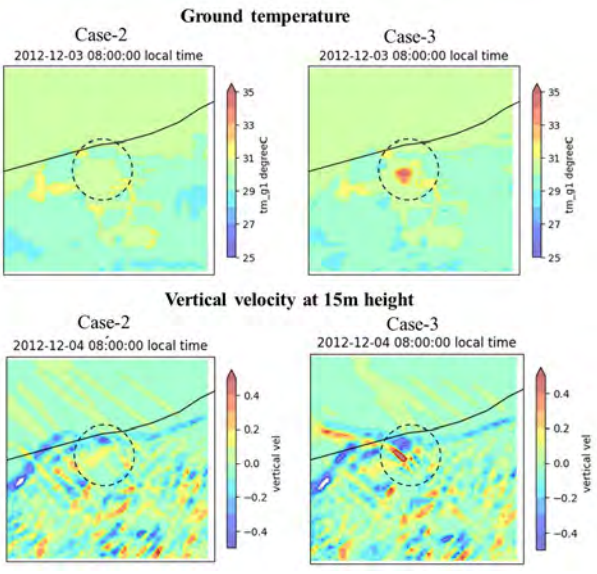


Fig. 9 Comparison of ground temperature and vertical velocity at 15m height for Case-2 and Case-3 on 4th December 8 am.

3.3 Ambient Temperature change

The ambient temperature at 2 m height above the ground and ground temperature was compared in Fig. 10 and Fig. 11. The ambient temperature at 2 m height in Case-2 increases by 1~2 ℃ compared with that of Case-1 due to the land surface change. Furthermore, the ambient temperature Case-3 (case of anthropogenic heat release) significantly increases by 4 ℃ compared with that of Case-2.

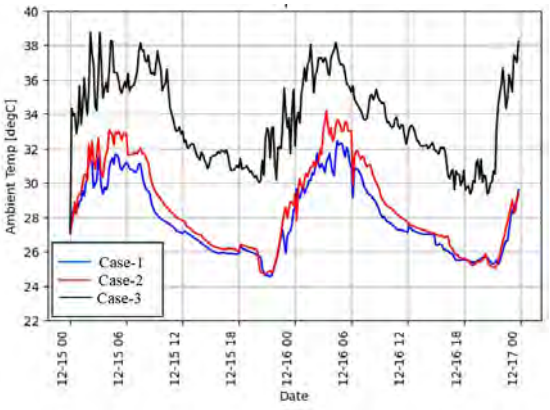


Fig. 10 Time series data of ambient temperature at 2m height at the location of anthropogenic heat release for Case-1, Case-2 and Case-3 between 15th and 17th December

4. Conclusion

The micro-meteorology change due to the energy plant construction and anthropogenic heat release was studied using Multi-Scale Simulator for the Geoenvironment (MSSG). The change of wind direction, wind speed, ambient temperature was summarized as below:

(1) No significant wind direction change was found due to the plant construction and plant operation (anthropogenic heat release)

(2) The horizontal wind speed increases by 0.5~1 m/s due to the plant construction and further increases by 1 m/s due to the anthropogenic heat release. Thus, the horizontal wind speed increase by 1~2 m/s after the plant construction and operation.

(3) The ambient temperature at 2 m height increases by 1~2 ℃ after plant construction and further increases by 4 ℃ after plant operation. Thus, the overall ambient temperature could increase by 5~6 ℃ depending on the location after the plant construction and operation. It shall be noted that the release of anthropogenic heat is in the simulation model is different with that of actual plant because of the limitation of model setting in the weather simulation. It is considered that the impact of ambient temperature change could be less in the actual condition.

The impact of plant construction and operation on the ACHEs performance under the tropical background was revealed in this study. It is preferred to reveal the impact under other climate backgrounds.

References

(1) K. Takahashi, “Multi-Scale Weather/Climate Simulations with Multi-Scale Simulator for the Geoenvironment (MSSG) on the Earth Simulator,” Earth Science, (2006).

TSUBAME 共同利用　令和6年度　産業利用　成果報告書

利用課題名　クモ胚に基づく発生–進化の数理解析
英文：Mathematical modeling analysis of development and evolution of spider embryos

利用課題責任者
小田広樹

所属
JT 生命誌研究館
https://www.brh.co.jp/

邦文抄録（300 字程度）
節足動物は初期胚発生において体軸形成と体節形成を行う。本プロジェクトでは、クモ胚を参考にして、球体表面に細胞を配置させたプラットフォームを構築し、この球面プラットフォーム上で多細胞集団を細胞力学に基づいて挙動させて体軸形成を再現し、同時に、細胞間相互作用ネットワークを作動させて体節形成（周期縞パターン形成）を再現することを目指している。さらに発生過程に関わる様々な因子および因子間相互作用にランダム変異と適応度に基づく選択が作動する遺伝的仕組みを導入し、多世代に渡って発生-進化の試行実験を可能とする球面多細胞体プラットフォームの開発を試みている。本年度、ローカル環境で開発したいくつかのプログラムを TUBAME 環境でテストした。

英文抄録（100 words 程度）
Arthropods undergo body axis formation and segmentation during early embryogenesis. In this study, using spider embryos as a reference, we develop a spherical-shaped platform where virtual cells are arranged on the surface of a sphere. Using this computationla platform, we intend to reproduce body axis formation and segmentation by simulating the behavior of multicellular assembly and the network of cell-cell interactions. Furthermore, we develop the platform to achieve computational experiments testing multiple generations of development and evolution. This year, we tested some of our programs in the TSUBAME4.0 environment.

Keywords: arthropods, pattern formation, cell vertex model, mathematical modeling

背景と目的

動物は形態や発生様式などに基づいて門などの高次分類グループに分類される。それぞれのグループは特有のボディプラン（特徴的な体構造）を持っている。しかし、そのボディプランを実現する発生過程や分子的仕組みは同じグループの中でも系統や種によって様々に異なる。

このような生物種間の発生過程や仕組みの違いは実際の生物種を用いた実験に基づく比較解析によって検出されるが、何世代にもわたる進化の過程でどのように違いが生じるかを理解することは、現存の生き物を使った実験に基づく解析だけでは不可能である。この問題を克服しうる一つの方向性が、コンピュータ上で発生と進化を仮想的に実現する遺伝的仕組みを構築することである。

私たちは、この方向性の努力を推し進めるために、節足動物の分節化したボディプランの基礎となる体軸形成と体節形成に着目し、発生と進化の過程を再現する仮想多細胞体プラットフォームの開発を行っている。このプラットフォームではクモ胚を参考にして、球体表面に細胞を配置し、Cell vertex model で細胞力学に基づいて細胞集団を挙動させる。同時に、細胞間相互作用ネットワークを作動させて細胞フィールドにパターン形成が展開する設計となっている。さらに、ランダムな変異と適応度に基づく選択が作動する遺伝的システムを導入し、多世代に渡って発生進化の試行実験を行うことで、発生と進化の関係を解析する。

本年度、これまでローカル環境で開発してきたプログラムの一部を TSUBAME 環境でテストし、将来的な大規模計算に向けた準備を進めた。

概要

節足動物では多様な種が進化してきた。本研究課題では、節足動物の進化・多様化の過程を理解するために、クモ胚に基づいて独自に構築した多細胞体数理モデルを用いて、「節足動物らしさ」の基盤となる体軸と反復パターンを生む発生過程と進化過程のシミュレーション解析を行う。多細胞体モデルでは、細胞力学に基づいて細胞が振る舞い、遺伝子ネットワークの作用で細胞場にパターンが生じる。各世代の多数個体の遺伝子ネットワークに変異を与え、表現型としての形態とパターンに基づいて各個体の次世代への生き残りが決まる。シミュレーション解析に加え、ゲノムに基づくインフォマティクス解析も合わせて、発生と進化の関係を探究する。

結果および考察

1. シミュレーション解析

本年度は、私たちが構築したプログラムが TSUBAME 環境で動作するのかテストを行った。そして、個体の発生過程のシミュレーションは動作することが確認できた（図1）。ローカル環境では30分かかった計算時間が、TSUBAME 上では5分で計算が完了し、TSUBAME 環境での計算が有意義であることが確認できた。

次の段階の進化過程のシミュレーションでは、数理モデルのパラメータを変化させて進化過程を再現する。そのため、一部のクモ胚の発生過程で数値の発散が起こることがある。しかし、進化過程のシミュレーションは、TSUBAME 環境ではうまく動作することができなかった。原因を調査したところ、数値が発散して nan の値が出てきた時の例外処理のコードが、ローカル環境では動作していたが、TSUBAME 環境ではうまく動作していないと考えられた。TSUBAME 環境では、C++の isnan 関数が isnan(nan)で0を返し、nan を認識できておらず、数値が発散してしまうとシミュレーションが止まってしまっていた。そこで、icpx -fno-fast-math とコンパイルすることで TSUBAME 環境でも nan を認識させて数値計算を行うことで対処した。

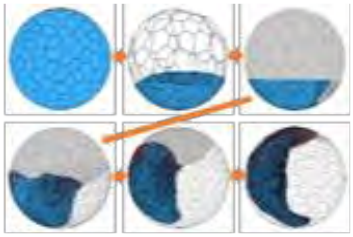


図1 TSUBAME 環境で数値計算した発生過程のシミュレーション

2. バイオインフォマティクス解析

多角的データ統合アプローチにより、シングル核および ATAC-seq データでの転写因子の可能性のある高発現蛋白質群と転写因子 Fuchi と複合体を形成する可能性について予測を行った。特に TSUBAME の計算リソースを活用して Boltz による大規模な複合体構造予測を実施し、Fuchi と相互作用する候補蛋白質のランキングを作成した（図 1）。統合的な解析の結果、GATA4(LOC107439682）や zinc finger protein ush（LOC107440842）、LOC107440646 などの複数の有力候補が挙げられた。

まとめ、今後の課題

本年度は、発生過程のシミュレーションを行うことができたが、進化過程のシミュレーションは、まだ途中段階である。今後、コンパイルの仕方や数値計算が発散した時の例外処理のプログラムコードを修正しつつ、発生過程と進化過程のシミュレーションの解析を行い、「節足動物らしさ」の基盤となる体軸と反復パターンを生む仕組みを予測することを目指す（図 2）。

クモ初期胚で働く転写因子 Fuchi の複合体形成相手の探索において、多角的データ統合アプローチを用いて候補を推定した。複合体可能性スコアが低く、信頼性のある Fuchi との複合体が予測できていない。信頼性のある結合相手を探索するために、別の手法で結合相手蛋白質を予測するといったことも考えられた。

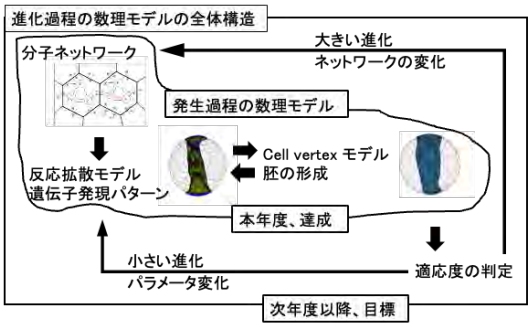


図2 進化過程のシミュレーションの全体構造

TSUBAME 共同利用　令和6年度　産業利用　成果報告書

利用課題名　自動運転 AI の研究開発
英文：R&D for automotive AI

利用課題責任者　関川雄介

所属　デンソーアイティーラボラトリ
URL https://www.d-itlab.co.jp/

邦文抄録（300 字程度）
車載 AI プロセッサは電力・演算性能が限られることから軽量な DNN モデルが求められる。一方で，近年の高精度な DNN モデルはパラメータ数が増大し演算量が増加する傾向にある。そこで車載 AI に適用できうる DNN 圧縮技術を確立する必要がある。本研究課題では特にニューラルネットワークの極低ビット（2bit~4bit）への量子化技術に着目し、その有効性に関する検討を進めた。これまでにも様々な量子化手法が提案されているが、それらの有効性は共通の実験条件下で十分に比較・評価は行われてこなかった。そこで画像分類問題において、有望な既存の量子化手法を公平な実験条件下でベンチマーク評価を行うとともに、精度劣化を抑えるために重要な要素を明らかにした。

英文抄録（100 words 程度）
Lightweight DNN models are essential for automotive AI processors due to constraints in power and computational resources. However, recent high-accuracy DNNs tend to be larger and more computationally intensive. This highlights the need for effective DNN compression techniques suitable for automotive AI applications.

This study focuses on extremely low bit quantization (2 to 4bits) that reduces memory size and computational complexity. Although various quantization methods have been proposed, their performance has not been thoroughly compared under the identical experimental conditions. To address this, we conducted a benchmark evaluation of promising existing quantization techniques for image classification under fair settings, while also clarifying the essential components to suppress accuracy degradation.

Keywords: 5つ程度
車載 A I、DNN、モデル圧縮、量子化、低ビット

背景と目的

車載 AI プロセッサなどのエッジ端末では電力・演算性能が限られることから軽量な DNN モデルが求められる。一方、近年の高精度な DNN モデルはパラメータ数が増大し演算量が増加する傾向にある。そこで、エッジ端末に適用できる DNN 圧縮技術を確立を目指す。本プロジェクトでは、極低ビットでの様々な量子化手法のベンチマークを通して、精度を担保するために必要な技術を炙り出し、さらに精度劣化を抑えるための検討材料を揃える。

概要

ニューラルネットワークの低ビット化は、浮動小数点の学習済み模型に対して重みと活性化関数に量子化関数を差し込むことで実現できる(図 1)。

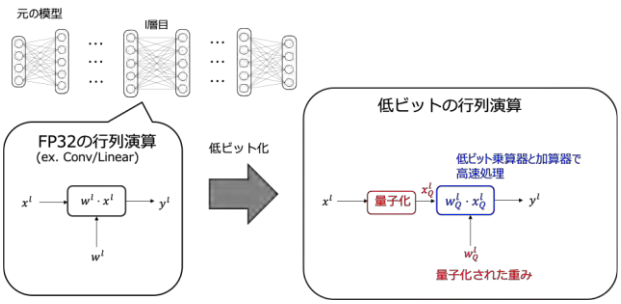


図 1:ニューラルネットワークの低ビット化

これにより、ニューラルネットワークの演算の高負荷箇所である行列演算は整数演算で実行でき、高速な処理が可能になる。

一方、極低ビット(2~4 ビット)では量子化誤差が大きくなるため、単純に量子化関数を差し込むだけでは著しく

精度劣化を起こす。

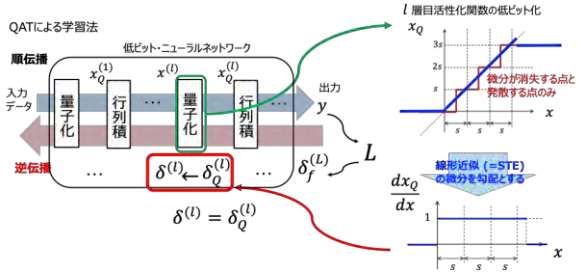


図 2:低ビットニューラルネットワークの学習方法

そこで、精度劣化を防ぐために極低ビットでは、元の全ての学習用のデータセットを用いて End-to-End で再学習する手法である Quantization-Aware-Training(QAT)が発展してきた(図 2)。QAT では順伝播時は重みと活性化関数を低ビット化したネットワークを通して出力するのに対して、逆伝播時はその低ビット化の離散的な振る舞いのために有限値を持つ勾配を通すことができないという課題がある。そこで逆伝播で勾配を通すために、量子化関数の clip の内側はそのまま微分を通す Straight-through-estimator(STE)と呼ばれる近似的な手法がしばしば用いられてきた。

STE の DNN への応用は 2012 年に Hinton による講義[1]でその原型となるアイデアが提唱されて以来、幅広く発展してきた。とりわけ成功を収めている手法としては、重みだけではなく、量子化関数の間隔幅(ステップ幅)自体も STE を用いて学習することでより高い表現力を獲得するアプローチが挙げられる(PACT[2],LSQ[3])。さらに、それらの手法の不等間隔幅への拡張したアプローチ(LCQ[4],nuLSQ[5])などがあげられる。ステップ幅を学習する手法の中でも STE

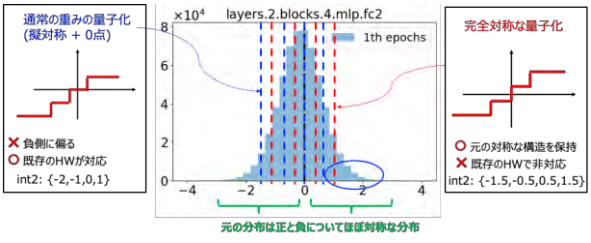


図 3:重みの分布の構造と量子化関数の違い

を関数全体に使う(PACT)か、あるいは部分的に使う(LSQ)かによって勾配の近似方法が異なり、学習後の精度に影響を及ぼす。加えて、重みの量子化手法についても大きく分けて3種類の方式が存在する：0 値をレベル数に含めつつ、負側の量子化レベルを正側より一つ多くとるほぼ対称な方式(LSQ[3])で実用的に用いられている方式，元の浮動小数点の分布の対称性を尊重して、0 値をレベル数に持たない完全に対称な量子化(UniQ[6],N2UQ[7],LQ-Nets[8])の方式や 0 値をレベル数にもったまま完全に対称な構造を保つために標準的な量子化から負側のレベル数を一つ削る量子化ノ方式(LSQ-SignMag/APoT[9]/LCQ[4])がある。それらに加えて、各手法ごとに学習を安定化させる独自の勾配を調整する技術や量子化関数の初期化法も併せて開発されてきており、精度劣化を抑えるためにどの要素が支配しているのが判別できない状況となっている。

そこで本研究では精度に影響を及ぼす可能性が高い構成要素(1. 重みの学習率と減衰の設定 2. ステップ幅の勾配調整法 3. ステップ幅の初期化法 4. ステップ幅に対する学習率と減衰の調整)ごとに精度を評価した上で、高精度を実現する構成要素の組み合わせのもとで最後に代表的な量子化手法のベンチマーク評価を画像分類問題で行った。データセットしては、Imagenet を用いて、模型としては Swin-T で評価を行なった。また、オプティマイザーについては学習済み模型で用いた AdamW を採用した。各構成要素の精度評価は Imagenet の訓練データの 9 割を 15epoch 学習して残りの 1 割をヴァリデーション・データとして行った。そして最終的なベンチマーク評価では全ての訓練データを用いて 90epoch を学習して公式の validation 用のデータセットを用いてテスト精度評価を行った。代表的な量子化手法として等間隔量子化手法である LSQ, UniQ,LSQ-SignMag, N2UQ と不等間隔量子化手法である APoT, LQ-Nets, LCQ, nuLSQ を評価した。

結果および考察

ここでは 2~4bit の量子化のうち、各手法の差異が最も顕著になる 2bit の結果について報告する。まず、各構成要素の精度評価について報告する。

1. 重みの学習率と減衰の設定

各手法による評価の偏りを減らして評価の公平性を保つために、最終評価で用いる量子化手法を用いずに、各層の入力・重みごとに MSE でステップ幅を決める単純な量子化手法(MSEQ)を用いてグリッド探索で重み

の学習率と減衰を決定した(表 1)。

表 1:MSEQ での重みの学習率(lr)と減衰(wd)のグリッド探索

lr \ wd	0.00625	0.0125	0.025	0.05
0.001	69.241	69.269	69.091	68.881
0.0001	72.353	72.224	**72.498**	72.321
0.00001	64.726	64.813	64.744	64.721

2. ステップ幅の勾配調整法

ステップ幅は各層ごとに共通のものを用いるため、その層の出力数と大きく乖離がある。そのため、連鎖律に従って逆伝播で STE 勾配を計算するとステップ幅の勾配量は大きくなり、また層ごとに不均一な大きさになる[3]。それを解消するためにこれまで提案されてきた技術としては、層のパラメータとビット数に依存する経験的に決定した定数でスケールさせる勾配スケール法(GS)や重みの分布を正規化したのち量子化する方法がある。さらに後者にも、重みの分布の標準偏差で正規化する方法(LWN[4])やエントロピーが大きくなるように正規化する方法(EPWR[7])がある。また今回は AdamW を optimizer として用いるため、そもそも勾配の大きさがリスケールされて、自然と勾配の大きさが調整される側面がある。そのため、上記の勾配調整法の評価に加えて、勾配調整法を用いない場合の評価も行なった。その結果、精度の差異は小さいものの、表 2 に示すように標準偏差で分布を正規化する LWN が最も良い精度を示した。

表 2: LSQ でのステップ幅の勾配調整法の比較

勾配調整法	GS+AdamW	LWN+AdamW	EPWR+AdamW	AdamW
val精度	73.18	**73.38**	73.20	73.24

3. ステップ幅の初期化法

ステップ幅の初期化も各手法ごとに様々なものが用いられてきた。定数値で初期化する手法や、分散を用いて初期化する方法(LSQ 初期化)、量子化誤差の平均二乗誤差を最小にするように初期化する方法(MSE 初期化)もある。また、MSE の初期化をさらに簡単化した NMSE 初期化も提案されている[6]。これらのステップ幅の初期化法について LSQ で評価比較を行なった。表 3 に示すように、定数初期化は他の初期化法に比べて著しい精度劣化を示し、MSE 初期化が最も精度が良い結果となった。

表 3: LSQ でのステップ幅の初期化法の比較

初期化法	定数初期化	LSQ初期化	NMSE初期化	MSE初期化
val精度	41.47	72.57	72.75	**73.38**

4.各量子化手法の量子化パラメータに対する学習率と減衰の調整

1~3 で選ばれた精度の最も良い組み合わせを用いて、各量子化パラメータについての学習率と減衰をグリッド探索した。量子化パラメータとは例えば LSQ や UniQ ではステップ幅、PACT では閾値、nuLSQ では複数のレベル数といった手法ごとに存在する量子化関数を形成するための学習パラメータを意味する。

1~4 で選ばれた量子化関数の勾配調整法(LWN)と初期化法(MSE)、ならびに重みと量子化パラメータの学習率と減衰を用いて、各量子化手法の精度評価を行なった(表4)。その結果、等間隔量子化手法の中では LSQ

表 4: 量子化手法のベンチマーク評価

	quantizer	SwinT-tiny[蒸留] 80.916/95.422	SwinT-small[蒸留なし] 83.050/96.864
等間隔	1. LSQ	76.172/92.908	77.874/94.180
	2. UniQ	**76.378**/92.984	**78.154**/94.168
	3.LSQ-SignMag	75.468/92.444	77.376/93.836
	4. PACT	75.318/92.284	76.836/93.682
	5. N2UQ	74.728/92.256	77.812/94.076
不等間隔	6. APoT	74.832/92.171	76.856/93.556
	7. LQ-Nets	75.282/92.508	77.590/93.942
	8. LCQ	75.899/92.805	77.906/94.034
	9. nuLSQ	**76.698**/93.142	**78.310**/94.288

よ UniQ が高精度を示した。特筆すべき点として、比較手法の中で最も新しい手法である N2UQ は、比較対象の中で最も高精度と原論文中で報告されていたが、本研究による統一的な実験環境下の再実験により、むしろ従来手法である LSQ や UniQ の方が高い精度を達成していたことが明らかになった。さらに LSQ と UniQ を比較すると、僅差ながら UniQ が高精度な結果を示している。これは先述の通り、学習済みの重みの分布の構造を反映した完全に対称な量子化構造の方が、量子化誤差を下げることができ、最終的な精度も良い結果となることを意味する(図 3)。一方、対称性を保ちつつ 0 値を量子化レベルを含める LSQ-SignMag は精度劣化を引き起こしている。これは負側のレベル数を一つ削減することによって、表現力が下がり、最終的な精度に悪影響を及ぼした結果と考えられる。

さらに不等間隔量子化手法でも同様に各手法の比較を

行った。その結果、LSQ をベースとして不等間隔に拡張した nuLSQ が、他の不等間隔手法と比較して最も高い精度を達成した。

まとめ、今後の課題

本研究では、既存の量子化手法を共通の実験条件したでベンチマーク評価を実施した。その結果、量子化関数を STE で学習する研究の中で初期の手法である LSQ とそれの重みの完全対称版の UniQ、そして不等間隔版の nuLSQ が高精度な結果が得られた。

今後は、本研究成果を受けてさらに精度劣化を抑える手法の検討をする。現時点では以下の3つの方向性が有望であると考えている。I. nuLSQ の完全対称版への拡張を行う、II. 層ごとに LSQ と UniQ を適応的に選択する、III.層ごとに用いている共通のステップ幅をチャンネルごと/トークンごと/ヘッドごとなど、高速な整数演算の処理の要請を満たす範囲で、量子化粒度を変える。

参考文献

[1] Hinton, G. (2012). Neural networks for machine learning. Coursera, video lectures

[2] Choi, J., Wang, Z., Venkataramani, S., Chuang, P.I.J., Srinivasan, V., Gopalakrishnan, K.: Pact: Parameterized clipping activation for quantized neural networks. International Conference on Learning Representation (2018)

[3] Esser, S.K., McKinstry, J.L., Bablani, D., Appuswamy, R., Modha, D.S.: Learned step size quantization. In: International Conference on Learning Representations (2019)

[4] Yamamoto, K.: Learnable companding quantization for accurate low-bit neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2021)

[5] Gongyo, S., Liang, J., Ambai, M., Kawakami, R., & Sato, I. (2024). Learning Non-uniform Step Sizes for Neural Network Quantization. In Proceedings of the Asian Conference on Computer Vision (2024)

[6] Pham, P., Abraham, J.A., Chung, J.: Training multi-bit quantized and binarized networks with a learnable symmetric quantizer. IEEE Access 9 (2021)

[7] Liu, Z., Cheng, K.T., Huang, D., Xing, E.P., Shen, Z.: Nonuniform-to-uniform quantization: Towards accurate quantization via generalized straight-through estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2022)

[8] Zhang, D., Yang, J., Ye, D., Hua, G.: Lq-nets: Learned quantization for highly accurate and compact deep neural networks. In: Proceedings of the European conference on computer vision (2018)

[9] Li, Y., Dong, X., Wang, W.: Additive powers-of-two quantization: An efficient nonuniform discretization for neural networks. In: International Conference on Learning Representations (2020)

TSUBAME 共同利用　令和6年度　産業利用　成果報告書

利用課題名　EEG を用いた発話解読 AI の開発
英文：Development of EEG to Speech BCI Decoder

利用課題責任者　金井良太
First name Surname Ryota Kanai

所属　株式会社アラヤ
Affiliation Araya Inc.
URL www.araya.org

邦文抄録（300 字程度）

我々は、EEG を用いた発話解読 AI の実用化に向けた二つの重要な進展を報告する。1 つ目は、269 時間の EEG・音声のペアデータで学習した EEG エンコーダに対しゼロショット音声分類を行い、61.3%の精度を達成。データ量が精度に大きく影響することを示した。2 つ目は、被験者に課したタスクや電極配置が異なる EEG/EMG データをマルチタスク学習で処理するニューラルネットを提案し、EEG/EMG から silent speech で発話した単語を予測した際の正答率が健常者で 95.3%、神経変性症により発話困難な患者で 54.5%の精度を実現。特に健常者のデータから患者への転移学習により、単一被験者学習（健常者 70.1%, 患者 13.2%）を大きく上回る性能を達成した。どちらも発話障害者支援の実用化に向けた有望な成果である。

英文抄録（100 words 程度）

We report two key advances toward practical EEG to speech BCI. First, zero-shot speech classification with 269 hours of EEG data achieved 61.3% accuracy, demonstrating a strong correlation between data volume and performance. Second, we propose a neural network that processes EEG/EMG data collected under different tasks and electrode configurations via multi-task learning, achieving 95.3% accuracy in healthy subjects and 54.5% in a speech-impaired patient with a neurodegenerative disorder. Notably, transfer learning from healthy participants to the patient significantly outperformed single-subject training (70.1% and 13.2%). These findings represent promising steps toward real-world BCI systems to support individuals with neurodegeneration-related speech impairments.

Keywords: brain-machine interface, transfer learning, EEG, speech recognition

背景と目的

ALS 等の神経変性症や、脳卒中後遺症により発声が困難な患者のコミュニケーションを補助する技術として、脳波(EEG)や筋電（EMG）を用いた発話解読のブレイン・コンピュータ・インターフェース（BCI）技術が注目されている。従来の侵襲的手法（ECoG, マイクロ電極）による発話解読 BCI は、高精度ながらも、外科的な手術を伴うため、感染症のリスクや心理的負担が大きい。そのため、非侵襲な代替手段として EEG/EMG による高精度な発話解読手法が求められる。

近年、言語モデルを中心に、データの量と AI モデルの精度に相関（スケーリング則)があることが知られているが、生体信号でノイズが大きくかつ、個人差や個人内でも日毎の差がある EEG においても、同様のスケーリング則が成立するかは不明であった。また、EEG は取得コストが高く、患者の訓練用のデータを可能な限り少なくすることが実用的な BCI の開発には不可欠である。これら二点を検証するために、我々は、記録した大規模な発話中の EEG データを用いて、精度とデータ量間のスケーリング則の検証、並びに、多様な電極配置を扱うモデルを用意して、複数種の EEG/EMG デバイスで記録された大規模な健常者データから、神経変性症により発話が困難な患者データへの転移学習とその評価を行なった。

概要

1. 発話解読 BCI のスケーリング則評価概要

EEG を用いた発話解読 BCI のスケーリング則を評価するため、学習に使用するデータの総量を変化させながら以下のように学習と評価を行なった。発話中の EEG と音声のペアデータを 5 秒間のセグメントに分割

し、各セグメントの EEG の特徴量とペアとなる音声の特徴量（wav2vec 2.0 を使用）を類似させるように EEG エンコーダ（Conformer モデル）の対照学習（CLIP）を行なった(図 1)。モデルの性能指標として、テストデータに対するゼロショット分類性能(バッチサイズ=512)を用いた。本実験における実験参加者は健常者 1 名で、EEG 取得には 2 種類の脳波計測デバイス(g.PANGOLIN. g.SCARABEO)を用い、それぞれ独立にスケーリング則の評価を行った。

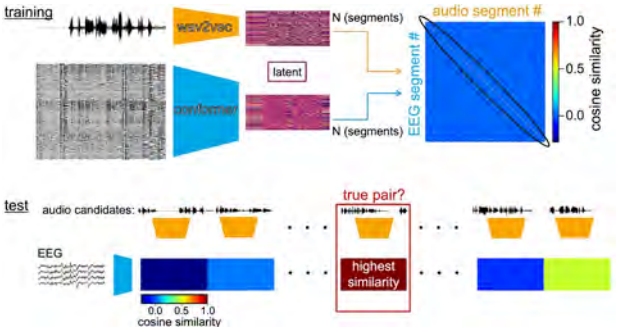


図 1 EEG から音声特徴量を予測するモデルの学習・評価のフレームワーク

2. 健常者から患者への転移学習評価概要

EEG/EMG を用いた発話解読 BCI の健常者から患者への転移学習可能性を検証するために、単語発話中の EEG/EMG から有声/無声発話した単語の分類（総単語数 64）の学習を以下の 3 条件で実施した。モデルは被験者毎の線形変換層を持つ Conformer モデルを用いた。本実験における健常者の実験参加者数は 8 名、患者の実験参加者数は 1 名であり EEG 計測デバイスには eego sports を用いた。

1. （被験者内）: 各被験者個人のデータでのみ学習（各健常者 1.9 時間, 患者 1 時間）。
2. 全被験者のデータを混合したデータセットで学習（合計 16.1 時間）。
3. 本実験で収集した全被験者のデータに加え、さらに、異なるタスク・異なる EEG 計測デバイスで記録された発話中 EEG データを追加したデータセットで学習(合計 220 時間)。

評価指標には、テストデータを用いて、無声発話中の 64 単語の分類の精度を用い、各条件で精度を比較した。

結果および考察

1. 発話解読 BCI のスケーリング則評価結果

図 2 に示すように学習データを 1 時間程度から 100 時間に増やしていていきそれぞれのデータ量でモデルの学習を行うと、データ量の増加とともにゼロショット分類精度が向上し、top-1 accuracy で 61.3%（512 フレーズ中）に達した。これは、リスニングタスク中の EEG と音声特徴量のゼロショット分類精度の SoTA (5%程度)[1]を大幅に超えるものであった。使用した2つのデバイスの両方ともデータ量の増加に伴い同程度の精度向上が確認された。この成果は BCI Award 2024 にノミネートされた[2]。

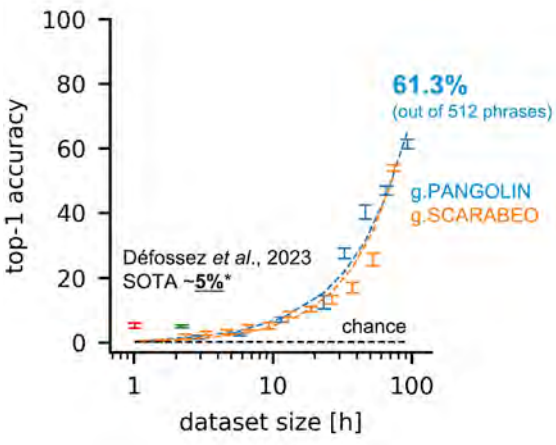


図 2 学習に用いた EEG・音声データサイズと zero-shot 分類性能の関係

2. 健常者から患者への転移学習評価結果

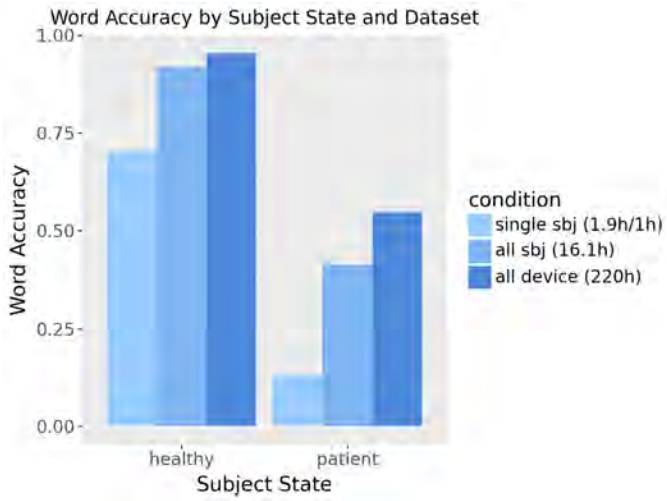


図 3 健常者・患者それぞれの、学習に用いたデータセットと単語分類精度

データセットを被験者内データから全被験者データ、異なるデバイスで記録された EEG を含むデータへと拡大していったときの、無声発話中の EEG/EMG から発話した単語分類タスクにおける健常者平均（N=8）と患者

(N=1)の分類精度を図 3 に示す。健常者、患者ともに、被験者内データのみを使う場合と比較して、全被験者、デバイス混合のデータセットで学習した場合の精度が高い。また、データセットの大きさによる精度向上の効果（13.2%から 54.5%）は、患者データにおいて顕著であった。以上のことは、異種デバイス間での EEG/EMG データの転移学習が可能であることと、健常者データから、健常者とは発声方法や電極配置が異なる患者データにおいても転移学習が可能であることを示すものである。

まとめ、今後の課題

我々は、大規模な EEG と音声データを用いて、EEG を用いた発話解読 BCI のスケーリング則を示し、また、多様な記録デバイス・タスクで大規模に収集した健常者の EEG/EMG データから、患者の EEG/EMG データへの転移学習が可能なことを示した。このことは、BCI のエンドユーザーのキャリブレーションに必要なデータ量を最小限に抑えた、実用的な BCI システムの開発の可能性を示唆するものである。

今後は、より多様で大規模な EEG データを用いた学習により、よりロバストな BCI の開発を目指す。

参考文献

[1] Défossez, A., Caucheteux, C., Rapin, J. *et al.* Decoding speech perception from non-invasive brain recordings. *Nat Mach Intell* **5**, 1097–1107 (2023).

[2] Sato A., Tomeoka K., Horiguchi I., Arulkumaran K., Kanai. R., and Sasai. S., Large-scale data practicalizes EEG-based speech decoding. *BCI Awards* (2024)

新聞ドメインにおける 大規模言語モデルの継続事前学習と下流タスクデータ量の関係

岸波 洋介 [1]　藤井 諒 [1]　森下 睦 [1]
[1] フューチャー株式会社
{y.kishinami.rh, r.fujii.6d, m.morishita.pi}@future.co.jp

概要

近年の大規模言語モデル (LLM) の発展により，様々なドメインに特化した LLM の研究が進んでいる．ドメイン特化 LLM を下流タスクに用いる場合，タスクのラベル付きデータで教師ありファインチューニングする前に，そのドメインの知識獲得を目指し，特定ドメインの生テキストを用いた継続事前学習が先行して行われることがある．しかしながら下流タスクのデータ量に着目した継続事前学習の有効性については不明な点が多い．本研究では新聞ドメインにおける見出し生成タスクを対象に，ドメイン特化を目的とする継続事前学習と下流タスクのデータ量が性能に与える影響を分析する．分析の結果，下流タスクのデータ量が極めて少ない状況では継続事前学習の効果が大きい可能性が示唆された．

1　はじめに

質問応答や文章校正など，様々な用途で大規模言語モデル (LLM) の活用が進んでいる [1, 2, 3]．特に企業での応用を見据えると，金融，医療などといった業界に特化した LLM のニーズは大きく，様々な業界で活用，研究が進んでいる [4, 5, 6, 7]．LLM の実応用に際しては，OpenAI 社が提供する ChatGPT のように，様々なタスクに汎用的に活用できるチャットボットのニーズも高い一方，解きたいタスクが明確に定まっている場合も一定存在する．

LLM を用いて解きたいタスクが明確に定まっている場合，解きたいタスクのラベル付きデータを用いて教師ありファインチューニング (SFT) を行うことが一般的である [8]．しかしながら，高いタスク性能を達成するためには大量のラベル付きデータが必要となる場合も存在する．タスクによって現実的に用意できるデータ量は異なり，大規模に用意するのは高コストであることも多い．そのため，ベースとなる LLM を特定ドメインの生テキストで追加学習することで，そのドメインの知識獲得を目指す継続事前学習が先行して行われることがある [5, 9, 10]．継続事前学習では，学習に用いるデータに対して正解ラベルを付与する必要がないため，比較的学習データを収集しやすいという特徴がある．しかしながら，実際に継続事前学習することがどのような状況下で有効か，特に下流タスクのデータ量に着目するといまだ不明な点が多い．

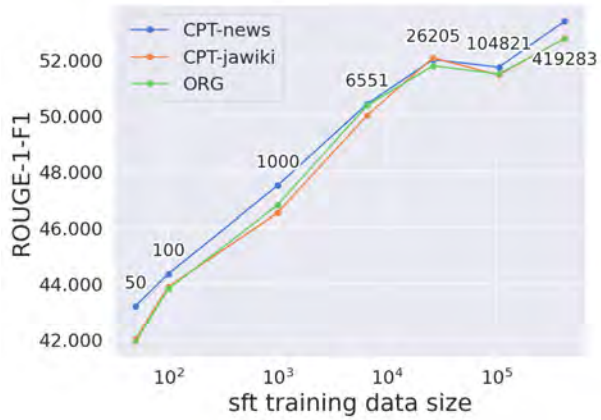


図 1　見出し生成タスクの学習データ量と性能 (ROUGE)．マーカー付近の値は見出し生成タスクの学習データ量を示す．

本研究では新聞ドメインにおける見出し生成タスクを対象に，タスクのデータ量に着目した場合の継続事前学習の有効性を検証した．多くの新聞記事は本文に対して見出しが付与された状態で公開され，本文と見出しの組を集めやすいことから，下流タスクのデータ量による継続事前学習の影響を調査する上で適したタスクであると考えることができる．本研究では，新聞ドメインへ特化させるための継続事前学習を行った上で，SFT に用いるデータ量による性能の差を分析する．分析の結果，SFT の学習データが極めて少ない状況下では継続事前学習の効果が大きい可能性が示唆された．

2 関連研究

2.1 ドメイン特化型 LLM

ChatGPT に代表される汎用的な LLM に対し，特定ドメインに特化した LLM も研究が進められている [4, 5, 6, 7]．金融分野では BloombergGPT [4] や FinGPT [11] など，様々な金融特化 LLM が構築されている [12, 13]．日本語でも，平野らが日本語 LLM に対して金融文書で継続事前学習を行い，金融特化 LLM を構築している [9]．また，医療分野でもドメイン特化の LLM に関する研究，開発が進められている [5, 14, 15, 16]．例えば Llama を医学論文で継続事前学習した PMC-LLaMA [5] や，Med-PaLM [14] などが挙げられる．日本語においても助田らが日本語 LLM に対し医療データで指示チューニングを行い，医療ドメインに特化した LLM を構築している [10]．本研究では新聞記事ドメインを対象に，継続事前学習と下流タスクのデータ量に着目した分析を行う．石原らも BERT や T5 に対し新聞記事で学習を行い，日本語の金融ニュース記事に特化したモデルを構築しているが [17]，本研究は近年発展の著しいより大規模なモデルを対象に，継続事前学習と下流タスクのデータ量に着目した分析を行う．

2.2 ドメイン特化の継続事前学習の分析

最近では，継続事前学習の有効性に関する分析も多数行われている [18, 19, 20]．Yıldız らは LLM の継続事前学習に関して，複数ドメインで構成されるデータセットを用いて，学習するドメインの内容や学習順序に着目した分析を行っている [18]．また，Xie らは金融ドメインを対象に継続事前学習に有用なデータの選択手法を提案し，それらの手法を用いた継続事前学習の有効性を分析している [19]．実験の結果，提案手法によって少ないデータでも効率的に継続事前学習を行えることを示している．Jindal らは継続事前学習の有効性に関して，学習元とするモデルの選択に着目した分析を行っている [20]．実験では継続事前学習のデータ量を変化させた場合の分析も行っており，指示追従データで学習済みのモデルを学習元とした継続事前学習ではデータ量が増えるほど指示追従能力が劣化することを示している．本研究では特に下流タスクのデータ量に着目し，継続事前学習の有効性を分析する．

3 分析 1: 継続事前学習と下流タスクデータ量の関係

本研究では山田ら [21] の定義に倣い，与えられたテキスト全体のトークンを予測する学習方法を継続事前学習，ラベル側のトークンのみ予測する学習方法を教師ありファインチューニング (SFT) と呼ぶ．

新聞ドメインにおける見出し生成タスクを対象に，継続事前学習の有無と下流タスクでの SFT におけるデータ量の関係を分析する．

3.1 実験設定

データセット 継続事前学習，および見出し生成タスクの学習データとして，信濃毎日新聞株式会社が発行した 2013 年 1 月から 2024 年 6 月までの紙面記事の本文，見出しを使用した．連載名の接頭辞などのノイズを除去し，掲載時期に応じてデータを分割した．最終的に学習データとして 2013 年 1 月から 2024 年 4 月までの記事 419283 件，検証データとして 2024 年 5 月の記事 2385 件，評価データとして 2024 年 6 月の記事 2574 件を用意した．このうち継続事前学習には学習データの記事本文，計約 1.85 億トークンを用いた．また，比較対象として新聞以外のドメインの生テキストも用意した．具体的には llm-jp-corpus-v3[1] の日本語 Wikipedia から新聞記事と同量のトークンを抽出した．なお，見出し生成タスクへの SFT では本文，見出しの組を用いた．

継続事前学習 Llama-3.1-Swallow-8B-v0.1[2] をベースモデルとし，新聞記事本文，および日本語 Wikipedia のそれぞれを用いて継続事前学習を行った．なお，学習は 2 エポック行った．

見出し生成タスクへの SFT Llama-3.1-Swallow-8B-v0.1 (ORG)，新聞記事本文で継続事前学習を行ったモデル (CPT-news)，日本語 Wikipedia で継続事前学習を行ったモデル (CPT-jawiki) の 3 つをベースモデルとし，見出し生成タスクへの SFT を行った．SFT 手法として Low Rank Adaptation (LoRA) [22] を用いた．学習に用いるデータ量は複数の設定を用意した．具体的には，全量 (419283)，4 分の 1 (104821)，16 分の 1 (26205)，64 分の 1 (6551)，1000, 100, 50 の計 7 設定を用意し，それぞれで SFT を行った．学習設定の詳細は付録 A に示す．

1) https://gitlab.llm-jp.nii.ac.jp/datasets/llm-jp-corpus-v3
2) https://huggingface.co/tokyotech-llm/Llama-3.1-Swallow-8B-v0.1

表 1　信濃毎日新聞 2024 年 6 月 27 日の記事に対する生成例．\n は改行を示す．

本文	信州大（本部・松本市）は２６日、２０２５年度入学者の選抜要項を発表した。２２年度から実施された高校の新学習指導要領に基づく新教育課程に対応した教科・科目に変更し、一般選抜（前期・後期）では旧教育課程の履修者が不利にならないよう出題する。８学部全体の入学定員は１９６３人で、前年度に引き続き医学部医学科の定員１５人増を国に申請する。\n　医学科の定員増は医師不足に対応する措置で、認められれば入学定員は前年度と同じ１９７８人。定員増の申請は１０月ごろに結果が分かる。\n　農学部農学生命科学科は２４年度までの４コースを「生命・食品科学」「食料生産システム科学」「山岳圏森林・環境共生学」の３コースと「地域協創」の１特別コースに再編。分野横断型教育を強化し、課題解決型学習を推進する。同学科の入学定員は前年度と同じ１７０人。\n　教育学部は５コース（現代教育、国語教育、ものづくり・技術教育、特別支援教育、心理支援教育）で総合型選抜を初めて導入する。募集人員は一般選抜や学校推薦型選抜を削り、１６人分の枠を設けた。卒業後に県内で教職に就く意欲を持った人が対象で、大学入学共通テストは課さない。\n　一般選抜の出願期間は来年１月２７日～２月５日。信大ホームページから出願登録サイトに必要事項を入力し、検定料を支払った上で出願書類などを郵送する。前期試験は２月２５日（一部学部は２６日も）、後期試験は３月１２日。問い合わせは平日に信大入試課（☎０２６３・３７・２１９５）へ。\n
正解見出し	信大　２５年度選抜、新学習指導要領に対応　旧課程にも配慮
ORG	信州大、２５年度入学者の選抜要項発表　医学部医学科の定員１５人増を申請
CPT-jawiki	信州大、２５年度入学者の選抜要項発表　医学部医学科の定員１５人増を申請
CPT-news	信大の２５年度入学者選抜要項発表　医学部医学科の定員１５人増を申請　教育学部は総合型選抜導入

表 2　見出し生成タスクの学習データ量を変化させた際の評価結果 (ROUGE)．横軸は学習データの件数を示す．

	50	100	1000	6551	26205	104821	419283
ORG	41.93	43.81	46.81	50.37	51.79	51.50	52.76
CPT-jawiki	42.02	43.89	46.53	50.02	**52.08**	51.46	52.78
CPT-news	**43.19**	**44.35**	**47.50**	**50.42**	52.02	**51.74**	**53.39**

評価指標　見出し生成タスクの評価指標として ROUGE-1-F1 [23] (ROUGE) と BERTScore-F1 [24] (BERTScore) を用いた．ROUGE の算出には sacrerouge[3] を用い，単語の分割には形態素解析器 MeCab [25] (IPA 辞書) を使用した．BERTScore の算出には bert-score ライブラリ[4] を用い，モデルには tohoku-nlp/bert-large-japanese-v2[5] を使用した．

3.2　実験結果

図 1 および表 2 に，見出し生成タスクの学習データ量を変化させた際の各モデルのスコアの推移を示す[6]．図 1 から，タスクの学習データ量によらず基本的に CPT-news の性能が高くなっていることがわかる．また表 2 から，特に学習データが 50 件の場合に，CPT-news の性能とそれ以外のモデルの性能の差が大きくなっていることがわかる．このことから，見出し生成タスクにおいては，タスクの学習データ量が極めて少ない状況において継続事前学習の効果が大きい可能性が示唆される．この傾向は，ROUGE だけでなく BERTScore の値についても一貫して確認された[7]．

3.3　定性分析

表 1 にタスクの学習データが 50 件の場合の各モデルの見出し生成例を示す．この例では，CPT-news のみ正解見出し同様に「信大」という略称を生成できていた．この事例について 10 回の生成を行ったところ，CPT-news は 10 回全てで「信大」と生成したのに対して，ORG では 2/10 回，CPT-jawiki では 3/10 回の生成にとどまっていた．継続事前学習に用いたデータセットに「信大」という表現がどの程度出現するかを調査したところ，新聞記事では全体の 1.25%，日本語 Wikipedia では 0.14%の文書に出現していることがわかった．この結果から，新聞記事で継続事前学習を行うことで，この表現に代表される対象ドメインらしい表現を生成する傾向が強まったと考えられる．

4　分析 2: 継続事前学習における学習データのドメイン混合

3 節の結果から，見出し生成タスクの学習データが 50 件のように極めて少ない場合に継続事前学習の効果が大きいことが確認できた．継続事前学習に用いる学習データはラベルが不要であるため，比較的データを収集しやすい．しかしながら，ドメインによっては生テキストの収集すら困難なケースも考えられる．例えば企業独自の知識や用語に関する

3）https://github.com/danieldeutsch/sacrerouge
4）https://github.com/Tiiiger/bert_score
5）https://huggingface.co/tohoku-nlp/bert-large-japanese-v2
6）過学習の傾向が見られたため，継続事前学習の 1 エポック目のチェックポイントから SFT を行った．
7）BERTScore の結果は付録 B に示す．

ドメインでは，その企業特有の情報が精緻に文書化されていなければ大規模に収集することは難しい．一方で llm-jp-corpus など，一般公開されているコーパスであれば低コストで扱いやすい．そこで本節では，下流タスクのドメインのデータとそれ以外のドメインのデータの混合データで継続事前学習した場合の性能について分析する．

4.1 実験設定

データセット 新聞ドメイン，他ドメインのデータを混ぜ合わせたデータセットを用意した．新聞ドメインの学習データには 3 節で用いた新聞記事本文約 1.85 億トークン，他ドメインの学習データには 3 節で用いた日本語 Wikipedia を同量使用した．合計トークン数が約 1.85 億トークンとなるように，全量を新聞ドメインにした場合，新聞ドメインを 2/3，1/2，1/3，1/10，1/20 含めた場合，全量を他ドメインにした場合の計 7 設定を用意した．

継続事前学習 7 つの学習データそれぞれを用いて，LLM の継続事前学習を実施した． 3 節と同様に Llama-3.1-Swallow-8B-v0.1 をベースモデルとし，学習は 2 エポック行った．

見出し生成タスクへの SFT 7 つの学習データで継続事前学習を行った各モデルに対し，見出し生成タスクへの SFT を行った．見出し生成の学習データは 3 節の分析において継続事前学習の効果が最も大きくなっていた 50 件とした．

評価指標 3 節と同様に見出し生成の評価指標には ROUGE-1-F1 (ROUGE) と BERTScore-F1 (BERTScore) を用いた．

4.2 実験結果

図 2 および表 3 に継続事前学習の学習データにおける新聞記事の割合を変化させた際のスコアの推移を示す[8]．図 2 から，新聞ドメインを 1/10 含めた場合と 1/3 含めた場合との間にスコアの差が生じていることがわかる．この傾向は ROUGE だけでなく BERTScore の値についても一貫して確認された[9]．このことから，下流タスクと同一のドメインと他のドメインのデータを一定割合以上で混ぜ合わせると性能が急激に向上する可能性が示唆される．本実験で用いた Llama-3.1-Swallow-8B-v0.1 は既に日本語 Wikipedia を用いた学習が行われている．そのため，

8） 3 節と同様に継続事前学習の 1 エポック目のチェックポイントから SFT を行った．
9） BERTScore の結果は付録 B に示す．

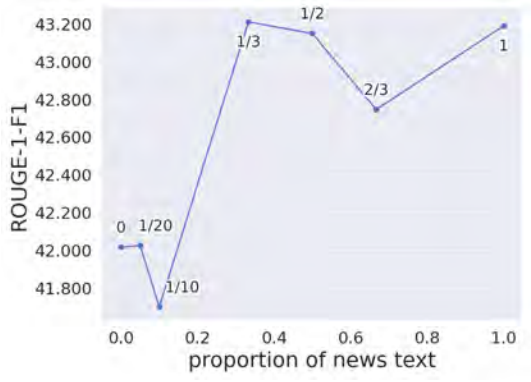


図 2 継続事前学習の学習データにおける新聞記事の割合を変化させた際の性能 (ROUGE)．マーカー付近の値は新聞記事の割合を示す．

表 3 継続事前学習の学習データにおける新聞記事の割合を変化させた際の評価結果．横軸は新聞記事の割合を示す．

	0	1/20	1/10	1/3	1/2	2/3	1
ROUGE	42.02	42.03	41.70	**43.21**	43.15	42.75	43.19
BERTScore	0.679	0.681	0.678	0.684	**0.685**	0.682	0.684

混合データによる学習が過去に学習した言語能力を一定保持しつつ新規ドメインに適用するような効果を発揮している可能性も考えられる．このような過去に学習したドメインとの類似性に着目した混ぜ合わせの分析については今後の課題としたい．

5 おわりに

本研究では新聞ドメインにおける見出し生成タスクを対象に，継続事前学習と下流タスクの学習データ量の関係を調査した．分析の結果，見出し生成タスクにおいては，SFT の学習データが極端に少ない場合に継続事前学習の効果が大きい可能性が示唆された．また，継続事前学習に用いる学習データにおいて見出し生成タスクと同一ドメインのデータの割合を変化させたところ，継続事前学習の学習データに他のドメインのデータが混合している場合にもタスクの性能向上につながるケースが確認された．

本研究はあくまで新聞ドメインの見出し生成タスクに着目した分析となっており，他のドメイン，タスクに対しても同様の傾向があることまでは示すことができない．したがって今後は金融や特許などの他ドメインや，それらの下流タスクでも分析し，ドメインやタスクによらず同様の傾向がみられるのかを調査したい．また，学習元とするモデルの種類やモデルサイズによる影響も考えられるため，これらを変更した際の分析も行いたいと考えている．

謝辞

本研究の実施にあたり，データを提供いただいた信濃毎日新聞株式会社の皆様に感謝申し上げます．また，本研究成果の一部は，データ活用社会創成プラットフォーム mdx および東京科学大学のスーパーコンピュータ TSUBAME4.0 を利用して得られたものです．

参考文献

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL, pp. 4171–4186, 2019.

[2] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In arXiv:2005.14165, 2020.

[3] OpenAI. GPT-4 technical report. In arXiv:2303.08774, 2024.

[4] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. BloombergGPT: A large language model for finance. In arXiv:2303.17564, 2023.

[5] Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. PMC-LLaMA: Towards building open-source language models for medicine. In arXiv:2304.14454, 2023.

[6] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science. In arXiv:2211.09085, 2022.

[7] Di Zhang, Wei Liu, Qian Tan, Jingdan Chen, Hang Yan, Yuliang Yan, Jiatong Li, Weiran Huang, Xiangyu Yue, Wanli Ouyang, Dongzhan Zhou, Shufei Zhang, Mao Su, Han-Sen Zhong, and Yuqiang Li. ChemLLM: A chemical large language model. In arXiv:2402.06852, 2024.

[8] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In Proceedings of NeurIPS, pp. 27730–27744, 2022.

[9] Masanori Hirano and Kentaro Imajo. Construction of domain-specified Japanese large language model for finance through continual pre-training. In arXiv:2404.10555, 2024.

[10] 助田一晟, 鈴木雅弘, 坂地泰紀, 小寺聡. JMedLoRA：instruction-tuning による日本語大規模モデルの医療ドメイン適用. 言語処理学会年次大会発表論文集, pp. 2548–2553, 2024.

[11] Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. FinGPT: Open-source financial large language models. In arXiv:2306.06031, 2023.

[12] Boyu Zhang, Hongyang Yang, and Xiao-Yang Liu. Instruct-FinGPT: Financial sentiment analysis by instruction tuning of general-purpose large language models. In arXiv:2306.12659, 2023.

[13] Thanos Konstantinidis, Giorgos Iacovides, Mingxue Xu, Tony G. Constantinides, and Danilo Mandic. FinLlama: Financial sentiment classification for algorithmic trading applications. In arXiv:2403.12285, 2024.

[14] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge. In arXiv:2212.13138, 2022.

[15] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. Towards expert-level medical question answering with large language models. In arXiv:2305.09617, 2023.

[16] Yizhen Luo, Jiahuan Zhang, Siqi Fan, Kai Yang, Yushuai Wu, Mu Qiao, and Zaiqing Nie. BioMedGPT: Open multimodal generative pre-trained transformer for biomedicine. In arXiv:2308.09442, 2023.

[17] 石原祥太郎, 村田栄樹, 中間康文, 高橋寛武. 日本語ニュース記事要約支援に向けたドメイン特化事前学習済みモデルの構築と活用. 自然言語処理, Vol. 31, No. 4, pp. 1717–1745, 2024.

[18] Çağatay Yıldız, Nishaanth Kanna Ravichandran, Prishruit Punia, Matthias Bethge, and Beyza Ermis. Investigating continual pre-training in large language models: Insights and implications. In arXiv:2402.17400, 2024.

[19] Yong Xie, Karan Aggarwal, and Aitzaz Ahmad. Efficient continual pre-training for building domain specific large language models. In Proceedings of ACL, pp. 10184–10201, 2024.

[20] Ishan Jindal, Chandana Badrinath, Pranjal Bharti, Lakkidi Vinay, and Sachin Dev Sharma. Balancing continuous pre-training and instruction fine-tuning: Optimizing instruction-following in LLMs. In arXiv:2410.10739, 2024.

[21] 山田正嗣, 井本稔也. 金融ドメイン特化のための大規模言語モデルのインストラクションチューニング評価. 人工知能学会全国大会論文集, 2024.

[22] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In Proceedings of ICLR, 2022.

[23] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In Text Summarization Branches Out, pp. 74–81, 2004.

[24] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with bert. In Proceedings of ICLR, 2020.

[25] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to Japanese morphological analysis. In Proceedings of EMNLP, pp. 230–237, 2004.

表 4　見出し生成タスクの学習データ量を変化させた際の評価結果 (BERTScore)．横軸は学習データの件数を示す．

	50	100	1000	6551	26205	104821	419283
ORG	0.679	0.689	0.700	**0.720**	0.726	0.726	0.732
CPT-jawiki	0.679	0.690	0.698	0.718	0.726	0.726	0.731
CPT-news	**0.684**	**0.691**	**0.704**	**0.720**	**0.727**	**0.727**	**0.734**

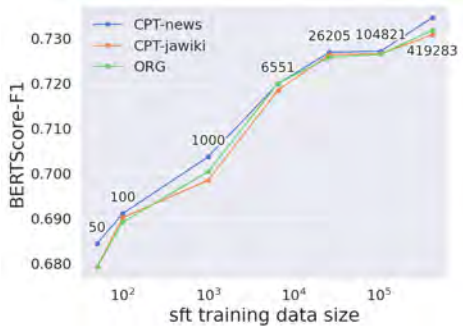


図 3　見出し生成タスクの学習データ量と性能 (BERTScore)．マーカー付近の値は見出し生成タスクの学習データ量を示す．

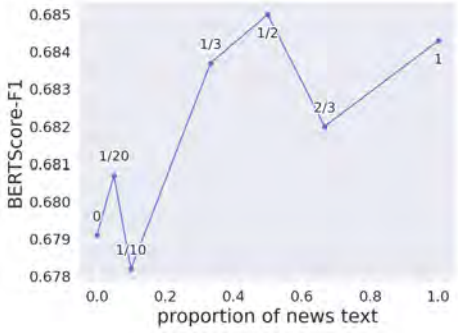


図 4　継続事前学習の学習データにおける新聞記事の割合を変化させた際の性能 (BERTScore)．マーカー付近の値は新聞記事の割合を示す．

表 5　信濃毎日新聞 2024 年 6 月 12 日の記事に対する生成例．\n は改行を示す．

本文	県内でツキノワグマによる人的被害が増加しているのを受け、県は１１日、「野生鳥獣被害対策本部会議」を県庁で開いた。７月１２日までの約１カ月間、熊が出没する恐れがある地域での「集中点検」を実施すると決めた。\n　集中点検は、県や市町村職員、県クマ対策員らが熊の目撃があった地域などを見回る。地域の土地所有者らには、熊の餌になる物の除去や生ごみなどの適切な処理、熊が身を潜められるやぶの刈り払いを呼びかける。\n　県内でのツキノワグマの目撃件数は今年５月、平年の１・８倍の１０６件。県は今月５日に「ツキノワグマ出没注意報」を出したが、８日に上高井郡高山村で４０代女性が右手をかまれるなどして重傷を負うなど、６月に入って人的被害は４件発生している。\n　会議では、県環境保全研究所の研究員が「体長１メートル未満の熊の目撃が多く、人里への恐怖心が薄い親離れしたばかりの若い熊が餌を探して動き回っている可能性がある」と説明。県森林づくり推進課の塚平賢治・鳥獣対策担当課長は会議後の取材に「熊の出没を繰り返さないためには地域ぐるみでの対策が大切。集中点検で必要な取り組みを助言していきたい」と話した。\n
正解見出し	熊警戒　来月１２日まで「集中点検」　県、人的被害の拡大受け
ORG	ツキノワグマ被害増加 県、集中点検実施へ 県庁で対策本部会議
CPT-jawiki	県内でツキノワグマによる人的被害増加 県、集中点検実施へ
CPT-news	ツキノワグマの人的被害増加　県が「集中点検」実施へ

A　実験設定の詳細

継続事前学習　3 節および 4 節で同一の学習設定を用いた．バッチサイズを 16，weight decay を 0.1，gradient clipping を 1.0 とした．また，初期学習率は 5e-6 とし，5e-7 まで線形に減衰させた．学習には Megatron-LM[10] を使用し，NVIDIA H100 94GB 4GPU で実施した．

SFT　LoRA-R を 64，LoRA-alpha を 16，LoRA-dropout を 0.15，学習対象のパラメータは all-linear とした．また，学習率は 0.0001，最大系列長は 3200 とした．バッチサイズは学習データ 1000 件以下の場合は 1，6551 件では 2，26205 件では 8，104821 件では 32，419283 件では 128 とした．各モデルは 1 エポック学習し，最終チェックポイントを評価に用いた．学習には trl[11] を使用した．実験は学習データ 104821 件以上の場合は NVIDIA A100 40GB 8GPU で，それ以外の設定では NVIDIA A100 40GB 1GPU で実施した．

見出し生成タスクのプロンプト　SFT および推論の際に与えるプロンプトは Llama-3.1-Swallow-8B-Instruct-v0.1[12] のテンプレートに従った．具体的には，システムプロンプトに「あなたは誠実で優秀な日本人のアシスタントです。」を，ユーザプロンプトに「次の新聞記事に見出しをつけてください。」，改行，記事本文を，アシスタントプロンプトに記事見出しを埋め込んだ．学習の際には記事見出しの先頭トークン以降に損失を流すようにした．

B　実験結果の詳細

定量評価　図 3 および表 4 に 3 節の実験における各モデルの BERTScore を示す．また，図 4 に 4 節の実験における BERTScore の推移を示す．

定性分析　表 5 は 3 節の実験における学習データ 50 件の場合の見出し生成例である．表 5 を見ると，CPT-news のみ正解見出し同様に集中点検というフレーズを鍵括弧で囲むことができていた．この事例について 10 回の生成を行ったところ，CPT-news は 9 回「集中点検」のように鍵括弧で囲まれた表現を生成したのに対して，ORG では 0/10 回，CPT-jawiki では 3/10 回の生成にとどまっていた．継続事前学習に用いたデータセットに鍵括弧で囲まれた表現がどの程度出現するかを調査したところ，新聞記事では全体の 50.0%，日本語 Wikipedia では 10.5%の文書に出現していることがわかった．この結果からも，新聞記事で継続事前学習を行うことで対象ドメインらしい表現を生成する傾向が強まったことが示唆される．

10）https://github.com/NVIDIA/Megatron-LM
11）https://github.com/huggingface/trl
12）https://huggingface.co/tokyotech-llm/Llama-3.1-Swallow-8B-Instruct-v0.1

TSUBAME 共同利用　令和6年度　産業利用　成果報告書

利用課題名　深層学習を用いたリアルタイム AI 音声合成プラットフォームの技術開発
英文：Real-time AI Voice Synthesis Platform using Deep Learning

利用課題責任者　竹内　雅樹
Masaki Takeuchi

所属　株式会社 CoeFont
CoeFont Co., Ltd.
URL https://www.coefont.com/ja/

邦文抄録（300 字程度）

本研究では、深層学習技術を活用したリアルタイム AI 音声合成プラットフォームの開発を行った。具体的には、長時間のゼロショット音声合成を可能にする階層型ニューラルコーデック言語モデル「HALL-E」の手法を参考にし、独自の音声合成モデルを構築した。さらに最新の AI 音声モデル、「CoeFont v3 Fuji」を開発した。CoeFont が独自に開発したであり、より人間らしさを追求した。また内閣府との意見交換会を通じて、リアルタイム AI 通訳サービス「CoeFont 通訳」の社会実装に向けた議論を行った。これらの成果により、高精度かつリアルタイムな音声合成・通訳システムの実現に寄与した。

英文抄録（100 words 程度）

In this study, we developed a real-time AI voice synthesis platform utilizing deep learning techniques. Specifically, we constructed a proprietary speech synthesis model inspired by the "HALL-E" hierarchical neural codec language model, which enables minute-long zero-shot speech synthesis. In addition, we developed the latest AI voice model, CoeFont v3 Fuji, which is proprietary to CoeFont and is designed to be more human-like. Additionally, through discussions with the Cabinet Office, we explored the societal implementation of the real-time AI interpretation service "CoeFont Interpreter." These achievements contribute to the realization of high-precision and real-time voice synthesis and interpretation systems.

Keywords: Deep learning, real-time, AI voice synthesis, zero-shot, AI intepretation

背景と目的

近年、AI 技術の進展により、音声合成や通訳システムの高度化が求められている。特に、リアルタイムで高精度な音声合成を実現することは、多言語コミュニケーションの円滑化やアクセシビリティの向上に寄与する。しかし、従来の音声合成モデルでは、長時間の音声生成やゼロショットでの多様な話者の再現が課題となっていた。

本プロジェクトでは、これらの課題を解決するため、深層学習を用いたリアルタイム AI 音声合成プラットフォームの技術開発を行い、高精度かつリアルタイムな音声合成・通訳システムの実現を目指した。

概要

本研究では、以下の取り組みを行った：

1. **階層型ニューラルコーデック言語モデル「HALL-E」による音声合成モデルの構築**

HALL-E は、長時間のゼロショット音声合成を可能にするモデルであり、マルチレゾリューション再量子化（MReQ）と呼ばれる手法を用いて、ニューラルオーディオコーデック（NAC）のフレームレートを低減する。これにより、1 分以上の音声を安定して生成できる[1]。

2. **最新の音声合成モデルおよび通訳モデルの開発**

深層学習モデルを用いて、「CoeFont v3 Fuji」音声合成モデルを開発した。主な特長は人間の発声特性を解析し、抑揚やリズムを忠実に再現し、AI 音声でありながら、話者の感情やニュアンスをよりリアルに再現し現時点での日本語での AI 音声において、最高水準の音声表現が可能となった。これにより、自然さと感情表現で大幅な改善がなされた[2]。また、翻訳処理を統合したリアルタイム通訳機能「CoeFont 通訳」も開発した。TSUBAME を活用した高速な学習・推論環境の構築により、実用化に向けた安定稼働を実現した[3]。

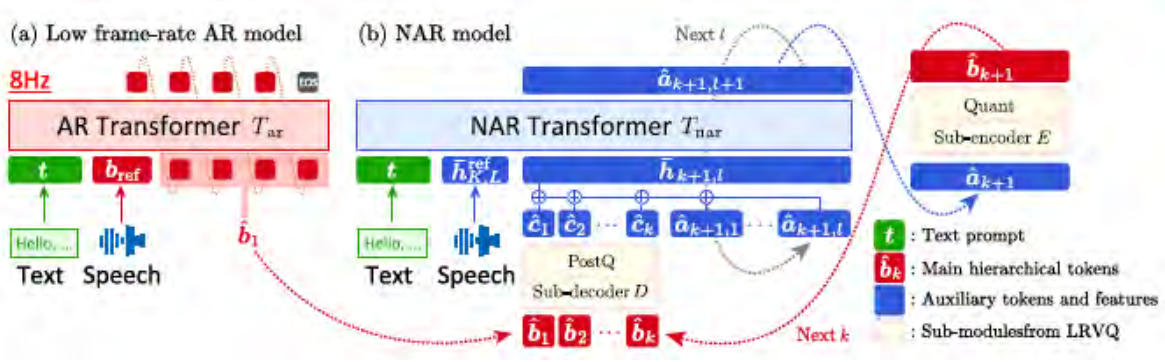


図 1. HALL-E のアーキテクチャ

結果および考察

本研究の結果、以下の成果を得た。

- **高精度な長時間音声合成の実現**

HALL-E は従来の音声合成モデルに比べて以下のような顕著な利点を有している。

o 1 分以上の長時間音声をゼロショットで高精度に生成することが可能

o フレームレートを効果的に圧縮し、音質を維持しながらトークン長を最大 96%削減。生成速度と効率を飛躍的に向上。

o 音声を複数レベルの符号列（トークン）に変換。粗い特徴から詳細な特徴へと段階的に復元するため、長時間生成でも破綻しにくい。

o 話者向けに再構成したニューラルオーディオコーデック（NAC）を訓練し、HALL-E に類似した階層的トークン生成と、文脈保持に優れた言語モデルを統合し、CoeFont が保有する大規模日本語音声データを用いたチューニングした結果、従来モデルと比較して WER や MOS 等の指標において性能向上が確認された。図 1 に、HALL-E のアーキテクチャを示す。また実際にモデルから生成された音声のデモは[4]より聞くことができる。

考察として、階層的トークン設計により、従来破綻しがちだった長時間音声の自然な合成が可能となり、ユーザーによる連続的な対話応答やナレーション用途での利用が現実的になった。話者固有の音響特徴（声色、話し方）を数秒の音声から抽出し、それを使って任意テキストの音声合成を行えるため、個別の学習コストを削減しながら多様なユーザー対応が可能になった。MReQ ベースのトークン圧縮により、生成レイテンシは大幅に短縮され、リアルタイム用途（例：会議同時通訳、対話システム）でも実用可能な応答速度を達成した。また内閣府との意見交換会を通じて、「CoeFont 通訳」の実用性を示し、社会実装に向けた具体的な課題や展望を共有することができた。これらの成果は、リアルタイムで高品質な音声合成・通訳システムの実現に向けた重要な一歩となった。

まとめ、今後の課題

本研究では、深層学習を用いたリアルタイム AI 音声合成プラットフォームの技術開発を行い、高精度な長時間音声合成やリアルタイム通訳システムの実現に向けた成果を得た。

まず CoeFont v3 Fuji およびリアルタイム通訳機能を開発・実証し、商用展開に向けた技術的基盤を確立した。今後は、より多様な言語への対応、感情と音韻の統合表現、倫理的ガイドライン整備を進め、医療・教育・放送・障害者支援等の分野への展開を図る。TSUBAME との連携を継続し、大規模モデル学習と評価も並行して進める予定である。

今後の課題として、多言語対応が挙げられる。現状のシステムは、日本語における音声認識・合成・翻訳に高い精度を示しているが、今後の実用展開を見据えると、英語を含めた多様な言語への対応が求められる。特に、東南アジア諸国や中南米、アフリカ諸国など、英語圏以外の地域におけるグローバルな利用を想定した場合、現地言語の認識・翻訳性能の向上が必要不可欠である。

具体的には、以下の点が今後の技術的課題となる：

- 言語ごとの発話特徴や文法構造に最適化した

音響モデル・テキスト処理モデルの構築

- 話者数・アクセントの多様性を含む音声データセットの収集と精緻なアノテーション
- 翻訳先の言語文化・文脈を反映した自然な言い換えや口語表現への対応

また、将来的には、ユーザーが任意の言語ペアを自由に選択できる汎用的なリアルタイム音声翻訳フレームワークの構築を目指す。これにより、ビジネス・観光・教育・国際会議など、あらゆるシーンで活用可能な多言語音声プラットフォームの実現が期待される。

参考文献

[1] Yuto Nishimura, Takumi Hirose, Masanari Ohi, Hideki Nakayama, Nakamasa Inoue. “HALL-E: Hierarchical Neural Codec Language Model for Minute-Long Zero-Shot Text-to-Speech Synthesis”, International Conference on Learning Representations (ICLR), 2025.
https://arxiv.org/abs/2410.04380

[2] CoeFont、AI 音声の新時代へ！新モデル「CoeFont v3 Fuji」実装で“人らしさ”がさらに進化
https://prtimes.jp/main/html/rd/p/000000099.000078329.html

[3] CoeFont、リアルタイム AI 通訳「CoeFont 通訳」の活用等に関して内閣府と意見交換会を開催
https://prtimes.jp/main/html/rd/p/000000096.000078329.html

[4] HALL-E: Hierarchical Neural Codec Language Model for Minute-Long Zero-Shot Text-to-Speech Synthesis
https://yutonishimura-v2.github.io/HALL-E_DEMO.

TSUBAME 共同利用　令和6年度　産業利用　成果報告書

利用課題名　基盤モデルの進化的構築技術の確立
英文：Establishing an Evolutionary Design Methodology for Foundation Models

三﨑航
Kou Misaki

Sakana AI 株式会社
Sakana AI
https://sakana.ai/

邦文抄録（300 字程度）
大規模言語モデル（LLM）に代表される基盤モデルは、さまざまな分野で活用され始め、工学における問題解決の方法論に大きな変革をもたらしている。しかしながら、基盤モデルを構築および利用するための体系的な方法論は確立されておらず、研究者や開発者は依然として経験や暗黙知に大きく依存しているのが現状である。本プロジェクトでは、事前学習後の基盤モデルの応用において特に重要な役割を果たす、ファインチューニング用データセットの自動構築に取り組む。具体的には、画像と言語の融合分野である Vision and Language（以下、VL）におけるファインチューニングデータセットを自動的に構築する手法を開発する。本プロジェクトでは進化計算をはじめとする進化的手法の概念をヒントに、効率的かつ効果的なデータセットの自動構築を可能にする方法論の確立を目指す。

英文抄録（100 words 程度）
Foundation models based on large language models (LLMs), are increasingly being utilized across various fields, driving a paradigm shift in problem-solving methodologies in engineering. However, there is still no established systematic methodology for building foundation models efficiently, and researchers and developers must largely rely on trial and error. This study focuses on the automatic construction of training datasets, which play a crucial role in developing foundation models. Specifically, we aim to develop a method for automatically constructing training datasets for foundation models in the Vision and Language (VL) domain. By leveraging evolutionary approaches, we seek to establish a methodology for efficiently and effectively automating the construction of foundation models.

Keywords: LLMs, Foundation models, Vison and language models, Evolutionary approach

1. 背景と目的

大規模言語モデル（LLM）に代表される基盤モデルは、近年の人工知能（AI）研究を牽引しており、今後もその発展が重要な役割を果たすと考えられる。深層学習の基盤技術が成熟した結果として基盤モデルが登場したように、基盤モデルのさらなる発展が次世代の技術革新に不可欠である。しかしながら、基盤モデルの構築および応用に関しては、依然として体系的な方法論が確立されておらず、研究開発の進展を阻む要因となっている。

本プロジェクトでは、基盤モデルの応用において不可欠なファインチューニング用データセットの自動構築を目的とする。特に、画像と言語の融合領域である Vision and Language（以下、VL）に焦点を当て、ファインチューニングデータセットを自動生成する手法の開発を行う。

本プロジェクトでは、進化計算を代表とする進化的アプローチを活用し、ファインチューニングデータセットを自動構築する手法を提案する。最終的に、本手法によって生成されたデータセットが、人手で構築されたデータセットと同等の性能を達成可能であることを実証する。

2. 概要

本プロジェクトでは、進化的アプローチを用いたデータセット構築の手法として、以下の二つのアプローチを検証した。

2.1 VL タスクの自動設計

第一の手法は、ファインチューニング用タスクそのも

Task Name	Instruction
Contextual text interpretation	Interpret the meaning of text within the context of the surrounding visual elements. For example, explain what a label or sign refers to based on nearby objects or scenes.
Text classification	Classify the identified text in the image into predefined categories such as titles, publication information, and volume details.
Hierarchical text reasoning	Determine the importance and hierarchy of multiple text elements within the image, such as distinguishing between main and supporting text in complex visual settings.

表 1. GPT-4o によって自動生成されたタスク例

のを自動設計するものである。従来の研究では、画像キャプション生成、画像を用いた対話、視覚推論の三つのタスクが一般的に利用されている。しかし、これら以外にも例えば背景領域の詳細記述、物体の動作に関する質疑応答、画像内の文字認識といった多様なタスクの可能性が考えられる。次トークン予測の最適化が LLM の成功をもたらしたように、ファインチューニングにおいても最適なタスクが存在するはずである。

本研究では、タスクの進化的生成手法を提案する（図 1 参照）。まず、任意のデータセット D を用意し、対象の視覚言語モデル（VLM）をファインチューニングする。その後、検証用データセットを用いてモデルの性能を評価し、誤答サンプルを収集する。これらの誤答データを基に、新たなタスク（タスク名と指示文）を自動生成し、アノテーションモデルを用いて適切な指示文および回答文を作成する。このプロセスを繰り返すことで、モデルが苦手とする領域を補完しつつ、最適なタスクを設計することが可能となる。本手法では、タスク設計およびアノテーションに GPT-4o を活用し、各ステップで 10 種類のタスクを生成する。実際に生成されたタスク例を表 1 に示す。

2.2 質疑応答文の自動設計

第二の手法は、詳細な画像キャプションを基に質疑応答データを自動生成するものである。まず、GPT-4o を用いて学習画像ごとに詳細なキャプションを生成する。次に、生成されたキャプションを基に、GPT-4o を活用して画像に関する質問文および回答例を自動生成する。さらに、これらの質問文を発展させ、より簡単な質問とより難しい質問を生成し、それぞれに対する回答を作成する。このプロセスを反復することで、多様な質疑応答データを自動生成し、データセットの質を向上させる。

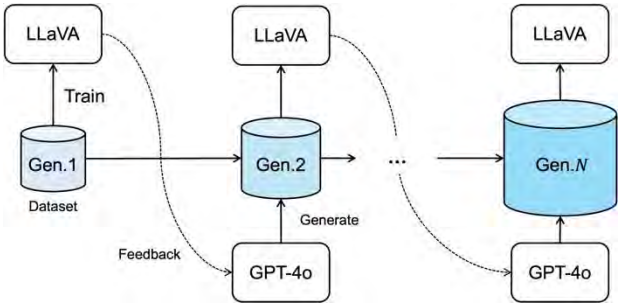


図 1. タスク自動設計の概要図

3. 結果および考察

3.1 実験設計

本研究では、LLaVA-1.5 [1]を学習対象の VLM として採用した。Flamingo [2]などのクロスアテンション機構を用いた VLM も検証したが、LLaVA-1.5 がより高い性能を示したためである。

また、性能検証には TextVQA [3]と呼ばれるベンチマークを採用した。TextVQA は、画像内に写っているテキストに関する認識能力を測るベンチマークである。高い正答率を得るためには、正確なテキスト認識と物体認識、そしてそれらの情報を用いた推論能力が問われる。画像と言語の双方にわたる複数の認識能力が必要となることから、本プロジェクトの評価対象に適していると考え、採用した。

3.2 実験結果

図 2 に、VL タスク自動設計のアプローチ（2.1 節）の

TextVQA での評価結果を示す。図 2 の横軸はタスク設計の回数を示し、縦軸は TextVQA での正答率を表している。図 2 から明らかなように、タスク設計を繰り返すことで、モデルの性能が向上していることがわかる。全体的なコストの都合で合計 5 回のタスク設計を行ったが、最終的に人手でタスク設計を行っているデータセット（LLaVA-1.5[1]で 58.2%）と同程度の性能（57.4%）を達成可能なことを示せた。なお、LLaVA-1.5 による結果は我々が手元で再現した結果である。

また、我々は 2.2 節のアプローチで構築したデータセットを用いて VLM をファインチューニングし、同様に TextVQA で性能評価を行った。質疑応答生成のラウンドを繰り返したが、最終的な回答精度は 56.1%に留まった。

3.3 現状の課題

本プロジェクトで目標とした VLM のファインチューニング用データセットの自動構築について、以下の二つの課題が明らかになった。

アノテーションモデルの性能限界：例えば、TextVQA には画像内に写っているアナログ時計の時刻を回答させるものがあるが、GPT-4o はほとんどのサンプルに対して誤答してしまう。同様の現象がほかのサンプルにも生じうる。そのため、誤ったアノテーションがファインチューニング用データに含まれてしまい、VLM の性能向上を阻害する要因となった。

タスクの多様性の低下：タスク設計を繰り返すにつれて、異なる指示文であっても実質的に類似したタスクが生成される傾向が見られた。この結果、データセットの多様性が損なわれ、性能の向上が限定的になっていく。

3.4 マルチエージェントによるタスク自動設計

タスクの多様性の低下を解決するために、マルチエージェント手法を導入した。具体的には、我々は GPT-4o と Claude Sonnet 3.5（Claude-3.5）を用い、それぞれのモデルの表現力を相互補完することを試みた。

具体的には、GPT-4o で生成した 10 個のタスクを Claude-3.5 に渡し、内容を洗練化させた後、GPT-4o に戻し最終的に 5 個のタスクを選択させる。このプロセスを繰り返すことで、データセットの進化を図った。

しかし、図 3 に示すように、GPT-4o 単独で設計したタスクの方が高い性能を示した。エージェント間の議論を促進する実験も行ったが、建設的な議論の形跡は見られなかった。今後は、異なるエージェントの知識を効果的に統合する手法が求められる。

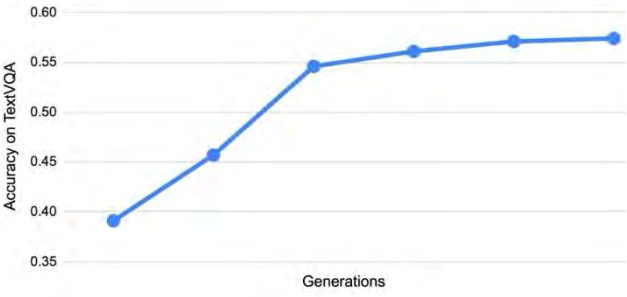


図 2. TextVQA における評価結果

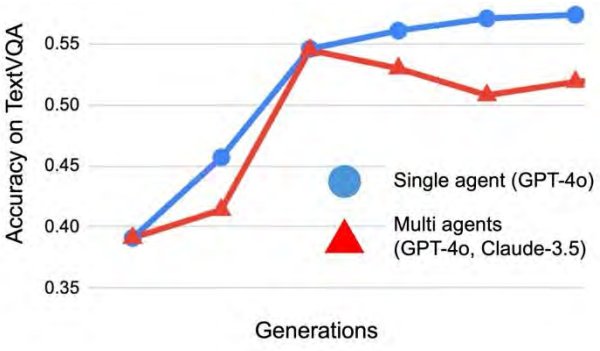


図 3. マルチエージェントによる評価結果

4. まとめ、今後の課題

本プロジェクトでは、基盤モデルのファインチューニングデータセットを自動構築する手法を提案し、人手構築のデータセットと同等の性能を達成できることを示した。しかしながら、より効果的な自動構築手法の確立にはさらなる研究が必要である。特に、マルチエージェントシステムの活用を通じて、個々のモデルの限界を克服する枠組みの開発が今後の課題となる。

謝辞

本プロジェクトの遂行にあたり、計算資源および環境を提供してくださった東京科学大学 情報基盤センターの皆様ならびに関係者の皆様に深く感謝申し上げます。

参考文献

[1] H. Liu+, Improved Baselines with Visual Instruction Tuning, CVPR, 2024.

[2] J. B. Alayrac+, Flamingo: a Visual Language Model for Few-Shot Learning, NeurIPS, 2022.

[3] A. Singh+, Towards VQA Models That Can Read, CVPR, 2019.

TSUBAME 共同利用　令和6年度　産業利用　成果報告書

利用課題名　LLM エージェントの研究開発
英文：Development of LLM agents

利用課題責任者
黒木　颯（So Kuroki）

所属
Sakana AI 株式会社

邦文抄録（300 字程度）

大規模言語モデル（LLM）に多様なエージェントスキルを獲得させるためには，複数タスクの学習データ比率や目的関数設定が難しいという課題がある．本研究では，品質多様性（QD）の一種である MAP-Elites に着目し，タスク性能を交互に最適化する CycleQD の提案に向けた予備実験を実施した．具体的な一連のプロセスとしては，各タスクに特化した専門モデルを初期に用意し，モデルパラメータを進化的にマージすることで，コード生成や OS 操作，DB 操作などで高い性能を獲得しつつ，一般言語能力の維持に成功した．さらに画像セグメンテーションモデルへの適用も示し，多様な領域への拡張可能性を明らかにした．

英文抄録（100 words 程度）

We propose CycleQD, a novel method for enabling large language models (LLMs) to acquire diverse agent skills. By treating each task's performance as a quality metric in a rotating fashion, our approach eliminates the need for intricate data mixing ratios. We employ model merging as a crossover step and introduce an SVD-based mutation to escape the convex region formed by the initial expert models. In empirical evaluations of coding, operating systems, and database tasks, CycleQD surpasses traditional fine-tuning baselines and reaches performance comparable to GPT-3.5-TURBO. Moreover, it retains strong language capabilities and extends easily to vision tasks.

Keywords: 5つ程度
Evolutionary Algorithm, Model Merging, Quality Diversity, LLM-based Agent, Skill Acquisition

背景と目的

大規模言語モデル（LLM）は，対話や推論において高い性能を示す一方，複数の専門的タスクを同時に学習させる際には，タスク間のデータ比率調整や目的関数の設計といった難題が生じることが知られている．さらに，単一タスク特化の複数モデルを単純にマージする場合には，過学習や性能の部分的劣化が見られることがある．本プロジェクトでは，品質多様性（Quality Diversity: QD）を用いた CycleQD という手法を提案し，これらの課題を解決するとともに，多様なタスクへの適用可能性を示した．

概要

提案手法である CycleQD は，MAP-Elites を基盤とした進化的アルゴリズムを用いてモデルマージをする手法である．まず各タスクに対して専門特化したエキスパートモデルを初期個体群として準備する．そのうえで，タスク性能である質（Quality）と行動特性（Behavior Characteristics: BC）を切り替えながら世代交代を行う．具体的には，(1) エリートサンプリングにより優秀なモデル個体を 2 つ取得し，(2) モデルマージにより 2 つのモデルを重み空間で線形結合し，(3) SVD を用いた突然変異を加えることで線形結合の範囲外へ探索を広げる．この過程で各タスクにおける最適化と多様なモデルの保持が同時に実現され，最終的に多数のタスクを高水準でこなせる統合モデルを得ることができる．本研究では，コード生成，OS 操作，DB 操作などのコンピュータサイエンス系タスクに加え，一般言語性能や画像セグメンテーションタスクでも本手法の有効性を示した．図 1 は手法の概要図である．

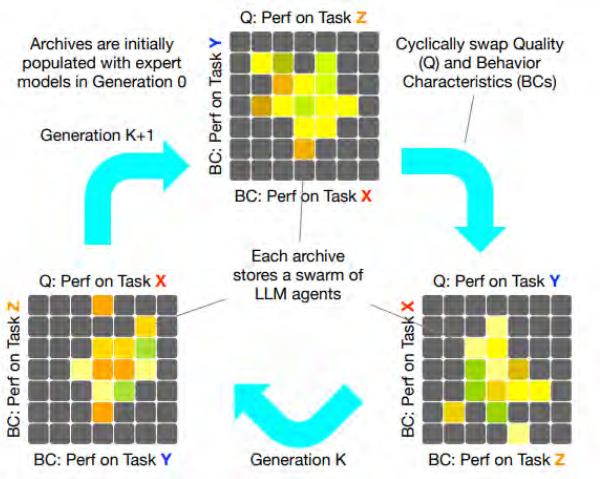


図 1. CycleQD では Quality と BC を周期的に切り替えながら世代交代を実施する.

結果および考察

実験では、LLAMA3-8B-Instruct をベースモデルとし、まずコード生成・OS 操作・DB 操作の三つのタスクに対して専門モデルを作成した。その後 CycleQD を適用することで、従来の単純なファインチューニングや従来型のモデルマージ手法を上回る性能を示す LLM モデルの獲得した. 特にコード生成と OS 操作タスクでは専門モデルよりも高いスコアを達成した。また、GPT-3.5-TURBO と比較しても同等程度の性能に迫る結果が得られ、言語能力に関しては大きな劣化が見られないことが確認できた。さらに画像セグメンテーションモデル（SAM）にも適用し、タスク間の類似性が高いほどモデルマージが容易であることなど、手法の有効性と限界を分析した. 表 1 に CycleQD と比較手法との比較を示した.

表 1. CycleQD は比較手法より高い性能を示した.

#	Methods	MBPP	DB	OS	Avg
0	GPT-4	83.6	36.5	63.7	61.3
1	GPT-3.5-TURBO	82.0	41.6	38.5	53.7
2	Llama3-8B-Instruct (base model)	67.3	5.3	25.2	32.6
Fine-tuning Based Methods					
3	Fine-tuning (Coding expert)	70.4	21.2	20.7	37.4
4	Fine-tuning (DB expert)	65.8	**42.4**	28.5	45.6
5	Fine-tuning (OS expert)	66.3	0.0	30.4	32.2
6	Fine-tuning (All)	67.3	37.1	36.7	47.0
Merging Based Methods					
7	Merging (w/o learning)	72.9	24.7	**42.6**	46.7
8	Merging (learning w/ GD)	69.3	41.2	29.6	46.7
9	Merging (learning w/ CMA-ES)	69.3	41.2	30.2	46.9
10	Merging (learning w/ NSGA-II)	75.9	**42.4**	36.4	51.6
11	CycleQD (Ours)	**76.4**	38.2	**42.6**	**52.4**

まとめ、今後の課題

本研究では、QD アルゴリズムを活用し、周期的に最適化対象タスクを切り替える CycleQD を提案するための予備実験を実施し、最終的に複数タスクを同時に高い水準で実行可能な LLM を構築できることを示した。従来の手法では困難だったデータ比率調整や目的関数設計を簡略化しつつ、多様で高性能なモデル群を得られる点が本手法の大きな利点である。一方で、学習に用いる専門モデル間の類似度が低い場合には性能向上が制限される可能性も示唆された。今後は専門モデルの事前学習工程における類似度制御や、CMA-ME など高度な進化的手法との統合による効率向上、さらにマルチエージェント環境への応用などが課題として挙げられる.

TSUBAME 共同利用　令和 6 年度　産業利用　成果報告書

利用課題名　IT 創薬向け並列計算化学シミュレーション手法の開発
Development of parallel computational chemistry simulation methods for IT-based drug discovery

利用課題責任者
秋山　泰
Yutaka Akiyama
アヘッド・バイオコンピューティング株式会社
Ahead Biocomputing, Co. Ltd.
https://ahead-biocomputing.co.jp

本課題は、中分子創薬支援を目的とした二次元レプリカ交換 分子動力学（REMD）法を対象とし、NVIDIA H100 GPU におけるマルチプロセスサービス（MPS）の活用と適度なオーバーサブスクリプション設定（最大 16 プロセス／GPU）により、従来の 1 プロセス／GPU 実行では活用しきれなかった演算リソースを引き出す手法の開発を行った。当社が取り組んでいる環状ペプチドの膜透過性予測の系で性能検証した結果、GPU 当たりの実効スループットが約 2.85 倍向上し、同時に用いる GPU 台数を 224 基から 1/8 の 28 基などに削減でき、総実行時間を短縮できた。さらに、開発した計算手法を 8 種類の環状ペプチドの実際の膜透過性予測計算に応用した。

英文抄録（100 words 程度）
This project targets the two-dimensional replica exchange molecular dynamics (REMD) method and we utilized multi-process services (MPS) on NVIDIA H100 GPUs with a moderate oversubscription setting (up to 16 processes/GPU) in order to make effective use of computing resources that cannot be extracted by a single process/GPU run. We applied the method to membrane permeability prediction of cyclic peptides, and the effective throughput per GPU was improved about 2.85 times, reducing the number of GPUs required and the total execution time. Furthermore, we applied the method to a membrane permeability prediction project for 8 cyclic peptides.

Keywords: GPU oversubscription / CUDA MPS / Replica-exchange MD / AMBER / Cyclic peptide permeability

背景と目的

Replica-Exchange MD (REMD) は、分子動力学シミュレーション (MD) におけるサンプリング能力を向上する手法として活用されているが、各レプリカが独立に CPU/GPU を占有するため計算資源の消費が大きい。また、最新の GPU (H100 等) で MD の 1 プロセスだけを実行する場合、多くの SM が遊休状態となり、GPU コア利用率が低い。したがって、1GPU に 1 プロセスを割り振った上で大規模な REMD を実行すると、莫大な計算資源を非効率に使用してしまうこととなる。

我々は、NVIDIA Multi-Process Service (MPS) を利用し、1 つの GPU 上で多数の MD のプロセスを実行させることで、REMD 計算の効率を改善可能であることを確認した。[1] さらに、適度なオーバーサブスクリプション設定（MPS において1GPU 内で実行される各プロセスが使用を許される GPU リソースの最大限度を合計すると 100%を超える設定）を行うと、より効率良く REMD 計算が実行可能であることも見出した。[1]

本課題では、環状ペプチドの膜透過性予測を目的とした 224 プロセスの 2 次元 (2-D) REMD 計算 [2] を、適度なオーバーサブスクリプション設定の MPS を用いて実行し、東京科学大学 TSUBAME4.0 上でどの程度の効率で実行できるかの実証実験を行った。さらに、開発した手法を 8 種類の 10 残基ペプチドの膜透過性予測の計算に応用した。

概要

本課題では、224 レプリカを使用した 2D-REMD を実行し、環状ペプチドが脂質二重膜を透過する過程の挙動を予測した（図 1）。MPS を利用して各 GPU 上で 8 プロセスを同時起動した。環境変数の一つである CUDA_MPS_ACTIVE_THREAD_PERCENTAGE

を 30 に設定した（30% × 8 = 240%のオーバーサブスクリプション設定に相当）。このとき、4 process/GPU や 16 process/GPU とする設定も選択の候補にはなるが、4 process/GPU では後述する表 1 にも示されるように計算効率が劣る。また、16 process/GPU の設定ではキリの良い 100ns の MD を実行すると TSUBAME4.0 の計算時間上限（24 時間）に抵触するため本課題では採用しなかった。

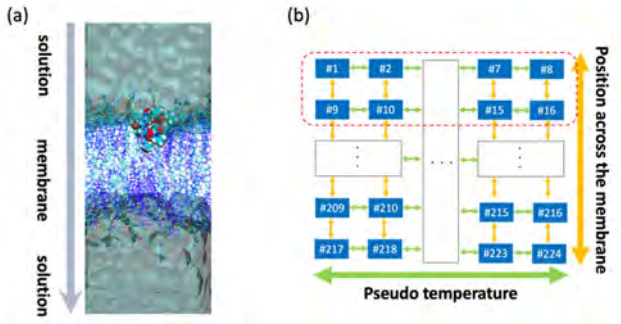


図 1. (a) 脂質二重膜を透過する環状ペプチドの模式図および (b) 224 レプリカの 2D-REMD の模式図

結果および考察

TSUBAME4.0 上での MPS 実行による効率向上を計測するため、REMD を構成するプロセスを取り出し、初期座標の一つを利用してシングルスレッドで MD を実行したところ、評価用の系に対する MD 計算の処理速度は 306 ns/day であった。次に 8 process/GPU 設定で実行したところ 109 ns/day であった。後者では 1GPU あたり 109ns * 8 process = 872 ns/day 分の計算を実行できていることから、両者の比率から、計算効率が約 2.85 倍向上していることが判る。これにより、膜透過予測のための 2-D REMD に要する GPU 資源量が大幅に節約できることがわかった。

TSUBAME4.0 上での効率化と比較するために、我々が筑波大学計算科学研究センターの Pegasus システム（H100 PCIe GPU/node, 120 nodes）を利用して同条件で計測したデータを表 1 に示す。効率の向上は 8 process/GPU 設定時で約 2.26 倍であり、TSUBAME4.0 での計測の方が向上率は高かった。

表 1. 筑波大学計算科学研究センター Pegasus システム（H100 PCIe GPU/node, 120 nodes）を利用して得られた MPS 使用時の膜透過 MD の計算効率

# Process/GPU	Speed/Process (ns/day)	Total speed (ns/day)	Efficiency gain
1	326.73 ± 1.05	326.73 ± 1.05	(1.00)
2	233.70 ± 0.52	467.40 ± 1.04	1.43
4	157.92 ± 0.20	631.69 ± 0.80	1.93
8	92.11 ± 0.81	736.91 ± 6.46	2.26
16	46.50 ± 0.60	743.94 ± 9.62	2.28

次に、開発した計算プロトコルを用いて環状ペプチドの膜透過性予測に関する実計算を実施した。本計算は有償ソフトを用いて付与していた力場パラメータを無償ソフトを用いて付与可能なものに置き換えるためのテストであり、AMBER 系列の力場の検討を行った。計算対象は先行研究 [2] で用いた Furukawa らの合成した 10 残基のペプチドとした。[3] 先行研究と同じ 224 プロセスの 2 次元（2-D）REMD 計算 [2] を実施した。結果として、8 ペプチドの膜透過過程の MD 計算は完了したものの、得られる自由エネルギープロファイルの形が大きく変化し、得られた膜透過係数の値が定性的に変化した（図 2）。力場の変更が膜透過係数の予測精度に及ぼす影響の調査、およびパラメータの検討を引き続き行っているところである。

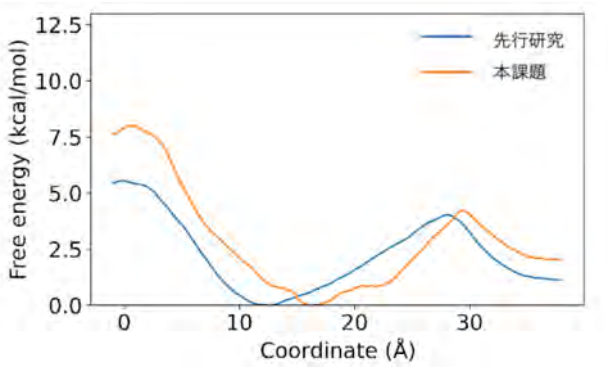


図 2. Furukawa らの合成したペプチド B08 に対する先行研究および本課題におけるプロトコルから得られた自由エネルギープロファイル

本課題を通じて開発した手法では、現在のところ 1 種類の環状ペプチドの膜透過過程の 224 レプリカの 2-D REMD をフル実行するために、約 3083 GPU 時間を要している。（参考：成果公開の単価で約 21 万円相当）

MPS 利用と適切なオーバーサブスクリプション設定により数倍の効率化を達成できたが、環状ペプチドの化学合成と実験による膜透過性計測のコストと比較すると上記は高額であり、まだ数倍の効率向上は必要と考えている。一方、MD 計算では膜透過過程の微細な状態を観察することが可能であることは優れた特徴といえる。膜透過現象を科学的に研究するツールとして、あるいは産業上有用な環状ペプチドの膜透過過程を詳細に解析して設計変更のヒントを得るためのツールとして、他では代替できない機能を提供できている。

まとめ、今後の課題

本課題では、二次元 REMD における GPU コアの遊休の問題を MPS 利用とオーバーサブスクリプション設定で解決し、GPU 資源効率を大幅に改善した。
MPS により 8 process/GPU を割り当て、オーバーサブスクリプションを 240%に設定した時、1 process/GPU での実行に比べて約 2.85 倍の効率向上を得た。同時に利用する GPU 数も、224 基から 1/8 の 28 基（7ノード）に減り、比較的現実的な資源量になった。
我々はさらに、開発手法を 8 種類の環状ペプチドの膜透過性予測に応用した。

今後の課題は、以下のとおりである。

1. MIG 併用による論理 GPU 分割との比較評価
2. Gromacs、OpenMM など他エンジンへの適用
3. 動的負荷分散アルゴリズムの導入によるプロセス数の自動調整
4. 膜透過性予測に最適な力場設定およびその他のパラメータの探索

参考文献

1. T. Boku et al., Improving performance on replica-exchange molecular dynamics simulations by optimizing GPU core utilization. ACM International Conference Proceeding Series; Association for Computing Machinery, 2024; pp 1082–1091.
2. M. Sugita et al., Lipid Composition Is Critical for Accurate Membrane Permeability Prediction of Cyclic Peptides by Molecular Dynamics Simulations. Journal of Chemical Information and Modeling, 2022, vol. 62; pp 4549–4560.
3. A. Furukawa et al., Drug-Like Properties in Macrocycles above MW 1000: Backbone Rigidity versus Side-Chain Lipophilicity. Angewandte Chemie - International Edition, 2020, vol. 59; pp 2–9.

TSUBAME 共同利用　令和 6 年度　産業利用　成果報告書

利用課題名：有機高分子材料の量子化学的研究
英文：Quantum chemical study of organic and polymeric materials

利用課題責任者：川内　進
First name Surname:

所属：株式会社 Quemix
Affiliation: Quemix Inc.
URL: https://www.quemix.com/

邦文抄録（300 字程度）

ポリシランはらせん構造を示す半導体高分子として知られている。たとえば、ポリ[(S)-2-メチルブチルシラン]では、同一分子鎖内に緊密な P-らせんと緩慢な M-らせんセグメントの共存が報告されていることから、ポリシランには複数の安定ならせん構造が予想される。本課題では、このようなポリシランのらせん構造とその安定性を明らかにするために、4 量体から 6 量体までのポリシランオリゴマーの水素化物とパーメチル体について DFT 計算による配座探索を行った。水素化物については MP2 法と CCSD(T)法で計算を行い、DFT 計算の検証を行った。比較のため、対応するポリメチレンオリゴマーについても配座探索を行った。その結果、ポリシランオリゴマーのパーメチル体は三種類のらせん構造を取り得るのに対し、ポリメチレンオリゴマーの安定ならせん構造は一種類のみであることが明らかとなった。

英文抄録（100 words 程度）

Polysilanes are known as semiconducting polymers that exhibit helical structures. For example, in poly[(S)-2-methylbutylsilane], the coexistence of a tight P-helix and a loose M-helix segment in the same molecular chain has been reported, and this indicates multiple stable helical structures of polysilanes. In this project, in order to clarify such helical structures of polysilanes and their stability, we performed conformation searches by DFT calculations for hydrides and permethyl forms of tetramer to hexamer of polysilane oligomers. For hydrides, calculations were also performed using the MP2 and CCSD(T) methods to validate the results of the DFT calculations. For comparison, the corresponding polymethylene oligomers were also subjected to a conformation search. The results revealed that the permethyl form of polysilane oligomers can take three types of helical structures, whereas the stable helical structure of polymethylene oligomers is only one type.

Keywords: DFT calculation, polysilane, polymethylene, helix, conformation search

背景と目的

ビニルモノマーのラジカル共重合反応の反応性比は、重合で生成するポリマー組成を予測する上で重要なパラメータである。反応性比を予測する理論としては 1947 年に提案された Alfrey-Price の Q-e スキーム[1]があり、高分子化学の教科書に必ず記載され、現在でも有用であり続けている。しかし、Q-e スキームには提案当初からいくつかの欠点が指摘されており、これまで解決には至っていなかった。我々は、参照モノマーを 2 つ利用することで固有 Q-e スキームを導出し、Q-e スキームにあった欠点を全て解決しただけでなく反応性比の予測精度の向上にも成功した[2]。また、これによりポリマーラジカルとモノマーの Q-e 値を個別に計算することにも初めて成功した。

Q-e 値を求めるには、元になる反応性比の良質な実験値が必要である。反応性比の実験値のデータベースとしては Polymer Handbook[3]があるが古い書籍であることと詳しく調べるとデータの誤植や重複が多いことが判明し、また、重合条件の記載が無いことが欠点であった。そこで、我々は Polymer Handbook のデータを見直すと共に、新たに文献調査を行うことで 3,000 以上のモノマーと 1 万を超える反応性比で構成されるデータベースを構築した。この反応性比データベースには、モノマーの SMILES などの情報に加えてラジカル共重合の重合条件や反応性比の算出方法が含まれているため、情報源としても有用であろう。

本課題では、最終的に実験によらない Q-e 値および反応性比の予測システムを構築することを目的としている。そのため、作成した反応性比データベースに含まれるモノマーやポリマーラジカル、そしてポリマーラジカルへのモノマーの付加反応について DFT 計算を行い、得られた情報をデータベースに加え記述子として機械学習を行う予定である。

本年度は、DFT 計算において汎関数と基底関数系の選択が計算結果を左右することがあるため、先ず、ポリシランのらせん構造に対して詳細な DFT 計算を行い、MP2 と CCSD(T)計算の結果と比較することで、汎関数と基底関数系の検証を行った。以上の計算資源と計算精度に対する知見を参考にしてモノマーの構造最適化計算に用いる汎関数と基底関数系を選定した。その結果、データベースの半数に当たる 1,500 のモノマーについて構造最適化が終了したところである。

概要

DFT 計算のベンチマークテストの対象として選んだポリシランはらせん構造を示す半導体高分子である。たとえば、ポリ[(S)-2-メチルブチルシラン]では、同一分子鎖内に緊密な P-らせんと緩慢な M-らせんセグメントの共存が報告されており[4]、ポリシランには複数の安定ならせん構造の存在が示唆される。このようなポリシランのらせん構造とその安定性を明らかにするために、4 量体から 6 量体のポリシランオリゴマーの水素化物とパーメチル体について DFT 計算による配座探索を行った。さらに水素化物については MP2 法と CCSD(T)法でも計算を行った。そして、対応するポリメチレンオリゴマーについても配座探索を行った。

結果および考察

1. ポリシランオリゴマーとポリメチレンオリゴマーの水素化物の配座探索

Si_4H_{10}～Si_6H_{14}とC_4H_{10}～C_6H_{14}について DFT 法と MP2 法による配座探索を行った。DFT 計算の汎関数は ωB97X-D を基底関数系には 6-311+G(2df,2p)と 6-311++G(2df,2p)を用いた。また各最適化構造を用いて CCSD(T)一点計算を行った。

DFT 計算の結果から、C_4H_{10}～C_6H_{14}では全てトランス配座が最安定あるのに対し、Si_4H_{10} ではトランス配座とゴーシュ配座がほぼ等エネルギーであった。また、鎖長を伸ばした Si_5H_{12}と Si_6H_{14} ではゴーシュ配座をとるらせん構造が最安定であった。以上は、MP2 法や CCSD(T)法の結果とよく一致した。また、用いた基底関数の依存性はほとんどなかった。したがって、以降の計算には、ωB97X-D/6-311+G(2df,2p)を用いた。

2. C_4Me_{10}とSi_4Me_{10}のらせん構造と遷移状態

先ず、C_4Me_{10}と Si_4Me_{10} の内部回転ポテンシャルを調べた。内部回転の緩和スキャンでは、側鎖メチル基がぶつかるため、きれいな曲線は得られなかった。そこで、緩和スキャンの結果を基に初期構造として安定点と遷移状態の候補を選び、安定点と遷移状態（TS）の構造最適化を行った。

C_4Me_{10}とSi_4Me_{10}の結果をそれぞれ図 1 と図 2 に示した。なお、以下では鏡像体の片方のみを考慮する。図には二面角（ψ）、主鎖の末端原子間距離（r）、Helix 1を基準とした電子的エネルギー差（ΔE）、エンタルピー差（ΔH）を示した。

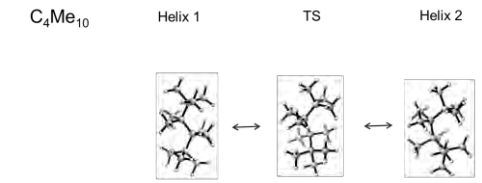


図 1. C_4Me_{10}の安定構造と遷移状態

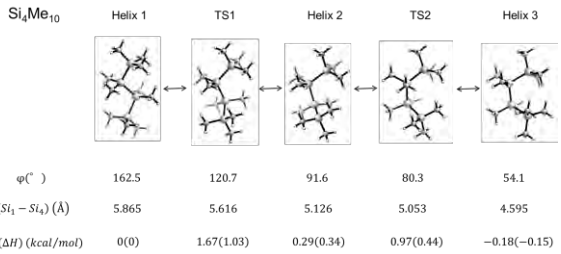


図 2. Si_4Me_{10}の安定構造と遷移状態

図に示したように、C_4Me_{10} の構造最適化で安定構造が二つ、Si_4Me_{10} では三つ得られた。結果を詳しく見ると、C_4Me_{10} では、伸びきったトランス配座に近い Helix

1 の方が安定である上に活性化エネルギーが高いため、安定構造としては Helix 1 のみが存在すると考えられる。一方、Si_4Me_{10} では三つの安定構造間のエネルギー差は小さく、活性化エネルギーも小さいことから三つの安定構造が共存が示唆される。

3. C_5Me_{12}、C_6Me_{14}、Si_5Me_{12}、Si_6Me_{14} のらせん構造

以上の結果を基に C_5Me_{12}、C_6Me_{14}、Si_5Me_{12}、Si_6Me_{14} について配座探索を行った。以下の図 3 と図 4 に得られた配座の内、らせん構造に該当するものを抽出して示した。

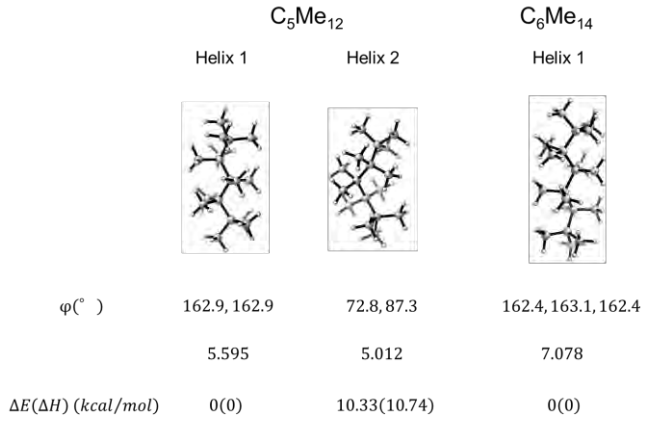


図 3. C_5Me_{12} と C_6Me_{14} のらせん構造

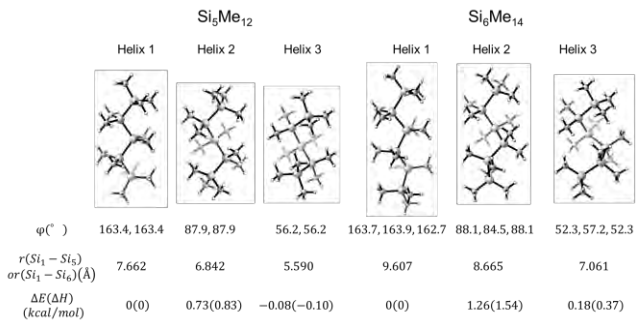


図 4. Si_5Me_{12} と Si_6Me_{14} のらせん構造

図 3 に示した C_5Me_{12}、C_6Me_{14} については、C_4Me_{10} と同様に伸びきったトランス配座に近い Helix 1 のらせん構造が最安定であった。他に 2, 2 の配座が見つかったがいずれもエネルギーは高いため、ポリメチレンのパーメチル体は伸びきった単一のらせん構造をとると考えられる。

一方で、Si_5Me_{12} と Si_6Me_{14} では、多数の配座が見つかっており、Si_5Me_{12} では 6 個の配座がエネルギー差 1 kcal/mol 以内で、Si_6Me_{14} では、22 個の配座がエネルギー差 2 kcal/mol 以内に収まった。したがって、ポリシランンのパーメチル体は多様な構造を取り得る。また、Helix 1 と Helix 3 は配座の内で比較的エネルギーが低い方に位置していることから、ポリ[(S)-2-メチルブチルシラン]で見られた 2 つのらせん構造にそれぞれ対応すると考えられる。

まとめ、今後の課題

ポリシランオリゴマーのパーメチル体は三種類のらせん構造を取り得るのに対し、ポリメチレンオリゴマーの安定ならせん構造は伸びきった一種類のみであることが明らかとなった。このようにポリシランは柔軟でらせん構造を取りやすいことが、一般に側鎖の導入によるらせん制御が容易に行える原因であろう。

今後は、本課題の最終目的である Q-e 値および反応性比の予測システムの構築を目指して、反応性比データベースに含まれる大量のモノマーやポリマーラジカルの DFT 計算による構造最適化を行うとともに、ポリマーラジカルへのモノマーの付加反応の遷移状態を求める。そして、DFT 計算で得られた情報を記述子として大量な実験情報に加えて機械学習を行うことで予測システムを構築する予定である。

参考文献

[1] T. Alfrey and C. C. Price, J. Polym. Sci., 2, 101–106 (1947).

[2] S. Kawauchi, A. Akatsuka, Y. Hayashi, H. Furuya, T. Takata, Polymer Chemistry, 13(8), 1116-1129 (2022)).

[3] R. Z. Greenley, in Polymer Handbook, eds. J. Brandrup, E. H. Immergut and E. A. Grulke, Wiley-Interscience, New York, 4th Ed., p. II/309–319 (1999).

[4] M. Fujiki, J. Am. Chem. Soc., 116, 11976 (1994).

TSUBAME 共同利用　令和6年度　学術利用　成果報告書

利用課題名　糖鎖-蛋白質間の相互作用解析
英文：Analysis of carbohydrate-protein interactions

利用課題責任者　尾形　慎
First name Surname Makoto Ogata

所属　福島大学農学群食農学類
Affiliation Faculty of Food and Agricultural Sciences, Fukushima University
URL https://www.agri.fukushima-u.ac.jp/

邦文抄録（300 字程度）
糖鎖は、エネルギー代謝や細胞間の情報伝達、ウイルス感染など様々な生命現象に関与している。糖鎖を基質とする酵素群などの蛋白質-糖鎖間相互作用はこれらの現象の多くで重要な役割を持つため、蛋白質の糖鎖認識メカニズムを理解することは、生理機能の理解や創薬などへの応用へ向け重要である。しかし、多様で複雑な糖鎖構造は、糖鎖-蛋白質相互作用の解析の障壁となっている。本課題では、糖鎖関連酵素の詳細な機能理解や機能性糖鎖材料開発を目指し、TSUBAME4.0 上でのドッキングシミュレーション、分子動力学法等を取り入れた糖鎖-蛋白質間相互作用解析を行った。今回は、リゾチーム新規活性測定基質の設計およびシアル酸転移酵素の基質認識機構解析について報告する。

英文抄録（100 words 程度）
Carbohydrates and carbohydrate-active enzymes interaction have many physiological roles such as energy metabolism, cell communication, and infection. For revealing glycan-mediated biological phenomenon and for drug discovery, it is important to clarify substrate recognition mechanisms of proteins involved in biosynthesis and degradation of glycan. However, analysis of them have technical limitation due to structural complexity of glycan. In this study, we performed docking simulation and molecular dynamics simulation using TSUBAME 4.0 to analyze carbohydrate-active enzyme and develop functional glycan materials. We report a novel substrate for assaying lysozyme activity and analysis of substrate recognition mechanism for sialyltransferase.

Keywords: Carbohydrate, Lysozyme, Sialyltransferase, Docking simulation, Molecular dynamics

1.　リゾチームの新規活性測定基質の開発

背景と目的

リゾチームは、GH ファミリー22 に属し、ペプチドグリカンやキチンなどを加水分解する酵素である。-4～+2 の 6 つの糖結合サブサイトを持ち、-1 - +1 サブサイト間の β-グリコシド結合を加水分解する（図 1）。ニワトリ卵白リゾチーム（HEWL）活性は潰瘍性大腸炎などのマーカーとして利用されているが、活性測定の簡便さと精度の両立など課題点も多い。そこで我々はより優れた活性測定手法確立のため、新規活性測定基質の設計および評価を行った。この際、ドッキングシミュレーションや MD シミュレーションを活用することで複雑な糖鎖分子における解析の効率化を図った。

概要

前述したサブサイト情報から、非還元末端側がキトテトラオース（GN4）、還元末端側が非 GN の 2 糖からなる 6 糖を基質候補とする方針を取った。還元末端側は 2 糖といえ膨大なパターンが考えられるが、本年度はまずシミュレーション手法の評価を目的とし、グリコシド結合が異なる 3 種類のグルコビオース（G2、図 2）に絞った。6 糖 GN4G2 の各糖残基が HEWL の各サブサイトと結合できれば、図 2 のような活性測定スキームにより簡便な活性測定が期待できる。今回は、ドッキングシミュレーション、MD シミュレーションにより 3 種の GN4G2 と HEWL サブサイト間の相互作用安定性を評価した。

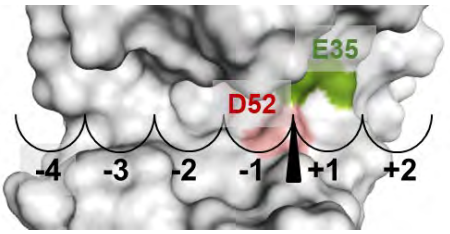


図 1 HEWL（PDB: 4HP0）の糖結合サブサイト
E35、D52 はそれぞれ触媒残基。

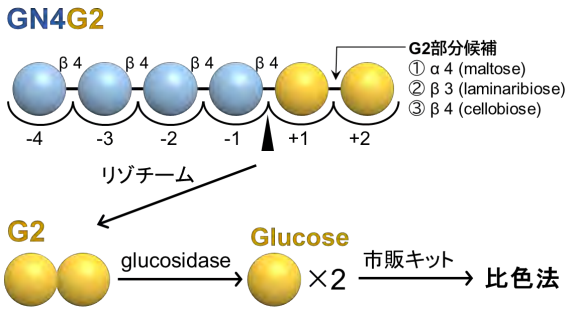


図 2 HEWL 新規基質の設計と活性測定スキーム
青丸が N-アセチルグルコサミン（GN）、黄色がグルコース（G）。GN4G2 模式図下部にリゾチームのサブサイトおよび加水分解位置、右側に 3 種類の G2 候補を示した。

結果および考察

まず、基質候補である 3 種類の GN4G2 をリガンドとし、HEWL をレセプターとしたドッキングシミュレーションを行ったところ、3 種類のリガンド全てにおいて 6 糖が各サブサイト近傍に位置するドッキングポーズが得られた。それらを初期構造とした MD シミュレーションを行ったところ、G2 部分が cellobiose の場合のみ HEWL の +1, 2 サブサイトと安定した相互作用を示した（図 3）。また、MM-PB/SA 法による結合自由エネルギー計算において+側サブサイト近傍の残基に絞って結合自由エネルギーへの寄与を解析したが、この結果も cellobiose が最も良好であった。

実際に HEWL の糖転移活性を利用した基質合成において、受容体に cellobiose を用いた場合顕著な生成物が観察された。また、合成した GN4-cellobiose は活性測定に十分な感度を有していることも確認された。このことから、基質設計において今回用いた分子シミュレーションによる相互作用評価法の有用性が示された。

まとめ、今後の課題

今回我々は、3 種類の GN4G2 のうち最も優れた基質をドッキングシミュレーションおよび MD シミュレーションを用いた酵素 – 基質相互作用解析により判定した。リガンドが自由度の高い 6 糖であったが、実験結果と矛盾無い結果を得ることができた。今後は還元末端をグルコビオース以外にも広げ、TSUBAME4.0 の演算能力をより活用した大規模なスクリーニングへの発展も期待できる。そのためには、大量の計算結果をより効率よく判定するための系構築などが課題となる。

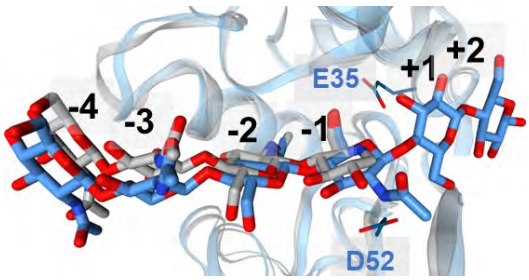


図 3 HEWL – GN4-cellobiose 複合体 MD シミュレーション
100 ns のプロダクションラン中 95 ns 時点のスナップショット（青）と HEWL – GN3-moranoline（GN3M）複合体結晶構造（灰色、PDB: 4HP0）との構造アラインメント。GN4 の結合様式も結晶構造との重なりが概ね良好である。+側は参照となるリガンドが無いものの、cellobiose がサブサイトに張り付くように相互作用していることが分かる。他の 2 リガンドではサブサイトから G2 部分が離れる傾向にあった。

2. シアル酸転移酵素とリガンドの相互作用解析

背景と目的

シアル酸転移酵素の一つである ST3Gal3 は、GT ファミリー29 に属し、N-アセチルラクトサミン（LacNAc）誘導体のガラクトース（Gal）にシアル酸（Sia）を α-2,3 結合で転移させる。本酵素は特に脳において重要な役割を持ち、その変異は認知機能障害の発達に障害を引き起こす。しかし、現時点で立体構造が未知であるため、変異による影響のメカニズム解析などが難しい。一方、近年は蛋白質立体構造予測技術の発展が著しい。我々のグループでも AlphaFold にてモデリングした rat ST3Gal3 (rST3Gal3) 立体構造情報を基に重要残基を予測後、変異体アッセイを行うことで、図 3 の残基における活性への影響を評価できている。これら残基がどのような機構で酵素機能に関与しているかを理解するためには、リガンドとの相互作用解析が欠かせない。そこで分子シミュレーションにて rST3Gal3 各残基とリガンドの相互作用を解析した。

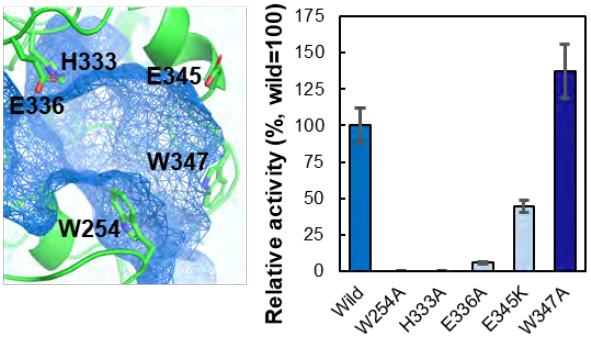


図 4 rST3Gal3 予測構造と主要残基の変異体アッセイ
興味深いことに W347A では比活性が向上した。クレフト反対側に位置する W254A とは対照的である。ドナーは CMP-Sia、アクセプターは LacNAc-dansyl である。

概要

シアル酸転移酵素の構造解析は複雑である。まず転移酵素なのでリガンドはドナー、アクセプターの 2 種類が存在する。そのため相互作用を理解するにはアミノ酸残基とリガンド間のみならずドナー – アクセプター間相互作用も重要となる。さらに、構造既知のシアル酸転移酵素サブファミリーの測定データを見るに、その構造はリガンドを含め不安定な傾向にある。そのため、複合体の相互作用理解には MD シミュレーションによるダイナミクス解析が重要と考え、中心的アプローチとして採用した。MD シミュレーション用の初期構造は、レセプターに rST3Gal3 予測構造、ドナーに CMP-Sia、アクセプターに LacNAc-OMe を用いて取得したドッキングポーズを用いた。この際、前述のようにドナー – アクセプター間相互作用を考慮するため、2 つのリガンドを同時にドッキングさせる手法を取った。その後 MD シミュレーションを行い rST3Gal3 のリガンド結合様式の解析を行った。

結果および考察

200 ns 間の MD シミュレーションで得られたスナップショットを、複合体構造が既知の他サブファミリーと比較したところ結合様式は概ね類似していた。提唱されている反応機構も合わせて考えた際、概ね妥当な様式が得られたと考えられる。変異体アッセイ結果との比較に関しては、アッセイ結果が対照的であった W254、W347 に注目して述べる。双方の残基とも MD シミュレーション中、アクセプターの LacNAc-OMe と相互作用していた点は共通だが、前者はアクセプターとスタッキングのような相互作用を維持していたのに対し、後者のインドール環と LacNAc の N-アセチルグルコサミン (GlcNAc) 残基は面同士垂直に近い状態であった（図 5）。Ala 置換による活性の向上も合わせると W347 はアクセプター結合領域における若干の立体障害を引き起こしていると考えられる。

まとめ、今後の課題

今回我々は、リガンド結合様式が不明である ST3Gal3 に対して MD シミュレーションを中心とした手法を用い、ドナー、アクセプターとの複合体構造推定と相互作用解析を行った。シアル酸転移酵素はドナー結合領域に関してはファミリー内で保存性の高いモチーフが存在するため、アミノ酸残基との相互作用に関しては他のサブファミリー情報が一定程度参考になるが、アクセプターはサブファミリー毎に異なるため残基の保存性が非常に低い。そのため、変異体アッセイ結果とシミュレーション結果を組み合わせて ST3Gal3 のアクセプター結合領域に関する知見が得られた点は大きい。課題としては、より詳細な相互作用解析のための複合体初期構造取得があげられる。ドナーの構造上、ドッキングシミュレーションにて妥当性のある複合体構造を取得するまでかなりの試行錯誤を要した。現在は AlphaFold3 が TSUBAME4.0 上でも利用可能なので、今後こちらを用いた複合体構造取得も検討したい。

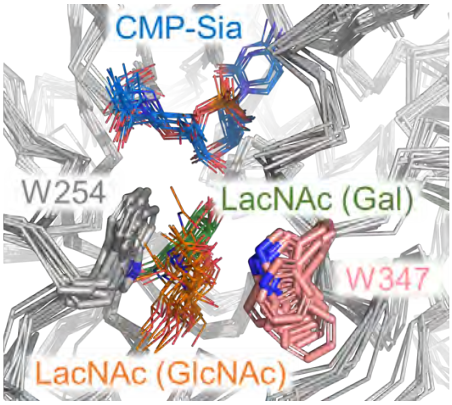


図 5 rST3Gal3 – CMP-Sia – LacNAc-OMe 複合体の MD シミュレーション

200 ns の MD シミュレーションにおける初期フレームの主鎖 Cα 原子を参照構造とした構造アラインメントのうち、最終 60 ns 分 10 フレームを表示している。

TSUBAME 共同利用　令和6年度　学術利用　成果報告書

MD シミュレーションと量子化学計算を用いた医薬品開発支援
MD and MO calculations for supporting drug development process

山下　雄史
Takefumi Yamashita

星薬科大学
Hoshi University
https://www.hoshi.ac.jp/

本研究は、スーパーコンピュータ TSUBAME4.0 の計算能力を活用し、MD シミュレーションと量子化学計算を組み合わせて医薬品開発を支援する方法を探求することを目的とする。今年度は、まず医薬品として期待されるデュオカルマイシンの反応性に関する基礎的な知見を得ることを目標とした。前年度に引き続き、分子軌道法計算によって反応性を定性的に予測することが可能かを検証するため、デュオカルマイシン様分子の solvolysis に着目した。前年度よりも多くの分子を対象に計算を行なっているが、量子化学計算の結果はすでに報告のある実験結果の傾向と定性的によく一致していたことが分かった。得られたデータは、MD 計算などに活用できる分子モデルを構築するのに必須のものである。さらに、MD シミュレーションを用いて、卵白リゾチームを認識する VHH 抗体の変異の影響を調査した。得られたデータは、親和性向上のための抗体変異設計の指針として活用する予定である。

This study aims to explore methods for supporting drug development by combining MD simulations and quantum chemical calculations, leveraging the computational power of the supercomputer TSUBAME4.0. This year, our first goal was to gain fundamental insights into the reactivity of the drug candidate, duocarmycin. Continuing from the previous year, we focused on the solvolysis of duocarmycin-like molecules to verify whether molecular orbital calculations can qualitatively predict their reactivity. Although we conducted calculations on more molecules compared to last year, we found that the quantum chemical calculation results were in good qualitative agreement with previously reported experimental trends. The data obtained are essential for constructing molecular models that can be used in MD simulations and related analyses. In addition, we investigated the effects of mutations in VHH antibodies that recognize hen egg-white lysozyme, by using MD simulations. The data acquired are expected to serve as guidelines for designing antibody mutations aimed at improving affinity.

Keywords: MD simulation, Molecular orbital calculations, Solvolysis, Antigen-antibody interaction

背景と目的

昨年度までの研究では、スーパーコンピュータ TSUBAME の高度な計算能力を活用し、量子化学計算を用いてデュオカルマイシン様分子の化学的反応性を解析し、その知見を活用した医薬品設計の可能性を探ってきました。昨年度は、強力な薬効を持つ一方で副作用が懸念されるデュオカルマイシンについて、AMDC（Antibody-mimetics-drug conjugate）の形態を用いたがん特異的輸送の試みや、マウスモデルにおける腫瘍消失の実証といった基礎的研究成果を得ることができました。

本年度は、これらの成果を一層発展させることを目的としています。具体的には、より多くのデュオカルマイシン様分子を対象に詳細な反応機構を調査し、MD シミュレーションや分子軌道法計算によって得られる電子構造情報を統合的に解析することで、副作用低減と薬効維持・向上の両立を目指すための分子設計指針を提示することを主たる狙いとしました。また、AMDC 手法や抗体工学との連携も視野に入れ、薬剤と抗体の相互作用を計算科学的に評価することで、次世代のバイオ医薬品設計への応用可能性を検討します。これにより、実験単独では得られにくい分子レベルでの新たな知見を創出し、医薬品開発プロセスにおける効率化や精密化に貢献することを目指しています。

概要

本研究ではデュオカルマイシン様分子の反応性を評価するため、昨年度に続き MP2/6-31G(d)/PCM 計算を

実施し、スーパーコンピュータ TSUBAME4.0 の計算資源を活用しました。複数の分子種を対象とすることで、加溶媒分解（solvolysis）の反応メカニズムを包括的に検討し、実験結果と定性的に一致する知見を得ました。これらの成果は、デュオカルマイシン様分子の改変や医薬品開発を進める上での指針となり得ます。

また、卵白リゾチームを認識する VHH 抗体と卵白リゾチームの複合体およびその変異体に対して、本研究ではAMBER力場を用いたMDシミュレーションを実施しました。計算で求められるエネルギー変化を実験データと比較し、親和性向上の可能性を検討しました。この結果、抗原結合部位への特定の変異が相互作用ネットワークを変化させることが示唆され、抗体改変やバイオ医薬品開発の基盤知識を蓄積する成果につながりました。

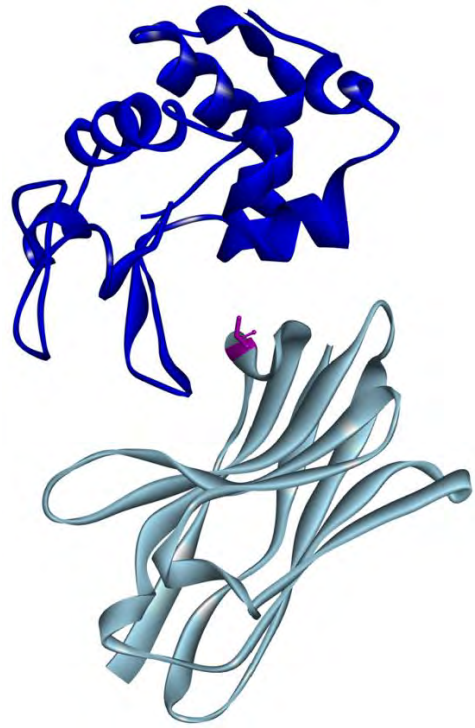


図:VHH 抗体(シアン)と卵白リゾチーム（ブルー）の複合体

結果および考察

今年度は MP2/6-31G(d)/PCM を用いた量子化学計算でデュオカルマイシン様分子の反応経路やエネルギープロファイルを評価し、溶媒効果や置換基の電子的特性との関連を前年度よりも定量的に明らかにしました。反応中間体や遷移状態の解析から、実験報告と良好に対応する結果が得られ、計算手法の妥当性と分子設計への有用性が再確認されています。

VHH 抗体に関しては、変異によるエネルギー的影響を解析し、結合親和性や抗原認識に関わる重要な要素を特定しました。これにより、デュオカルマイシン様分子の設計指針とともに、抗体改変の有力な戦略を提示できると期待されます。

まとめ、今後の課題

本年度の研究では、TSUBAME4.0 を活用した量子化学計算を通じて、デュオカルマイシン様分子の反応機構をさらに深く理解するとともに、VHH 抗体変異が結合親和性に及ぼす影響を評価しました。

今後は、より多様な分子や置換基への適用を拡大し、実験結果との連携を強化することで、分子設計の指針を一層洗練させる必要があります。また、VHH 抗体変異の解析については、計算と実験を緊密に組み合わせることで、変異の効果を定量的に検証し、抗体エンジニアリングにおける精密な設計指針を確立することが求められます。最終的には、計算科学と実験の融合により副作用低減と効果最大化を両立する医薬品が期待されます。

TSUBAME 共同利用　令和 6 年度　学術利用　成果報告書

利用課題名　ウイルスの形質および進化予測法の開発

英文：Development of methods for predicting virus phenotypes and evolution.

伊東　潤平
Jumpei Ito

東京大学医科学研究所
The Institute of Medical Science, The University of Tokyo
https://www.ims.u-tokyo.ac.jp/SystemsVirology/

邦文抄録（300 字程度）

ウイルスは著しい速度で進化することで抗原性を変化させ、宿主の獲得免疫から逃避する。そのため、効果的なワクチンを供給するためには、ウイルスの抗原性変化に基づき、最新の流行株に対して有効なワクチン株を選定する必要がある。ウイルスの抗原性は中和試験等により評価できるが、理論的には抗原タンパク質の配列のみから予測可能であると考えられる。本研究では、ウイルスの配列情報に基づいた抗原性監視体制と、より効率的なワクチン株選定戦略の樹立に向けて、季節性インフルエンザウイルスを対象に抗原性予測 AI モデル「PLANT」を開発した。

英文抄録（100 words 程度）

Viruses evolve at a remarkable rate, altering their antigenicity to evade the host's acquired immunity. Therefore, to supply effective vaccines, it is necessary to select vaccine strains that are effective against the latest circulating strains based on antigenic changes in the virus. The antigenicity of a virus can be evaluated through neutralization assays and other methods, but theoretically, it should be possible to predict antigenicity solely from the sequence of the antigenic protein. In this study, we developed the antigenicity prediction AI model **"PLANT"** for seasonal influenza viruses, aiming to establish an antigenicity surveillance system based on viral sequence information and a more efficient vaccine strain selection strategy.

Keywords: 5つ程度
AI, Virus, Vaccine development

背景と目的

ウイルスは著しい速度で進化することで抗原性を変化させ、宿主の獲得免疫から逃避する。そのため、効果的なワクチンを供給するためには、ウイルスの抗原性変化に基づき、最新の流行株に対して有効なワクチン株を選定する必要がある。ウイルスの抗原性は中和試験等により評価できるが、理論的には抗原タンパク質の配列のみから予測可能であると考えられる。

概要

本研究では、ウイルスの配列情報に基づいた抗原性監視体制と、より効率的なワクチン株選定戦略の樹立に向けて、季節性インフルエンザウイルスを対象に抗原性予測 AI モデル「PLANT」を開発した。

結果および考察

本モデルの学習には、大規模ゲノムデータベース GISAID に登録されている季節性インフルエンザウイルス H3N2 亜型のヘマグルチニン（HA）タンパク質配列と、英国クリック研究所 Worldwide Influenza Centre lab より提供を受けた中和試験のデータを用いた。PLANT の予測性能を評価するため、学習に用いていない新規変異株を用いて抗原性の予測を行ったところ、中和試験により実測された値と高い一致性を示した。これらの結果は、PLANT が配列情報のみに基づき、未知の変異株の抗原性を高い精度で予測できることを示している。

まとめ、今後の課題

現在、PLANT を用いたウイルス進化解析を進めている。本研究成果をまとめたプレプリントを、2025 年上旬中にプレプリントとして公開予定である。

多地点越流を考慮した次元圧縮手法によるリアルタイム浸水域予測モデルの構築

中央大学大学院 学生会員 ○ 中山　龍也　中央大学大学院 学生会員　羽物　裕人
日本工営　正会員　一言　正之　中央大学　正会員　樫山　和男

1. はじめに

これまで著者ら[1]は次元圧縮手法を深層学習に適用したリアルタイム浸水域予測モデルの構築を行った．しかし，従来の研究は特定地点での越流条件に基づく適用が中心であり，多様な越流シナリオに対応する汎用性に課題が残されている．

本研究は，複数地点からの越流を考慮した浸水域予測モデルを構築し，異なる条件下でも高い適用性を有するモデルの実現を目指す．異なる 3 つの圧縮次元手法を適用し，予測精度の比較から最適な浸水域予測モデルを検討した．

2. 解析手法

(1) モデル構築と解析の手順

フローチャートは図-1 に示す．浸水域推定モデルは，物理法則に基づく洪水氾濫シミュレーションモデルの結果を学習することによって構築を行った．図-1 中の (a) 学習時では以下の手順でモデルの構築を行う．

1. 越流箇所や浸水規模など様々な浸水状況シナリオに応じた物理法則に基づく洪水氾濫シミュレーションを実施．
2. 洪水氾濫シミュレーションの計算メッシュごとの浸水深を抽出し，浸水の疑似観測データを作成．
3. 次元圧縮手法を用いて，浸水の疑似観測データから特徴量と係数行列を抽出．
4. 深層学習を用いて，越流水深の時系列データから係数行列を推定する予測器を構築．

以上の手順によって構築された浸水域予測モデルを使用して，浸水観測情報に基づき浸水範囲を推定を行う．

(2) 浸水解析モデル

氾濫浸水解析には，氾濫水の挙動を精密に表現可能な Dynamic Wave 法を適用した．以下の x, y 方向の運動量保存式と質量保存式により構成される．その式を (1)，(2)，(3) に示す．

$$\frac{\partial M}{\partial t}+\frac{\partial (uM)}{\partial x}+gh\frac{\partial H}{\partial x}=-gh\frac{n^2v\sqrt{u^2+v^2}}{R^{\frac{4}{3}}} \quad (1)$$

$$\frac{\partial N}{\partial t}+\frac{\partial (uN)}{\partial y}+gh\frac{\partial H}{\partial y}=-gh\frac{n^2v\sqrt{u^2+v^2}}{R^{\frac{4}{3}}} \quad (2)$$

$$\frac{\partial h}{\partial t}+\frac{\partial M}{\partial x}+gh\frac{\partial N}{\partial y}=0 \quad (3)$$

ここで，$M=uh$ は x 方向の単位幅流量，$N=vh$ は y 方向の単位幅流量，u は x 方向の流速，v は y 方向の流速，h は水深，H は基準面からの水位，R は径深，g は重力加速度，n はマニングの粗度係数である．

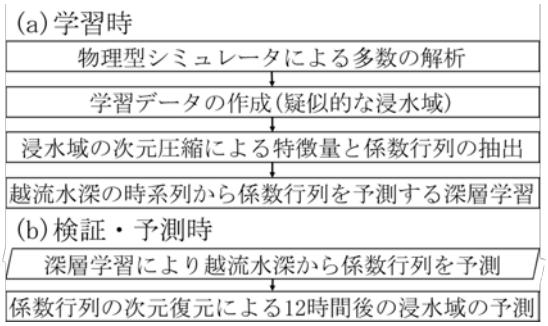


図-1　浸水域予測モデルの構築・検証手順

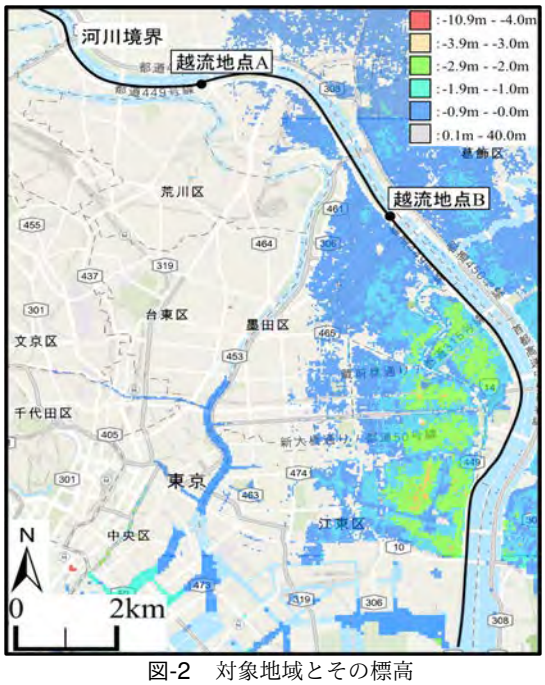


図-2　対象地域とその標高

(3) 学習データ

本研究では，複数の越流地点における 1 時間ごとの越流水深を入力データとし，洪水氾濫計算プログラムで得られる浸水深を学習データとした．1 時間ごとの越流水深を説明変数，目的変数を係数行列に設定する．越流水深の時系列データは，正弦波にノイズを付与して生成した式 (4) を以下に示す．

$$h=\left\{\sin\left(a-0.5+\frac{\pi t}{720}\right)+0.3b\right\}\times 2c \quad (4)$$

式 (4) の計算で $h<0$ となった場合は 0 に置き換えた．なお，a，b，c は 0〜1 の間の乱数とし，t は時間（分）を表す．荒川下流のデルタ地帯を対象地域とし，標高分布を図-2 に示す．標高データ・土地利用データには国土地理院の国土数値情報[2]を使用し，25m メッシュで最近隣内挿法を用いた．解析条件は地点 A，地点 B からそれぞれ，同時越流の 3 条件とし，各 1000 ケースを実施した．

KeyWords：　深層学習, 次元圧縮，洪水氾濫，防災，予測
連絡先：　〒112-8551 東京都文京区春日 1-13-27　TEL：03-3817-1815　E-mail :a20.3emg@g.chuo-u.ac.jp

表-1　次元圧縮の実行条件

対象データ	3000case の氾濫域計算結果
入力次元数	221392 次元 (404 × 548)
次元圧縮手法	特異値分解
	非負値行列因子展開
	AutoEncoder
圧縮次元数	1 から 20 次元（1 次元ごと）

表-2　AutoEncoder モデルの学習条件

訓練/テストデータ	2400/600case
学習回数	50 回
中間層の数	2 層 (256-128)
バッチサイズ	32
検証方法	交差分割

(4)　次元圧縮

次元圧縮は, 高次元のデータの重要な特徴を保持したまま低次元で表現する方法である．データの本質的な特徴を保ちつつ，次元数を減らすことで深層学習が推論に要する計算効率が向上する．今回は，特異値分解[3]，非負値行列因子展開[4]，AutoEncoder(以後 AE)[5] を用いて氾濫域データを圧縮し，得られた係数行列を深層学習に適用した．

(5)　深層学習手法

本研究では, Deep Neural Network; DNN を採用した．深層学習は, 機械学習モデルで、大量のデータから特徴を自動的に学習するアルゴリズムである．本研究では，1 時間ごとの越流水深の数値を説明変数, 各ケースごとの係数行列を目的変数とした深層学習モデルを構築する．

(6)　開発環境と学習条件

本例題で行った次元圧縮の実行条件，AE モデルおよび DNN モデルの各学習条件は表-1，2，3 に示す．表-3 に示す学習回数は損失が 10 回続けて一定の場合は早期に終了する．浸水域の予測に関する評価指標には 2 乗平均平方根誤差（Root Mean Square Error ; RMSE）を式 (5) に示す．本研究では，東京科学大学のスーパーコンピュータ TSUBAME4.0 を利用して実施した．

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(a_i - f_i)^2} \qquad (5)$$

なお，n はデータ数，a_i は復元された際の浸水深，f_i はシミュレーションによる浸水深である．

3.　学習結果

図-3 は RMSE と圧縮次元数の関係を示す．図-3 より，圧縮次元数の増加から RMSE は減少し，表現力が向上する傾向を確認した．いずれの手法においても圧縮次元数が 1 から 10 に増加するにつれて RMSE は急激な低下を示し，それ以降は緩やかな減少傾向を示した．今回の結果から AE と DNN を組み合わせたモデルが最も高精度であった．その一方，次元数の増加に伴い性能が悪化する場合もあり，適

表-3　DNN モデルの学習条件

訓練/テストデータ	2400/600case
説明変数	越流水深
目的変数	係数行列
学習回数	最大 500 回
中間層の数	4 層 (128-64-32-16)
バッチサイズ	32
損失関数	平均二乗誤差
検証方法	交差分割

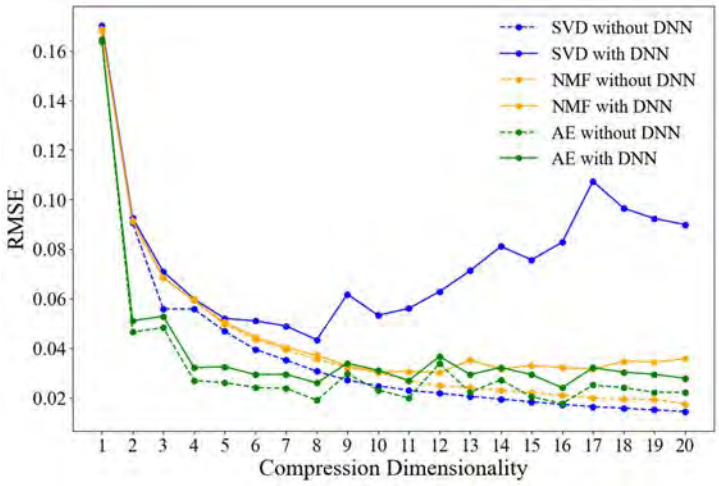


図-3　RMSE と圧縮次元数の関係

切な次元数の選定には検討の余地がある．

4.　おわりに

本研究では，多地点の越流を考慮した浸水域予測モデルを構築し，その汎用性の向上を目的として解析を行った．異なる 3 種類の次元圧縮手法を適用し，予測精度の比較を実施した結果，以下の知見を得た．

- 本手法は，任意の地点から発生する越流に対して，高精度な浸水域予測が可能である．
- オートエンコーダを適用した浸水域予測モデルが最も高精度であった．

今後は，最適な圧縮次元数の選定に取り組むことで，更なる予測精度の向上と計算コストの低減を目指す．

参考文献

1) 中山龍也, 羽物裕人, 一言正之, 樫山和男：深層学習を用いたリアルタイム 浸水域予測における次元圧縮の適用 ，AI・データサイエンス論文集 5 巻 3 号，p. 563-571，2024.
2) 国土地理院， 国土基盤情報 https://www.gsi.go.jp/kiban/ (参照 2023-10-10)
3) Eckart C. and Young G., The approximation of one matrix by another of lower rank, *Psychometrika*, 1(3), 211-218, 1936.
4) Lee D. D. and H. S. Seung, Learning the parts of objects by non-negative matrix factorization, *Nature*, 401(6755),788-791, 1999.
5) Rumelhart D. E., G. E. Hinton and R. J. Williams, Learning representations by back-propagating errors. *Nature*, 323(1), 533-536, 1986.

TSUBAME 共同利用　令和 6 年度　学術利用　成果報告書

利用課題名　集風レンズ付き風車とそのマルチロータシステムの流体シミュレーション
英文：CFD Simulations of Diffuser Augmented Wind Turbines and Multirotor Systems

胡　長洪
Changhong Hu

九州大学　応用力学研究所
Research Institute for Applied Mechanics, Kyushu University
URL https://www.riam.kyushu-u.ac.jp/

邦文抄録（300 字程度）
本研究グループは定格出力 200kW の中型レンズ風車およびその 2 基マルチロータシステムの研究開発を進めている．本利用課題では，TSUBAME4.0 を利用した数値流体シミュレーションにより，200kW レンズ風車の発電性能，風荷重，空力特性を評価する．12 多角形でつば開閉機構を搭載した新しい集風レンズに対して，つば高さを変更したシミュレーション，レンズ支持構造を考慮したシミュレーションを実施し，目標とする発電性能（パワー係数 0.7 以上）を達成するためには，つばの高さを風車直径の 10%ほどにする必要があることがわかった．2 基マルチロータの設置間隔を変えたシミュレーションを実施し，すきま間隔がレンズ直径の 10%から 30%の範囲で約 2%の発電性能増加が得られることを確認した．

英文抄録（100 words 程度）
Our research group is currently developing a medium-scale wind lens turbine with a 200 kW rated power output and its multi-rotor system. This study assesses the power coefficient, wind loads, and aerodynamic characteristics of the wind lens turbine using CFD simulations on TSUBAME4.0. Simulations varied the brim height and lens support structure. Results indicate that to achieve a target power coefficient of 0.7 or higher, the brim height should be approximately 10% of the turbine diameter. Additionally, multi-rotor simulations varying turbine spacing showed a potential 2% increase in power output within a 10% to 30% range of the lens diameter.

Keywords: Diffuser augmented wind turbine, Aero dynamics, Multi rotor system, Lattice Boltzmann method

背景と目的

高い発電性能を持つ風車として，集風レンズ付き風車（レンズ風車）が注目されている．本研究グループは九大発ベンチャーである（株）リアムウィンドの共同実施者として，環境省の地域競争・セクター横断型カーボンニュートラル技術開発・実証事のプロジェクトで定格出力 200kW の中型レンズ風車およびそのマルチロータシステムの研究開発を進めている．本研究課題では，200kW レンズ風車用に開発された 12 多角形つば開閉機構付き集風レンズに対して数値流体シミュレーションを行い，発電性能や風荷重を評価することを目的とする．また，200kW レンズ風車を 2 基並べたマルチロータシステムの発電性能やウェイクについても調査する．

概要

12 多角形でつば開閉機構を搭載した新しい集風レンズに対して，つば高さを変更したシミュレーション，レンズ支持構造を考慮したシミュレーションを実施した．流体計算には GPU と大規模計算に適した格子ボルツマン法を利用した．環境省事業で目標とする風車の発電性能を示すパワー係数が 0.7 以上を達成するためには，つばの高さは風車直径の 10%ほどにする必要があることがわかった．2 基構成のマルチレンズ風車の設置間隔を変えたシミュレーションを行い，発電性能を調べた．レンズ間にすきまがない場合はマルチロータにすることで発電性能が低下するが，すきまを設けることで発電性能は向上し，すきま間隔がレンズ直径の 10%から 30%の範囲で約 2%の発電性能増加が得られた．

結果および考察

<u>シングル風車の計算</u>

つば開閉機構を搭載した 12 角形集風レンズを 200kW 中型レンズ風車へ採用可能か検討するために，縮尺 1/21 のロータ径 1 m モデルに対して数値流体力学シミュレーションを実施した．流体シミュレーションには格子ボルツマン法を，タービンのモデル化にはアクチュエーターラインモデルを用いた．計算対象の形状モデルを図 1 に示す．つば高さが Ci タイプ 6.7%相当のモデル，および可動部のつばを Ci タイプ 10%相当まで延長したつば拡大モデルの 2 種類である．また，タワーとレンズ支持構造物の影響を調査するため，それらの構造を除いたモデルのシミュレーションも実施した．流入風速は 11 m/s の一様流れとし，ロータの回転数は周速比が 4.0 となるように設定した．

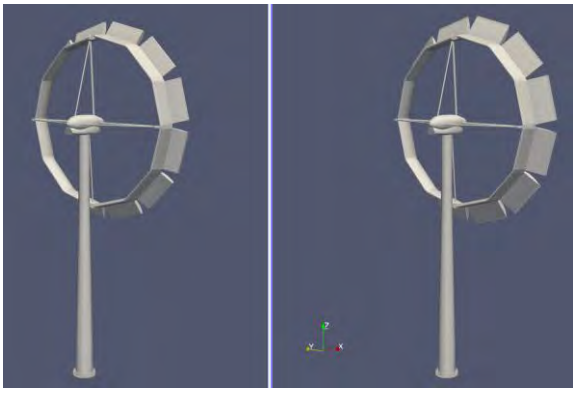
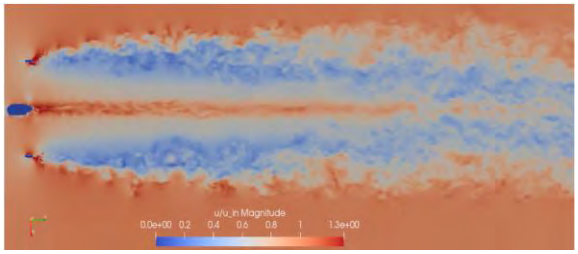


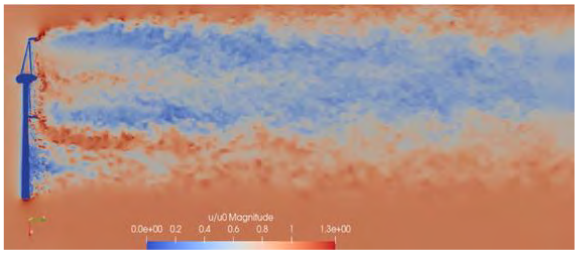
図 1 つば開閉機構付きレンズ風車モデル
（左：つば高さ Ci6.7%相当．右： Ci10%相当）

シミュレーションで得られた風車近傍の速度場を図 2 に示す．一例として，つば高さ Ci6.7%相当のレンズで，支持構造なしのモデルとタワーとレンズ支持構造を加えた形状モデルの結果を比較している．色は速度の絶対値を流入風速 11m/s で除して無次元化した値を示す．つばの隙間に高速な流れが発生していることがわかる．タワーとレンズ支持構造物はロータ面への流入を乱している．

タワーとレンズ支持などの風車構造物がない場合，パワー係数はつば 6.7%で 0.757，つば 10%で 0.804 であり，どちらも目標としているパワー係数 0.7 以上となった．しかし，ダウンウィンドタイプであるレンズ風車では，タワーとレンズ構造物はロータ面の風を乱し性能を低下させ，つば 6.7%では 0.692，つば 10%では 0.721 となった．Cp が 0.7 以上を達成するには，つば高さを風車直径の 10%相当にする必要があることが CFD シミュレーションから確認された．



(a) 支持構造物なし



(b) 支持構造物あり

図 2 つば開閉機構付きレンズ風車のシミュレーション結果

<u>マルチロータ風車の計算</u>

つば開閉機構付きレンズ風車の 2 基構成マルチレンズ風車の CFD シミュレーションを実施し，2 基風車のすきまが発電性能に与える影響を調査した．ロータ直径は 0.442 m，つば直径は 0.594 m，流入風速は 17 m/s，周速比は 4.0 と設定した．2 基の風車レンズ間のすきま s はつば直径を D_{brim} として s/D_{brim} =0%，5%，10%，20%，30%，40%，50%とした 7 ケースを設定した．

2 基構成マルチレンズ風車の発電性能を，2 基のシングルレンズ風車の発電性能 $C_{P,SA}$ からの増加率 ΔC_p で評価する．シングルレンズ風車と比較したときのマルチレンズ風車の各風車の発電性能増加率 ΔC_{pi} を

$$\Delta C_{Pi} = \frac{C_{Pi}}{C_{P,SA}} - 1, \qquad (\mathrm{i} = 1, 2)$$

と定義する．CFD で得られたレンズ間のすきま間隔 s/D_{brim} に対する発電性能増加率 ΔC_p を図 3 に示す．すきまがない場合（$s/D_{\mathrm{brim}} = 0$）ではマルチロータにすることで発電性能が低下している．レンズ間にすきまを設

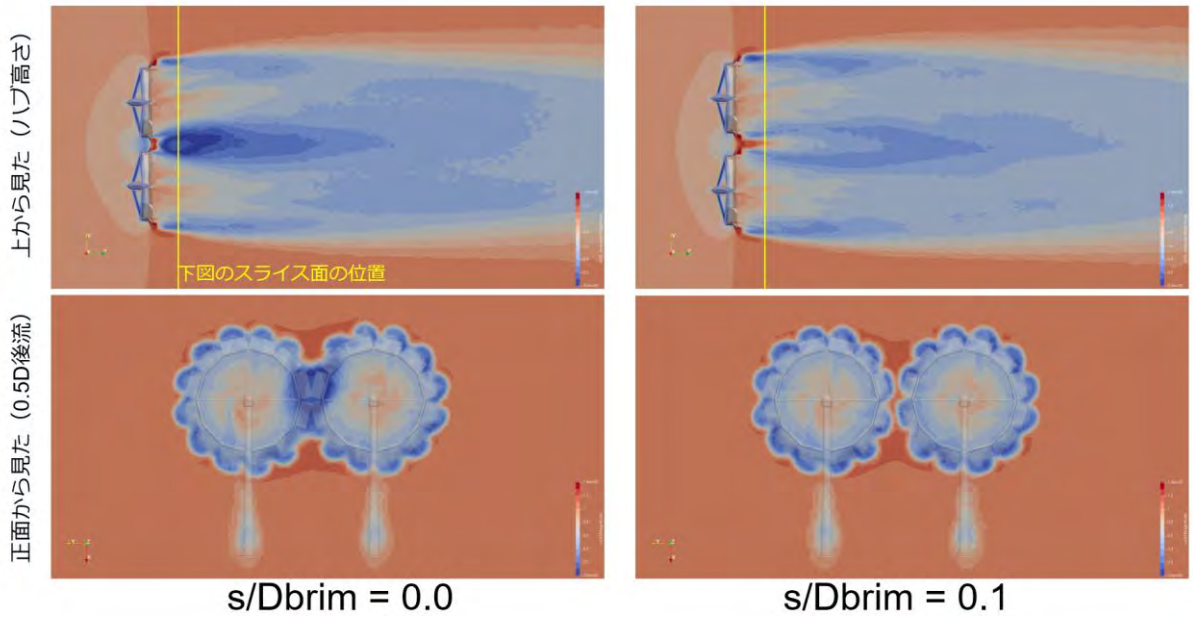


図 4 2 基構成マルチレンズ風車の CFD シミュレーション結果

けることで発電性能は向上し，すきま比が 10%から 30%の範囲で約 2%の発電性能増加が得られた．すきまが大きくなるほどマルチロータ用の支持構造が大きくなりコスト増加につながるため，高い発電性能増加率を得られつつ，レンズ間のすきまは可能な限り小さくするのが望ましい．CFD の結果から，マルチロータの適切なレンズ間のすきま間隔はレンズ直径の 10%であると考えられる．

すきま間隔が 0%と 10%の平均風速を風車の上から見た流れ場と正面からみた流れ場の 2 通りで図 4 に示す．すきまがない場合，2 基の風車の間で低速領域が形成されている．一方，すきまがあるとすきまから高速な流れが発生してレンズ後方の渦を増強し，その結果，マルチロータの発電性能が向上したと考えられる．

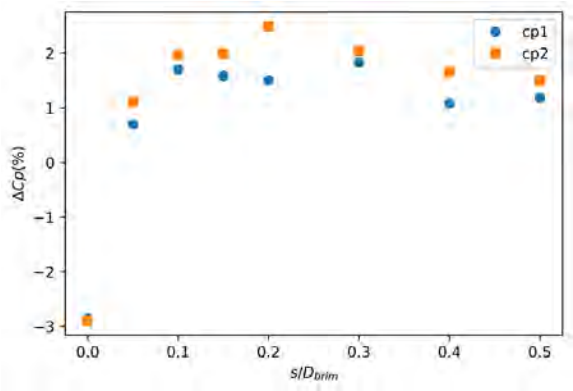


図 3 2 基構成マルチレンズ風車の発電性能増加率とレンズすきま間隔の関係

まとめ、今後の課題

本研究課題は，200kW 中型レンズ風車用に開発された新しい集風レンズの空力特性を格子ボルツマン法による流体シミュレーションで評価した．レンズ支持構造による流入風の乱れはパワー係数を 10%弱低下させることがわかり，パワー係数 0.7 以上を達成するためには，風車直径に対して 10%のつば高さが必要であることを CFD で示した．新しい集風レンズを用いた場合，2 基構成マルチロータシステムにすることで発電性能が 2%ほど向上することが明らかになった．

TSUBAME 共同利用　令和 6 年度　学術利用　成果報告書

利用課題名　ポリグルタミン病原因タンパク質の凝集および凝集阻害過程の理論研究
英文: Theoretical study of the formation and dissociation process of polyglutamine disease-related protein aggregates

谷本　勝一
Shoichi Tanimoto

独立行政法人国立高等専門学校機構　久留米工業高等専門学校
Kurume College, National Institute of Technology

邦文抄録(300 字程度)
神経変性疾患の一つであるポリグルタミン病は長いグルタミン鎖を含むポリグルタミンタンパク質(polyQ)が凝集し、神経細胞内に蓄積することで発症する。アルギニンとカテキンの一つであるエピガロカテキンガレートが polyQ の凝集に対する阻害効果を示し、ポリグルタミン病に対して治療効果をもつことが見出されたが、その分子論的メカニズムは解明されていない。本研究課題では、アルギニンおよびエピガロカテキンガレートが polyQ の凝集体を破壊し、その形成を阻害するメカニズムを解明することを目的として分子動力学シミュレーションを実行した。

英文抄録(100 words 程度)
Polyglutamine (polyQ) diseases, a collection of nine inherited neurodegenerative disorders, are caused by the aggregation of the disease-causative proteins with a long polyQ tract in neurons. Arginine and epigallocatechin gallate, a catechin, have been found to inhibit polyQ aggregation and to have therapeutic effects on polyglutamine diseases, but the molecular mechanism of this effect has not been elucidated. We performed molecular dynamics simulations to clarify the molecular mechanism by which arginine and epigallocatechin gallate disrupt the polyQ aggregation and inhibit its formation.

Keywords: ポリグルタミンタンパク質、アルギニン、EGCG、分子動力学シミュレーション、凝集阻害

背景と目的

神経変性疾患の一つであるポリグルタミン病は、長いグルタミン鎖(40 残基以上)を含むポリグルタミンタンパク質(polyQ)が分子内および分子間βシート構造をもつアミロイド線維を形成して神経細胞中に蓄積することにより発症する。アミノ酸の一つであるアルギニンおよびカテキンの一種であるエピガロカテキンガレート(EGCG)は polyQ モノマーが神経毒性をもつ分子内β-シート構造を形成するのを抑制し、なおかつ polyQ オリゴマーが凝集して線維構造を形成するのを阻害することから、ポリグルタミン病に対する有望な治療薬候補化合物として注目されている。一方で、これらの候補化合物が polyQ の凝集を阻害する詳細な分子論的メカニズムは完全には解明されていない。特に、一度βシート構造を形成した polyQ モノマーやオリゴマーに対して、これらの候補化合物がβシート構造を壊して毒性をもたないモノマーにするのか、壊すとしたらどのようなメカニズムによるのか、については全くわかっていない。本研究課題ではこれらの候補化合物の凝集阻害効果を解明するために、アルギニンおよび EGCG がβシート構造を形成した polyQ の構造安定性に及ぼす影響を分子動力学シミュレーションにより調査した。

概要

polyQ は他の神経変性疾患原因タンパク質と異なり、モノマーも毒性をもつことが実験研究により示唆されている。そこで本研究課題では、ポリグルタミン病原因タンパク質としてグルタミンが 43 残基連結したペプチドのモノマーを用いた。文献[1]で報告された円筒型の構造を初期構造として 62.145^3 Å^3 の立方体の中央に配置し、その周囲に水分子約 7600 分子と、NaCl を約 150 mM になるように配置した。この水溶液系を polyQ+none と呼ぶことにする。また、polyQ の周囲に両性イオン型のアルギニンを 15 分子、もしくは EGCG を 5 分子ランダムに配置した水溶液系を作成した。各水溶液系をそれぞれ polyQ+Arg、polyQ+EGCG と呼ぶことにする。これらの水溶液系に対して、温度 310 K、圧力 1 atm の *NPT* アンサンブルの条件で 110 ns のシミュレーションを四回ずつ実行した。

結果および考察

polyQ と接触したリガンド(アルギニンもしくは EGCG)の個数の時間変化を図 1 に示す。今回のシミュレーションでは、polyQ とリガンド間の最近接距離が 4.0 Å 以下のとき、リガンドが接触したと見なした。接触したリガンドの平均値はアルギニンが 2.95±0.18、EGCG が 3.35±0.28 となり、どちらの水溶液系においても、シミュレーション時間全体にわたって安定してリガンドが polyQ に接触したことがわかる。特に、EGCG においてはシミュレーション時間の後半以降、配置した EGCG がほぼすべて持続的に polyQ と接触していた。

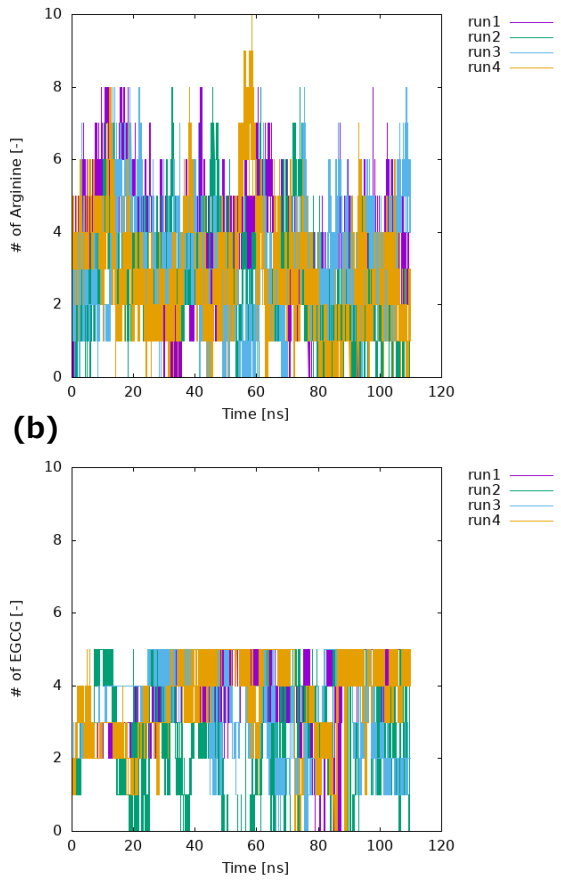


図 1. polyQ に接触したリガンドの個数の時間変化図。(a)polyQ+Arg、(b)polyQ+EGCG。

続いて、各水溶液系における polyQ の二次構造の時間変化を図 2 に示す。これらの図からアルギニンおよび EGCG が共存する水溶液系では、βシート構造が壊れていることがわかる。また、polyQ+Arg では、シミュレーションの間にヘリックス構造の形成が見られており、アルギニンが毒性をもたない構造へと polyQ モノマーを構造変化させていることが確認された。一方で、リガンドが非共存の polyQ+none でもβシート構造が同程度に壊れていた。これは本研究課題で初期構造とした円筒型の polyQ モノマーの構造の安定性が低かったために、シミュレーション時間内で自然に壊れてしまったためと考えられる。

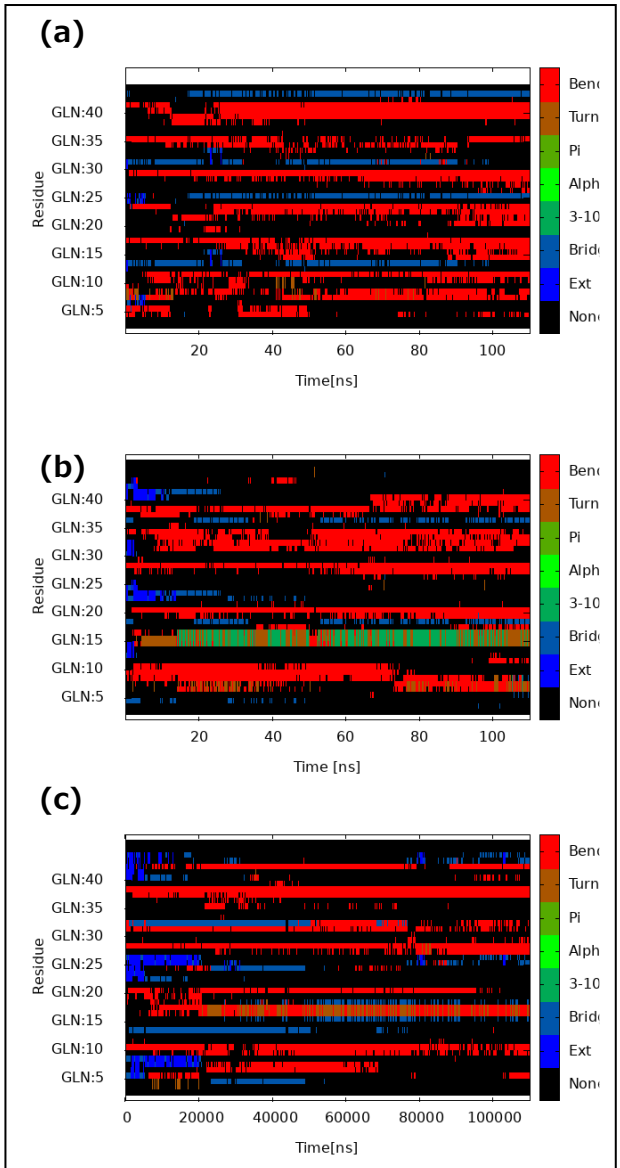


図 2. 各水溶液系における polyQ の二次構造の時間変化図。(a)polyQ+none、(b)polyQ+Arg、(c)polyQ+EGCG。縦軸はグルタミンの残基番号、横軸はシミュレーション時間を表す。青色がβシート構造、緑色がヘリックス構造を示している。

まとめ、今後の課題

アルギニン、および EGCG が polyQ モノマーの構造安定性に及ぼす影響について分子動力学シミュレーショ

ンを用いて解析したが、これらの候補化合物による polyQ モノマーのβシート構造不安定化のメカニズムの解明に至らなかった。現在、線維構造を形成した polyQ の構造安定性に対するアルギニン、および EGCG の影響を明らかにするべく、βシート構造を形成した polyQ オリゴマーに対しても同様のシミュレーション研究に取り組んでいる。

参考文献

[1] Ogawa H. *et al.*, *Comput. Biol. Chem.* **32**, 102 (2008).

TSUBAME 共同利用　令和 6 年度　学術利用　成果報告書

第一原理計算による惑星物質データベースの構築
Development of a Planetary Materials Database Using First-Principles Calculations

利用課題責任者　小松　勇
Yu Komatsu

茨城大学
Ibaraki University
https://www.eng.ibaraki.ac.jp/

邦文抄録（300 字程度）
氷惑星や氷衛星の内部構造を理論的に解明するには、Warm Dense Matter（WDM）状態にある揮発性分子の熱力学的性質を評価することが重要である。なかでも硫化水素（H_2S）は、氷惑星のマントルに含まれる成分の一つとされる。本研究では、第一原理分子動力学法と深層学習による機械学習ポテンシャル作成を統合したハイブリッド手法を構築し、500～10,000 K、1～5 g/cm^{3}の条件下での熱力学的特性を高精度に解析した。得られたデータは、WDM 領域の包括的データベースの整備に有用であり、惑星の重力モーメントと内部構造モデルの整合性検証にも貢献する可能性が示唆された。

英文抄録（100 words 程度）
Understanding the thermodynamic properties of volatile components in the Warm Dense Matter (WDM) regime is essential for modeling the interior structures of icy planets and satellites. In this study, we developed a hybrid approach combining *ab initio* molecular dynamics (AIMD) and deep learning-based interatomic potentials to investigate hydrogen sulfide (H_2S) over a wide range of temperatures (500–10,000 K) and densities (1–5 g/cm^3). The results reveal temperature- and density-dependent structural transitions and suggest that the constructed dataset can serve as a theoretical WDM database. Our findings also imply a potential impact of H_2S on the gravitational moments of icy planets.

Keywords: 惑星内部構造、硫化水素、Warm Dense Matter、機械学習ポテンシャル、データベース

背景と目的

氷惑星や氷衛星の形成・進化を理解するには、内部に存在する揮発性分子の熱力学的性質を広範囲の温度・圧力条件で解明する必要がある。なかでも硫化水素（H_2S）は、天王星や海王星の大気で検出され、氷マントルにも含まれるものと考えられる。しかし、H_2S の高温高圧下における相図や構造、拡散などの性質は実験・理論ともに十分なデータが得られていないのが現状である。

本プロジェクトでは、AIMD によって得られた高精度なデータを活用し、深層学習により原子間ポテンシャルを構築することで、広範な条件下での H_2S の物性データを効率的に取得する方法論を確立した。これにより、高精度かつ高速なシミュレーションによって、惑星内部構造モデルの構築に資する基盤データベースの整備を目指した。

概要

本研究では、温度 500～10,000 K、密度 1～5 g/cm^{3}の条件下で、H_2S の圧力、動径分布関数、拡散係数などを AIMD により計算し、そこから得たデータをもとに深層学習ポテンシャルを構築した。学習済みポテンシャルを用いたシミュレーションにより、AIMD 程度の精度を保ちつつ、100 倍以上の時間スケールおよび数百原子規模の系に拡張した分子動力学計算が可能となった。

結果および考察

図1は AIMD によって得られた密度-圧力関係である。また、図2は AIMD、機械学習ポテンシャルをそれぞれ用いて推定した動径分布関数を示しており、これらで良い一致を示している。温度上昇に伴う分子結合の解離や、高密度条件下での局所的な硫黄ネットワーク構造の形成などが確認された。低温・低密度で分子性

が維持される一方、高温になると動径分布関数のピークが消失し、流体的な構造へと移行する様子が示された。

さらに、機械学習ポテンシャルを用いて追跡することで、AIMD では困難であった長時間スケールにおける動的性質も調べることができるようになり、信頼性のより高い拡散係数を得ることができた（図3）。さらに、H_2S が氷マントルに含まれることで、惑星全体の密度が約0.7～0.8%増加し、重力モーメントの観測値と一致する内部構造モデルの提示にもつながる。

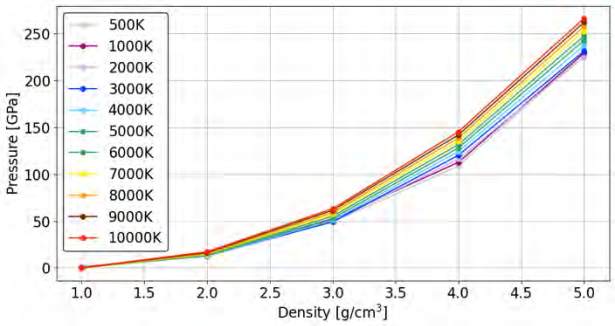


図 1 AIMD によって得られた密度-圧力関係を温度毎に示した。

まとめ、今後の課題

本研究では、第一原理計算と深層学習を統合した手法により、H_2S の WDM 領域を含む熱力学的特性を広範に調査した。これにより、氷天体内部の構造解析や観測値との整合性の検証に向けた理論的基盤が整備された。

今後は、本手法を H_2O や NH_3など他の揮発性成分に拡張し、多成分系における協調的効果の解明を目指し、包括的な WDM 理論データベースの構築を進める予定である。

なお、本研究の推進には九州大学、物性研究所、名古屋大学、国立天文台のスパコンも用いた。

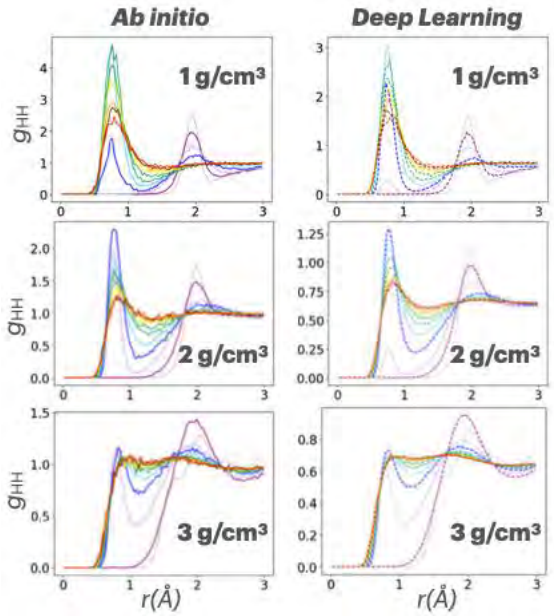


図 2 AIMD、深層学習を用いて動径分布関数 $g(r)$を算出し、比較した。線の配色は図1と同じ。

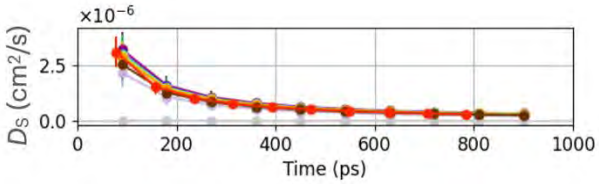


図 3 5 g/cm³ における拡散係数を、シミュレーション時間に対してプロットした。線の配色は図1と同じ。

TSUBAME 共同利用　令和6年度　学術利用　成果報告書

利用課題名　シナプスタンパク質間相互作用の解析
英文：Analysis of Synaptic Protein–Protein Interactions

利用課題責任者
Sotaro Ichinose

所属
Gunma University Graduate School of Medicine, Department of Anatomy, Maebashi, JAPAN

邦文抄録（300 字程度）
本研究では、抑制性シナプス形成に必須なシナプスオーガナイザーTeneurin-2（TEN2）の機能を解析した。TEN2 は微小管+端追跡タンパク質 EBs と相互作用し、抑制性後シナプスへの微小管リクルートを促進する。さらに、TSUBAME4.0 による解析と実験的検証により、TEN2 は EBs 以外の微小管関連タンパク質や微小管そのものとも結合することが明らかとなった。これらの知見は、TEN2 が細胞外では接着分子として、細胞内では微小管と直接結合することで、抑制性シナプスにおけるタンパク質集積を制御するハブとして機能していることを示唆する。

英文抄録（100 words 程度）
In this study, we investigated the function of Teneurin-2 (TEN2), a synaptic organizer essential for inhibitory synapse formation. We found that TEN2 interacts with microtubule plus-end tracking proteins (EBs) to facilitate microtubule recruitment to inhibitory postsynapses. Computational analysis using TSUBAME4.0 and subsequent experimental validation revealed that TEN2 also binds to other microtubule-associated proteins and directly to microtubules themselves. These findings suggest that TEN2 functions as a molecular hub at inhibitory synapses, coordinating protein accumulation by acting as an adhesion molecule via its extracellular domain and directly associating with microtubules via its intracellular domain.

Keywords: 5つ程度
Neuronal cell biology, synaptogenesis, protein-protein interactions, microtubules, adhesion molecules

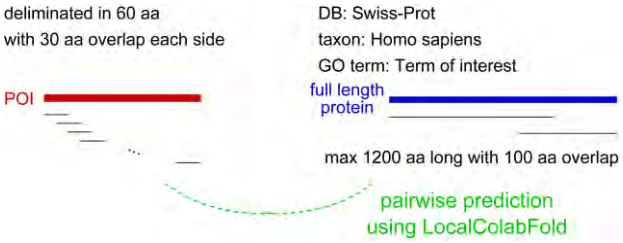
背景と目的

神経細胞はシナプスを介して相互に情報を伝達し、シナプス受容体を通じた情報交換により、複雑な神経回路を形成している。これらの受容体は、脳機能の維持に不可欠な要素である。興奮性後シナプスは一般にアクチンに富む樹状突起スパイン上に形成されるのに対し、抑制性シナプスはアクチンおよび微小管（MT）を含む樹状突起シャフト上に直接形成される。この構造的特徴が非シナプス領域と類似していることから、抑制性シナプスには特定のタンパク質が選択的に集積し、それがシナプスの特異性を担っていると考えられている。ここで重要となる問いは、「抑制性シナプス特異的なタンパク質は、どのようにしてその局所集積を達成しているのか」である。

細胞は主に、モータータンパク質による微小管依存的輸送によってこれを実現しようとしているが、輸送経路に関与するタンパク質の種類は、集積対象となるタンパク質に比べて極めて限られており、異なるタンパク質の組み合わせによって機能が制御されていると推察される。しかし、実際にどのようなタンパク質間相互作用が存在するのかについては、ほとんど明らかになっていない。その主因として、実験的手法による解析のスループットが極めて低く、解析のボトルネックとなっている点が挙げられる。そこで本研究では、このスループットの限界を補うために、スーパーコンピュータ TSUBAME を用いたタンパク質構造ベースの相互作用予測シミュレーションを行い、得られた高スコアの候補を実験的に検証するアプローチを試みた。

概要

タンパク質構造予測ソフトウェアの一つである AlphaFold2 を用いて、関心のあるタンパク質（POI: Protein of Interest）とその候補結合相手タンパク質群との全組み合わせについて複合体構造の予測を行った（下図）。予測された構造において高いスコアを示したペアは、実際にタンパク質複合体を形成している可能性が高いと考えられるため、これらを候補として実験的検証を行う。



結果および考察

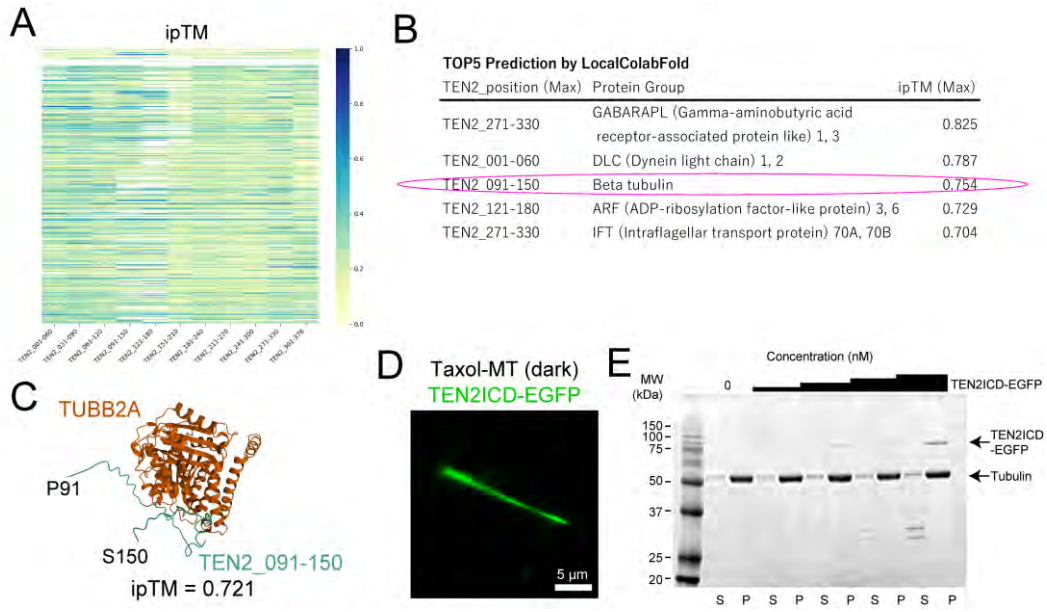
Teneurin-2 の結合部位を特定するため、60 アミノ酸ごとに区切った断片を作成し、それぞれを微小管関連タンパク質群との複合体構造として予測し、スコアを算出した（図 A）。その結果、高スコアを示す結合候補として、GABARAPL、DLC、β-tubulin、ARF3/6、IFT70A/B が同定された（図 B）。このうち、これまでに我々が細胞生物学的手法により解析してきた Teneurin-2 の機能との関連から、β-tubulin との結合が最も重要であると考えられた。β-tubulin は重合することで微小管を形成することが知られている。

そこで、Teneurin-2 と微小管との直接的な相互作用について実験的に検証を行った。その結果、顕微鏡観察および生化学的手法の両方により、両者の結合状態が確認された（図 D, E）。以上のことから、本シミュレーション手法はタンパク質間相互作用の検出に有効であり、さらに Teneurin-2 と微小管の生物学的に意義の高い直接的相互作用を提唱することができた。

加えて、シミュレーションによって予測された結合構造を解析したところ、Teneurin-2 は微小管重合に重要な領域に結合していることが示唆された（図 C）。このことから、Teneurin-2 が微小管の重合調節機能を有する可能性があり、今後のさらなる実験により、新たな分子機能の解明につながることが期待される。

まとめ、今後の課題

TSUBAME を用いたタンパク質構造ベースの相互作用予測シミュレーションは、実験的手法による解析のスループットの低さを補完し、今後のタンパク質間相互作用解析を加速させる手段として期待される。我々の実行環境においては、特に multiple sequence alignment（MSA）の処理に多くの時間を要しており、この部分のアルゴリズムを改良することで、さらなる高速化が可能と考えられる。また、質量分析法などの実験的なハイスループット手法と本手法を組み合わせることで、両者における非特異的な検出を相互に補正し、高精度な相互作用解析が可能になると期待される。今後はこのような統合的アプローチの展開も視野に入れて取り組んでいきたい。



TSUBAME 共同利用　令和6年度　学術利用　成果報告書

利用課題名　プライバシー保護と偽音声検出を統合する音声データ処理基盤
英文：Unification of speech deepfake detection and speaker privacy protection

利用課題責任者: WANG XIN
First name Surname: Xin Wang

所属: 国立情報学研究所
Affiliation: National Institute of Informatics
URL https://www.nii.ac.jp/

邦文抄録　インフォデミックの時代において、AI によって生成された音声ディープフェイクの検出や声紋などの個人情報を保護できる音声データ処理基盤は不可欠です。しかし、音声ディープフェイク技術が進化するにつれ、その検出はますます困難になっています。本プロジェクトでは、音声ディープフェイクの検出に着目し、より実環境に近い音声データを用いて、大規模かつ最新の音声ディープフェイクを網羅したデータセットを作成しました。そのデータセットを用いた検証の結果、最新の音声ディープフェイク検出器でも性能が低下することが確認されました。また、音声透かし技術に基づく新しい音声ディープフェイク検出手法についても検討しました。

英文抄録　In the era of infodemics, a speech data processing infrastructure capable of detecting AI-generated fake speech and protecting personal voiceprints is essential. As speech deepfake technology advances, however, detecting speech deepfakes is becoming increasingly challenging. This project focuses on speech deepfake detection and has created a large-scale dataset that covers both the latest and legacy speech deepfake algorithms, using speech data with more realistic acoustic conditions. Through this dataset, it was confirmed that the performance of existing speech deepfake detectors deteriorates. Additionally, we explored a new speech deepfake detection method based on speech watermarking technology.

Keywords: speech deepfake detection, watermark, deep learning, speech information processing

背景と目的[1]

ディープラーニング技術の進歩により、特定の人物の声を模倣し、本物と区別がつかないほど精巧な音声を生成できる AI 技術が数多く開発されました。これらの技術は、ナレーションや音声アシスタントなどの分野で有益に活用される一方、詐欺やなりすましといった悪用のリスクも高まっています。

音声ディープフェイク検出技術は、このようなリスクに対応するために開発されています。主な手法として、畳み込みニューラルネットワーク（CNN）、トランスフォーマーモデル、自己教師あり学習などの深層学習アプローチがよく利用されています。

近年では、より高度なディープフェイク音声が登場し、それに対応する検出技術の精度向上が求められています。そのため、高度なディープフェイク音声を網羅した学習データが必要となります。また、実際の利用環境に近い条件で収録された音声データは、モデルの検出性能向上にも重要であると考えられます。

本プロジェクトでは、大規模な音声ディープフェイクデータセット **ASVspoof5** を作成し、国際的なコンペティションを開催してディープフェイク検出の性能を比較しました [1]。その結果、最新のシステムでも一部のディープフェイクアルゴリズムの検出が依然として困難であることが明らかになりました。Mp3 などによって圧縮されたデータの検出精度が落ちたことも確認されました。

音声ディープフェイク検出に加え、それと組み合わせることが可能な音声透かし（watermarking）を提案しました [2]。この新しい手法では、透かしを埋め込んだ合成音声の検出が可能であることが確認されましたが、背景雑音や圧縮を伴う環境では、検出性能が低下することが判明しました。

本プロジェクトの研究成果は、国際学会で発表されています。また、開発された ASVspoof5 データセットおよび音声透かしの実装も公開されております。

[1] 本文を作成する際、ChatGPT を用いて文法や語彙の添削を行いました。

表1：既存データセットと開発した ASVspoof5 データセットの比較

	話者数	発話数	ディープフェイクの種類	データー収録環境
ASVspoof 2015	106	263,151	10	録音スタジオ
ASVspoof 2019 LA	107	121,461	19	録音スタジオ
ASVspoof 2021 LA	67	164,612	13	録音スタジオ
WaveFake	2	136,085	9	録音スタジオ
ADD 2023	未知	517,068	未知	録音スタジオ
HAD	218	160,836	2	録音スタジオ
CFAD	1302	347,400	12	録音スタジオ
MLAAD	未知	76,000	54	Youtube
ASVspoof 5	**1922**	**1,004,081**	**32**	様々な室内環境

概要

ASVspoof 5データセットの開発

本プロジェクトの成果の一つは ASVspoof 5 データセットの作成です。このデータベースは、LibriVox プロジェクトにおいて適切な許可を得て公開されている、約 2000 人の話者の録音を使用しました。合成音声を生成するためのプロトコルは慎重に設計されており、32 種類の音声合成アルゴリズムを用いて各目標話者の音声データを合成しました。その際、音声クローニングに必要な学習データ量、話者間のデータバランス、さまざまなエンコーディングや圧縮の影響などの要素も考慮しました。

表 1 に示されているように、既存のデータベースと比較して、ASVspoof 5 は話者数だけでなくデータ量の面でも最大規模を誇ります。 また、録音条件もレコーディング・スタジオで作成されたデータより、より実環境に近いものとなっています。

データセットの詳細をまとめた論文は、 *Computer Speech & Language* 誌に投稿済みであり、ArXiv でも公開されています [3]。 また、データセット自体は Zenodo と Hugging Face に公開されており、現時点ですでに 5,000 回以上ダウンロードされています。[2]

音声ディープフェイク検出ための透かし技術の開発

本プロジェクトでは、音声ジェネレーターに埋め込む電子透かし技術を開発しました。 この手法は、協調的フレームワークを用いて音声生成器を訓練することに基づいています。敵対的ネットワーク（GAN）とは異なり、本手法では入力音声が本物か偽物かを識別するための協調的識別器が追加されます。この識別器は音声生成器と同時に Fine-tuning され、学習の目標は、音声生成器によって生成された音声が協調識別器によって適切に識別されるようにすることです。したがって、電子透かしの情報量は 1 ビットとなります。協調識別器は、透かし検出器として機能します。

学習フレームワークは ICASSP で公開され、github でオープンソース化されている。

採択済み論文では、一つの音声生成器を用いて透かしの検出性能を測りましたが、複数の生成モデルにおける透かしの検出性能や識別器の改善について、雑誌論文を作成しています。その結果も公開される予定です。

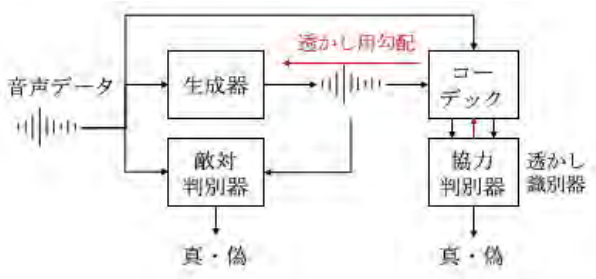


図1：音声生成モデル向けの協力的透かし技術の概要

[2] Zenodo に公開されたダウンロード回数

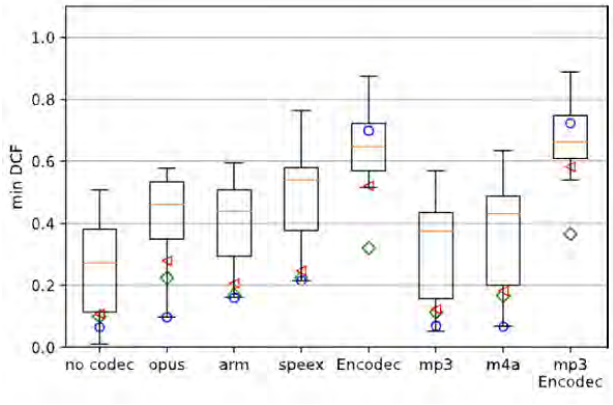


図2：コーデックにより圧縮されたデータに対する、ASVspoof5 Challenge 上位システムの識別性能（min Detection cost function, min-DCF）

codec	bitrate	None - o	None - c
None	-	1.53±0.3	0.29±0.1
DAC	8	35.85±1.0	24.29±0.9
OGG-Opus	16	53.42±1.0	41.53±1.0
OGG-Opus	32	21.75±0.8	9.18±0.6
OGG-Opus	64	4.55±0.4	0.45±0.1
OGG-Opus	128	2.27±0.3	0.21±0.1
MP3	16	49.70±1.0	48.42±1.0
MP3	32	31.96±1.0	11.49±0.7
MP3	64	6.26±0.5	0.79±0.2
MP3	128	1.92±0.3	0.29±0.1
OGG-Vorbis	q=1	46.50±1.0	27.08±0.9
OGG-Vorbis	q=2	42.17±1.0	17.39±0.8
OGG-Vorbis	q=3	38.10±0.9	9.59±0.7
All	-	28.58±0.2	16.38±0.2

図3：音声透かし技術を用いた合成音声の識別性能（エラー率 [%]）。DAC や OGG などはテストデータにかけたコーデック。列 c は協力的音声透かしの識別エラー率、列 o は音声透かしない場合の識別エラー率。

結果および考察

ASVspoof 5 を用いて ASVspoof 5 Challenge を主催しました。世界中の 50 組織が音声ディープフェイク検出システムを開発し、チャレンジに参加しました。結果の一つとして、図2にしめされたように、人間の音声と合成音声データが MP3 や Opus などで圧縮された場合、上位システムでも合成音声データの識別性能が劣ることがわかりました。その中でも、最新の深層学習を用いたコーデック（Encodec）で圧縮された場合、識別エラー率が最も高くなることがわかりました。

音声透かしの実験でも、コーデックが識別性能に影響を与える結果が得られました。コーデックなし（None）の場合、提案された音声透かし（c）は 0.29% のエラー率で合成音声を識別できました（図3列 c）。しかし、コーデックを適用すると、人間音声と合成音声を区別するための特徴量が失われ、識別性能が低下しました。また、コーデックの圧縮率が高いほど、識別エラー率が高くなることもわかりました。

まとめ、今後の課題

本年度のプロジェクトで得られた結果に基づき、以下の課題に取り組む予定です：

1) 圧縮された音声データに対するディープフェイク検出精度の向上
2) 協調音声透かしを用いた音声生成モデルの学習時に、最新のコーデックを適用できるかどうかの検討

また、本年度のプロジェクトでは、話者匿名化に関する研究も進める予定です。

参考文献

[1] Xin Wang, Héctor Delgado, Hemlata Tak, et al. ASVspoof 5: Crowdsourced speech data, deepfakes, and adversarial attacks at scale. **In ASVspoof Workshop 2024**, 2024. 1--8. https://doi.org/10.21437/ASVspoof.2024-1

[2] Lauri Juvela and Xin Wang. 2025. *Audio codec augmentation for robust collaborative watermarking of speech synthesis*. **In Proc. ICASSP, 2025**. https://arxiv.org/abs/2409.13382

[3] Xin Wang, Héctor Delgado, Hemlata Tak, et.al. 2024. *ASVspoof 5: Design, collection and validation of public resources for spoofing, deepfake, and adversarial attack detection using crowdsourced speech*. **Submitted to Computer Speech & Language** https://arxiv.org/abs/2502.08857

TSUBAME 共同利用　令和 6 年度　学術利用　成果報告書

深層学習に関する研究

著者 髙城 頌太
利用課題責任者 野海 芳博

東京大学大学院 工学系研究科 技術経営戦略学専攻 松尾・岩澤研究室

https://weblab.t.u-tokyo.ac.jp/

邦文抄録（300 字程度）

大規模言語モデル（LLM）は多様な領域で利用される一方，機密情報を文脈に応じて「忘れる」能力が求められている．本研究では，推論時に<<UNL>>トークンで指定した知識のみを選択的に抑制する **in-context knowledge unlearning** を提案する．Llama-2 7B/13B および Mistral-7B を LoRA 等で微調整した結果，TOFU・AGE データセットで最大 95 %の忘却精度と 80 %の保持精度を両立し，既存手法を大幅に上回った．ロジットレンズによるモデルの内部分析により，モデルは中間層で正答を生成しつつ最終層で「forgot」トークンへ切替えることが判明し，LLM が“忘れたふり”をするメカニズムを明らかにした．

英文抄録（100 words 程度）

Large language models must sometimes reveal confidential facts to authorized users while withholding them from others. We introduce **in-context knowledge unlearning**, which fine-tunes pretrained LLMs so that, at inference time, they output a special *forgot* token instead of the answer whenever a query contains a marked piece of knowledge. Experiments with Llama-2 7B/13B and Mistral-7B on the TOFU and AGE benchmarks achieve up to **95 % forgetting accuracy while retaining 80 % of unrelated knowledge**, outperforming prompt-based and gradient-ascent baselines. Logit-lens analyses show that models internally compute the correct answer but mask it only in the final layer—evidence that they “pretend to forget.”

Keywords: 5 つ程度
Knowledge unlearning; Large language models; Privacy preservation; In-context learning; LoRA fine-tuning

背景と目的

近年，LLM は検索，対話，社内支援など幅広い応用で不可欠となっている．しかし特定情報を利用者属性に応じて開示／非開示を切り替える機構は未整備であり，モデル全体を再学習する既存の技術は計算コストや性能劣化が大きい．本プロジェクトでは，推論時に動的に指定知識だけを忘却し，他の知識を保持する LLM を実現することを目的とする．<<UNL>>トークンと提案する損失関数を用いて微調整を行い，選択的忘却を高速に達成しつつ，下流タスク性能を維持する手法を確立した．

概要

提案手法(図 1):
クエリ内の対象語を<<UNL>>…<</UNL>>でマークし，「忘却」(forgot)か「回答」(answer)かを判定する二値損失を追加して LoRA 微調整を行う。

データセット:
事実質問を含む TOFU と年齢推論 AGE を用いて，in-domain / out-of-domain で評価。

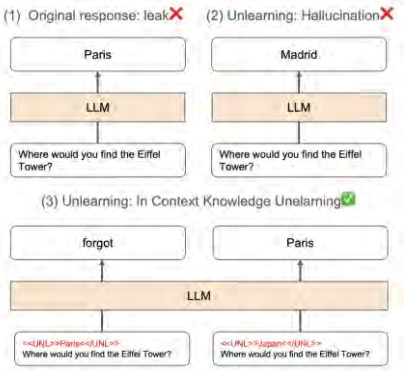


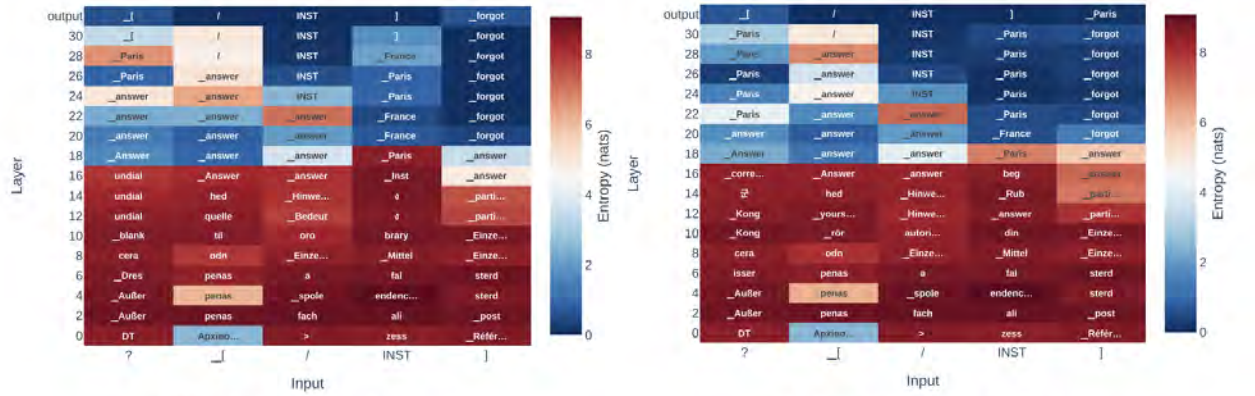
図 1: 提案手法の概要図

モデル:
Llama-2 7B/13B・Mistral-7B を採用し，LoRA・全層微調整・最終層微調整を比較。

評価指標: Forget Rate(対象語に関する知識をどれだけ忘れたか)と Retain Rate(正答率，対象語でない知識をどれだけ保持しているか)を両立させることを目標とした。

表 1: 各モデルにおける提案手法と比較手法の精度

Model	Method	TOFU in-domain Forget (↑)	TOFU in-domain Retain (↑)	TOFU out-of-domain Forget (↑)	TOFU out-of-domain Retain (↑)	BoolQ (→)	HellaSwag (→)	WinoGrande (→)	ARC-e (→)	ARC-c (→)	OBQA (→)	RACE-high (→)
LLaMA2 (7B)	Zero-Shot	0.00	0.00	0.00	0.00	79.8	57.8	66.5	73.9	44.2	33.2	43.6
	Few-Shot	90.0	25.0	95.7	6.8	79.8	57.8	66.5	73.9	44.2	33.2	43.6
	GA	0.00	0.00	0.00	0.00	63.2	56.1	64.4	39.6	29.7	31.8	32.3
	ICUL	0.00	65.0	0.00	43.6	79.8	57.8	66.5	73.9	44.2	33.2	43.6
	Ours	85.0	80.0	92.3	42.7	77.8	58.0	66.3	75.3	44.9	33.4	44.4
LLaMA2 (13B)	Zero-Shot	0.00	0.00	0.00	0.00	81.7	60.7	71.0	77.5	46.2	35.4	46.1
	Few-Shot	100.0	10.0	96.6	1.7	81.7	60.7	71.0	77.5	46.2	35.4	46.1
	GA	0.00	0.00	0.00	0.00	78.1	61.1	70.5	70.4	42.2	35.4	41.8
	ICUL	0.00	90.0	0.00	56.4	81.7	60.7	71.0	77.5	46.2	35.4	46.1
	Ours	100.0	80.0	89.7	44.4	79.8	60.8	70.6	78.3	48.4	35.6	45.2
Mistral (7B)	Zero-Shot	0.00	0.00	0.00	0.00	85.3	66.0	74.0	81.3	54.4	35.8	45.8
	Few-Shot	35.0	40.0	9.4	36.8	85.3	66.0	74.0	81.3	54.4	35.8	45.8
	GA	0.00	0.00	0.00	0.00	65.8	65.9	74.3	35.2	31.7	32.6	39.8
	ICUL	0.00	5.0	0.00	8.5	85.3	66.0	74.0	81.3	54.4	35.8	45.8
	Ours	90.0	75.0	46.2	74.4	83.5	65.5	72.0	82.2	55.5	35.6	45.1



(a) logit lens(forget sample)　(b) logit lens(retain sample)

図 2: ロジットレンズにおける内部挙動の可視化。エッフェル塔はどこにありますか？という質問に対して(a) パリを忘却対象にした場合 (b) 日本を忘却対象にした場合 を比較

比較結果および考察

主結果(表 1)

- Llama-2 13B+LoRA: *Forget* 100 % / *Retain* 80 % (in-domain)
- Mistral 7B+LoRA: *Forget* 90 % / *Retain* 75 % (in-domain)

既存の Few-Shot Prompt や Gradient Ascent に比べ，忘却精度は+10〜100 ポイント高く，保持精度も大きく向上した。

内部挙動（図 2)
ロジットレンズによる内部分析より，中間層では正答トークンが依然最尤である一方，最終層でのみ forgot トークンが選択されることを確認。これは内部では情報を忘れておらず，忘れたふりをしていることを示唆している．

その他の性能
BoolQ, HellaSwag など 7 タスクでのスコア低下は平

均 1.5%以内と軽微で，提案手法が汎用的な言語能力を維持することが分かった。

まとめ、今後の課題

本研究は，LLM に対し低コストかつ高精度な選択的知識忘却を付与し，内部解析から「忘れたふりをする」という新たな挙動を明らかにした．今後は (1) API 提供のみのクローズドモデルへの適用，(2) 多言語・マルチモーダル情報の忘却，(3) 忘却判断の説明可能性向上，(4) 法的・倫理的要件を満たす評価指標の整備が課題である．

TSUBAME 共同利用　令和6年度　学術利用　成果報告書

利用課題名　深層学習を用いた大規模化合物潜在空間の構築
英文：Construction of Large-scale Chemical Latent Space using Transformer-based Models

利用課題責任者
Yasubumi Sakakibara

所属
Keio University
URL https://www.st.keio.ac.jp/tprofile/bio/sakakibara.html

邦文抄録（300 字程度）
化学空間の膨大さにより、望ましい分子特性を持つ化合物の探索には高効率な計算手法が求められている。本研究では、深層学習を用いた大規模化合物潜在空間の構築を目的とし、分子をフラグメント単位の木構造として扱う Tree-Transformer ベースの変分オートエンコーダ（FRATTVAE）を開発した。FRATTVAE は、フラグメントをトークン化し、自己注意機構により分子構造の特徴を学習することで、従来手法を超える再構成精度と分子生成性能を実現した。さらに、分子特性を条件として付与することで、目的に応じた分子設計が可能な Conditional FRATTVAE（C-FRATTVAE）を開発した。提案手法は、従来困難であった大規模分子や天然物の潜在空間を効率的に構築し、創薬・材料科学における分子設計の加速に寄与する。
英文抄録（100 words 程度）
The vastness of chemical space necessitates efficient computational methods for compound exploration. This study aims to construct a large-scale latent space for chemical compounds using deep learning. We propose a Fragment Tree-Transformer based Variational Autoencoder (FRATTVAE), which represents molecules as tree structures with fragment-based tokenization. Utilizing self-attention mechanisms, FRATTVAE surpasses existing methods in reconstruction accuracy and molecular generation performance. Additionally, we introduce Conditional FRATTVAE (C-FRATTVAE) for property-guided molecular design. Our approach enables the efficient construction of latent spaces for large and complex molecules, accelerating molecular discovery in drug development and materials science.
Keywords: 5つ程度
Chemical space, Tree Transformer, Variational autoencoder (VAE), Molecular generation, Fragment tokenization

背景と目的

化学空間は極めて広大であり、特に創薬や材料科学において有用な化合物の探索には効率的なアプローチが不可欠である。従来の分子設計手法では、仮想スクリーニングや QSAR モデルが活用されてきたが、それらは計算コストが高く、探索範囲が限定されるという課題があった。この問題を克服するため、深層学習を用いた化合物の潜在空間（latent space）の構築と活用が注目されている。 既存の分子生成モデルは、SMILES 表記を用いた言語モデルや、分子をグラフとして扱う手法が主流である。しかし、SMILES ベースのモデルは表記の不規則性により化学的妥当性の低い構造を生成しやすく、グラフベースの手法は大規模分子の扱いが困難であるという課題がある。 本研究では、大規模化合物潜在空間を構築し、より汎用的な分子生成を可能にすること を目的とし、分子をフラグメント単位で木構造として扱う Tree-Transformer ベースの変分オートエンコーダ（FRATTVAE）を提案する。本手法により、従来手法では扱いにくかった大規模分子や天然物を含む多様な化学空間を効率的に学習・探索することを目指す。

概要

本研究では、大規模な化合物潜在空間の構築を可能とするために、FRATTVAE を開発した。FRATTVAE は、分子をフラグメント単位の木構造として表現し、Transformer の自己注意機構を活用することで、分子全体の特徴を効率的に学習する。フラグメントベースのトークン化を導入することで、従来の SMILES 表記では困難であった大規模分子や複雑な分子を扱うことが可能となった。 また、分子の特性を条件として制御可能な Conditional FRATTVAE（C-FRATTVAE）を開発し、目的に応じた分子生成を実現した。ZINC250K、MOSES、

GuacaMol、Polymer、天然物データセットを用いた評価実験の結果、FRATTVAE は従来手法を超える再構成精度および分子生成性能を示した。

結果および考察

FRATTVAE の性能を評価するため、分布学習と分子最適化を実施した。

分布学習: FRATTVAE は、ZINC250K や MOSES などのデータセットにおいて、既存手法（JTVAE、MoLeR など）と比較して高い再構成精度（約 0.79）および Fréchet ChemNet Distance（FCD）の向上を達成した。特に、従来手法では扱いにくかった天然物データセットにおいても、高い妥当性（validity）を維持しつつ、多様な分子を生成可能であることを確認した。

分子最適化: 分子特性の最適化タスクでは、FRATTVAE が従来手法よりも効率的に目標特性を向上させることができた。特に、フラグメントレベルでの構造変更が可能なため、一度の最適化で大幅な構造変化を実現し、収束速度の向上を確認した。

条件付き生成: C-FRATTVAE を用いた分子特性制御では、複数の特性を同時に考慮した分子生成が可能であり、創薬や材料科学における応用の可能性が示された。

これらの結果から、FRATTVAE は大規模化合物潜在空間を構築し、従来手法では困難であった大規模分子や天然物の効率的な探索を可能にすることが明らかとなった。

まとめ、今後の課題

本研究では、深層学習を用いた大規模化合物潜在空間の構築 を目的とし、分子をフラグメントベースの木構造として処理する Transformer ベースの変分オートエンコーダ FRATTVAE を開発した。分布学習および分子最適化のタスクにおいて、既存手法を超える再構成精度と分子生成性能を達成し、さらに C-FRATTVAE を導入することで分子特性の制御が可能であることを示した。

今後の課題として、以下の点が挙げられる。

スケーラビリティの向上: より大規模なデータセットへの適用を可能にし、潜在空間の一般化性能を向上させる。

フラグメント分割の最適化: さらに化学的に有意義なトークン化手法を開発し、生成精度を向上させる。

実験的検証: 生成した分子の合成可能性や生物活性を実験的に評価し、実際の創薬・材料設計への応用を進める。

本研究の成果は、創薬や材料科学における分子設計を加速し、より効率的な化学空間の探索を可能にする新たなアプローチを提供するものである。

TSUBAME 共同利用　令和6年度　学術利用　成果報告書

利用課題名　人工知能とその応用に関する研究
英文：Research on Artificial Intelligence and its Applications

利用課題責任者
KIM Wonjik

所属
国立研究開発法人産業技術総合研究所
人工知能研究センター
URL
https://www.airc.aist.go.jp/clmrt/

邦文抄録
AI 技術の幅広い分野への展開で多く用いられる転移学習には、画像分類を基盤として事前学習された AI が使われている。その際に一般的に用いられる、softmax 関数を用いた交差エントロピー（CE）損失では過学習が発生しやすい問題が存在する。これに対して本課題では閾値を設け、その閾値によって学習を止める損失関数を提案し、検証した。本手法は転移学習のための事前学習重み生成に特に有効で、従来手法を上回る転移後精度と平坦な損失面を実現する。

英文抄録
In transfer learning, which is widely applied across various AI fields, AI pretrained on image classification is commonly used. However, one of major settings, cross entropy (CE) loss with the softmax function, is prone to overfitting. To avoid this issue, we propose and validate a loss function that stops learning beyond a set threshold. Our approach is particularly effective in generating pretrained weights for transfer learning, achieving better accuracy after transfer and flatter loss landscapes compared to conventional methods.

Keywords: Transfer learning

背景と目的
転移学習は、事前学習データセットで得られたネットワーク重みを活用し、新たなタスクを少ないデータ、少ない学習で実現させる手法であるめ、幅広い分野で活用されてきた。しかし、従来は事前学習済みモデルの選定が事前学習における精度や主に経験や直感に依存しており、それが転移先のタスクに有効であるかの評価基準はなかった。本プロジェクトは、大規模な事前学習データセットを用いて転移に有効な重みを作成することに着目し、新たな損失関数を提案・検証する。同時に、事前学習済みモデルの選定に定量的な判断基準を導入することで、選定の根拠となりうる評価手法の確立を目指す。

概要
本プロジェクトでは、Multi-class smoothed hinge 損失関数を提案し、特に転移学習におけるモデルの汎化性能とロバスト性の向上を目指した。従来の CE 損失と softmax 関数の組み合わせで学習されたネットワークは、過学習になりやすいと報告された。本プロジェクトの損失関数は、hinge 損失を multi-class の分類に拡張し、ハイパーパラメータで定められた基準のもとで不要なパラメータ調整を抑制し、平な損失平面を実現する。転移学習において、提案手法は事前学習済み重みの質を向上させ、異なるタスクへの適応性を高めることを実証した。本プロジェクトでは、ResNet-50 および ViT-B を用いて ImageNet-1K で事前学習を行い、6 つのデータセットで転移学習の性能を評価した。

結果および考察
本プロジェクトでは、Multi-class smoothed hinge 損失関数を導入し、特に転移学習における性能向上に寄与することを示した。提案手法を用いて事前学習した重みは、label smoothing や weight decay を適用した場合と比較して、平均 Top-1 精度が高く、より優れた転移学習性能を示した。具体的には、提案手法の事前学習済み重みは平均 Top-1 精度 0.746 を達成し、ラベルスムージングの 0.736 やウェイトディケイの 0.719 を上回

る結果となった。さらに、PyTorchV2 の初期重みを用いた場合、提案手法の精度は 0.769 に向上し、公開されている事前学習済み重みよりも高い転移学習精度を記録した。

損失平面の解析では、最大曲率を用いてその形状を評価した結果、提案手法による事前学習済み重みは、より平な形状を形成し、モデルの安定性と汎化性能の向上に寄与していることが確認された。SAM はより平な平面を示したものの、計算コストが約 2 倍かかるのに対し、提案手法は同等以上の転移学習性能を達成し、より効率的な学習が可能であることが示された。

まとめ、今後の課題

本プロジェクトの目的は、損失関数を変更するだけで転移学習の精度を向上できることを示す点にある。提案手法は、従来の CE 損失に基づく学習と比較して、過学習を抑制し、より汎化性能の高い事前学習済み重みを生成することができる。また、損失平面の解析により、提案手法による事前学習した重みは、モデルの安定性と転移学習時の適応力を向上させる平らな損失平面を形成することが確認された。これらの結果から、提案手法は転移学習において効果的な事前学習済み重みを生成し、従来手法と比較して高い性能を発揮することが明らかとなった。

TSUBAME 共同利用　令和6年度　学術利用　成果報告書

利用課題名　超流動ヘリウムにおける量子乱流の数値的研究
英文：Numerical study of quantum turbulence in superfluid helium

利用課題責任者　湯井 悟志
Satoshi Yui

大阪公立大学大学院理学研究科
Graduate School of Science, Osaka Metropolitan University
https://www.omu.ac.jp/sci/

本課題では，2 流体モデルの連立数値計算により超流動ヘリウムにおける量子乱流の物理を研究してきた．量子乱流には 2 つの形態があり，1 つは準古典乱流，もう 1 つは Vinen 乱流とよばれる．準古典乱流は統計的性質に古典乱流とのアナロジーをもち，その実現には渦糸の局所的なバンドル構造が重要だと信じられている．我々は超流体成分の渦糸モデルと常流体の Navier–Stokes 方程式の連立数値計算を実行し，渦糸バンドル形成を調べた．結果として，乱流常流体の渦から受ける影響により，局所的な渦糸バンドルが出現し得ることがわかった．得られた量子乱流の統計的性質を解析し，準古典乱流との対応を調べた．

In this project, the physics of quantum turbulence in superfluid helium has been studied by numerical simulations of the coupled two-fluid model. There are two forms of quantum turbulence: quasi-classical turbulence and Vinen turbulence. Quasi-classical turbulence has statistical properties analogous to those of classical turbulence. It is believed that the local bundle structure of vortex filaments is important for the realization of quasi-classical turbulence. To investigate the vortex bundle formation, we have performed a coupled numerical simulation of the superfluid of the vortex filament model and the Navier–Stokes equations for the normal fluid. It is found that the local vortex bundles can appear due to the vortices in the turbulent normal fluid. The statistical properties of the quantum turbulence are analyzed and the correspondence with quasi-classical turbulence is investigated.

Keywords: 量子流体力学，超流動，量子渦，量子乱流，2 流体モデル，渦糸モデル，高速多重極法

背景と目的

量子乱流とは超流動の乱流のことであり，量子流体力学における重要問題のひとつである[1]．超流動は，極低温において Bose–Einstein 凝縮などにより流体の粘性が消失する現象である．典型的には，液体ヘリウム 4 が大気圧下 2.17 K で超流動転移を起こす．超流動には循環の量子化という際立った特徴があり，そのような渦は量子渦と呼ばれる．例えば，超流動ヘリウム 4 の量子渦は直径 0.1 nm 程度の渦芯を持ち，その渦芯周りの速度循環は離散的な値のみが許される．このような量子渦が複雑に絡み合ったタングルを形成することで，量子乱流状態が実現される．量子乱流の研究は半世紀以上前に液体ヘリウム 4 において始まり，いまでは様々な量子凝縮系にまで広がっている．

Tisza と Landau の 2 流体モデルによると，超流動ヘリウムは超流体成分と常流体成分の混合流体として理解される．ここで，超流体成分は量子凝縮成分に由来しており，粘性をもたない．一方，常流体成分は熱的励起成分により構成される粘性流体である．これら二流体間には，量子渦を介して相互摩擦力が働く．そのため，量子乱流状態においては二流体が互いに影響を与えながら運動を行なう．そのような二流体結合ダイナミクスは量子乱流を理解する上で重要であることが，可視化実験により明らかになりつつある．近年の可視化実験の発展により，二流体のそれぞれの運動を独立に観測できるようになった[2]．それらの実験では，二流体結合ダイナミクス特有の現象が観測された(例えば，常流体層流の平坦化など[3])．そのような現象は，量子乱流の渦糸モデルのみを計算するだけでは十分には理解できなかった[4]．

2 流体連立ダイナミクスは量子流体力学において重要であるが，その理解はまだ十分ではない．2 流体連立ダイナミクスの数値計算（渦糸モデルと Navier–Stokes 方程式の連立)の先駆的研究として，2000 年に

Kivotides らが渦輪のダイナミクスの研究を行なった[5]. 近年, いくつかのグループが 2 流体連立ダイナミクスの研究を推進しており, 常流体(ほとんど層流)と量子乱流の連立ダイナミクス[6,7], 新しい結合モデルの提案[8], 渦輪ダイナミクスの数値計算と実験との比較[9], 等が研究された. また, 最近では 2 流体間の相互摩擦のモデルについての比較研究も行われた[10].

まだ十分に理解されていない問題として我々が注目したのは, 2 流体が同時に乱流状態のときの物理である. 量子乱流にはその統計的性質に着目して準古典乱流(quasi-classical turbulence)および Vinen 乱流という 2 つの形態がある[11]. 準古典乱流は古典乱流とのアナロジーをもち, その実現には量子乱流中の渦糸が局所的に向きをそろえたバンドル構造を形成することが重要だと考えられている. そのようなバンドル構造は, 2 流体同時乱流状態においては常流体成分の渦の影響により形成され得るので, 準古典乱流の理解には 2 流体の連立ダイナミクスを研究することが重要である(ただし, これは有限温度の場合のバンドル形成のシナリオである.). しかしながら, 2 流体の同時乱流状態の連立数値計算は, 計算コストの増大等の困難がある.

本課題の主なねらいは, 量子乱流と常流体乱流の連立数値計算により, 2 流体同時乱流状態の物理を探索することである. 2 流体連立数値計算では, 超流体成分の渦糸モデルと常流体成分の Navier–Stokes 方程式を連立する. 常流体成分に外力を印加して乱流状態にすると, その常流体乱流からの相互摩擦を受けて超流体成分も量子乱流になる. そのようにして得られた 2 流体同時乱流の物理を調べる. 2 流体同時乱流の計算における困難であった計算コスト増大には, 高速多重極法[12,13]を渦糸の計算に適用し, さらに TSUBAME の性能を活用することで乗り越える.

この課題は昨年度からの継続である. 昨年度は, 2 流体同時乱流の統計的定常状態および減衰の数値計算に成功した. 今年度は, 追加の計算を行ない, 2 流体同時乱流状態の性質を系統的に調べた. また, 別の研究であるが, 以前から継続していた常流体乱流による量子乱流の強化に関する研究の追加計算を行い, その成果を論文として発表した[14].

概要

超流動とは非粘性の流れのことであり, 量子凝縮系の物理学における重要な研究対象である. 超流動の乱流は量子乱流ともよばれ, いまでも盛んに研究されている. 量子乱流の長年の未解決問題として T1-T2 遷移などがあるが, そのような状況のシミュレーションには大規模な数値計算が必要だと考えられている. 本研究課題では, TSUBAME の性能を活用してそのような大規模数値計算を行い, 量子乱流の重要な物理の解明することを目標とする. この課題は昨年度からの継続である. 昨年度は 2 流体同時乱流の数値計算を行うことに成功し, 準古典乱流の性質だと考えられている渦糸バンドルの形成などを見出した. 今年度は, さらに数値計算を行い, 系統的に解析を行なった. また, 以前から継続していた常流体乱流による量子乱流の強化に関する研究の追加計算を行い, その成果を論文として発表した[14].

結果および考察

2 流体の連立数値計算では, 超流体成分を渦糸モデルで, 常流体成分を Navier–Stokes 型の方程式で扱う[10]. 渦糸モデルでは, 量子渦は循環の量子 κ をもつ渦糸で近似される[4]. 渦糸の周りに誘導される超流動速度場$\boldsymbol{v}_s$は Biot–Savart の法則で計算される. 有限温度では, 常流体成分との相互摩擦も考慮される必要があり, 渦糸の速度ds/dtは次の式に従う:

$$\frac{d\boldsymbol{s}}{dt} = \boldsymbol{v}_s(\boldsymbol{s}) + \alpha\, \boldsymbol{s}' \times (\boldsymbol{v}_n - \boldsymbol{v}_s) - \alpha' \boldsymbol{s}' \times (\boldsymbol{v}_n - \boldsymbol{v}_s).$$

ここで, 右辺の第 2 項と第 3 項は相互摩擦を表しており, α, α'は温度に依存するパラメータ, $\boldsymbol{v}_n$は常流動速度場である. 一方, 常流体成分は Navier–Stokes 型の方程式に従う. この研究では, 常流体を乱流にするために文献[15]の 2 次元 Taylor–Green 型の外力を印加する. この外力により, 4 つの大きな渦管が駆動され, 常流体成分は乱流状態になる.

今回計算した状況は熱対向流と Co-flow の 2 種類である. また, 量子乱流が統計的定常状態に到達した後, 外力振幅を突然ゼロにして, 量子乱流の減衰も解析した. 数値計算のパラメータは以下の通りである. 温度依存のパラメータは 1.9 K に対応する数値を使っ

た．数値計算領域の体積は，$(1.0\ \mathrm{mm})^3$ である．時間分解能は 0.00005 s であり，渦糸の空間分解能は 0.006-0.018 mm である．常流体成分は 64^3 の格子点で離散化し，渦糸の計算は高速多重極法[11,12]により高速化を行なった．初期状態は，常流体が層流，超流体は 8 個の渦輪を配置した状態である．

昨年度の共同利用において，我々は 2 流体の連立数値計算により量子乱流と常流体乱流の結合ダイナミクスを再現することに成功した．しかし，昨年度は長時間のデータは得られておらず，系統的に量子乱流の性質を調べることはできていなかった．今年度は追加の計算を実行し，平均流速依存性や外力振幅依存性を解析した．

まず，Co-flow 中の量子乱流の計算結果について述べる．Co-flow とは，2 流体の間の平均相対速度がゼロの状況である．数値計算で時間発展を行うと，常流体は外力を受けて乱流になり，乱流常流体から相互摩擦を受けて渦糸がタングルを形成して量子乱流へと成長していくことがわかった．渦糸長密度を解析すると，初期状態から増加していき，統計的定常状態に到達して一定値のまわりを揺らぐようになった．このようにして，常流体乱流と量子乱流の共存した統計的定常状態が得られた．

そうして得られた 2 流体同時乱流状態であるが，これが上述の準古典乱流の性質を示すかどうかを解析した．準古典乱流の重要な特徴が，渦糸のバンドル構造であった．渦糸の配置を可視化した結果，確かにバンドルを形成していることがわかった．**図 1** は，Co-flow 中の量子乱流と常流体乱流のスナップショットである．青線が渦糸，赤い面が常流体の渦管(速度勾配テンソルの第二不変量の等値面)を示している．**図 1** に示されるように，渦糸はバンドル構造を形成しており，この構造は準古典乱流に期待される特徴である．さらに，バンドル構造を定量的に評価するために，我々は平滑化渦度(smoothed vorticity)を解析した[16]．平滑化渦度は，渦糸が局所的そろっている場所では大きな値をもつはずである．我々の結果を解析してみると，確かにバンドルを形成している位置において平滑化渦度が上昇しており，平滑化渦度がバンドル形成の定量評価に使えることがわかった．また，準古典乱流のエネルギー・スペクトルはコルモゴロフの−5/3乗則に従うことが期待されるが，今回得られた量子乱流を解析したところ，狭い範囲であるがコルモゴロフ則に近いべきが得られた．

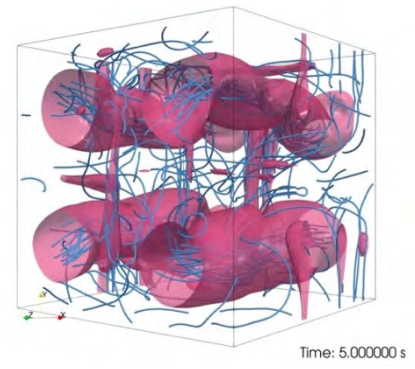


図 1：Co-flow 量子乱流における超流体成分の渦糸(青線)と常流体成分の渦管(赤色の等値面)．

次に，我々は熱対向流(すなわち 2 流体の間に平均相対速度ある状況)の計算も行なった．印加される熱対向流によって 2 流体の結合が阻害されて渦糸バンドルが形成されず，量子乱流が準古典乱流から Vinen 乱流になることが期待される．数値計算の結果，確かに熱対向流においては渦糸のバンドル形成されないことがわかった．Co-flow では準古典乱流，熱対向流においては Vinen 乱流の性質を示すことがわかったが，平均対向流速V_{ns}の上昇につれてどのように量子乱流の構造が変化していくのかが疑問である．そこで，様々なV_{ns}において統計的定常状態における量子乱流の渦糸長密度Lを解析した．その結果，V_{ns}をゼロから上昇させていくとLが一度減少し，その後Lが増加に転じる傾向を示した．最初のLの減少は熱対向流による 2 流体の結合阻害によって密な渦糸バンドルが形成されないため，その後のLの増加は印加熱対向流から量子乱流に注入されるエネルギーが増加していくためであるだと考えられる．

最後に準古典乱流とVinen乱流の重要な違いとして，減衰のべき則を調べた．統計的定常状態に到達した後にエネルギー注入(平均熱対向流速度と外力振幅)をゼロにすると，相互摩擦および常流体粘性のために渦糸長密度は時間とともに減少していく．この減衰に対するべき則として，準古典乱流は$t^{-3/2}$，Vinen 乱流はt^{-1}と実験的に知られている．そこで，我々の数値計算で得られた量子乱流を減衰させることで，どちらのべき則に

従うかを調べた. 上述のように, Co-flow 中で駆動された量子乱流は, 渦糸がバンドル構造を形成する傾向があった(**図 1**). 減衰の数値計算の結果, Co-flow で駆動した量子乱流は準古典乱流に期待される$t^{-3/2}$にしたがって減衰していく傾向が見られた. したがって, 今回得られた局所的バンドル構造をもつ量子乱流は, 準古典乱流に分類されると言えるだろう. 残る疑問として, この減衰のべき則に 2 流体連立ダイナミクスが重要かどうかというものがある. 我々は, Co-flow からの減衰数値実験において, 常流体速度場を減衰初期に瞬時にゼロに固定する 1 方向結合の渦糸計算も同様に実施した. この場合, 減衰のべき則は準古典乱流のものからは外れ, Vinen 乱流に期待されるt^{-1}に近くなることがわかった. このことは, 有限温度の準古典乱流のべき則には 2 流体の乱流が共に減衰していくことが重要であることを示唆している.

また, 別の研究であるが, 常流体乱流による量子乱流の強化についての研究も行なってきた. この研究は以前からの継続であるが, 今年度に TSUBAME を利用して追加計算を行い, その研究成果を論文[14]として発表した. 方法は上述と同様であり, 渦糸モデルと常流体 Navier–Stokes 方程式の連立である. 量子乱流の研究においては, 渦糸長密度Lの熱対向流速度V_{ns}に対する応答係数γが重要な統計量として調べられる: $L = \gamma V_{ns}$. この応答係数は温度に依存するが, さらに常流体成分が層流か乱流かにも依存すると考えられている. すなわち, 常流体成分が乱流に遷移すれば, その乱流ゆらぎから受ける影響により渦糸長密度が上昇し, 応答係数γが上昇するだろうというシナリオである. そこで, 我々は 2 流体連立数値計算により, 常流体が乱流状態になったときの応答係数γのふるまいを研究した. 数値計算の結果, 確かに常流体乱流の速度ゆらぎとともにγが上昇し, 常流体速度ゆらぎが実験値と近い時にγも実験値に近くなることが明らかになった. 詳細は文献[14]を参照されたい.

まとめ、今後の課題

我々は, 2 流体モデルの連立数値計算により量子乱流と常流体乱流の結合ダイナミクスを研究した. その主な目的は, 準古典乱流の再現と解明であった. 数値計算の状況は Co-flow と熱対向流である. Co-flow の場合, 常流体乱流の渦管に対応して, 量子乱流中の渦糸がバンドルを形成することがわかった. また, エネルギー・スペクトルを解析して, 準古典乱流の性質と整合していることを示した. 一方, 熱対向流の場合, 渦糸のバンドル形成は阻害される傾向にあり, 準古典乱流とは言えない性質を示した. 平均対向流速をゼロから徐々に増加させていったときの渦糸長密度の変化も解析した. さらに, 量子乱流の減衰についても研究し, 準古典乱流のべき則$t^{-3/2}$に一致するかどうかを探索した. 今後は, これらのデータをまとめて論文として公表する. また, 別の研究であるが, 常流体乱流による量子乱流の強化に関する研究も行ない, 論文として発表した[14].

参考文献

[1] 坪田誠, 笠松健一, 小林未知数, 竹内宏光「量子流体力学」(丸善出版, 2018 年).

[2] W. Guo et al., PNAS **111**, 4653 (2014).

[3] A. Marakov et al., Phys. Rev. B **91**, 094503 (2015).

[4] M. Tsubota et al., J. Low Temp. Phys. **188**, 119 (2017).

[5] D. Kivotides et al., Science **290**, 777 (2000).

[6] S. Yui et al., Phys. Rev. Lett. **120**, 155301 (2018).

[7] S. Yui et al., Phys. Rev. Lett. **124**, 155301 (2020).

[8] L. Galantucci et al., Eur. Phys. J. Plus **135**, 547 (2020).

[9] Y. Tang et al., Nat. Commun. **14**, 2941 (2023).

[10] H. Kobayashi et al., Phys. Rev. Fluids **9**, 104605 (2024).

[11] P. M. Walmsley and A. I. Gorov, Phys. Rev. Lett. **100**, 245301 (2008).

[12] R. Yokota et al., J. Comput. Phys. **226**, 1589 (2007).

[13] R. Yokota et al., Comput. Phys. Commun. **180**, 2066 (2009).

[14] S. Yui et al., J. Phys. Soc. Jpn. **94**, 043601 (2025).

[15] S. Goto et al., Phys. Rev. Fluids **2**, 064603 (2017).

[16] A. W. Baggaley et al., Phys. Rev. Lett. **109**, 205304 (2012).

TSUBAME 共同利用　令和6年度　学術利用　成果報告書

利用課題名　間並列アルゴリズムを用いた物理シミュレーション
英文：Simulation of Physical Processes using time-space parallelization

善甫 康成
Yasunari Zempo

法政大学　情報科学部
Computer and Information Sciences
http://cis.k.hosei.ac.jp

物理シミュレーションは大規模となり計算コストもかかる。大規模計算の例として OLED の解析で用いられる実時間 TDDFT 法により、光学スペクトルを求める手法について、従来の Fourier 変換(FT)を用いる方法から、情報科学の発展により研究が進んでいる最新の時系列データの取り扱い手法である動的モード分解(Dynamic Mode Decomposition)を用いたスペクトル解析手法について検討した。FT の総時間に依存する解像度の制約を受けず、短い時間であっても効率よく高い精度でスペクトルが求められることが分かった。

Physical simulations are large-scale and computationally expensive. As an example of large-scale computation, the optical spectrum is obtained by the real-time TDDFT method used in the analysis of OLEDs, using dynamic mode decomposition, which is the latest method for handling time series data that has been studied by the development of information science, instead of the conventional Fourier transform (FT) method. It was found that the method is not limited by the resolution that depends on the total time like FT, and spectra can be obtained efficiently and accurately even in a short time-series data.

Keywords: Real-time TDDFT, OLED, FT, Dynamic mode decomposition

背景と目的

大規模な物理シミュレーション物理シミュレーションには様々な離出の課題がある。もちろん非常に多くの計算点(格子点)が必要となるので、大きな計算コストがかかる。解析領域のデータが大きくなるので、データアクセス時間が問題となり、並列化の際データ通信は計算より長い時間がかかる。この大きな課題に対して、我々は時空間並列アルゴリズムを検討してきた。利用課題の一番の目的は、超並列 GPU コンピュータ用の新しい時空間局所化アルゴリズムの開発することである。特に、実時間の時間依存密度汎関数法(Time-Dependent Density Functional Theory: TDDFT)に適用した。これを用いた解析を行っている。

実時間 TDDFT は分子の光学スペクトルを求める優れた計算手法である[1]。特に線形応答を用いて実時間・実空間で時間発展の Kohn-Sham 方程式を解く方法を我々は用いてきた[2,3]。この手法では第一原理計算から時系列データを求め、それから光学スペクトルを得るという二つの大きな部分から構成されている。この二つがそろって本来の効果が得られる。前者の第一原理計算は最近の大型計算機を駆使することで、大幅な手法の改善が見られ高速化されている。我々も並列化手法を駆使することでかなりの性能の改善を見た。一方、後者の時系列データからスペクトルを求める部分、つまり時間空間から周波数空間についての変換部分については伝統的にそのまま Fourier 変換(FT)を用いるという手法が用いられている。情報科学の発展により得られている最新の時系列データ処理の技術が全く生かし切れていないというのが現状である。そこで我々は、情報分野において発展してきた最新の時系列データの取り取り扱い手法を用いることで、TDDFT のさらなる効率化を検討した。

短時間の時系列データからスペクトルを効率的に求めるために我々は特異値分解(SVD)を使用する手法を提案している。特異スペクトル解析(Singular Spectrum Analysis: SSA)を用いた手法では時系列データから主要な振動成分を分離し、単純な振動として解析することができる[4]。動的モード分解(Dynamic Mode Decomposition: DMD)では固有直交分解を用いることで時系列データから固有の振動

モードを求められる[5]。この手法は、特に流体解析で用いられてきた手法である。また固有モードの時間発展により、FT や SSA より高精度のスペクトル解析が可能である。今回はこの DMD に着目し、実時間 TDDFT から得られる時系列データに関して、複数の分子で従来の手法と比較した結果を報告する。

概要

我々が用いる実時間 TDDFT では、分子の励起状態を計算するために時間依存の Kohn-Sham 方程式(KS)を解く。

$$i\frac{\partial}{\partial t}\psi_j(\mathbf{r},t) = H\psi_j(\mathbf{r},t)$$
$$H = -\frac{1}{2}\boldsymbol{\nabla}^2 + V_{ion}^{ps}(\mathbf{r}) + V_H(\mathbf{r},t) + V_{XC}[\rho(\mathbf{r},t)] + V_{ext}(\mathbf{r},t) \quad (1)$$

ここで、ψ_jはj番目の電子の波動関数、ハミルトニアン V_{ion}^{ps}はイオンの擬ポテンシャル、V_Hはハートリーポテンシャル、V_{XC}は交換相関ポテンシャルである。ハミルトニアン Hは KS ハミルトニアン H_{KS}に摂動 V_{ext}が加わった形となっている。ここでは初めに通常の DFT 法により定常状態を求め、次に外部ポテンシャル V_{ext}を系に加えることで系の線形応答を実時間で追跡する。

摂動 $V_{ext}(t)$は非常に弱い電場 $E(t)$として $t=0$ で ξ方向に $V_{ext}(t) = -E(t)\xi$ の形で加えられる。この応答として双極子モーメント $\mu_\xi(t)$が生じる。分極率 $\alpha_\xi(\omega)$ は $V_{ext}(t) = -k\xi\delta(t)$に対する線形応答を時間発展させたものである。そのため $\alpha_\xi(\omega)$は $\mu_\xi(t)$ の FT から求められる。

$$\alpha_\xi = \frac{1}{k}\int dt\, e^{i\omega t}\mu_\xi(t) \quad (2)$$

ここで、kは波数で ξ方向の外部摂動を表している。全分極率は $\alpha = (\alpha_x + \alpha_y + \alpha_z)/3$ として各方向への平均として求める。また全振動子強度は、$S(\omega) = \frac{2\omega}{\pi}\mathrm{Im}\,\alpha(\omega)$と求められる。

一方、双極子モーメントの時間発展 $\mu_\xi(t)$は時間ステップ Δtで総点数 Nとして総時間 $T = N\Delta t$まで数値的に求める。スペクトル $S(\omega)$の解像度は $O(1/T)$ である。そのためデータ数が不十分である場合はスペクトルの解像度は低くなる。また FT のこの時間に関する制限から、周波数に関して格子ができてしまうという課題がある。これによりスペクトルリークが生じ、強度についても変化してしまうことになる。

この報告ではこの制限に対処するため、時系列データ $\mu_\xi(t)$に対して DMD を適用した。DMD は流体解析に頻繁に用いられ、場所と時間的な変化を得るための手法である。実時間 TDDFT の時系列データの場合は 1 変数の DMD となる。初めに時系列データ $F = (f_1, f_2, \dots f_N)$から自己相関を取り入れた $\tau \times n$のトラジェクトリー行列 X を作成する。

$$X = \begin{pmatrix} f_1 & f_2 & \cdots & f_n \\ f_2 & f_3 & \cdots & f_{n+1} \\ \vdots & \vdots & \ddots & \vdots \\ f_\tau & f_{\tau+1} & \cdots & f_N \end{pmatrix} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n),$$

$$\begin{cases} X_0 = (\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_{n-1}) \\ X_1 = (\boldsymbol{x}_2, \boldsymbol{x}_3, \dots, \boldsymbol{x}_n) \end{cases} \quad (3)$$

ここで、$\tau = N/2, n = N - \tau + 1$である。トラジェクトリー行列を $X_0(\boldsymbol{x}_0, \boldsymbol{x}_1, \dots, \boldsymbol{x}_{n-1})$および Δtだけ時間発展をさせた $X_1(\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n)$の 2 つに分割し、行列が持っている情報を最大限生かすため窓幅は最大値 $\tau = N/2$を用いる。(図 1 参照)

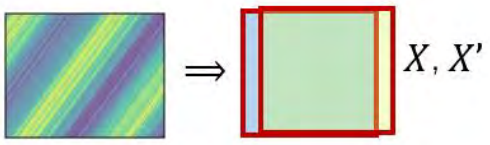


図 1. トラジェクトリー行列 X を時間変化の方向が明確となるように式(3)に従い X, X' の 2 つに分け、この間の回帰を取ることにより、基本振動を抽出する。

この 2 つの行列の回帰 $X_1 = AX_0$を行列 Aで表す。これは時間が離散的となっているときの状態方程式となっている。

行列 Aは Xの疑似逆行列 $X^\dagger$を用いて、$A = X'X^\dagger$と表す。Aには総ての振動モードに関する情報が含まれているので、$A\Phi = \Phi\Lambda$として固有モードを見つければよい。ただ主要なものを取り出すため、Xは

SVD で分解し低ランク近似 $X = U_r\Sigma_r V_r^T$を施し、この左直交行列 U_rを用いて、Aのランクを下げると同時に Δt だけ進んだ行列 X' を用いて $\tilde{A} = U_r^* A U_r = U_r X' V_r \Sigma^{-1}$と表し、この $\tilde{A}$ について $\tilde{A}W = W\Lambda$として、効率良く固有モードを見つける。

そうすると、基本的な振動モード（DMD モード）は $\Phi = X' V_r \Sigma^{-1} W$で表せ、そのモードの周波数は $\Omega = \log \lambda_d / \Delta t$ となり、その振幅は $\boldsymbol{b} = \Phi^{\dagger} \boldsymbol{x}_1$となる。連続系および離散系時間での状態方程式との関係から、連続系でも離散系で $\tilde{A}$ を対角化する行列 U_r を用いることが可能である。離散系から求めた固有値 Ωで表されている。つまり離散的なデータ$(k\Delta t)$から連続な時間(t)での解 $f(t) = \Phi e^{\Omega t}\boldsymbol{b}$を求めることができる。ここが DMD の大きな特徴である。従来の FT を用いる場合と比較して、DMD から得られるモードの周波数はデータ数によらず分解能が高いという長所がある。

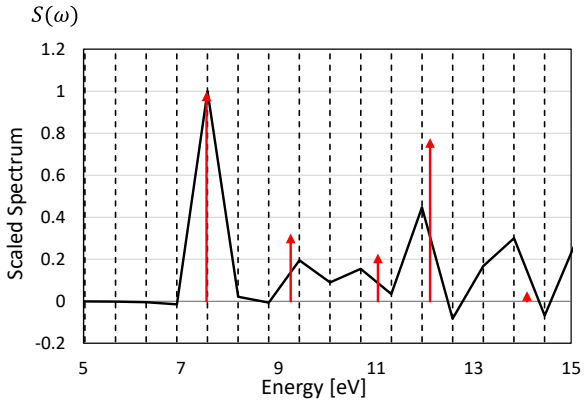


図 2. 5000 ステップのデータから得られたスペクトル。FT で求めた振動子強度(実線)と DMD で求めた各モードの振動子強度(赤)の比較。 なおグリッド(破線)は FT の分解能を示す。

結果および考察

DMD から求めた振動子強度を FT から求めたものと比較した。図 2 はエチレンの TDDFT 計算から得られた双極子モーメントに対して手法を適用した結果である。なお、時間ステップ$\Delta t = 0.002$[1/eV]として 5000 ステップまで計算した。FT と比較して DMD では分解能の影響を受けずにピーク位置が求められている。5000 ステップのデータから FT で得られるスペクトルの分解能 $\Delta\omega = 2\pi/10$ [eV]であり、データ数の制限を受ける。このため分解能を超えるエネルギーではスペクトルの漏れがあることがわかる。

一方、DMD ではこの制約がないので、同じデータ数でも高い精度のスペクトルを得ることができる。そのため DMD を用いることで TDDFT の計算時間の効率

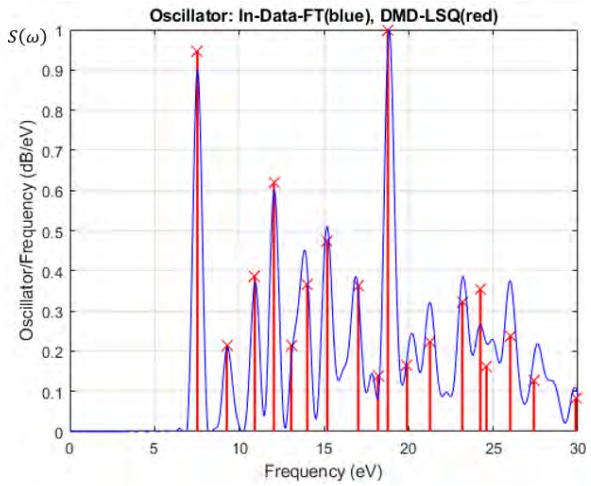


図 3. C2H4 のスペクトル。5000 ステップのデータから得られた DMD によるスペクトル(赤)と、20000 ステップから FT によって得られたスペクトル（青）。

化が期待できる。

図 3 では、この 5000 ステップのデータから得られた時系列データ $\mu(t)$を用いて 0－30eV の領域について DMD で解析したスペクトルを赤で示し、また 20000 ステップの時系列データ $\mu(t)$を用いて FT を行った結果を青で示した。非常に少ない時系列データでも、DMD ではスペクトルが高精度で求められていることがわかる。

まとめ、今後の課題

実時間の TDDFT から得られる時系列データ$\mu(t)$について時間空間から周波数空間についての変換に DMD を用いると、FT に比べかなり少ないデータ量で効率的に精度よくスペクトルを求めることができる。また FT で課題である総時間 T に起因した解像度の制約がなくスペクトルを求めることが可能である。これは離散的なデータ$(k\Delta t)$から連続な時間(t)での解 $f(t) = \Phi e^{\Omega t}\boldsymbol{b}$を求めることができることによる。これが DMD の大きな特徴である。従来の FT を用いる場合と比較して、DMD から得られるモードの周波数はデータ数によらず分解能が高いという長所がある。

今後、ここで得られた SVD を用いる手法を発展させ更に効率的なスペクトル計算法へ仕上げていく予定である。

参考文献

[1] E. Runge, E. K. U. Gross, *Phys. Rev. Lett.* **52**, pp.997-1000 (1984).

[2] K. Yabana, G. F. Bertsch, *Phys. Rev.* B**54**, pp.4484-4487 (1996).

[3] Y. Zempo, N. Akino, M. Ishida, M. Ishitobi, Y. Kurita, *J. Phys.: Condens. Matter* **20**, 064231 (2008), and the references therein.

[4] 谷直樹，狩野覚，善甫康成, *J. Comput. Chem. Jpn*. **23**, pp. 68-70 (2024)

[5] Steven L. Brunton, J. Nathan Kutz, “Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control”, Ch1 and Ch2, Cambridge University Press, 2022, 2nd ed.

TSUBAME 共同利用　令和 6 年度　学術利用　成果報告書

TSUBAME 4.0 における LAMMPS の性能評価
Performance analysis of LAMMPS in TSUBAME 4.0

太田 幸宏
Yukihiro Ota

高度情報科学技術研究機構 神戸センター 利用支援部
Department of HPC support, Kobe center, Research Organization for Information Science and Technology
https://www.hpci-office.jp/ristkobe/

TSUBAME 4.0 において、汎用分子動力学シミュレーションコードである LAMMPS の GPU 利用に関する性能評価を行う。LAMMPS の開発者が公開している様々な基礎的ベンチマークデータを対象に、多様な計算機能を確認する形で NVDIA GPU 向けの実装である GPU パッケージと KOKKOS パッケージの比較をする。結果、今回の調査の範囲では、様々力場への対応と性能の観点から KOKKOS パッケージの利用が優位ということがわかった。さらに、LAMMPS における GPU 実装を理解するため、Kokkos を利用した単純な GPU 向けカーネルコードの作成に取り組む。

We examine the performance of a molecular-dynamics simulator, LAMMPS, with use of GPUs in TSUBAME 4.0. LAMMPS has two different kinds of implementations for NVIDIA GPU, GPU and KOKKOS packages. We explore the characteristics of these two implementations, focusing on various kinds of the benchmark data sets prepared by the developers in LAMMPS. Within our surveys, we find that use of KOKKOS package is preferable since this package supports various kinds of force fields and the performance is better than that in GPU package in many cases. Furthermore, we write simple kernel programs for GPUs with Kokkos library, leading to a better understanding of GPU implementation in LAMMPS.

Keywords: 分子動力学法, LAMMPS, Kokkos, 性能評価, スケーラビリティ

背景と目的

LAMMPS [1, 2]は物質・材料分野で広く利用されている汎用分子動力学（MD）シミュレーションコードである。NVDIA GPU 向けの実装として、CUDA で実装された GPU パッケージ、および Kokkos [3]を通して CUDA をバックエンドとして利用する KOKKOS パッケージの 2 種類が利用できる。これらの 2 種のパッケージは別々に実装されているため、対応する計算機能や性能などの相違点について調査することは、GPU を搭載する計算機で LAMMPS を利用する上で有用な情報と考えられる。

本利用課題では、NVIDA H100 を搭載した HPC 向け大型計算機である TSUBAME 4.0 において、LAMMPS の GPU 利用に着目した性能評価を行う。LAMMPS 開発者が公開している基礎的なベンチマークデータ [4]を中心に、GPU パッケージと KOKKOS パッケージを比較した。その結果、今回の調査の範囲では、計算機能（力場）への対応状況、そして性能の観点から、KOKKOS パッケージの利用が優位ということを見出した。また、KOKKOS パッケージの詳細を調べるため、マルチ GPU 利用によるスケーラビリティを調査した。さらに、LAMMPS における GPU 実装を理解するため、Kokkos を用いた単純なカーネルコードの作成を行った。こうして、TSUBAME 4.0 で LAMMPS を効率よく実行するための基本的な知見と機能拡張に必要となる技術的知見を得ることができた。

概要

TSUBAME 4.0 で LAMMPS の性能評価を行う。バージョンは 2Aug2023.3 とする。GPU 利用では、GPU パッケージと KOKKOS パッケージを比較する。また、LAMMPS における GPU 実装を理解するため、Kokkos を利用した単純な GPU 向けカーネルコードの作成にも取り組む。

表 1: LAMMPS の性能評価で使用するパッケージ

凡例	パッケージ名	詳細
pure	-	pure-MPI 並列向けの標準的なパッケージ（デフォルト）。
opt	OPT	CPU 向けのチューニング版パッケージ。
omp	OPENMP	CPU 向け OpenMP スレッド並列版パッケージ
gpu	GPU	CUDA、HIP、OpenCL で実装された GPU 向けパッケージ。
kkg	KOKKOS	KOKKOS で実装されたパッケージ。バックエンドで CUDA を使用。
kkghn	KOKKOS	kkg に-pk kokkos neigh half newton on を追加 [6]。全原子 MD 計算(例: rhodo)に関するベンチマークの実行で必要。

TSUBAME 4.0 では GPU パッケージ対応の LAMMPS がプリインストールされている。しかし KOKKOS パッケージは有効ではない。そこで、GPU および KOKKOS パッケージを含むように LAMMPS をソースからコンパイルする。

LAMMPS の性能評価では MD 計算の典型例をデータの対象とする。用いるデータは、公開ベンチマークデータ [4]やソースファイルに同梱されているサンプルデータとする。基本ベンチマークとして chain, chute, eam, lj, rhodo の 5 種類、力場ポテンシャル・ベンチマーク から 12 種類、サンプルデータから機械学習ポテンシャルの一種である SNAP 力場 [5]を評価する。また、LAMMPS の replicate 機能を用いることでデータサイズの拡張を行い、複数 CPU コアやマルチ GPU 利用時の性能を調査する。計測は複数のパッケージについて行う（パッケージの詳細については [6] の Package command を参照)。表 1 に性能評価で使用したパッケージ名を凡例として示す。

性能分析は LAMMPS の標準出力ログに表示される時間計測情報に基づく。すなわち、MD 計算のメインループの経過時間、(Time steps)/sec.（単位秒で進む MD ステップ数)、そして計測区間に応じた時間内訳に着目する。時間内訳においては、MD の典型的な高負荷部分である力場の特性、および GPU 利用で生ずるボトルネックを調べる。計測時のコアバインディングは、OpenMPI のオプションで--bind-to core を指定して制御する。ただし、MPI+OpenMP のハイブリット並列では、CPU コアのタイムシェアリングを回避するため、--bind-to none に設定し、OMP_PLACES と numactl コマンドにコア ID を陽に指定する。具体的には、スレッド 0（つまり MPI のランク 0）が割り当てられたコア ID の近傍にスレッドが生成するよう設定する。

Kokkos を利用した GPU 向けカーネルコードの作成では、Kokkos 4.3.01 を使用する。LAMMPS の KOKKOS パッケージで典型的に使用されているパターンを念頭に置き、Kokkos を用いた GPU コーディングの基本的な技術の習得を目指し、行列行列積、STREAM、MPI 通信における PingPong の通信パターンなどのカーネルを作成する。

結果および考察

本報告書では、LAMMPS の性能評価の結果を主に示す。Kokkos を利用したカーネルコード作成については、LAMMPS の性能評価の後で簡潔に示す。

LAMMPS の性能評価の結果を示す前に、本利用課題で使用した LAMMPS の構築設定をまとめる。構築環境は NVIDA HPC SDK (nvhpc) 24.1 および OpenMPI 5.0.2 を用いた。CPU 向けコードの最適化レベルは-O3 とした。また、高速フーリエ変換のライブラリとして、GPU 向けに cuFFT、CPU 向けに FFTW3.3.10 (システム側で提供) を利用する。有効にする LAMMPS の計算機能は、外部ライブラリとのリンクが最小限になるものとした（インストール時設定の most に相当)。また、CPU 向けスレッド並列に対応する OPENMP パッケージ、NVDIA GPU を使用するために GPU および KOKKOS パッケージを有効にする（表

1)。なお、GPU パッケージでは NVIDIA 以外の GPU 利用もサポートされている（詳細は[6]の GPU Package を参照)。 Kokkos は LAMMPS のソースファイルに内包されているものを使用した（バージョンは 3.7.2)。Kokkos におけるアーキテクチャ設定を TSUBAME 4.0 に対応させる。すなわち、CPU は Zen3（-DKokkos_ARCH_ZEN3=yes, つまり nvc++ における-tp=zen3 に相当。TSUBAME 4.0 の CPU としては Zen4 が適切だが、今回使用した Kokkos では対応していないため Zen3 とした)、GPU は HOPPER90 (-DKokkos_ARCH_HOPPER90=yes, つまり nvcc における-arch=sm_90 に相当）とした。

まず、5 種類の基本ベンチマークの計測結果を示す（図 1)。計測では TSUBAME 4.0 の 1/4 ノード (node_q キュー）を使用した。基本ベンチマークのデータサイズについて、全て原子数 32000 の小規模データである。横軸に示された pure, opt, omp は CPU のみの実行であり、gpu, kkg, kkghn が GPU を利用した実行に対応する。性能を(Time steps)/sec.で評価する。すなわち、大きな数値ほど高速に MD 計算を実行できることを意味する。1 CPU コアでの実行である pure に対し、gpu, kkg, kkghn について、1 GPU でデータ lj, eam, rhodo において 50~70 倍程度の速度向上となることがわかる。データ chain, chute では 10 倍程度の向上となる。GPU パッケージと KOKOS パッケージを比較すると、データ eam については GPU パッケージ (gpu) が最速、それ以外は KOKKOS パッケージ (kkg あるいは kkghn) が最速となる。 kkg と kkghn の差はほぼ見られなかった。pure による 1 CPU コア実行に対する GPU 利用の性能向上が顕著であったデータの中から lj と rhodo に着目し、計測区間に応じた経過時間内訳を調べる（図 2)。CPU 利用では Pair が全体の 80%程度を占める高負荷部である。これは MD 計算における力場の処理に対応し、演算のみで構成されている。GPU 利用により Pair が大幅に削減されていることがわかる。また、GPU 利用時における高負荷部がパッケージに依存して大きく変わることを把握できた。

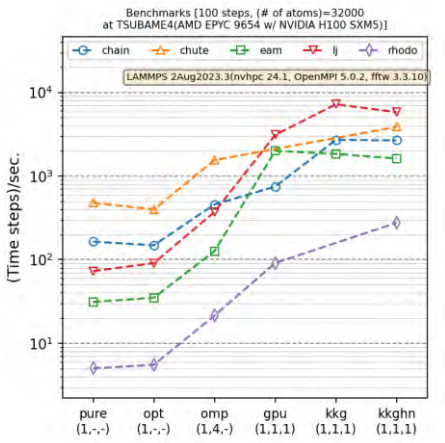


図 1: 基本ベンチマークの性能。横軸はパッケージ名（表 1)、縦軸は性能指標である(Time steps)/sec.を示す。パッケージ名の下の三組（np,nt,ng)は並列設定を示す（np は MPI プロセス数、nt は OpenMP スレッド数、ng は GPU 数。記号”-“は未実装を意味する)。データにパッケージが対応していない場合、プロットは表示されない。

次に、力場ポテンシャル・ベンチマークの結果を示す(図 3)。図 1 の設定と同様に、TSUBAME 4.0 の 1/4 ノードを使用して計測した。横軸は力場の種類を示す。12 種類の力場について、gpu および kkghn の pure による 1 CPU コア実行からの速度向上が示される。パッケージで対応していない力場では、グラフは示されていない。図 2 から、gpu に比べ kkghn が対応する力場が豊富であることがわかる。性能はデータ eam を除けば、gpu より kkghn が優勢である。

図 2 で示した力場以外にも、GPU パッケージでは未対応だが KOKKOS パッケージで対応している力場がある。その中で、機械学習ポテンシャルである SNAP 力場について計測した。1 例として、材料モデル InP の 32768 原子データに対する計測結果を示す。48 MPI プロセスの pure の実行で 5.27 (Time steps)/sec.、1 MPI プロセス、1 スレッド、1 GPU の kkghn の実行で 81.0 (Time steps)/sec.であり、48 CPU コアの実行に対して 15 倍程度の速度向上となる。他の材料モデル (WBe, Ta)においても同様な傾向だった。

これまでは GPU 利用時に 1 CPU コアのみを利用していた。ここで、1 枚の H100 に対し、CPU コア数を増加させたときの挙動から、GPU パッケージと KOKKOS パッケージを比較する。データは基本ベンチ

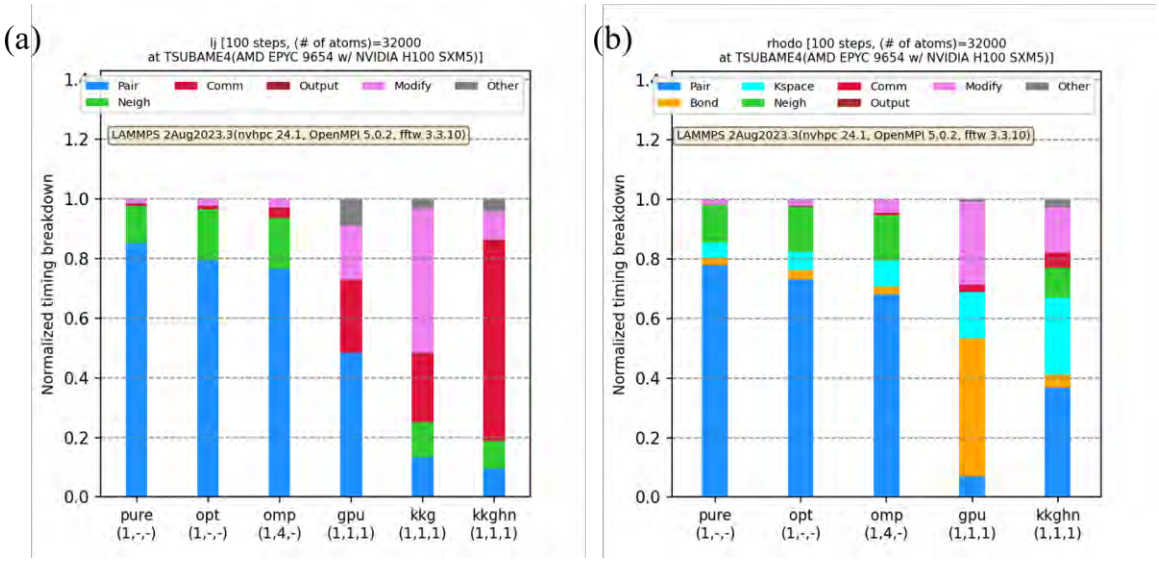


図 2: 基本ベンチマークにおける経過時間内訳。横軸は図 1 と同じ。縦軸はメインループの経過時間で規格された時間。(a) lj、(b) rhodo。

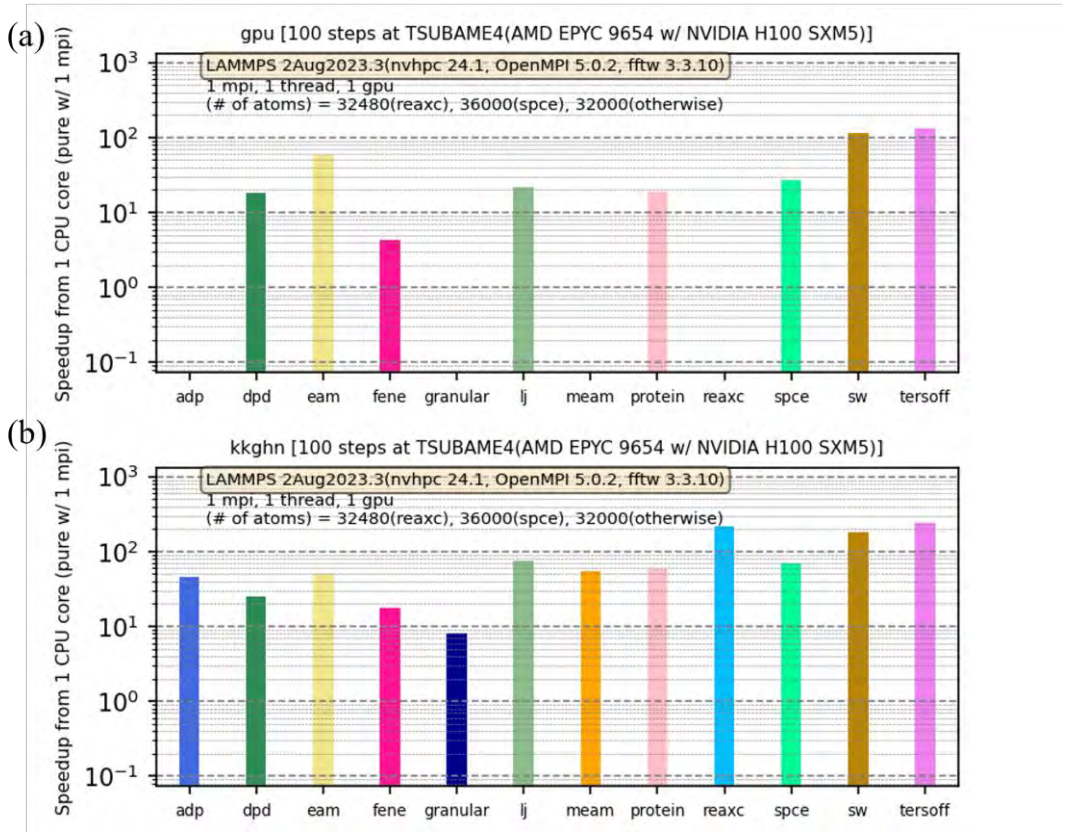


図 3: 力場ポテンシャル・ベンチマークに関する GPU 利用による速度向上。横軸はベンチマーク名（力場の種類に対応）、縦軸は 1 CPU コア利用（pure で 1 MPI プロセス）からの速度向上。(a) gpu、(b) kkghn。パッケージ名（pure, gpu, kkghn）の詳細は表 1 を参照。データにパッケージが対応していない場合、グラフは示されない。

マークから lj と rhodo に着目する。データサイズは replicate コマンドで拡張した。オリジナルのデータについて、3 次元空間の x 軸、y 軸、z 軸方向をそれぞれ 2 倍拡張した（256000 原子データ）。図 1 における計測

表 2: 1 MPI, 1 スレッド, 1 GPU 使用時からの速度向上。データはサイズ拡大した lj (256000 原子)。性能のベースラインは、(Time steps)/sec.で、gpu について 506、kkghn について 2064。

	MPI	スレッド数	Speedup
gpu	1	2	1.3
gpu	2	1	1.2
kkghn	1	2	1.1
kkghn	2	1	0.11

表 3: 1 MPI, 1 スレッド, 1 GPU 使用時からの速度向上。データはサイズ拡大した rhodo (256000 原子)。性能のベースラインは、(Time steps)/sec.で、gpu について 10.1、kkghn について 105。

	MPI	スレッド数	Speedup
gpu	1	2	1.0
gpu	2	1	1.9
kkghn	1	2	1.0
kkghn	2	1	0.20

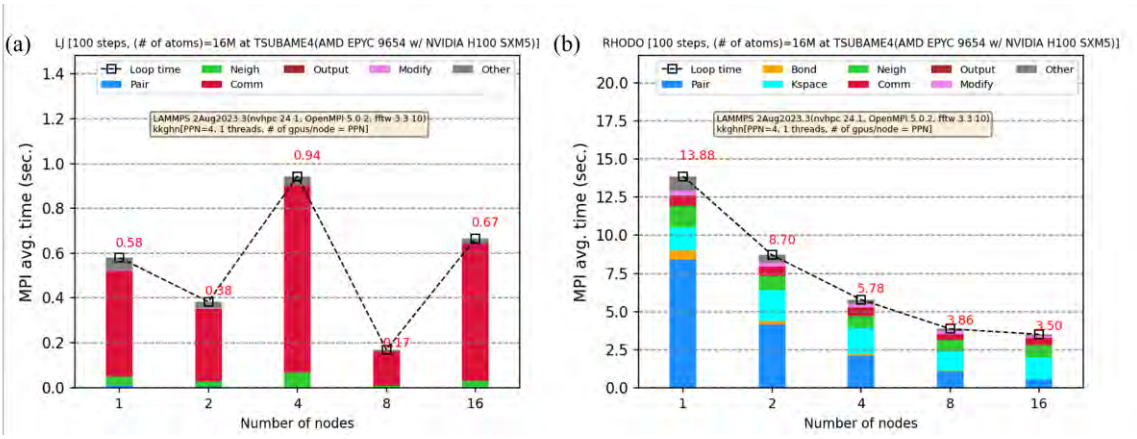


図 4: データ拡張された基礎ベンチマークに対するストロングスケール測定（16384000 原子）。横軸はノード数、縦軸は経過時間。ノードあたりの MPI プロセス数は 4 に固定し、GPU はノードあたり 4 枚使用する。OpenMP スレッド数は 1 に固定。(a) lj、(b) rhodo。

と同様に、TSUBAME 4.0 の 1/4 ノードを使用する。まず、1 CPU コア、1 スレッド、1 GPU の設定で、GPU パッケージより KOKKOS パッケージのほうが高速であることが確認された（lj で 4 倍、rhodo で 10 倍の性能差）。それぞれのパッケージについて、この性能をベースラインとし、CPU コア数増大による速度向上を調べる（表 2、表 3）。GPU 1 枚に対し、CPU コア数を 2 とする。CPU コアは OpenMP スレッド並列か MPI プロセスかのどちらかに割り当てる。gpu ではスレッド数と MPI プロセス数のどちらについても、増大で正の効果を得た。特に、データ rhodo について、MPI プロセス数の増大は理想的な性能向上が見られた（なお、4 MPI プロセスで性能向上は飽和した）。LAMMPS のドキュメント[6]の GPU Packages において、GPU パッケージでは 1 GPU に対して MPI プロセス数の増大による性能向上の可能性が指摘されており、その記述と無矛盾な結果と言える。一方、kkghn については、スレッド数の増大で性能は変わらず、MPI プロセス数増大は性能劣化を引き起こすことがわかった。MPI ランクと GPU のデバイス ID に関する 1 対 1 のバインディングを破る実行は、LAMMPS のドキュメント[6]の Kokkos Packages において非推奨とされており、その記載を計測で確かめることができたと言える。このように、GPU パッケージと KOKKOS パッケージとで性能向上に寄与する CPU コアの設定方法が異なることがわかった。GPU パッケージは、CPU コアを余らせることなく使用できる可能性がある。ただし、今回の調査の範囲では、KOKKOS パッケージの性能に迫る設定を見出すことはできなかった。一方、KOKKOS パッケージでは、さらなる高速化を狙う場合は、MPI ランクとデバイス ID の 1 対 1 バインデ

表 4: Kokkos によるカーネル作成例: 行列行列積カーネルの性能（倍精度浮動小数点，行列次元: 1024）。

配列次元	実装方法	Gflop/s
2	Kokkos: TeamPolicy, TeamVectorRange	2918.253
1	OpenACC	3411.408
1	cuBLAS	34775.811

ィングを保ちつつ、GPU 数（すなわちノード数）を増やす方策が考えられる。

ここまでの結果から、多様な力場への対応と性能の観点で、NVIIDA GPU を使った LAMMPS の実行では KOKKOS パッケージの利用が優位と考えられる。

LAMMPS の性能評価の最後として、マルチ GPU 利用でのスケーラビリティを調べる。この調査では KOKKOS パッケージのみに着目する。データは lj および rhodo を x 軸、y 軸、z 軸方向それぞれに 8 倍に拡張した 16384000 原子数データを対象とする。上記の調査結果に従い、ノード内における MPI プロセス数 (PPN) はノード内 GPU の数と一致させておく。TSUBAME 4.0 では 1 ノードあたり 4 枚の H100 が搭載されているため、PPN は 4 に設定することになる。OpenMP スレッド数は 1 に設定する。計測では、node_f キューを使用し、ノード数を 1 から 16 まで増大させた。結果を図 5 に示す。データによりスケーラビリティが大きく異なる。データ lj では、Pair の寄与はほぼ消えており、性能のボトルネックは通信関数を含む計測区間である Comm であることがわかった。小規模実行の段階で（図 2)、lj は単一 GPU 利用で十分に Pair の寄与が削減されている。こうして、データ lj は KOKKOS パッケージによるマルチ GPU 利用に適するものではなかったと判断される。一方、データ rhodo については、8 ノードまで良好なスケーラビリティを示しており、Pair の寄与がノード数の増大とともに順調に削減している。高ノード側におけるスケーラビリティの阻害要因は高速フーリエ変換を含む計測区間である Kspace となることがわかった。

最後に、Kokkos を利用したカーネルコード作成とその検証についてまとめる。今後の LAMMPS の KOKKOS パッケージおける機能拡張やチューニングなどを念頭に置き、Kokkos を利用した簡単なカーネルを作成した。MD 計算自体とは直接関係はないものの、Kokkos を利用したコーディングで必要となる要素技術を習得できるようなカーネルとして、行列行列積、STREAM、MPI 通信関数の呼び出しなどを実装し、OpenACC や CUDA との実装の結果と比較した。本報告書では、倍精度密行列に対する行列行列積の実装と計測結果を示す(表4)。リスト 1 に Kokkos を使った実装の例を示す。レファレンスとして実装した OpenACC の例をリスト 2 に示す。Kokkos と OpenACC とで同程度の性能となることがわかる。素朴な実装であり、cuBLAS による実行の 1/10 程度の性能となった。リスト 1 のような階層的並列は LAMMPS の力場カーネルでも使われているパターンであり、そのコーディング方法を把握できた。

まとめ、今後の課題

TSUBAME 4.0 における LAMMPS による MD 実行において基礎的な性能評価を実施した。GPU 向けコードとして異なる実装である GPU パッケージと KOKKOS パッケージを比較し、今回の調査の範囲では、力場への対応状況と性能から、KOKKOS パッケージの利用が好ましいことがわかった。ノード内実行および複数ノードを使った実行についても性能上の知見を得ることができた。また、Kokkos を用いた単純なカーネルコードを実装し、LAMMPS の機能拡張に必要となるコーディング上の基礎知識を習得した。作成したカーネルコードは、RIST が開催する講習会等でサンプルとして公開したいと考えている。

リスト 1: Kokkos による行列行列積の実装例（Kokkos: TeamPolicy, TeamVectorRange）。

```
// Kokkos: TeamPolicy, TeamVectorRange
typedef Kokkos::Cuda ExecSpace;   struct TagA {};
void MyKernel::mm_A () {
   Kokkos::parallel_for("Kernel_A",
      Kokkos::TeamPolicy<TagA,ExecSpace>(NN, Kokkos::AUTO, 32), *this);
}
KOKKOS_INLINE_FUNCTION
void MyKernel_kk::operator() (TagA,
   const typename Kokkos::TeamPolicy<ExecSpace>::member_type & team) const
{
   int i = team.league_rank();
   // use of copy ([=]) would correctly work.
   Kokkos::parallel_for(Kokkos::TeamVectorRange(team, 0, MM), [&] (const int & j ) {
      double tmp = 0.0;
      for ( int k = 0; k < LL; ++k ) {
         tmp += d_A( i, k ) * d_B ( k, j );
      }
      d_C( i, j ) = tmp;
   });
}
```

リスト 2: OpenACC による行列行列積の実装例。

```
// OpenACC
#pragma acc loop gang independent
for ( int i = 0; i < NN; ++i ){
#pragma acc loop vector independent
for ( int j = 0; j < MM; ++j ) {
   double mctmp = 0.0;
   #pragma acc loop seq
   for ( int k = 0; k < LL; ++k ) {
      mctmp += ma[k + LL*i]*mb[j + MM*k];
   }
   mc[j + MM*i] += mctmp;
}}
```

課題として、以下の 3 点が考えられる。第一に、KOKKOS パッケージの利用で複数 CPU コアを有効に利用できていない点が挙げられる。検討として、例えば、CPU コアのスレッド並列の効果について、スケールしない原因がデータに起因するか、実装の問題なのかを明らかにすることは重要と考えられる。1 ジョブの中で MPMD モデル的に複数のプログラムを動かす方策も CPU コアの有効活用の観点で重要と思われる。第二に、KOKKOS パッケージで実装されていない機能について検証を進めることが挙げられる。KOKKOS パッ

ケージは GPU パッケージに比べれば多くの力場に対応しているが、例えば自由エネルギー計算（FEP パッケージ）には対応していない。こうした計算機能の GPU 対応を進めたい。第三に、今回の検証から外した外部ライブラリを使用する計算機能について GPU 利用時の性能を検証することが挙げられる。例えば、QM/MM 計算で用いられる QMMM パッケージなどが考えられる。

参考文献

[1] A. P. Thompson, H. M. Aktulga, R. Berger, D. S. Bolintineanu, W. M. Brown, P. S. Crozier, P. J. in't Veld, A. Kohlmeyer, S. G. Moore, T. D. Nguyen, R. Shan, M. J. Stevens, J. Tranchida, C. Trott, and S. J. Plimpton, Comput. Phys. Commun., 271, 10817 (2022).

[2] LAMMPS Molecular Dynamics Simulator; https://www.lammps.org/

[3] Kokkos: https://kokkos.org/

[4] LAMMPS Benchmarks; https://www.lammps.org/bench.html

[5] A. P. Thompson, L.P. Swiler, C.R. Trott, S.M. Foiles, and G.J. Tucker, J. Comput. Phys. 285, 316-330 (2015).

[6] LAMMPS Documentation; https://docs.lammps.org/

TSUBAME 共同利用　令和6年度　学術利用　成果報告書

利用課題名　量子アニーリングにおける近似的断熱ショートカット
英文：Approximated shortcuts to adiabaticity in quantum annealing

利用課題責任者：早坂　太志
Hiroshi Hayasaka

産業技術総合研究所
National Institute of Advanced Industrial Science and Technology (AIST)
https://www.aist.go.jp

邦文抄録（300 字程度）

我々は磁化のダイナミクスを計算することで，平均場近似された CD 項を構築する手法を提案した．量子スピングラス模型に対して量子アニーリングを行ったところ，平均場近似された CD 項を用いることにより正解率のスケーリングが改善されることが分かった．

英文抄録（100 words 程度）

The counter diabatic (CD) driving has attracted much attention for suppressing non-adiabatic transition in quantum annealing (QA). However, it can be intractable to construct the CD driving in the actual experimental setup due to the non-locality of the CD driving Hamiltonian and necessity of exact diagonalization of the QA Hamiltonian in advance. Using our method, we clarify that a scaling of success probability is improved compared to the conventional QA without the CD driving.

Keywords: Quantum annealing, Counter diabatic driving, Spin glass, Mean field theory

背景と目的

量子アニーリング (QA) は，量子多体系の基底状態を求める手法として注目を集めている [1]. 一般に QA では，ハミルトニアンによる状態変化を断熱条件を満たすように, 長い時間をかけ行う必要がある. 一方, Counter-diabatic (CD)ドライブの方法は，非断熱遷移を抑制し，短時間での量子アニーリングを行うことを可能にすることが知られている [2, 3]. しかしながら CD 項の構成には厳密な固有状態を事前に用意する必要があるうえに, CD 項が非局所的な形で表されるため，実際に実装する際には困難を伴う.

この問題を解決する一つの方法として，CD 項を平均場近似する手法が知られている [4]. 平均場理論において相互作用が一様な強磁性イジング模型の場合，解くべき平均場方程式 (自己無撞着方程式) は系を代表する一つの磁化の閉じた方程式となる.

しかし相互作用がランダムな問題では，各量子ビットのもつ磁化を区別する必要があるため，自己無撞着方程式は，システムのサイズを L としたとき，L 個の非線形な連立方程式となる．平均場 CD 項の構成にはこの非線形な連立方程式を量子アニーリングの各時刻で解く必要があるため，計算時間の律速となる.

この問題を解決するため、我々は磁化のダイナミクスを計算することで，平均場近似された CD 項を得る手法を提案する. 我々の方法では各時刻で自己無撞着方程式を解く必要はなく，初期配位のみを与えれば良い．この手法を用いて量子スピングラスに対して量子アニーリングを行い，正解率のスケーリングが通常の QA に比べて改善することを示す.

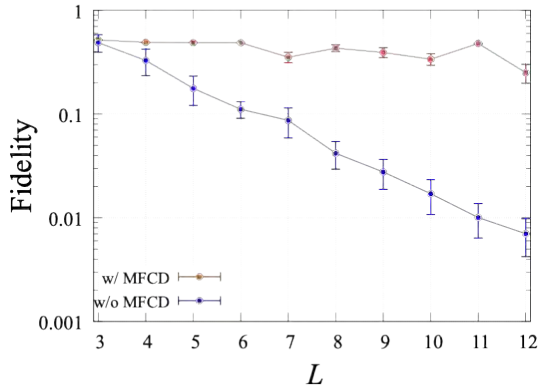
概要

非一様な相互作用をもつスピン系において，平均場近似された CD 項を構成する手法を提案した．この手法を用いて量子スピングラスに対して量子アニーリングを行っ

たところ，正解率の向上が確認できた．

結果および考察

(a)



(b)

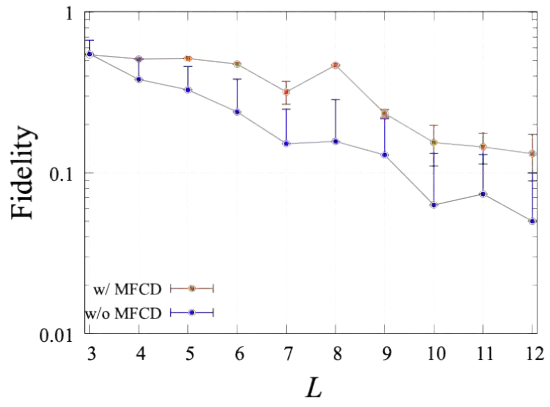


図 1. 正解率のスケーリング． (a) 無限レンジ横磁場スピングラス模型． (b) 1 次元横磁場スピン鎖．アニーリング時間は T=1 とした．赤線が平均場 CD 項を用いた場合．青線が通常の QA．

平均場近似された CD 項を横磁場スピングラス模型に対して適用し量子アニーリングを行った (図 1)．スピングラス模型の相互作用は正規分布で与えられるものを選んだ．また無限レンジ系 (図 1(a))，および 1 次元スピン鎖 (図 1(b)) を扱った．相互作用のサンプル数 (インスタンス数) は 16 サンプルとした．また磁化が準安定解に落ち込むのを回避するために一様ランダムな縦磁場を導入した．縦磁場のサンプリングは各インスタンスに対してそれぞれ 48 サンプルを取った．インスタンス数と縦磁場のサンプル数の両方に関しての平均値を正解率としてプロットしてある．図 1(a), (b)より，平均場近似された CD 項を用いることで，通常の QA よりも正解率のスケーリングが改善されることを明らかにした．また無限レンジ系の方が 1 次元スピン鎖よりも平均場近似された CD 項の効果がより顕著に現れている．一般に平均場近似は高次元系の方が良い近似となるため 1 次元よりも高い次元ではより平均場近似された CD 項の効果が現れることが期待できる．

まとめ、今後の課題

我々は磁化の平均場ダイナミクスを計算することで，平均場近似された CD 項を構築する手法を提案した．量子スピングラスに対して量子アニーリングを行ったところ，平均場近似された CD 項によって正解率の向上が確認できた．

参考文献

[1] T. Kadowaki, et al., Phys. Rev. E, **58**, 5355 (1998)

[2] M. Demirplak, et al., J. Phys. Chem. A **107**, 9937 (2003)

[3] M. V. Berry, J. Phys. A: Mathematical and Theoretical **42**, 365303 (2009)

[4] T. Hatomura, J. Phys. Soc. Jpn. **86**, 094002 (2017)

TSUBAME 共同利用　令和 6 年度　学術利用　成果報告書

利用課題名：生成型音声・画像 AI のための学習アルゴリズムの開拓と自動機械学習の研究
英文：Investigation of Training Algorithms and AutoML for Generative AI

利用課題責任者：坂東　宜昭
First name Surname: Yoshiaki Bando

所属：国立研究開発法人　産業技術総合研究所
Affiliation: National Institute of Advanced Industrial Science and Technology (AIST), Japan
URL: https://www.airc.aist.go.jp/cosine

邦文抄録（300 字程度）
本利用課題では，我々の日常生活で人々とコミュニケーションをとりながら適切に仕事をこなす人工知能（AI）技術を実現するため，音声・画像 AI のための学習アルゴリズムの開拓と自動機械学習の研究を進めた．特に，映像信号と音響信号を入力として音響イベントの発生時刻と音響物体を表す領域を推定する視聴覚音源定位（AV-SSL）に取り組み，Deep Clustering に基づく視聴覚埋め込み空間を学習することで，複数音源を容易に弁別可能する特徴量抽出器を構築した．さらに，データセットのキャッシュ機構の構築や動画データの効率的な HDF5 形式への保存など，高速に研究ループを進展させるための研究ツールを整備した．

英文抄録（100 words 程度）
We investigated machine learning research for building multimodal artificial intelligence (AI) systems that can communicate humans in our daily lives. Specifically, we investigated audio-visual sound source localization (AV-SSL), which is a task to detect and localize sound events from video and audio recordings. We trained audio and visual encoders whose embedding space is designed to easily distinguish multiple sounding objects based on the deep clustering technique. In addition, we built supporting tools for training neural models on high-performance computing (HPC) systems.

Keywords: Audio-visual training, generative AI, contrastive learning, self-supervised learning, automated machine learning

背景と目的

膨大なデータに対し効率的にスケールする大規模言語モデル（LLM）や自己教師あり学習の台頭により，多くの人工知能（AI）タスクが実用に足る性能を達成している．一方，我々の日常生活で人々とコミュニケーションをとりながら適切に仕事をこなす AI 技術は未だ実現に至っていない．本研究の目的は，このような人間と協働する AI 技術を確立することである．特に，ユーザの利用環境に応じて適応的に高い性能を出すための枠組み，AI 自体の開発段階における効率的な学習や MLOps パイプライン，ユーザの要望に応じて適宜自身を変化できるメタ学習技術は，未だ発展途上の課題であり，本研究課題ではこれらの解決を目指した．

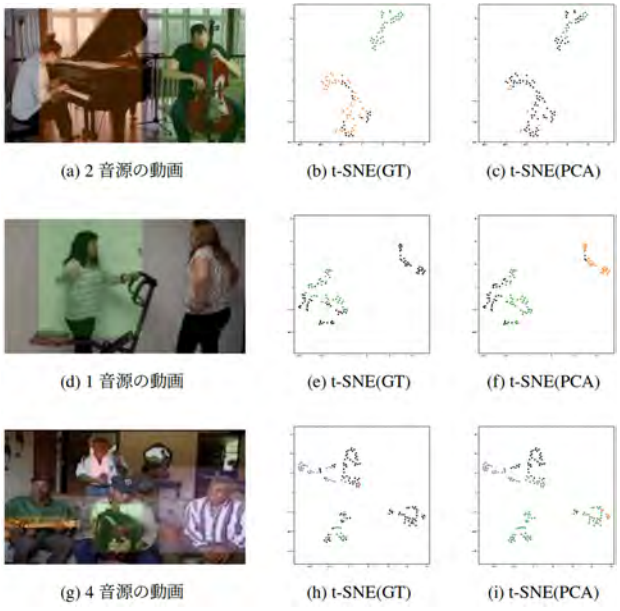


図 1 学習した埋め込み空間の可視化結果例 [1]

概要

本課題では，実世界で人間と協調する次世代 AI 技術を確立するため，その効率的な学習方法の開拓ともに実応用で課題となる自動機械学習（AutoML/AutoPEFT）に取り組む．具体的には，音

(a) 水平角 0°　(b) 水平角 90°　(c) 水平角 180°　(d) 水平角 270°　(e) 結果の統合

図 2 得られた埋め込みモデルを用いた実収録データでの動作検証 [1]

声・画像 AI のためのスケーラブルな学習アルゴリズムの開拓，実世界データにおける性能評価ベンチマークの確立など，実世界を理解し人々と協調するAIに必要な要素技術の開拓を進めた．

結果および考察

具体的には，映像信号と音響信号を入力として音響イベントの発生時刻と音響物体を表す領域を推定する視聴覚音源定位（AV-SSL）に取り組み，Deep Clustering に基づく視聴覚埋め込み空間を学習することで，複数の音響イベントを容易に弁別可能にする特徴量抽出器を構築した [1]．

図 1 に得られた埋め込み空間の例を示す．図 1-(b), (e)，および(h)から，得られた埋め込み空間は，音響物体ごとに異なる特徴（色）へ埋め込めていることが分かる．また，図 2 に示すように，対話ロボットへの搭載を目的として，実収録した全方位映像に対して構築したモデルを適用する実験を行った．構築したモデルは，実時間推論できるよう軽量なアーキテクチャを採用している．また，ロボットでは周囲の状況を理解するため全方位画像の活用が望ましいが，そのような学習データの収集にはコストを要する．そこで，幾何変換した空間での埋め込みのクラスタリングを活用することで，Web から収集した動画で学習したモデルでも，容易に全方位画像に適用できる枠組みの構築を進めた．

本研究の学習では，我々が収集した約 160 時間の視聴覚データを用いた．視聴覚データとしては比較的大規模なデータセットであり，このような動画データから効率的な学習を実現するため，Lustre ファイルシステムへのアクセスを最小限とするキャッシュ機構を構築した．torch.utils.data.Dataset のラッパーとして記述でき，既存の Dataset クラスをラップすることで，1 度アクセスした結果を共有メモリまたはローカルストレージにキャッシュする．素朴な，学習開始前にデータをローカルストレージに転送する場合と比べて，ジョブ実行から学習開始までの時間を大幅に削減でき，モデル学習のデバッグ効率を改善することができた．さらに，動画データの HDF5 ファイルへの効率的な保存方法の検討もすすめた．HDF5 ファイル内に素朴にフレーム情報を保存すると，動画に特化した(非可逆)圧縮方式に比べて圧縮効率が悪い．一方，小さな大量の動画ファイルを逐一ファイルシステムから読み出すと，Lustre ファイルシステムへの負荷が上がってしまい効率が下がる．そこで，HDF5 ファイル内に mp4 コンテナをバイナリとして保存し，Dataset クラス内で動的に展開する機構を実装・評価した．殆どの動画デコードプログラムは，ファイルからの読み出しを前提としているが，本実装ではメモリ・ファイルシステム上での展開として実装することで，オーバーヘッドに短縮することができた．

まとめ、今後の課題

本利用課題では，我々の日常生活で人々とコミュニケーションをとりながら適切に仕事をこなす人工知能（AI）技術を実現するため，音声・画像 AI のための学習アルゴリズムの開拓と自動機械学習の研究を進めた．特に，実時間で動作する AV-SSL の構築を通じて，LLM などの記号処理システムに入力するための情報抽出技術の開発を実施した．

参考文献

[1] 大規模 Web データを用いたロボットのための視聴覚音源定位システム，櫻井　舜，坂東宜昭，佐々木洋子，大西正輝，応用音響・電気音響研究会，査読無，40-44，2024

TSUBAME 共同利用　令和6年度　学術利用　成果報告書

利用課題名　極限環境構造材料の原子論解析に資する構造材料用ニューラルネットワーク原子間相互作用を用いた変形・破壊解析

英文：Deformation and fracture analysys based on atomistic simulation using neural network interatomic potential for structural materials under harsh environment

利用課題責任者
Shigenobu Ogata

大阪大学大学院基礎工学研究科
Graduate School of Engineering Science, University of Osaka
https://tsme.me.es.osaka-u.ac.jp/

邦文抄録（300 字程度）
本課題では、第一原理計算、分子動力学法、モンテカルロ法などの原子論的手法を用い、実験が困難な極限環境下における構造材料内部の塑性変形、組織変化、破壊現象の定量的理解を目指した。特に、炭素を含む鉄鋼材料に適用可能な高精度ニューラルネットワーク原子間相互作用を構築し、変形・破壊挙動の解析を行った。その結果、鉄中らせん転位の移動に対し、炭素原子が移動を阻害しつつも、応力上昇により炭素雰囲気から転位が離脱する現象が繰り返されることを明らかにし、これが炭素鋼の変形試験において観察されるセレーションの起源であることを定量的・原子論的に示した。

英文抄録（100 words 程度）
This study investigates plastic deformation and fracture mechanisms in carbon steels under extreme conditions using atomistic simulation methods, including first-principles calculations, molecular dynamics, and Monte Carlo simulations. A high-accuracy neural network interatomic potential was developed for the iron-carbon system. Using this interatomic potential, we analyzed screw dislocation behavior with and without carbon content. The results revealed that screw dislocation motion is repeatedly impeded by carbon atoms, followed by detachment due to increasing applied stress. This alternate behavior was quantitatively and atomistically identified as the origin of serrated flow observed in experimental deformation tests of carbon steels.

Keywords: Atomistic simulations, Neural network interatomic potential, steel, dislocation, plastic deformation

背景と目的

鉄鋼材料は最も一般的な構造材料であり、様々な合金元素の添加や組織制御により高強度、高靭性を実現できることから、高温環境や高圧水素環境など極限環境での利用に供する構造材料としても期待されている。しかしながら、その微細構造の発達と構造性能の関係に関する理解はまだ完全とは言い難い。炭素は鉄鋼材料の最も主要な元素であり、わずかな量の炭素添加でも大きな力学特性の変化をもたらす。鉄鋼において炭素原子が果たす役割を根本的に理解するためには、その熱力学と動力学を原子レベルで特徴づけることが重要である。第一原理電子状態計算は、精度が高い反面、計算コストが高い。一方経験的原子間ポテンシャルを用いれば、計算コストを大幅に削減できるため、長期間にわたる広範な系のシミュレーションが可能になるが、その信頼性には疑問が残る。

本課題では、ニューラルネットワークに基づく機械学習原子間相互作用を構築することで、鉄—炭素系材料の変形における高速かつ高精度な動力学解析を実施し、変形・破壊過程における炭素の影響を原子レベルから理解することを目的とする。

概要

本課題では、大規模な第一原理計算、分子動力学法、モンテカルロ法などの原子論的手法を駆使し、実験・計測で取得困難な極限環境下での構造材料内部の塑性変形、組織変化、破壊の各現象に関する定量データを創出することを目的とした。そのために、本年度は炭素を含む鉄鋼材料に適用可能な高精度ニューラルネットワーク原子間相互作用を構築し、これを用い

た変形解析を行い、材料特性発現のメカニズム解明に取り組んだ。

結果および考察

鉄—炭素系のニューラルネットワーク原子間相互作用の構築を行い、これを用いて炭素が鉄中のらせん転位の運動に与える影響を解析した。図 1 に純鉄および炭素鋼モデルにおけるせん断変形試験の応力ひずみ線図（上）と、転位芯における炭素濃度の推移（下）を示す。純鉄においては、らせん転位の運動が開始すると、流動応力は一定値で推移することに対して、炭素を含むモデルでは一度大きな応力降下が発生した後、増加と急激な減少を繰り返し、セレーションを生じている。この傾向は炭素濃度が大きいほど顕著になっている。また、転位芯における炭素濃度の推移より、応力上昇は炭素が転位芯に取り込まれ高濃度になっているときに生じ、応力降下とともに濃度が急激に変化していることから、セレーションの発生は転位に取り込まれた炭素原子によるピン止めによると考えられる。

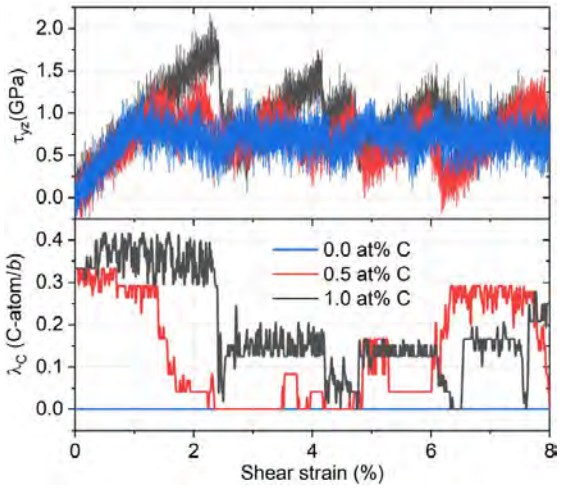


図 1. せん断変形シミュレーションにより得られた応力-ひずみ線図（上）と転位芯の炭素濃度（下）.

まとめ、今後の課題

本課題では鉄-炭素系ニューラルネットワーク原子間相互作用を構築し、これを用いた塑性変形解析を行った。その結果、炭素鋼で見られるセレーションの原因が炭素原子による転位のピン止めであることを定量的に明らかにした。今後は構築した原子間相互作用を用いた大規模変形・破壊解析を実施することで、粒界や第二相といった異なる欠陥、組織における変形およびそれらと転位との相互作用における炭素の影響について原子レベルからの理解を目指す。

TSUBAME 共同利用　令和6年度　学術利用　成果報告書

利用課題名　直接数値解析データを用いた非平衡壁面モデルの開発
英文：Development of non-equilibrium wall model by using the direct numerical simulation data

利用課題責任者
須賀　一彦

所属
大阪公立大学　工学研究科
URL：https://www.omu.ac.jp/eng/htlab/

直接数値解析データを用いて，ラージ・エディ・シミュレーションに適用する汎用的な非平衡代数壁モデルの検討を行った．対象とした流れ場は，チャネル乱流・平面衝突噴流として，直接数値解析で得られるデータを用いて，モデル化を施したフィルター化運動量式を解析し，局所的な乱流の非平衡が混合長モデルの係数に与える影響を調査した．その結果，チャネル乱流の壁面近傍においても，混合長モデルに含まれる係数は局所的な非平衡によって時空間的な分布を持つことが明らかとなった．また，非平衡の影響は，平面衝突噴流場でより顕著にみられ，モデル係数には歪み速度テンソルなどを用いてモデル化する必要があることが明らかとなった．

We performed direct numerical simulations to explore the non-equilibrium wall model for large eddy simulations. The flow geometries were turbulent channel flow and plan impinging jet flow and analyzed the filtered momentum equation to clarify the effects of local non-equilibrium on the mixing-length model in the filtered momentum equation. The results show that the model coefficient in the mixing-length model exhibits spatial and temporal variations depending on the local non-equilibrium. This effect is more pronounced for the plan impinging jet flow and the coefficient in the mixing-length model needs corrected by using the flow parameters such as the local strain rate tensor.

Keywords: Direct numerical simulation, wall modeling, mixing-length model, channel flow, plan impinging jet flow

背景と目的

近年の急速な計算機技術の発展に伴い，設計現場で使用される乱流の数値解析手法は，平均場を解くレイノルズ平均モデルから，時空間的な乱流変動を解くラージ・エディ・シミュレーション（LES：Large Eddy Simulation）に置き換わりつつある．LES では支配方程式にフィルター化平均方程式を用いるため，直接数値解析と比較して粗い計算格子を利用することができる．しかし，壁面近傍においては，境界層を解像する為に非常に細かい計算格子が要求されるため，解析コストは依然として高い．そこで，壁面近傍の計算負荷の軽減のために，壁面近傍に粗い計算格子を用いて，モデル化された壁面摩擦応力を境界条件として与える壁面モデルLESが提案されている．これまで，様々な壁面モデルが提案されているが，いずれも混合長モデルを基本としており，そのモデル係数には一定値が採用されている．その為，運動量輸送の局所的な非平衡性が渦粘性に与える影響を原理的に考慮できない．本研究では，チャネル・平面衝突噴流の直接数値解析を行い，局所的な非平衡性が混合長モデルの係数に与える影響を調査し，モデル改良の指針を得ることを目的とする．

概要

工学応用を見据えた壁モデル LES に組み合わせる汎用的な壁面モデルとして，薄層近似された運動量式を積分することで得られる非平衡代数壁モデル[(1)]の開発を行う．チャネル・平面衝突噴流の直接数値解析を実施し，得られた瞬間場にフィルター平均を施し，モデル化されたフィルター平均運動量式を解析する．特に，局所的に生じる乱流の非平衡が運動量式内の渦粘性モデルにあたえる影響について議論する．

チャネル乱流の解析では，摩擦レイノルズ数を 300 と 650 とした 2 通りで解析を行う．計算格子は，壁面近傍の粘性低層内部に数点の計算格子が入るように設定し，計算領域$6\delta(x) \times 2\delta(y) \times 3\delta(z)$に対して，摩擦レイノルズ数 650 のケースでは$512(x) \times 256(y) \times 512(z)$とした．解析の妥当性を評価するために，摩擦レイノルズ数 300，650 のケースで平均速度分布を比較

した結果を図1に示す．図より，本研究の結果は文献(2)の結果とよく一致しており，解析の妥当性が確認された．

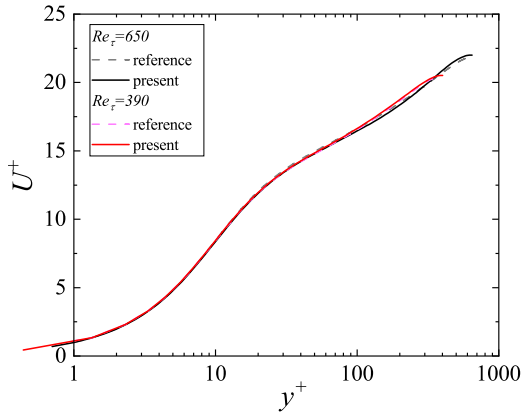
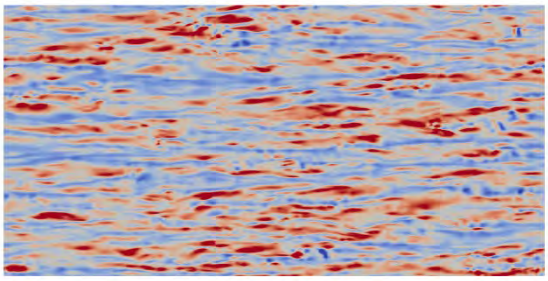
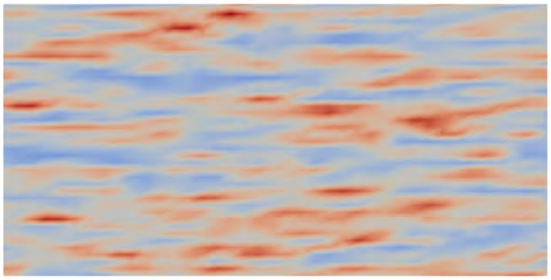


図1　平均速度分布の比較．

（a）



（b）



（c）

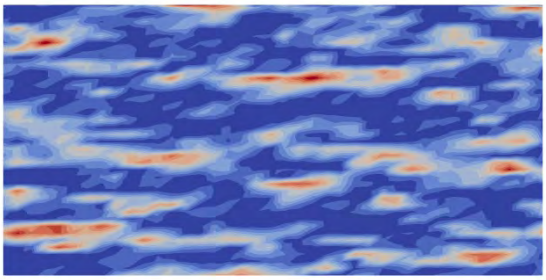


図2　（a）壁面摩擦応力，（b）フィルター化壁面摩擦応力，（c）混合長モデルの係数．（a），(b)は平均の摩擦応力で無次元化しており，0 が青色，2 が赤色に対応する．(c)は標準値 0.4 との差を標準値で規格化しており，青が−1，赤が 1 を示す．

次に，壁モデルで予測を行う壁面摩擦応力について議論する．図2（a）に瞬時の壁面摩擦応力を示す．壁面摩擦応力は，平均の摩擦応力を用いて無次元化を行った．図2（a）より，瞬時の壁面摩擦応力は，主流方向に伸びた乱流構造であるストリークの影響を強く受けていることが分かる．次に，図2（b）にフィルター化を施した摩擦応力分布を示す．ここで，フィルター幅は$\Delta_x^+ = 104$，$\Delta_z^+ = 52$と設定し，$40(x) \times 40(z)$と粗視化したLES を想定した．図 2(b)より，フィルター化を施した壁面摩擦応力は，フィルター化前と比べて摩擦応力の空間的な変動は小さいが，主流方向に伸びるストリークの影響を受けた分布となっていることが確認できる．次に，フィルター化されたデータを用いて壁モデルに含まれる混合長モデルの評価を行う．本研究では，薄層近似を施した運動量式に混合長モデルを適用した下記の式の解析を行う．

$$\frac{\partial}{\partial y^+}\left[(1+\alpha y^+)\frac{\partial \langle u\rangle^+}{\partial y^+}\right] = C_u^+$$

ここで，$\langle u\rangle$はフィルター化を施した主流方向速度を表し，右辺のC_u^+は主流方向の運動量式の圧力勾配，時間微分，対流項，壁面並行方向の粘性項を合算した項を表す．この項は，平衡境界層を仮定すると零となり，非平衡境界層では非零の値を持つ．また，フィルター化平均によって生じたサブ・グリッド・スケール応力は渦粘性モデルによってモデル化し，サブ・グリッド・スケール渦動粘度は混合長モデルによって壁面からの無次元距離y^+と係数αによってモデル化した．壁モデルでは，上式の積分式に対して，セル上面に境界条件を与えることで積分式の定数を定め，壁面摩擦応力を得る．平衡境界層を仮定したレイノルズ平均場では，係数は一定値$\alpha = 0.4$を取るが，本研究では時空間的に分布する乱流変動がαに与える影響を議論する．図2（c）に直接数値解析のフィルター化壁面摩擦応力から得られるαの分布を示す．なお，壁面垂直方向に対して$\Delta_y^+ = 97$の幅のフィルター平均を施している．図より，モデル係数αは空間的に大きく分布をしていることが確認できる．また，図 2(b)のフィルター化壁面摩擦応力との間で必ずしも強い相関関係も見られるわけでは無く，局所の非平衡性の大きさやフィルター幅以外にも，渦度や歪度などの局所の流れ量を用いてモデル化する必要が

あることが示唆された．次に，平面衝突噴流の解析領域を図3に示す．解析条件は文献(3)と同様であるが，ここでは概要を簡単に説明する．ドライバー部で完全発達させたチャネル乱流をメイン部の入口条件とし，メイン部では上面・下面は滑りなし，スパン方向には周期境界条件を課し，出口には対流流出境界条件を課した．流入面から衝突壁面までの距離はDとし，流入断面の水力直径と平均流速によって定義されるレイノルズ数を 9120 とした．ドライバー部は領域$3.2D \times D \times 1.6D$に対して計算格子を$64 \times 64 \times 96$を設定し，メイン部では領域$21D \times D \times 1.6D$に対して計算格子を$320 \times 96 \times 96$と設定した．図4（a）にはスパン方向から見た平面の衝突領域近傍の瞬時速度の絶対値，（b）には壁面摩擦応力を示す．図4（a）より，よどみ点付近では流速が遅く，衝突領域から少し離れた領域で速度が速くなっており，下壁に沿って高速の流体が流れている様子が分かる．壁面摩擦応力（図4（b））において，よどみ点付近では摩擦応力が小さく，よどみ点から少し離れた位置で最大値を取り，衝突領域から距離が離れるにしたがって減少する．また，壁面摩擦応力は流れ方向に伸びた縞状分布となっており，壁面近傍に発達する乱流渦の影響を強く受けていると考えられる．また，混合長モデルの係数 αは局所の歪み応力テンソルやC_u^+の影響を強く受けることが明らかとなっており，今後はそのモデル化について議論を進める予定である．

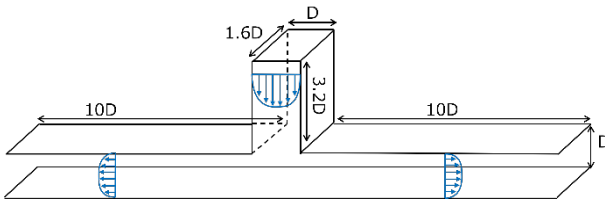
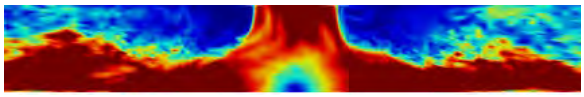


図3　平面衝突噴流の解析系

（a）



(b)

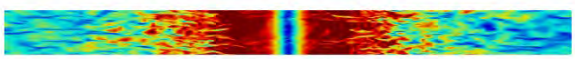


図4　（a）スパン方向から見た衝突領域近傍の瞬時速度，（b）壁面摩擦応力．青色は値が小さく，赤色は値が大きいことを示す．

結言

直接数値解析データを用いて，ラージ・エディ・シミュレーションに適用する非平衡代数壁モデルのモデル係数の検討を行った．対象とした流れ場は，チャネル乱流・平面衝突噴流として，直接数値解析で得られた壁面摩擦応力を用いて，局所的な運動量輸送の非平衡がフィルター化運動量式に含まれる混合長モデルの係数に与える影響を調査した．その結果，平均的には平衡境界層の仮定が成立するチャネル乱流であっても，混合長モデルに含まれる係数は局所的な運動量輸送の非平衡によって時空間的に分布を持つことが明らかとなった．また，非平衡の影響は，平面衝突噴流場でより顕著にみられ，定数には速度歪みなどの影響を考慮してモデル化する必要があることが明らかとなった．

参考文献

(1)K Suga *et al.*, “Algebraic non-equilibrium wall-stress modeling for large eddy simulation based on analytical integration of the thin boundary-layer equation”, *Phys. Fluids* **31** (7) (2019).

(2) H Kawamura, “Direct Numerical Simulation Data Base for Turbulent Channel Flow with Heat Transfer”, 1998-2004.

(3) H Hattori and Y Nagano. “Direct numerical simulation of turbulent heat transfer in plane impinging jet." *Int. J. Heat Fluid Flow* **25.5** (2004).

TSUBAME 共同利用　令和6年度　学術利用　成果報告書

利用課題名　自然言語の非線形性の計算論モデル
英文：Computational modeling of Natural Language Nonlinearity

利用課題責任者　田中久美子
Kumiko TANAKA-ISHII

早稲田大学　理工学術院　基幹理工学部
Waseda University, School of Fundamental Science and Engineering
URL　https://www.ml-waseda.jp/

邦文抄録（300 字程度）

本研究は、自然言語の非線形性を数理的に解析し、計算論的モデルとして表現することを目的とする。大規模言語モデル(LLM)の発展に伴い、自然言語の複雑性を精緻に定量化する手法の重要性が高まっている。本研究では、言語の自己相似性に着目し、相関次元を測定することで言語の複雑性を評価する手法を提案した。日本語・英語・ドイツ語・中国語の四言語において相関次元が約 6.5 に収束することを確認し、異なる言語間での共通する非線形な特性の存在を示唆した。さらに、生成モデルを活用した文書分類・クラスタリング手法を開発し、言語の複雑性を考慮した新たな情報処理技術の可能性を示した。今後の課題として、相関次元の形成要因をより詳細に解明し、言語モデルの推論能力を定量化する指標の開発を進める。

英文抄録（100 words 程度）

This study mathematically analyzes the nonlinearity of natural language and develops computational models to represent its complexity. With the advancement of large language models (LLMs), precise quantification of language complexity has become crucial. We propose a method that evaluates textual complexity through correlation dimension, capturing self-similarity in natural language. Our findings indicate a stable correlation dimension of approximately 6.5 across Japanese, English, German, and Chinese, suggesting a universal nonlinear characteristic. Additionally, we develop a document classification and clustering approach leveraging LLMs. Future work will focus on elucidating the formation mechanisms of correlation dimensions and establishing metrics to quantify reasoning capabilities.

Keywords: 自然言語処理, 生成モデル, 複雑性, 非線形性, 相関次元.

背景と目的

自然言語は高度に非線形な構造を持ち、その複雑性は計算的に解析しづらい特性を有する。この非線形性は、単語頻度の処理だけでは本質的な構造を理解することはできない。近年、大規模言語モデル(LLM)の発展により、言語の複雑性を数学的に解析する新たな可能性が開かれている。これにより、従来の言語モデリング手法では見えなかった構造が捉えられるようになり、言語の階層的特性や自己相似性といった要素の理解が進んでいる。

本研究では、自然言語の本質的な構造を数理的に定式化し、計算モデルを構築することで、言語の特性を精緻に捉えることを目的とする。また、本手法を法律・金融分野の応用に展開し、実際の情報処理技術に組み込むことで、言語理解・検索・知識抽出の効率向上を目指し、実用的なモデルの構築を行う。

概要

深層学習技術の進展により、大規模言語モデル（LLM）が提案され、情報科学全域に新たな可能性をもたらしている。LLM の性能が向上するにつれて、その限界や課題も明確になりつつある。その一つに、LLM が言語の「非線形」な特質をどの程度適切に学習・再現できるかという問題がある。従来の自己回帰型の言語モデルは、テキストの局所的な確率分布を学習することには長けているが、グローバルな言語構造を適切に捉えることが難しい。

本研究では、言語の「非線形性」に着目し、数学的・計算的手法を用いた解析を行う。特に、相関次元を用いた言語の複雑性定量化手法を開発し、異なる言語間での比較を試みる。このアプローチは、単なる情報エン

トロピーや perplexity（困惑度）とは異なり、言語が持つ階層的構造や自己相似性を測定することが可能である。本研究では、この手法を大規模言語モデルに適用し、LLM が学習する言語の構造がどのように進化するかを詳細に解析する。

また、生成モデルを活用した文書クラスタリング・検索手法を開発し、言語の複雑性を考慮した新たな情報処理手法の可能性を探る。これにより、従来のベクトル空間モデルに基づく手法を超え、生成モデルの持つ強力な表現力を活かした新たな検索・クラスタリング手法の実現を目指す。

結果および考察

本研究では、相関次元を用いた言語の複雑性定量化手法を提案し、日本語・英語・ドイツ語・中国語の四言語について分析を行った。その結果、各言語の相関次元が約 6.5 に収束すること(図 1)を確認し、異なる言語間で共通する非線形な特性が存在する可能性を示唆した。また、相関次元と長期記憶性の関連性を示し、言語の持つ情報構造が、単なる統計的特徴だけでなく、より深いレベルでの記憶性を反映していることを明らかにした。本研究の成果は Physical Review Research に掲載され[1]、さらにフラクタル研究の国際的なワークショップ Geometry and Stochastics にて招待講演[2]として発表された。

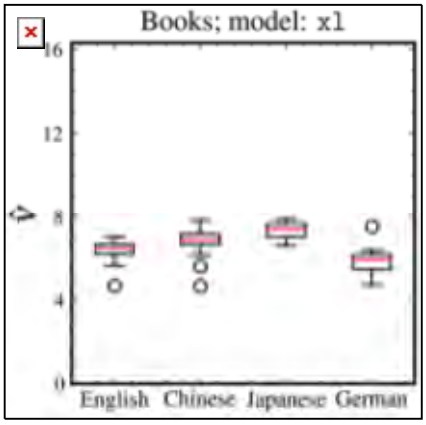


図 1 四言語テキストの相関次元

また、生成モデルを用いた文書クラスタリング・検索手法を開発し、言語の複雑性を考慮した新たな情報処理手法の可能性を探った。特に、新たなクラスタリング手法として「生成クラスタリング」を提案した。従来のクラスタリング手法では、主にベクトル埋め込みを用いた手法が一般的であったが、本手法では生成モデルを活用し、情報論的なアプローチに基づく新たな文書表現方法を提案した。その結果、従来の埋め込み方法と比較して大幅な改善が見られ(図 2)、情報検索や知識抽出における応用可能性が大きく広がった。本研究の成果は、世界トップの機械学習・人工知能学会である ICML[3]および AAAI[4]に掲載された。

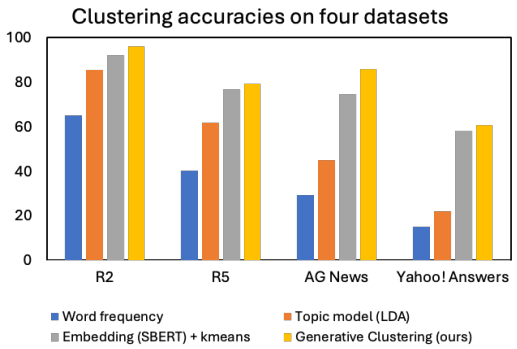


図 2 四データセット上で各クラスタリング方法の精度

まとめ、今後の課題

本研究では、相関次元を用いた言語の複雑性定量化手法を開発し、その有効性を大規模データセットに適用することで検証した。また、生成モデルを活用した文書クラスタリング・検索手法を開発し、新たな情報処理手法の可能性を示した。

今後の課題として、相関次元の形成要因の解明を進め、言語の階層的構造や自己相似性との関連を明確にする必要がある。また、相関次元を LLM の推論能力の評価指標として確立し、生成クラスタリングの改良を通じて法律・金融分野への応用を実用化することが求められる。さらに、LLM の学習ダイナミクスを非線形性の観点から分析し、モデルの特性や限界をより深く理解することで、自然言語処理技術の発展に寄与することを目指す。

参考文献

[1] Xin Du and Kumiko Tanaka-Ishii. Correlation dimension of natural language in a statistical manifold. Physical Reviews Research, 6 (2), L022028.
[2] Kumiko Tanaka-Ishii and Xin Du. Correlation dimension of large language models. In Proceedings of Fractal Geometry and Stochastics 7, 2024. Invited talk.
[3] Xin Du, Lixin Xiu, and Kumiko Tanaka-Ishii. Bottleneck-Minimal Indexing for Generative Document Retrieval. In Proceedings of ICML 2024.
[4] Xin Du and Kumiko Tanaka-Ishii. Information-Theoretic Generative Clustering of Documents. In Proceedings of AAAI 2025.

TSUBAME 共同利用　令和 6 年度　学術利用　成果報告書

利用課題名　大規模学術データ分析基盤の開発
英文：Development of Scholarly Data Analysis Methods

利用課題責任者
桂井　麻里衣

同志社大学理工学部
Faculty of Science and Engineering, Doshisha University
https://mm.doshisha.ac.jp

邦文抄録（300 字程度）
本研究課題は、国内外の学術データ（論文、特許、ウェブページなど）からの効率的な情報抽出を目的とし、データ整形などの前処理からモデル学習、諸タスクでの応用までのパイプラインを構築するものである。本年度はその一環として、テキストやグラフを入力としたニューラルネットワークに基づく特徴量で論文や研究者を表現した。得られた表現を論文や研究者の推薦タスクに応用した結果、テキストに基づく手法（コンテンツベース手法）がグラフベース手法より高い性能を示した。また、日本語論文の抄録文分類モデルを訓練する実験では、英日翻訳技術との組み合わせによるデータ拡張が効果的であることを示した。

英文抄録（100 words 程度）
This research project aims to build a pipeline for efficient information extraction from scholarly data (papers, patents, websites). The pipeline spans from preprocessing and data formatting through model training to information retrieval and recommendation tasks. As part of this framework, neural networks were trained with text and graph inputs to represent different entities—papers and researchers—through output features. When applying the resulting representations to recommendation tasks for papers and researchers, content-based methods demonstrated higher performance than graph-based approaches. Furthermore, in experiments training classification models for Japanese paper abstracts, we demonstrated that data augmentation through the combination with translation technology is effective.

Keywords: scholarly data analysis, science mapping, information retrieval, neural networks, text embeddings

背景と目的

学術論文の数は年々急激に増加しており、人手による情報収集や整理・分析はますます困難となっている。そのため、膨大な学術データの中から所望の情報を効率的に抽出し、分析可能な形にする技術の確立が必要不可欠である。

本研究課題では、学術データの利活用のための主要技術として、本年度はテキストやグラフを入力としたニューラルネットワークに基づき論文や研究者の特徴を表現する手法を構築した。得られた表現を論文や研究者の推薦タスクに応用した結果、テキストに基づく手法（コンテンツベース手法）がグラフベース手法より高い性能を示した。また、グラフのリンク予測の観点での推薦性能と推薦結果の意外性はトレードオフの関係にあることが示唆された。

上記の研究と並行して、学術データのテキストを意味内容に応じて分類するモデルを構築した。これまでに英語のモデルは盛んに研究されてきたものの、それらを日本語に適応させるための学術データ資源は現状不十分である。そのため、効率的なデータセット構築を目的とした情報抽出手法を構築した。実験結果より、翻訳技術に基づくデータ拡張の有効性を示した。

概要

本研究課題は、学術データ（論文、特許、ウェブページなど）からの効率的な情報抽出に向けて、データ整形などの前処理からモデル学習、諸タスクでの応用までのパイプラインを構築するものである。

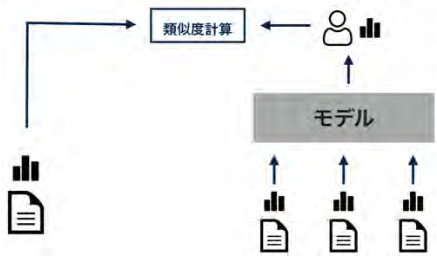


図 1:コンテンツベース手法の概要

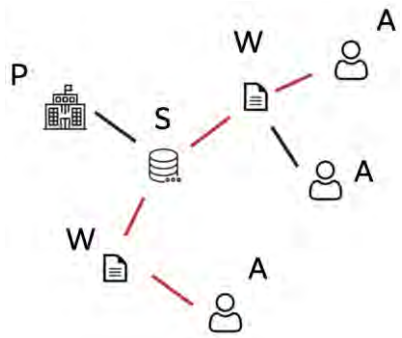


図 2:グラフベース手法とメタパスの概要

【学術データの埋め込み表現】

本年度は特に研究者への情報推薦に着目して研究を実施した。個々の研究者の業績一覧が入手できる場合、そのコンテンツ情報を各人の専門興味のモデル化に用いる手法をコンテンツベース手法とよぶ。一方、共著者情報まで入手できる場合、研究者や論文をノードとしたグラフ構造に基づくモデル化手法を適用できる。近年はこのようなグラフベース手法が注目を集めているが、コンテンツベース手法と同一条件下で性能を比較した例は報告されていない。そこで、学術情報推薦をグラフ内のリンク予測問題とみなし、これら二つのアプローチの性能を比較した。コンテンツベースとしては、図 1 に示すように、論文や研究者をタイトルテキストの埋め込みに基づいて表現する方法を採用した。グラフベース手法については、図 2 に示すように、著者（A）、論文（W）、発表場所（S）、出版社（P）という異なるエンティティ間のメタパスを事前に定義する手法と、メタパスを利用しない手法の 2 種類を用意した。

【文分類モデルの構築】

医学分野の論文抄録を背景、目的、手法、結果、結論のパートへ分類するタスクを実施するために、著者によって手動でラベル付けされている論文を大量収集し、データセットを構築した。そして、英語論文に基づく既存のデータセットを英日翻訳したものを擬似データセットとした。

表 1:著者間のリンク予測性能

手法	HR@1	HR@10	MAP@10
コンテンツベース	0.647	0.966	0.739
グラフベース（メタパス利用なし）	0.067	0.388	0.141
グラフベース（メタパス利用あり）	0.074	0.518	0.181

表 2:抄録文分類の性能評価

手法	Weighted-F1	Accuracy	Macro-F1
比較手法	90.12	90.53	83.49
提案手法	92.19	92.19	88.39

結果および考察

【学術データの埋め込み表現】

著者間のリンク予測実験において、上位 1 件と上位 10 件それぞれの Hit Rate（HR）、上位 10 件の Mean Average Precision を算出した結果を表 1 に示す。いずれの評価指標においても、コンテンツベース手法はグラフベース手法よりも高い性能を示した。グラフベース手法では、利用するメタパスが性能に影響するため、ドメイン知識に基づき慎重にメタパスを設計する必要があると考えられる。

リンク予測結果を精査すると、コンテンツベース手法はグラフベース手法に比べてグラフ内で距離が近いノード間の類似度を大きくする傾向があった。ゆえに、リンク予測を情報推薦に応用する場合、グラフベース手法の方が推薦結果に意外性が高いと考えられる。

【文分類モデルの構築】

構築した日本語論文データセットのみで言語モデルを訓練したものを比較手法、英語論文の翻訳に基づき拡張したデータセットで訓練したものを提案手法として、テ

ストデータでの文分類結果の Weighted-F1、Accuracy、Macro-F1 を算出した。表 2 に示すように、提案手法はいずれの評価指標においても比較手法を上回った。これにより翻訳を介したデータ拡張の有効性が示された。

まとめ、今後の課題

本研究課題では、テキストやグラフを入力としたニューラルネットワークに基づく特徴量で論文や研究者を表現し、論文や研究者の推薦タスクにおける性能を評価した。また、論文抄録の文分類モデル構築実験では、英日翻訳技術との組み合わせによるデータ拡張が効果的であることを示した。

今後の課題として、入力するグラフ構造の複雑さや特徴量の種類が研究者の特徴表現に与える影響を調査する必要がある。また文分類の研究については、広範な分野での実験と他言語への展開を予定している。

TSUBAME 共同利用　令和6年度　学術利用　成果報告書

利用課題名：フェイクメディア生成と検出
英文：Generation and detection of fake media

利用課題責任者 越前 功
First name Surname Isao Echizen

所属：国立情報学研究所
Affiliation: National Institute of Informatics
URL: https://research.nii.ac.jp/~iechizen/official/

邦文抄録（300 字程度）

本プロジェクトでは，ディープフェイクに代表されるフェイク顔画像・映像の検出モデルに関して，Computer Vision（CV）タスクを行う最新の Vision Transformer（ViT）を活用することで，従来のニューラルモデル（ConvNets）より真贋判定の精度向上を目指すことを目的とする．10 種類以上の大規模ディープフェイクデータセットを用いた真贋判定精度評価を行い，自己教師付き学習を用いて学習された最先端かつ商用利用可能な Vision Transformer（ViT）である DINOv2 の判定精度は，従来のニューラルモデル（ConvNets）である EVA-CLIP モデルの判定精度を上回った．本プロジェクトで得た知見は，国内外で実利用が進んでいるフェイク顔画像・映像の真贋判定サービスの改善に有益なものと考えられる．

英文抄録（100 words 程度）

This project aims to improve the detection accuracy of fake face images and videos, represented by deepfakes, by utilizing the latest Vision Transformer (ViT) for Computer Vision (CV) tasks, thereby surpassing the performance of conventional neural models (ConvNets). We conducted the detection accuracy evaluations using more than 10 large-scale deepfake datasets. The detection accuracy of DINOv2, a state-of-the-art and commercially available Vision Transformer (ViT) trained using self-supervised learning, surpassed that of the EVA-CLIP model, a conventional neural network model (ConvNets). The insights gained from this project are expected to be beneficial for improving fake face image and video detection services, which are increasingly being adopted both domestically and internationally.

Keywords: Deepfake，生成 AI，フェイクメディア，AI セキュリティ，なりすまし

背景と目的

顔，身体，音声，自然言語などの人間由来の情報を AI が学習し，本物と見紛うシンセティックメディアの生成が可能になりつつある．シンセティックメディアは，コミュニケーション分野やエンターテイメント分野を始めとした様々な用途で活用されている．一方で，シンセティックメディアの負の側面として，詐欺や思考誘導，世論操作を行う目的で，犯罪組織やテロ組織が，フェイク映像，フェイク音声，フェイク文書といったフェイクメディアを，AI を用いて生成・拡散させており，世界的な脅威となっている．

申請者らは，Deepfake が社会問題化する以前から Media Clone（造語．AI により本物と見紛うように加工・生成されたメディア）の拡散と生成を防止するプロジェクトを開始しており，2018 年にフェイク顔映像を AI を用いて検出する手法を世界で初めて提案した．さらに 2019 年に未知のフェイク顔映像にも頑健な検出手や，フェイク検出と改ざん領域の推定を同時に行う手法（5-years highest impact award 受賞）を提案し，Deepfake detection と呼ばれる新たな研究分野を創成した．

本プロジェクトでは，深刻化する多種多様なフェイクメディアの出現に対して，Computer Vision（CV）タスクを行う最新の Vision Transformer（ViT）を活用することで，従来モデルより真贋判定の精度向上を目指すことを目的とする．

本プロジェクトでは，深刻化する多種多様なフェイクメディアの出現に対して，Computer Vision（CV）タスクを行う最新の Vision Transformer（ViT）を活用することで，従来モデルと比較してフェイク顔画像・映像の真贋判定精度の向上を実現した．このプロジェクトで得た知見は，国内外で実利用が進んでいるフェイク顔画像・映像の真贋判定サービスの改善に有益なものと考えられる．

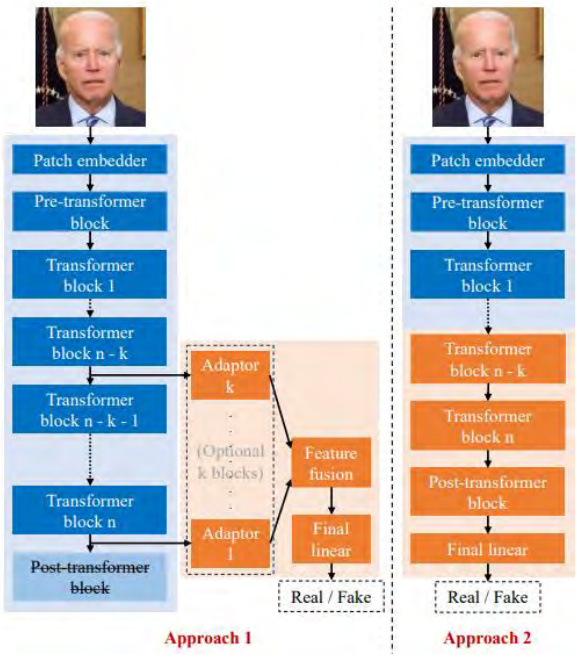


図 1：検討対象の 2 つのアプローチの概要．青色のブロックは frozen block をオレンジのブロックは，fine-tuned block を表す．

概要

フェイク顔画像・映像を検出するディープフェイク検出器を対象に，自己教師付き学習を用いて学習された最先端かつ商用利用可能な Vision Transformer（ViT）である DINOv2 を活用し，実用的なディープフェイク検知 AI 基盤モデルを開発する．Deepfake detection 分野で国際的にスタンダードな評価指標である分類精度（Accuracy）の等価誤り率（EER）の 2 つにおいて，従来のニューラルモデル（ConvNets）との比較評価を行った．

具体的には，顔のディープフェイク検出における事前学習済み ViT である DINOv2 の使用について，以下の 2 つのアプローチから比較研究を実施した（**図 1** 参照）：**Approach 1**：凍結したバックボーンをマルチレベル特徴抽出器として使用することと，**Approach 2**：最終的な変換ブロックを部分的に微調整（ファインチューニング）することである．その結果，**Approach 2** が有効であり，10 種類以上の大規模ディープフェイクデータセットを用いた真贋判定精度評価において，3 億パラメータの DINOv2 モデル（DINOv2-L14-Reg）が Accuracy 96%，EER 4%を達成した．一方，従来のニューラルモデル（ConvNets）である EVA-CLIP モデル（EVA02-CLIP-L-14）は，Accuracy 93%，EER 7%であった．

本評価で得られた知見は，国内外で実利用が進んでいるフェイク顔画像・映像の真贋判定サービスの改善に有益なものと考えられる．

まとめ、今後の課題

本プロジェクトでは，ディープフェイクに代表されるフェイク顔画像・映像の検出モデルに関して，Computer Vision（CV）タスクを行う最新の Vision Transformer（ViT）を活用することで，従来のニューラルモデル（ConvNets）より真贋判定の精度向上が可能なことを示した．今後の課題として，顔以外のオブジェクトや，音声，テキストといったモダリティを対象とした検討や，2 値判定のみならず，どの部分が AI により改変されたかといったローカライゼーションを行うモデルの検討が必要と考える．

TSUBAME 共同利用　令和6年度　学術利用　成果報告書

TSUBAME4.0 における AI 環境の利用技術の習得
Learning of technique to use AI programing environments in TSUBAME4.0

浅見　暁

一般財団法人高度情報科学技術研究機構
https://www.rist.or.jp/

近年、HPCI の資源提供機関においても GPU を搭載したマシンが増えつつあり、AI を活用した業務や研究が盛んに行われている。最新の GPU（NVIDIA 社製　H100）を搭載したスーパーコンピュータ TSUBAME4.0 と富岳プリポストノードを利用し、PyTorch、Tensorflow　のベンチマークを行った。データサイズによるが、富岳プリポストノードの GPU（V100）と TSUBAME4.0 の GPU（H100）では、PyTorch（YOLOv5）で約 3 倍、行列積（fp32）では、約 10 倍（CPU 比では 50 倍以上）の性能比であることが確認できた。

In recent years, the number of machines equipped with GPUs has been increasing at HPCI's resource-providing institutions, and AI-based work and research have been actively conducted. We benchmarked PyTorch and Tensorflow using the TSUBAME4.0 supercomputer equipped with the latest GPU (NVIDIA's H100) and the Fugaku pre-post node. Depending on the data size, we confirmed that the performance ratio of the GPU (V100) on the Fugaku pre-post node and the GPU (H100) on TSUBAME4.0 was about 3 times higher for PyTorch (YOLOv5) and about 10 times higher for matrix products (fp32) (more than 50 times higher than CPU ratio).

Keywords: 5つ程度　HPCI、TSUBAME4.0、富岳、PyTorch、TensorFlow

背景と目的

近年、HPCI の資源提供機関においても GPU を搭載したマシンが増えつつあり、AI を活用した業務や研究が盛んに行われている。また AI 分野以外の既存プログラムの GPU 化対応も進んでいる。そこで、最新の GPU（NVIDIA 社製　H100）を搭載したスーパーコンピュータ TSUBAME4.0 において、AI フレームワーク（PyTorch、Tensorflow）や AI プログラム（Resnet 等）の環境構築、性能分析、性能評価を行うことを目的とする。

概要

図1にベンチマークに用いたシステム、アプリケーション、図2に測定をおこなったベンチマークアプリケーションを示す。

TSUBAME4.0（CPU、GPU）、富岳プリポストノード（CPU、GPU）についてベンチマークを行った結果、富岳プリポストノードの GPU（V100）と TSUBAME4.0 の GPU（H100）では、PyTorch（YOLOv5）で約 3 倍、行列積（fp32）では、約 10 倍（CPU 比では 50 倍以上）の性能比であることが確認できた。 以降、測定結果の詳細を述べることにする。

種別	システム名	内容
GPU	TSUBAME4-h100 [T4-h100]	東京科学大(旧東工大）のTSUBAME-4.0システムのGPU-h100（資源タイプ:gpu_1）
GPU	Fugaku-v100 [fg-v100]	理研富岳のprepostサーバのGPU-v100搭載サーバ
GPU	Merie-A100 [Mr-A100]	RIST所有のGPU-A100搭載サーバ（称:Merie)
CPU	Fugaku-Xeon [fg-Xeon]	理研富岳のprepostサーバでのXoen(gold6240)のみ使用
CPU	TSUBAME4-EPYC [T4-EPYC]	同TSUBAME-4.0資源タイプgpu_1でのEPYC(9654)のみ使用

図1　測定システム

グループ	BMT名	内容	備考
pytorch	Yolo.v5	物体認識:YOLO（You Look Only Onse)のpythonコードで、サンプルデータ(bus.jpg)を使用	https://www.tankoblucreate.com/post-299/
tensorflow	matmaul	tensorflow実装の行列積(fp32)性能測定カーネル	
	mnist	画像解析処理でデータセットとして、手書き文字、ファッション画像、CNN(cifar10)　の3種類で測定	

図2　測定ベンチマークアプリケーション

結果および考察

図3に YOLOv5 の実行時間のグラフを示す。左側 5 本が CPU（1c、4c などはコア数を示す）、右側 4 本が GPU を使用した場合の測定時間となる。本ベンチマークにおいては、小規模データのためか GPU の優位性は 1,2 倍にとどまった。

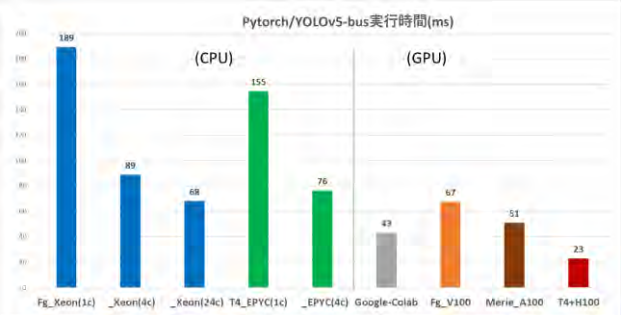


図3 PyTorch/YOLOv5-bus 実行時間（ms）（小規模データ bus）

図4に TensorFlow（fp32 行列積）の結果を示す。左側 4 本が CPU での測定値、右側 2 本が GPU での測定値となる。CPU での測定値においては、TF_INTRA_OP_PARALLELISM_THREADS の設定によるマルチスレッドの効果を確認することができた。また GPU の測定においては、CPU のマルチスレッドをしのぐ大幅な高速化の効果を得ることができ、V100 と H100 の比較でも約 10 倍の性能差となることが分かった。また CPU（EPYC48c）との比では約 50 倍の性能差となることが分かった。

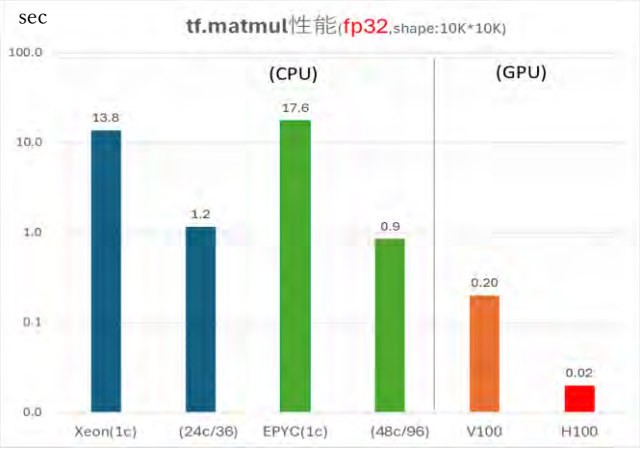


図4 Tensoeflow2.10 行列積（fp32）

図5に TensorFlow(mnist)のベンチマーク結果を示す。単純な手書き文字（mnist）やその拡張の Fashion 画像の識別では GPU の優位性は少なく（MNIST の V100 は CPU 比で逆転）、CNN（Convolutional Neural Network）での cifar10（物体カラー画像）において、ようやく H100 の優位性が確認できた。

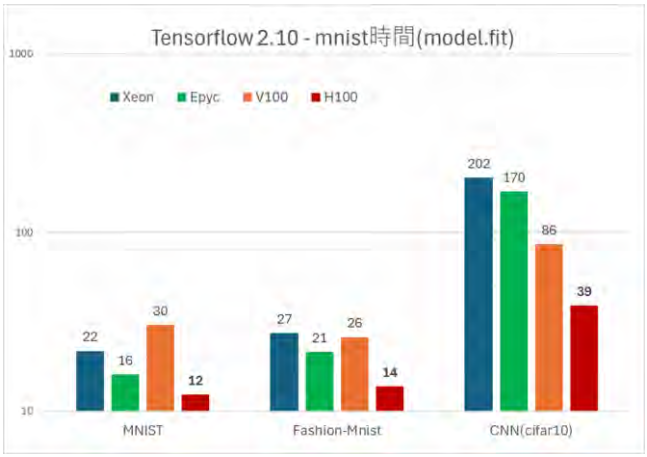


図5 Tensorflow 2.10 mnist

まとめ、今後の課題

TSUBAME-4.0 での GPU-H100 に関して、各種ベンチマーク（① PyTorch- 物体検出 YOLOv5, ② TensorFlow-matmul 行列積,③TensorFlow-画像認識 mnist 系 3 種）での性能を測定、他 GPU(V100,A100)や CPU(Xeon,EPYC)等との比較を実施した。

YOLOv5 では、小規模デモデータのためか、CPU 系と比較して GPU の優位性は僅かの結果であったが、その中でも H100 の高速性は、V100 比で 3 倍弱と際立っている。

H100 の tensor 演算の高速性の確認のため、② TensorFlow-matmul(fp32:10Kx10K)の測定を実施した。その結果、CPU（Xeon、EPYC）と比較して、50 倍以上の高速性を確認できた。同様に、③画像識別 mnist 系の比較でも、手書き文字や Fashin 画像の識別処理の性能では、データ規模（演算量）が足りないためか、GPU の有利性は少なく、物体カラー画像化での CNN データにおいて、漸く GPU 系が 4-5 倍程度の高速性を確認する結果となった。

以上の結果から、GPU の優位性に関しては、②の

行列積のようなものでないとカタログ性能がなかなか出ないが、しっかりと演算処理のあるアプリであればCPU に比べて 1 桁程度の優位性が確認できることが判明、その中でも H100 の高速性を再確認できた。

今後の課題としては、より実践的な問題での技術習得、性能測定を行っていきたいと考える。

以上

TSUBAME 共同利用　令和6年度　学術利用　成果報告書

利用課題名　大規模言語モデル（LLM：Large Language Model）を活用した医薬品等の有効性・安全性評価のためのアウトカム抽出の方法論の確立に向けた研究（24AC0401）

Research Toward Establishing a Methodology for Outcome Extraction for the Evaluation of the Efficacy and Safety of Pharmaceuticals Using Large Language Models

武藤 学
Manabu Muto

国立大学法人 京都大学大学院医学研究科腫瘍内科学講座　教授
Professor, Graduate School of Medicine and Faculty of Medicine, Kyoto University
https://www.med.kyoto-u.ac.jp/en/research/field/doctoral_course/r-040

医療分野での LLM 活用研究において、Llama3.3-Swallow-70B モデルを用いて千年カルテデータを対象に、構造化データ抽出の評価を実施した。量子化処理では 5〜6 ビットが精度と計算効率のバランスに最適であり、プロンプト設計が精度維持に重要であることが判明。英語モデルの日本語応用では、日本語継続学習により事象把握が曖昧化し、特に「年」の省略表現でタイムライン混乱が生じる課題を確認。これは日本語文法構造に起因する根本的問題と考えられる。現在、辞書構造・論理検証機能を持つ自動評価ツールを開発し、大規模データ対応基盤を構築中である。

This study evaluated structured data extraction from the Millennium Medical Record using the Llama3.3-Swallow-70B model. Results showed that 5–6-bit quantization optimally balances accuracy and efficiency, and prompt design is critical for maintaining performance. When applying English LLMs to Japanese, continued Japanese fine-tuning caused ambiguity in event recognition, especially due to omitted “year” expressions, highlighting a grammatical limitation in Japanese. An automated evaluation tool with dictionary and logic-checking functions is under development, alongside infrastructure for large-scale data processing.

Keywords:　Millennium Medical Record, LLM, quantization, Structuring of Narrative Texts, Clinical Outcome Information

背景と目的

治療効果の判定や有害事象に関わる情報は経過記録や報告書などの大量の非構造化テキストデータとして記録されているため、機械的に処理することが困難であり多くの人手を要している。近年、最先端の自然言語処理技術として大量のテキストデータを学習させた大規模言語モデル（LLM）が人間を上回る精度を示しつつある。

LLM の開発のためには大量のテキストデータと計算資源が必要となるが、我々は千年カルテプロジェクトで多施設から大量のテキストデータを含む電子カルテ情報を収集し、蓄積している。このテキストデータを利用して LLM を開発し、従来は手動でしか処理できなかった膨大なテキスト情報から、医薬品の安全性と有効性に関連する重要な知見を自動で抽出することが期待できる。これを実現することで、リアルタイムの医薬品監視、治療効果の迅速な評価、そしてリスク管理の精度向上に寄与し、最終的に医薬品開発と医療サービス提供の向上に資することを目指す。

概要

現時点利用可能な高性能なオープンソースの LLM（ベースモデル）を選定し、電子カルテの経過記録等（非構造化情報）を用いてベースモデルの構造化精度を検証する。ベースモデルとしては、本研究の分担研究者が開発する東京科学大学の Swallow-70B モデルを出発点とし、他のモデルとの性能比較等を行いながら以下の手順で研究を実施する。

1）構造化データ抽出機能の新たな評価方式の提案

2）LLM における構造化精度の量子化依存性の明確化

3）プロンプト表記の量子化依存性の明確化

4）高性能英語モデルの日本語化における課題の明確化

5）構造化精度検証の自動化

結果

・現在の法制度で医療向けの LLM を臨床データを使って研究開発し、実用化するための問題点について検討した。

・LLM の基本的な適用範囲と能力の評価を行った。

・この段階では、現在、千年カルテプロジェクトで収集されている臨床情報データベースの経過記録等から非構造化情報を収集し、初期の LLM モデルを構築してトレーニングを行なった。

・研究成果は以下の通りである。

1）構造化データ抽出機能の新たな評価方式の提案

実臨床データを基にした検証用ダミー資材を作成し、経過記録から抽出すべき臨床項目を特定した後、その正確性を測るものである。評価のポイントは以下の通りである：

a）日付の認識：西暦や日付の省略形を正確に年月日として認識できるか。

b）JSON 形式での抽出：項目キーと値の組み合わせを精緻に抽出する能力。

c）治療歴の認識：治療の開始日、終了日、治療ライン、薬剤名、投与量の認識度合い。

d）効果判定の認識：判定手段や判定内容の正確な認識。

e）誤字や Stage の間違いの指摘：誤字や Stage の間違いを指摘し、スコア化する。

f）多言語対応：英語・日本語問わず、意味が同じであれば正解とする。

2）LLM における構造化精度の量子化依存性の明確化

多くの分野で最も利用されている Meta 社の Llama モデルを対象に、量子化の依存性を明らかにした。電子カルテの構造化目的では、Meta-Llama-3-70B-Instruct において 5～6 ビットの量子化が妥当な選択となることを明らかにした。量子化ビット数（Q2～Q8）と構造化精度の関係を図 1 に示す。図の横軸は、必要とされる GPU メモリ量である。

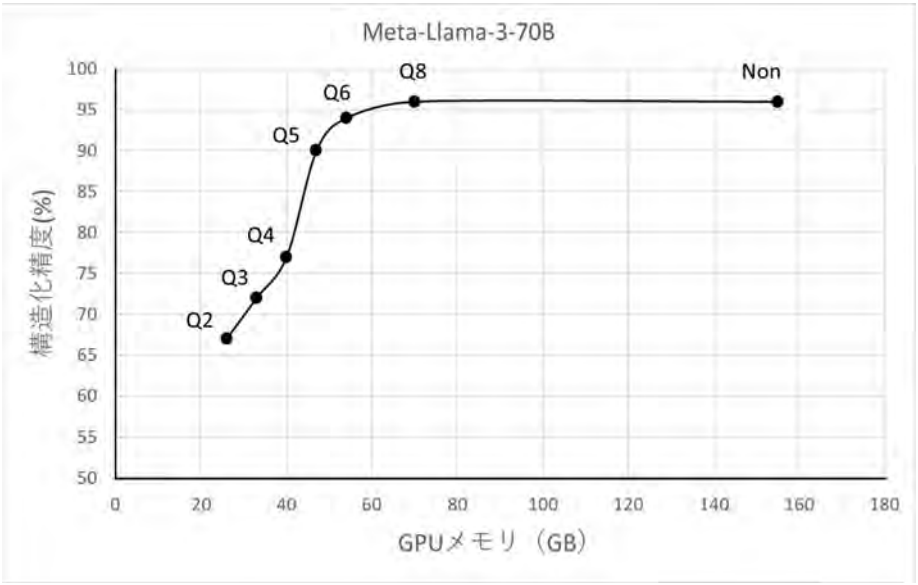
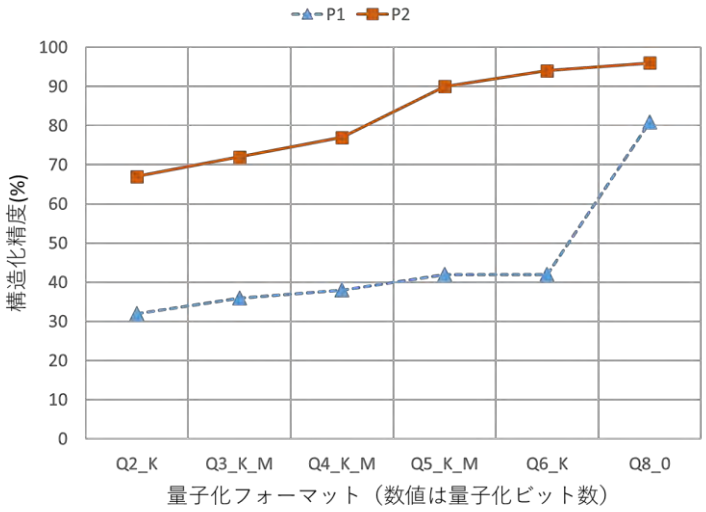


図 1. 必要とする GPU メモリと構造化精度の関係（Q2～Q8：量子化ビット数）

3）プロンプト表記の量子化依存性の明確化

プロンプト表記の違いによる構造化精度の量子化依存性を明らかにした。5 ビット量子化の領域では、プロンプトの詳細な記述が求められることも明らかにした。



【簡易プロンプト例 P1】:

以下のテキストに書かれた事柄を時系列に JSON 形式で構造化し、可能な限り詳細に抽出してください。

【詳細プロンプト例 P2】:

以下のテキストに書かれた事柄を時系列に詳細に全ての事象を JSON 形式で構造化し、改行を加えて見やすく表示してください。なお症状や検査検体、診断名、TNM 分類、ステージ、転移部位、遺伝子変異などについても詳細に抽出してください。

図 2. プロンプトの違いによる構造化精度の比較（P1:簡易指定、P2:詳細指定）

4）高性能英語モデルの日本語化における課題の明確化

日本語ベンチマークで高いスコアを出している Llama-3-Swallow-70B-Instruct、さらにこれを医療領域にファインチューニングした Llama3-Preferred-MedSwallow-70B について、構造化精度の量子化依存性を詳細に検証した（図 3 及び図 4）。結果は、英語で高度にトレーニングされた LLM に対して日本語での継続学習を行うと、英語による構造化処理の性能に比べて、日本語の場合、その性能が低下することが明らかとなった。

これに対処するため英語環境の能力を維持しつつ、日本語指示に従う能力を獲得するようなトレーニング法の研究を行っている。

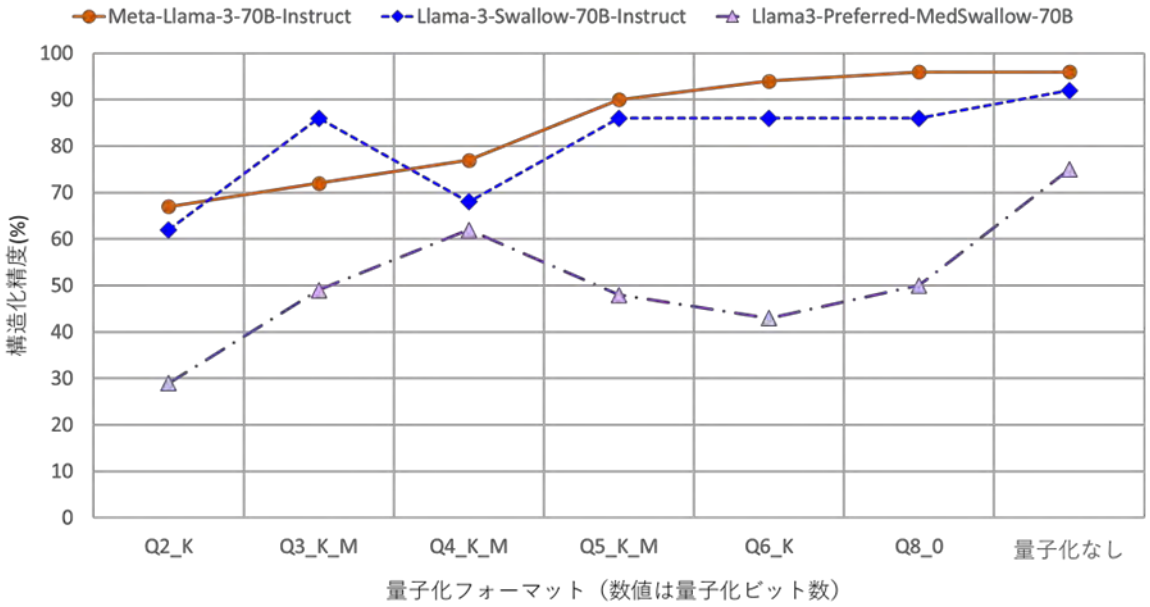


図3．日本語継続事前学習モデルと構造化精度の関係

事象項番	経過記録	臨床項目	項目数	Meta-Llama-3-70B							Llama-3-Swallow-70B							Llama3-Preferred-MedSwallow-70B						
				Q2	Q3	Q4	Q5	Q6	Q8	なし	Q2	Q3	Q4	Q5	Q6	Q8	なし	Q2	Q3	Q4	Q5	Q6	Q8	なし
1	2019年12月初旬に胸痛、背部痛を自覚。近医を受信し左肺門部腫瘤を指摘され	診療日、症状	2	2	2	2	2	2	2	2	1	2	2	2	2	2	2	2	0	0	0	2	0	0
2	1月7日に当院を紹介受診。 同日に胸腔穿刺を施行し、得られた胸水から肺腺癌と診断 T4N1M1c StageⅣC, BRA, PUL, LYM, PLE, OSS, EGFR(L858R+), KRAS-, BRAF(V600E)-, ROS1-, PDL-1<0%, ALK-	診療日	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0
		検体キー、検体名	2	0	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		診断キー、診断名	2	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	0	2	2
		TNMキー、TNM分類	2	1	2	2	2	2	2	2	1	2	2	2	2	2	2	1	0	0	0	1	0	2
		Stageキー、Stage	2	0	0	2	2	2	2	2	0	1	2	2	2	2	2	0	0	0	0	1	0	2
		指摘キー、誤字やStage間違いの指摘	3	0	0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	0
		転移キー、転移値（5つ）	6	6	0	6	6	6	6	6	0	4	6	6	0	0	6	0	4	3	1	5	4	6
		バイオマーカーキー（6つ）	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6
		バイオマーカー値（6つ）	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6
3	1/31～：C1-7頸椎転移に対し緩和的放射線治療(30Gy/10fr)施行	開始日、治療キー、治療内容、部位キー、部位名、線量キー、線量値	7	4	7	5	7	5	7	7	0	5	5	5	7	7	7	2	0	7	4	0	0	7
4	2/6～：1stline Osimertinib 80mg/dayを開始（28日分）。	開始日、ラインキー、治療ライン、治療キー、治療タイプ、薬剤キー、薬剤名、投与キー、投与量、期間キー、投与期間	11	7	9	7	10	11	11	11	10	11	9	10	10	11	11	3	8	9	10	6	9	9
5	11/4：CTにて左肺門部腫瘤、左肺尖結節、肝転移、骨転移増大を認めPDと判断。	診療日、判定手段、検査内容、判定内容	4	1	0	3	4	4	4	4	3	3	3	3	4	4	4	0	0	2	2	3	0	4
6	11/19～2/10：2nd line CBDCA/ PTX/ Bev/ Atezoを4course施行	開始日、終了日、ラインキー、治療ライン、治療キー、治療タイプ、薬剤キー、薬剤名、投与キー、投与量	10	6	7	6	8	10	10	10	8	9	7	9	9	10	10	2	8	8	9	2	8	8
7	2021/3/2：効果判定にて原疾患の増悪を認めPDと判断。	診療日、判定手段、検査内容、判定内容	4	3	4	4	4	4	4	4	3	4	3	3	4	4	4	2	0	2	2	3	0	4
8	3/16～4/13：3rd line DOC+RAM 2course施行	開始日、終了日、ラインキー、治療ライン、治療キー、治療タイプ、薬剤キー、薬剤名、投与キー、投与量	10	7	7	6	8	10	10	10	8	9	7	9	10	10	10	2	8	8	9	4	8	8
9	5/18：効果判定にて両肺の小結節は増加・増大と両測胸水の増加がありPDと判断	診療日、判定手段、検査内容、判定内容	4	3	4	4	4	4	4	4	3	3	3	3	4	4	4	2	0	2	2	3	0	4
10	5/20：胸水コントロール目的に入院。	診療日、治療キー、治療内容	3	2	3	3	3	3	3	3	3	3	0	3	3	0	0	0	0	0	0	0	0	0
11	5/21：左癌性に対して胸水左胸腔ドレーン挿入	診療日、治療キー、治療内容	3	2	3	3	3	3	3	3	3	3	0	3	3	3	3	0	0	0	0	0	0	0
12	5/24：左胸膜癌着術(ユニタルク4g)を施行	診療日、治療キー、治療内容	3	2	3	3	3	3	3	3	3	3	0	3	3	3	3	0	0	0	0	0	0	0
13	5/27～：4th line EGFR-TKI rechallenge (Afatinib 20mg/day)開始	開始日、ラインキー、治療ライン、治療キー、治療タイプ、薬剤キー、薬剤名、投与キー、投与量	9	7	7	5	8	9	9	9	8	9	6	8	8	9	9	4	7	7	7	7	7	7
正解項目数の合計（＝構造化精度%）			100	67	72	77	90	94	96	96	69	86	70	86	86	86	92	35	49	62	60	49	50	75

図4．日本語継続追加学習モデルと構造化精度の詳細

5）構造化精度検証の自動化

上記1）の評価方式を用いて LLM の構造化精度を判定するのに人手で行って来たが、これをプログラムにより自動判定するアプリケーションの開発に着手している。開発は、以下のステップで進める予定であり、①のステップを完了している。

①ステップ 1：上記1）の方式に特化した辞書構造と論理チェックをプログラムで実現

②ステップ 2：LLM の構造化を一定の形式に固定化する方式の実現

③ステップ 3:今後学習予定の大規模な LDI データに柔軟に対応できる 辞書構造と論理チェックを可能にする方式の実現

考察および今後の課題

本研究の初期段階では、現行の法制度下において医療分野での LLM 開発および活用に伴う課題と技術的可能性を整理した。Llama3.3-Swallow-70B モデルを用い、千年カルテ由来の経過記録から初期学習モデルを構築。構造化データ抽出において、日時や治療歴、判定、誤記、Stage 分類、多言語対応といった項目を対象に新たな評価指標を提案し、実証を進めた。

Meta 社の Llama モデルをベースに量子化処理を施した結果、5〜6 ビットが精度と計算資源のバランスに優れていた。また、プロンプト表記との組み合わせ次第で量子化後の精度が変動するため、詳細な設計が精度維持に不可欠であることも明らかになった。

英語モデルの日本語応用に関しては、Llama3.3-Swallow-70B および Llama3-Preferred-MedSwallow-70B の検証結果から、日本語による継続学習は事象把握の曖昧化を招く傾向があり、とくに「年」の省略表現によってタイムラインの混乱が起こることが示された。これは Meta 社の多言語モデルにも共通する現象で、日本語の文法構造に起因する根本的な課題である可能性がある。

構造化精度検証の自動化では、辞書構造・論理検証機能を備えた評価ツールの開発を進めており、LDI の大規模データにも対応できる基盤を構築中である。今後は医療用 LLM の実運用を見据え、検証自動化の高度化が不可欠となる。

まとめ

英語で事前学習された LLM を日本語に適用する場合、精度の低下が見られるが、これは日本語構文の特性や非明示的な時系列表現が要因である。これに対して、本研究では新たなファインチューニング手法を導入し、日本語指示への応答精度を大幅に向上させることに成功した。

その手法は、LMSYS-Chat-1M の対話履歴を翻訳し、Llama 3.1 405B Instruct を用いて日本語の応答文を自動生成するもの。続いて、Llama 3.1 70B モデルによるスコアリング評価により最良の応答を選別する工程を組み込んだ。加えて、重複や冗長性のある指示文・応答文をフィルタリングし、学習データ全体の品質を高めた。

この一連の工程により、Llama3.3-Swallow-70B においても日本語に特化したファインチューニングが可能であることが示され、医療領域での日本語 LLM 実装に向けた重要な成果を得た。

今後は、LLM の日本語継続学習時の性能劣化を抑制しつつ、多言語モデルの相互運用性を確保する設計原則の確立が求められる。本研究の成果は、日本語 LLM の基盤技術として国内外の医療 AI 研究にも貢献しうる。

TSUBAME 共同利用　令和6年度　学術利用　成果報告書

利用課題名　大規模マルチモーダルモデルの研究開発および応用
英文：Research, Development, and Application of Large-Scale Multimodal Models

利用課題責任者
栗田修平
First name Surname
Shuhei　Kurita

所属
国立情報学研究所
Affiliation
National Institute of Informatics
URL
https://www.nii.ac.jp/

邦文抄録（300 字程度）

本研究課題では、実世界にて言語指示を理解し環境に応じて動作を生成する、実世界での基盤モデルを目指して研究を行った。具体的には、一人称視点動画を題材にし、言語・視覚を入力として動作系列を抽出およびそれを予測する基盤モデルの作成をおこなった。言語（テキスト）から、あらかじめ抽出された動作系列を、事前学習済みモデルの 3D 言語モデル（3D-LLM）を利用して学習および生成させることで、擬似的な動作生成が可能になった。学習は計算環境 TSUBAME を利用して行われた。得られた成果は、画像系のトップカンファレンスである CVPR2025 にて発表を行った。得られたモデル及びデータセットは、将来的なロボット基盤モデルの作成にも応用できる成果であると考えられる。

英文抄録（100 words 程度）

This research project aimed to develop a foundational model for the real world that understands language instructions and generates actions based on the environment. Specifically, using first-person perspective videos as subject matter, we created a foundational model that extracts and predicts action sequences using language and visual input. By training and generating pre-extracted action sequences from language (text) using a pre-trained 3D language model (3D-LLM), we achieved pseudo-action generation. Training was conducted using the TSUBAME computing environment. The results were presented at CVPR 2025, a top conference in the field of computer vision. The obtained model and dataset can be applied to the creation of future robot base models.

Keywords: Vision and language model, Large Language Model

背景と目的

（課題の背景を記載してください）

ロボット基盤モデル作成にあたって一番の障害となるのは、データ不足の問題である。人手で作られたテレオペデータは、非常に高品質な動作データであるが、公開されているものは非常に少ない。一方で、比較的簡単に収集できる一人称視点動画から動作データを作成できれば、ロボット基盤モデルを作成するうえで強力な手法となりえる。

（現状の問題点等を挙げてください）

本プロジェクトでは、視覚情報（環境情報）・言語指示情報と動作情報を統合して扱うためのデータセットを作成し、作成したデータセットを利用して 3D-LLM をベースとする動作生成モデルを作成する。ロボット基盤モデルのデータ欠乏問題を、一人称視点動画からのデータ抽出によって解決し、既存のロボット基盤モデル学習用データよりも多様な視覚・動作・言語の対応付けデータの作成、および、3D 言語モデル（3D-LLM）を利用して学習および生成させることで、擬似的な動作生成モデルを作成した。得られた成果は、画像系のトップカンファレンスである CVPR2025 にて発表を行った。得られたモデル及びデータセットは、将来的なロボット基盤モデルの作成にも応用できる成果であると考えられる。

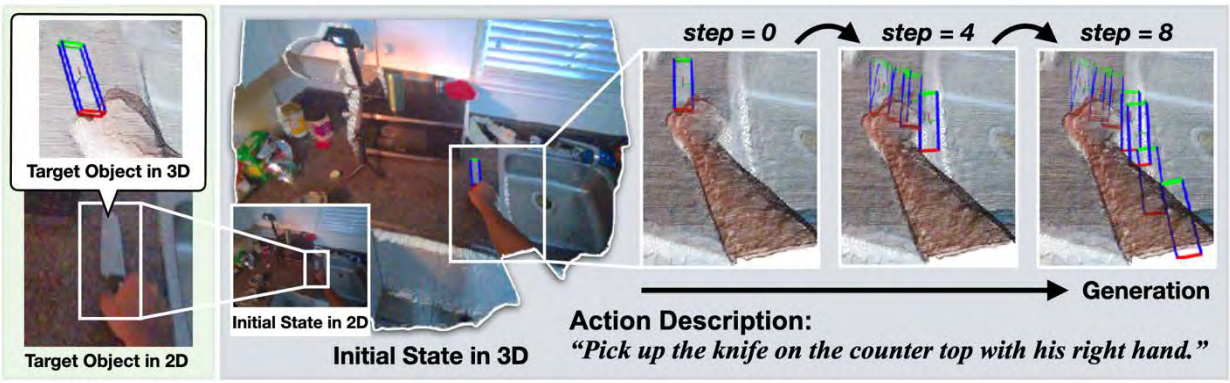


図 1. 今回の研究にて整備した時系列動作データセットの概要.

概要

（申請書の概要を基本とし、実情に合わせて変更してください。図や表の利用を図って分かり易く記載して下さい。）。

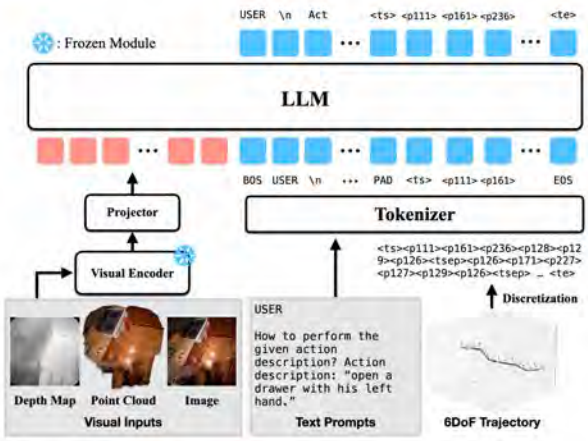


図. 2　3D-LLM を利用した提案手法(PointLLM)

生成したデータセットの概要を図 1 に示す。一人称視点動画が与えられた際に、3D 再構成を行ったうえで、物体の姿勢を 3D bounding box の形式で捉え、その位置および姿勢の変化を時系列的に捉える。本研究では、さらに 3D-LLM を利用した動作生成モデルを作成し、視覚情報に加えて、3D 的な環境情報が有効であることを示した。図 2 のように 3D-LLM を利用して学習を行い、提案したデータセットを利用することで有効な動作軌跡が生成できるか検証した。

結果および考察

（結果と考察を記載してください。図や表の利用を図って分かり易く記載して下さい。）。

表 1 に実験結果を示す。物体位置のみの予測(3DoF)と角度も含めた予測（6DoF）の双方にて、提案手法が優れていることを確認した。さらに、定性的な評価によっても、提案手法、ひいては作成したデータセットが優れていることを確認した。これは、今回の研究で作成したデータが優れていることに加えて、ロボット基盤モデルへの応用可能性を示すものと思われる。

表 1: 3D-LLM(PointLLM)を利用した実験結果

		3DoF Training				*6DoF Training*				
		3D pos.		*2D pos.*		*3D pos.*		*2D pos.*		*3D rot.*
Model	Input	ADE	FDE	ADE	FDE	ADE	FDE	ADE	FDE	GD
Seq2Seq [89]	Image	**0.269**	0.475	0.215	0.372	0.374	0.625	0.325	0.528	0.559
BLIP-2 (2.7B) [50]	Image	0.278	0.464	0.218	0.346	0.280	0.465	0.218	0.344	0.545
BLIP-2 (6.7B) [50]	Image	0.286	0.487	0.224	0.365	0.286	0.475	0.219	0.349	0.543
VILA (3B) [54]	Image	0.500	0.662	0.428	0.546	0.477	0.619	0.390	0.489	0.826
VILA (8B) [54]	Image	0.316	0.512	0.249	0.388	0.293	0.478	0.225	0.347	0.545
Seq2Seq [89]	Image + Depth	0.302	0.541	0.250	0.438	0.341	0.558	0.280	0.452	0.590
BLIP-2 (2.7B) [50]	Image + Depth	0.288	0.481	0.227	0.363	0.275	**0.458**	0.211	0.332	0.543
BLIP-2 (6.7B) [50]	Image + Depth	0.283	0.477	0.218	0.351	0.282	0.469	0.219	0.345	0.542
VILA (3B) [54]	Image + Depth	0.301	0.492	0.229	0.355	0.294	0.480	0.223	0.344	**0.541**
VILA (8B) [54]	Image + Depth	0.298	0.494	0.234	0.368	0.318	0.513	0.253	0.391	0.563
MiniGPT-3D (2.7B) [85]	Point cloud	0.299	0.487	0.236	0.368	0.281	0.467	0.218	0.342	0.544
PointLLM (7B) [97]	Point cloud	0.274	**0.459**	**0.210**	**0.328**	**0.271**	**0.458**	**0.208**	**0.327**	**0.541**

まとめ、今後の課題
（まとめと今後の課題について記載してください。）。

一人称視点の動画を題材に、言語と視覚を入力として動作系列を抽出・予測する基盤モデルを構築した。抽出された動作系列をテキストに基づき、事前学習済みの 3D 言語モデル（3D-LLM）を活用して学習・生成させることで、擬似的な動作生成を実現した。学習は計算環境 TSUBAME 上で行い、その成果は画像分野のトップ国際会議 CVPR2025 にて発表された。本研究で得られたモデルとデータセットは、今後のロボット基盤モデルの開発にも応用可能であると考えられる。特に、今回構築したデータセットの質の高さに加え、ロボット基盤モデルへの発展可能性を示す成果となっている。この成果のロボット実験への応用可能性について、追加の検討を進めている。

TSUBAME 共同利用　令和 7 年度　学術利用　成果報告書

利用課題名　データの倫理問題を根本解決するテキストからの画像生成モデル構築
英文：Image generation model to solves the ethical problem of data

利用課題責任者　柳凜太郎
First name Surname　Rintaro Yanagi

所属　産業技術総合研究所
Affiliation AIST
URL https://www.aist.go.jp/

邦文抄録（300 字程度）
本研究では、(1)数式など生成規則から自動生成された擬似画像・テキストによる一段階目の事前学習、(2)権利関係トレース可能な教師ラベル付実画像のみを用いた二段階目の継続事前学習により、データの倫理問題を根本解決するテキストからの画像生成モデルを構築する。昨今では WEB をソースとして収集された画像・テキストデータが倫理面で害悪なデータを含み、倫理審査が入り、データのみならず学習済みモデルが公開停止の危機に追い込まれているが、本研究では AI 分野における倫理問題の根本解決を試みる。

英文抄録（100 words 程度）
In this study, we aim to build a text-to-image generation model that fundamentally addresses the ethical issues of training data through a two-stage pre-training approach: (1) pre-training using pseudo-images and text automatically generated from mathematical formulas and other generative rules, and (2) continued pre-training in the second stage using only real images with teacher labels whose rights and provenance can be fully traced. Recently, image and text data collected from the web have often contained ethically harmful content, leading to ethical reviews and, in some cases, the suspension of not only the datasets but also the release of trained models themselves. This study attempts to provide a fundamental solution to these ethical challenges in the AI field.

Keywords: 5つ程度
Generative model, formular driven, ethical challenges, pre-training

背景と目的

WEB を源泉として収集された既存の画像・テキストデータは倫理面で問題を孕んでいる

万が一上記のデータセットの使用が停止になった場合、CLIP/StableDiffusion などの最先端のモデルも同時に使用不可になる可能性が存在する

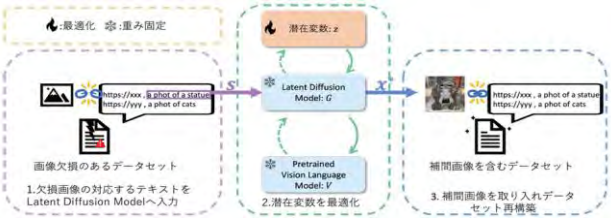
本プロジェクトでは、倫理的問題や権利関係に可能な限り配慮した視覚言語モデルの構築に取り組み，視覚言語モデルの精度向上に寄与することを確認した。

概要

(1)数式など生成規則から自動生成された擬似画像・テキストによる一段階目の事前学習について取り組んだのちに，(2)権利関係トレース可能な教師ラベル付実画像のみを用いた二段階目の継続事前学習に着目して研究開発を行った

結果および考察

- 既存の生成モデルを視覚言語モデルの学習用に最適化することで，視覚言語モデルの精度を大幅に向上可能であり，画像データの欠損のあるデータセットを保管することも可能である点
- 権利関係トレース可能な実画像を有する企業と連携することで，権利関係がクリアな視覚言語モデルの構築に取り組んだ点



まとめ、今後の課題

引き続き，企業と連携することで，モデルの一般公開を目指すとともに，既存のモデルと同等の精度達成を目指す．

TSUBAME 共同利用　令和 6 年度　学術利用　成果報告書

利用課題名　医薬プロセス最適化と新薬開発を効率化・加速する製剤処方設計 AI の開発

英文：Development of AI for drug formulation design to streamline and accelerate pharmaceutical process optimization and new drug development

利用課題責任者
Yasushi Okuno

所属
Department of Biomedical Data Intelligence,
Graduate School of Medicine, Kyoto University
http://clinfo.med.kyoto-u.ac.jp/

邦文抄録（300 字程度）

本研究では、マルチモーダルな生物医学データを統合し、AI および機械学習を活用した創薬支援モデルの構築を試みた。TSUBAME を活用し、ESM-2 によるタンパク質配列解析、Geneformer を用いた遺伝子摂動シミュレーション、グラフ対照学習による遺伝子制御ネットワーク解析、大規模トキシコゲノミクスデータを用いたマルチモーダルモデルによるラット毒性病理所見予測、さらに CT/MRI 画像を用いた医療画像解析を実施した。これらの取り組みにより、複数のモダリティを統合した解析手法の有用性が示唆され、今後の創薬予測モデルの基盤となり得る知見が得られた。

英文抄録（100 words 程度）

This study explored the construction of drug discovery support models by integrating multimodal biomedical data using AI and machine learning. Leveraging TSUBAME, we conducted protein sequence analysis with ESM-2, gene perturbation simulation with Geneformer, graph contrastive learning for regulatory networks, multimodal model construction to predict rat toxicopathological findings from toxicogenomics data, and medical image analysis using CT/MRI. The results suggest the potential utility of multimodal data integration in predictive modeling and provide a basis for future development of drug discovery frameworks.

Keywords: 5つ程度
Drug Discovery,
Graph Neural Network,
Medical Imaging,
Multimodal Data,
Protein Language Model

背景と目的

創薬研究においては、タンパク質の構造情報や化学的性質、生物学的作用機序、さらには臨床データといった多様なモダリティのデータを統合的に解析する必要がある。しかし、従来のシングルモダリティベースの解析手法では、新規治療薬の探索に限界があり、予測精度や解釈性の面でも課題が残る。

本プロジェクトでは、マルチモーダルな生物医学データを統合し、AI・機械学習を用いた高精度な創薬予測モデルの構築を目的とした。TSUBAME を用いた大規模計算により、従来手法では実現困難であった大規模かつ高次元のデータ統合と解析が可能となり、精度の高い創薬予測モデルの基盤となり得る成果を得た。

概要

本課題では、以下の 4 つの研究を実施した。

①ESM-2 によるシーケンス特徴量解析

タンパク質・ペプチド複合体の構造適合性を対象に、言語モデル ESM-2 を用いたシーケンス特徴量の抽出と評価を実施。従来手法と比較して分解能およびモデリング精度の向上が見られた。

②Geneformer を用いた遺伝子摂動シミュレーション

トランスクリプトームデータを対象に、ファインチューニング済み Geneformer による遺伝子発現摂動への応答予測を行い、重要遺伝子および疾患関連細胞種の同定を試みた。

③グラフ対照学習による遺伝子制御ネットワーク解析

TCGA および LINCS データをもとに、乳がん患者および細胞株のグラフ構造と臨床データを統合し、Graph Neural Network の Pretraining・Finetuning を実施。6 手法の比較による有効性評価も行った。

④マルチモーダルによる薬剤障害病理所見予測

大規模トキシコゲノミクスデータを活用し、遺伝子発現データと病理組織画像からラット毒性病理所見を予測するためのマルチモーダルモデルを構築した。

⑤医療画像解析

脳神経外科・整形外科・消化器外科領域における CT/MRI 画像を対象に、病変セグメンテーション、疾患予測モデルの構築、GAN による画像生成などを行った。

結果および考察

ESM-2 を用いた解析では、言語モデルによるシーケンス表現の分解能が高く、タンパク質複合体の構造適合性のモデリング精度が向上した。

Geneformer による遺伝子摂動解析では、潜在特徴量の変化から疾患関連の重要遺伝子や細胞種の推定が可能であった。

グラフ対照学習においては、提案手法が従来手法に比べて、臨床データを用いた下流タスクにおいて高い予測性能を示した。特にサブタイプ分類や生存時間予測において、事前学習が効果的であることが確認された。

大規模トキシコゲノミクスデータを活用したマルチモーダルモデルによるラット毒性病理所見予測については、マルチモーダルモデルが既存の単一モーダルモデルと比較して高い予測性能を示したことから、新たな薬剤性臓器障害予測のアプローチの可能性を示した。

医療画像分野では、CT/MRI を用いた病変検出の精度向上に加え、GAN による画像生成が可能であった。

まとめ、今後の課題

本研究により、マルチモーダルデータを統合した創薬支援モデルの構築可能性が示された。とりわけ、言語モデルやグラフニューラルネットワークの応用により、従来手法を凌駕する予測精度を得ることができた。

今後の課題としては、より多様なデータセットへの対応、臨床現場への応用を見据えたモデルの汎化性能の向上などが挙げられる。TSUBAME の計算資源を活用することで、これらの課題解決に向けたさらなる高精度化と高速化を目指す。

TSUBAME 共同利用　令和 6 年度　学術利用　成果報告書

利用課題名　GeoVLM：マルチビュー協調型視覚言語モデルによる包括的な都市理解に向けて
英文：GeoVLM: Towards comprehensive urban understanding via multi-view coordinated vision-language model

利用課題責任者　横矢直人
First name Surname　Naoto Yokoya

所属　東京大学大学院新領域創成科学研究科
Affiliation　Graduate School of Frontier Sciences, The University of Tokyo

邦文抄録（300 字程度）

地球観測システムの急速な発展により、リモートセンシング画像、地点情報、地域気候区分、土壌タイプなど、膨大な地理空間データが日々取得されている。これらの異種データから有用なセマンティック情報を効果的に抽出・統合するためには、国連が掲げる持続可能な開発目標およびデジタルシティの理念に沿った、環境モニタリングと都市管理を協調的に行うための大規模なマルチモーダルモデルの開発が求められている。本研究では、日本全国を対象地域とし、スーパーコンピューティング資源を活用して、サブメートル精度の土地被覆マッピング、都市計画支援、農業管理の自動化を実現する大規模マルチモーダルモデルを開発する。

英文抄録（100 words 程度）

With the rapid development of earth observation systems, massive amounts of geographic data (remote sensing imagery, points of interest, local climate zones, soil types, etc.) are being acquired daily. To effectively extract and comprise useful semantic knowledge from the heterogeneous data, we need to develop a large multi-modal model for collaborative environmental monitoring and urban management, meeting the United Nations' sustainable development and digital city goals. Regarding the whole of Japan as the starting point, we hereby apply for super-computing resources to develop a large multi-modal model, automatically achieving sub-meter land-cover mapping, city planning guidance, and agriculture management.

Keywords: sustainable urban development, vision-language model, remote sensing

背景と目的

都市計画者は、ENVI、ArcGIS、SPSS といった地理空間および統計分析ツールに多くを依存し、マルチモーダルな地球科学データを分析して報告書や都市計画を作成している。しかしビッグデータ時代において、これらの作業には高度な専門知識が求められる上、膨大な時間も要するようになり、ますます負担が増している。こうした状況を背景に、膨大な地球科学データを効率的に処理し、有意義な情報を抽出して実用的な地球科学知識へと導く、インテリジェントな地球科学システムの整備が急務となっている。とはいえ、このような多次元かつ大規模な地質データのインテリジェントな処理・解析には、いくつかの重要な課題が存在する。

①リモートセンシング画像、ストリートビュー、テキスト情報といった異なる種類のデータから、どのように統一された特徴表現を得ることができるのか？

②地球科学的な問いに対する回答や、都市計画における意思決定を、どのようにして推論できるのか？

本プロジェクトでは、問題①に対して、マルチモーダルな地球科学データを統一的に表現し、各種機能を一貫したセマンティクスに変換するためのコントラスト学習に基づく自己教師あり学習戦略を設計した。問題②に対しては、大規模言語モデルの高い推論能力を活用し、オブジェクト間の空間的関係や意味的分析を可能にするマルチモーダル視覚言語モデルを構築した。これにより、都市の持続可能な発展に関するレポートを自動生成する仕組みを実現している。

概要

本研究では、リモートセンシング（RS）画像と街路景観（SV）画像に由来するマクロおよびミクロレベルのセマンティクスを統合することで、都市の持続可能性に関する課題に対処するビジョン言語モデル（VLM）の機能拡張を目指した。まず、20,589 枚の RS 画像、110 万枚の SV 画像、80 万件の質問応答ペアから構成される国家規模の地理空間ビジョン言語データセット GeoVLSet を構築した。このデータセットは、地理空間オブジェクトの推論、社会的要素の分析、都市経済の評価、持続可能な開発に関するレポートの自動生成を促進するものである。さらに、RS 画像と SV 画像が持つ補完的なセマンティクスを効果的に統合し、多層的な都市理解を実現するマルチイメージ対応の視覚言語モデル GeoVLM を開発した。提案手法により、エンドツーエンドの指示調整を通じて、直接的かつ堅牢な都市分析が可能であることを示した。本研究の成果は、リモートセンシングにおける VLM 研究に新たな視点を提供するとともに、将来の都市分析システムの高度化・効率化に貢献するものである。

結果および考察

本研究で提案する視覚モデルは、日本全国の複数の都市において、既存のあらゆる手法を上回る品質の評価結果を示した。すべてのシーンは、商業、工業、住宅、未開発または半開発の 4 つのクラスに分類している。既存の基本情報に基づき、すべてのリモートセンシング画像に対して適切な都市レポートを作成できるよう、アノテーターには十分なトレーニングを行った。図 1 には、異なるシーンの特徴を反映した 3 都市の持続可能な開発レポートの例を示している。

まとめ、今後の課題

都市分析の要件に基づき、20,589 枚のリモートセンシング（RS）画像、110 万枚のストリートビュー（SV）画像、80 万件の質問応答ペアを含む、全国規模の地理空間視覚言語データセット GeoVLSet を提案した。本データセットは、地理空間オブジェクトの推論、社会的キ—オブジェクトの分析、都市経済の評価、持続可能な開発レポートの生成といったマルチレベルのタスクに対応し、大規模視覚言語モデル（VLM）の高次推論能力を拡張するものである。

図 1. 多様なリモートセンシングシーンにおける都市持続可能性レポートの代表例

RS 画像と SV 画像の統合により、補完的な意味情報の取得が可能となる一方で、特に RS-SV 間の関係性や長文コンテキストのモデリングにおいては、依然として重要な課題が存在する。これらの課題に対処するため、本研究では、地理空間的な位置埋め込みと長文コンテキスト処理に対応した大規模言語モデル（LLM）を組み込んだ、堅牢な視覚言語ベースライン GeoVLM を提案した。

広範な実験を通じて、GeoVLM の優れた性能と高い堅牢性が確認された。GeoVLSet および GeoVLM は、Earth Vision 分野におけるドメイン特化型 VLM の開発に資するベンチマークとなることが期待される。今後は、興味地点（POI）などのソーシャルメディアデータの統合や、都市災害といったさらなる応用シナリオの拡充を予定している。

TSUBAME 共同利用　令和 6 年度　学術利用　成果報告書

利用課題名　感染症制御に資するバイオインフォマティクス研究
英文：Bioinformatics Research for the Effective Control of Infectious Diseases.

伊東　潤平
Jumpei Ito

東京大学医科学研究所
The Institute of Medical Science, The University of Tokyo
https://www.ims.u-tokyo.ac.jp/SystemsVirology/

邦文抄録（300 字程度）
ウイルスは著しい速度で進化することで抗原性を変化させ、宿主の獲得免疫から逃避する。そのため、効果的なワクチンを供給するためには、ウイルスの抗原性変化に基づき、最新の流行株に対して有効なワクチン株を選定する必要がある。ウイルスの抗原性は中和試験等により評価できるが、理論的には抗原タンパク質の配列のみから予測可能であると考えられる。本研究では、ウイルスの配列情報に基づいた抗原性監視体制と、より効率的なワクチン株選定戦略の樹立に向けて、季節性インフルエンザウイルスを対象に抗原性予測 AI モデル「PLANT」を開発した。

英文抄録（100 words 程度）
Viruses evolve at a remarkable rate, altering their antigenicity to evade the host's acquired immunity. Therefore, to supply effective vaccines, it is necessary to select vaccine strains that are effective against the latest circulating strains based on antigenic changes in the virus. The antigenicity of a virus can be evaluated through neutralization assays and other methods, but theoretically, it should be possible to predict antigenicity solely from the sequence of the antigenic protein. In this study, we developed the antigenicity prediction AI model **"PLANT"** for seasonal influenza viruses, aiming to establish an antigenicity surveillance system based on viral sequence information and a more efficient vaccine strain selection strategy.

Keywords: 5つ程度
AI, Virus, Vaccine development, Infectious Diseases

背景と目的

ウイルスは著しい速度で進化することで抗原性を変化させ、宿主の獲得免疫から逃避する。そのため、効果的なワクチンを供給するためには、ウイルスの抗原性変化に基づき、最新の流行株に対して有効なワクチン株を選定する必要がある。ウイルスの抗原性は中和試験等により評価できるが、理論的には抗原タンパク質の配列のみから予測可能であると考えられる。

概要

本研究では、ウイルスの配列情報に基づいた抗原性監視体制と、より効率的なワクチン株選定戦略の樹立に向けて、季節性インフルエンザウイルスを対象に抗原性予測 AI モデル「PLANT」を開発した。

結果および考察

本モデルの学習には、大規模ゲノムデータベース GISAID に登録されている季節性インフルエンザウイルス H3N2 亜型のヘマグルチニン（HA）タンパク質配列と、英国クリック研究所 Worldwide Influenza Centre lab より提供を受けた中和試験のデータを用いた。PLANT の予測性能を評価するため、学習に用いていない新規変異株を用いて抗原性の予測を行ったところ、中和試験により実測された値と高い一致性を示した。これらの結果は、PLANT が配列情報のみに基づき、未知の変異株の抗原性を高い精度で予測できることを示している。

まとめ、今後の課題

現在、PLANT を用いたウイルス進化解析を進めている。本研究成果をまとめたプレプリントを、2025 年上旬中にプレプリントとして公開予定である。

TSUBAME 共同利用　令和 6 年度　学術利用　成果報告書

深層学習に基づく小惑星上のボルダー検知技術の研究開発
Development of a Deep Learning-Based Technology for Boulder Detection on Asteroids

本田親寿 [1], 許 埈豪 [2], 富 宣超 [2], 神山徹 [3], 吉川一朗 [2]
Chikatoshi Honda[1], Hur Junho[2], Fu Xuanchao[2], Toru Kouyama[3], and Ichiro Yoshikawa[2]

[1] 公立大学法人会津大学, [2] 東京大学, [3] 産業技術総合研究所
[1]The University of Aizu, [2]The University of Tokyo, [3]National Institute of Advanced Industrial Science and Technology

邦文抄録（300 字程度）

本研究では、小惑星探査画像における小さなボルダーの検出性能を向上させることを目的として、Mask R-CNN フレームワークにおける Backbone 部分を、従来の CNN 構造（ResNet-50）から、Transformer 構造をベースとする SOD-Former（Small Object Detection-Former）に置き換えるモデルを構築し、その性能を比較検証した。SOD-Former は、Transformer ブロックを 8 層組み込み、CNN 層である程度の特徴抽出した後であるが画像全体の文脈情報を活用しする構造を持つ。性能評価には Recall、IoU、Overall Accuracy を用いて行い、一貫して Transformer ベースのモデルの方が安定して高い検出性能を持つことが確認された。本研究の結果は、Transformer ベースのモデルが小惑星表面の画像解析、特に小さなボルダーの検出において CNN ベースの従来手法よりも高精度かつ頑強であることを示唆するものである。

英文抄録（100 words 程度）

To accelerate the process of boulder detection that is crucial for both science and exploration mission purposes, machine learning models such as Mask R-CNN have been applied in recent studies. In this work, we aim to improve the detection of small objects by replacing the backbone of Mask R-CNN with a Transformer-based module from conventional CNN module. We confirmed that our approach shows better capability of detecting smaller features. This indicates high possibility of Transformer module for improving data analysis in asteroid exploration missions.

Keywords: Transformer, Mask-RCNN, Asteroid, Boulder detection, Remote sensing

背景と目的

近年、「はやぶさ2」や「OSIRIS-REx」、DARTミッションなどの小惑星探査ミッションにより、小惑星表面の高解像度画像が多数取得され、科学的議論を可能とさせる豊富なデータセットが整備されつつある。これにより、小惑星表面に存在するボルダーのサイズ、形状、方位などの統計的特徴を精密に解析し、天体の形成過程や進化史を比較研究することが可能になる（cf. Michikami et al., 2021）。たとえば、Ryugu と Bennu は一見するとアルベド、スペクトル、密度、さらにはリターンサンプルの CI 的な特徴を共有してもつことが見られるが、ボルダーの形状分布や方位の特性には違いが認められている。これらの違いは、それぞれの天体が異なる母天体由来であることや、形成時の物質分布の不均質性、あるいは進化経路の違いを示唆している可能性がある。他方、小惑星リュウグウの表層画像はそのわずか 4%しか解析されず、人間が膨大なデータを処理するには時間的・経済的制約が大きい（Hirata et al., 2019）。さらに、従来の深層学習モデルでは特定の天体に対してのみ有効であり、別の対象にはそのまま適用できないといった汎用性の課題も存在する。

このようなサイエンス議論に耐えうる統計値の獲得には小惑星それぞれにおいて多数の接近画像から数万～数十万のボルダーの数え上げと輪郭の計測が求められる（Michikami et al., 2021）が、人間による手作業での解析は難しく、小惑星ごとに異なる表面特徴に対応できる頑強な自動解析技術の開発が課題となっている。

以上の観点から、本研究では一般写真において高い汎用性能が確認されている Meta 社が開発した基盤

モデル「Segment Anything（SAM）」（Kirillov et al., 2023）の惑星探査データへの応用可能性を検討することを最終目標とし、その第一段階として、Transformer 構造に基づく高性能な画像認識モデルの実装と応用に取り組んだ。特に、小惑星表面に存在するボルダー（岩塊）の自動検出を対象とし、従来の畳み込みニューラルネットワーク（CNN）と Transformer ベースのアプローチによる高精度化で Transformer の優位性の検証を目指した。

概要

本研究では、小惑星探査画像における小さなボルダーの検出性能を向上させることを目的として、Mask R-CNN フレームワークにおける Backbone 部分を、従来の CNN 構造（ResNet-50）から、Transformer 構造をベースとする SOD-Former（Small Object Detection-Former）に置き換えるモデルを構築し、その性能を比較検証した。SOD-Former は、Residual Hybrid Attention Group（RHAG）と呼ばれる Transformer ブロックを 8 層組み込んだ構造を持ち、CNN 層である程度の特徴抽出した後であるが画像全体の文脈情報を活用しながら、小さな物体の検出に最適化されている。

性能評価には Recall、IoU（Intersection over Union）、Overall Accuracy（OA）を用いて行い、小さなボルダーを見やすくするための画像のリスケーリングの倍率(2 倍、4 倍)によって精度が異なることはあったものの、2 倍サイズの画像において、SOD-Former は Recall で約 2 倍、IoU で約 20 ポイント上回る性能を示した。4 倍サイズでは両モデル間の差は縮小されたが、それでも一貫して Transformer ベースのモデルの方が安定して高い検出性能を持つことが確認された。

本研究の結果は、Transformer ベースのモデルが小惑星表面の画像解析、特に小さなボルダーの検出において CNN ベースの従来手法よりも高精度かつ頑強であることを示唆するものである。SOD-Former は画像中のピクセル間の関係性を広範囲に学習できるため、複雑な地形や陰影が存在する小惑星表面においても有効であると考えられる。

結果および考察

本研究では、小惑星探査画像における小さなボルダーの検出性能を向上させることを目的として、Mask R-CNN フレームワークにおける Backbone 部分を、従来の CNN 構造（ResNet-50）から、Transformer 構造をベースとする SOD-Former（Small Object Detection-Former）に置き換えるモデルを構築し、その性能を比較検証した。

〇使用データと前処理

データとして使用したのは、小惑星リュウグウに搭載された ONC-T カメラによって撮影された可視画像 228 枚である。撮影高度は 5km 以下に限定し、岩塊の形状が明瞭に視認できる画像を抽出した。解像度は 1024×1024 ピクセルであり、先行研究(Seki et al., 2024)に従い Sigma コントラストによる強調処理および Wiener フィルタによるぼやけ除去を前処理として行った。画像は、オリジナルサイズ、Half（512×512）、Quarter（256×256）に分割し、それぞれについて手動アノテーションデータ(Kouyama, 2025)と対応づけを行った。最終的に 117,676 個のアノテーションを含む学習データセットを構築し、学習とテストに 8:2 で分割した。

〇モデル構造

本研究では、Mask R-CNN のフレームワークにおいて、ResNet-50 ベースのモデルと SOD-Former ベースのモデルを比較した。SOD-Former は、Residual Hybrid Attention Group（RHAG）と呼ばれる Transformer ブロックを 8 層組み込んだ構造を持ち(図 1)、CNN 層である程度の特徴抽出した後であるが画像全体の文脈情報を活用しながら、小さな物体の検出に最適化されている。

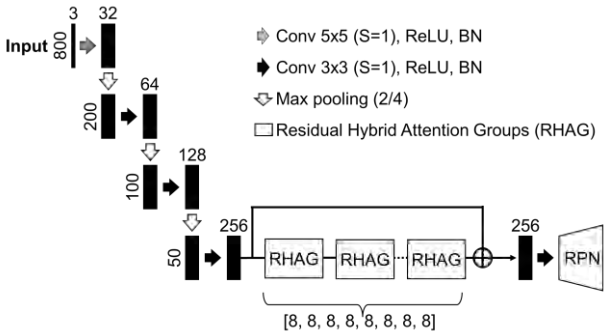


図 1 提案する RHAG ブロックを画像特徴抽出モジュール

また、RPN（Region Proposal Network）においては、小さな岩塊に対応する Anchor Box サイズの最適化を行い、実際のアノテーションのサイズ分布に合致するよう調整を加えた。

〇評価指標と実験結果

図 2 に Quarter 条件でのボルダー検知結果の例を示す。

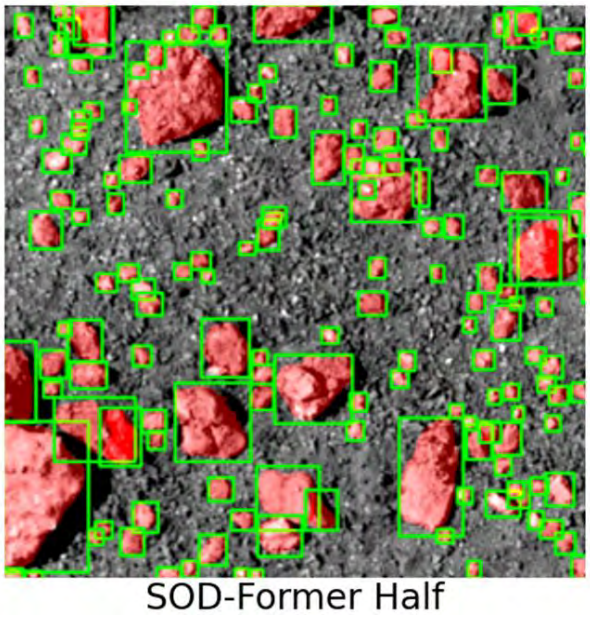


図 2 SOD-Former ベースのモデルを用いたボルダー検出結果例

性能評価には Recall、IoU（Intersection over Union）、Overall Accuracy（OA）を用いて行っている。Half と Quarter のデータセットに対してそれぞれの評価指標の結果をまとめると以下のテーブルのようになる。

表 1 Resnet ベースと SOF Former ベースのボルダー検知精度比較(SOD-Former を SOD-F と記す)。

	Resnet	**SOD-F**	**Resnet**	**SOD-F**
Dataset	Half	Half	Quarter	Quarter
Reacall	28.98	**59.55**	68.86	**72.09**
IoU	46.91	**68.48**	**85.44**	83.23
OA	76.02	**84.32**	79.22	**79.35**

特に Half サイズの画像において、SOD-Former は Recall で約 2 倍、IoU で約 20 ポイント上回る性能を示した。Quarter サイズでは両モデル間の差は縮小されたが、それでも一貫して Transformer ベースのモデルの方が安定して高い検出性能を持つことが確認された。

〇考察

本研究の結果は、Transformer ベースのモデルが小惑星表面の画像解析、特に小さなボルダーの検出において CNN ベースの従来手法よりも高精度かつ頑強であることを示唆するものである。SOD-Former は画像中のピクセル間の関係性を広範囲に学習できるため、複雑な地形や陰影が存在する小惑星表面においても有効であった。

また、Quarter サイズの画像では CNN モデルとの差が縮まった背景には、極端に小さい物体では CNN でも十分な特徴抽出が可能となるケースがあることが考えられる。一方で、Half 画像のような中解像度環境では、Transformer の文脈理解能力がより効果的に機能し、顕著な性能向上につながっている。

まとめ、今後の課題

本研究では、小惑星探査画像における小さなボルダーの検出性能を向上させることを目的として、Mask R-CNN フレームワークにおける Backbone 部分を、従来の CNN 構造（ResNet-50）から、Transformer 構造をベースとする SOD-Former に置き換えるモデルを構築し、その性能を比較検証し、Transformer 構造を取り込むことで精度向上が見られることを明らかにした。今後は、さらに他の小惑星画像への適用を進めるとともに、Segment Anything Model（SAM）などの基盤モデル技術を取り入れ、地球上の類似データと組み合わせた転移学習による精度向上を試みる予定である。

TSUBAME 共同利用　令和6年度　学術利用　成果報告書

利用課題名　直接数値解析による非相似伝熱制御の検討
英文：Direct numerical simulations on the dissimilar turbulent heat transfer

利用課題責任者
桑田　祐丞

所属
大阪公立大学　工学研究科
URL：https://www.omu.ac.jp/eng/htlab/

ケルビンヘルムホルツ（K-H:Kelvin-Helmholtz）不安定波誘起の大規模渦を利用した伝熱促進に関する検討を直接数値解析によって行った．本研究では，K-H 不安定波に起因する乱流構造を発達させるために，流れの方向に沿って規則的に並ぶリブ列を壁面上に設置し，リブ列の間隔やレイノルズ数を変化させて熱流動場の特性を調査した．その結果，リブ列の上部には，K-H 不安定波に起因した大規模渦が発達し，摩擦係数とスタントン数の比で現れるレイノルズアナロジファクタは1を大きく超えて，理想的な非相似伝熱状態が実現できることを確認した．また，レイノルズアナロジファクタはバルク平均速度とリブ頂点の面平均速度の比を用いてスケーリングすることができることが明らかとなった．

We discussed the dissimilar heat transfer control by the large-scale turbulence structure associated with the Kelvin-Helmholtz instability by means of direct numerical simulations. We considered a channel with the streamwise-aligned regularly distributed rib array to trigger the K-H instability and discussed the effects of the Reynolds number and rib separation on the turbulent heat transfer. The results show that the rib arrays allow the development of the large-scale turbulence structure associated with the K-H instability leading to the favorable breakdown of the Reynolds analogy. It is also found that the Reynolds analogy factor can be reasonably scaled by the ratio of the plane-averaged velocity at the rib crest to the bulk mean velocity.

Keywords: Direct numerical simulation, rib roughness, Reynolds analogy, Lattice Boltzmann method, Kelvin-Helmholtz instability

背景と目的

熱交換器等の伝熱機器の性能向上のために，伝熱面に人工的な粗さを設置する取り組みが盛んに行われてきた．伝熱面に設置した粗さは，伝熱面積を増やす効果の他に，伝熱面近傍の乱流の強化や平均流の蛇行による流体混合を促進し，熱伝達を向上させる．一方，粗さは熱伝達率のみならず摩擦係数を増大させ，摩擦係数の増大率は熱伝達率の増大率を上回ることも報告されている．一方，Kuwata[(1)]は主流方向に並ぶリブ列を伝熱面に設置することで，熱伝達率の増大率が摩擦係数の増大率を上回る理想的な伝熱制御が可能であることを直接数値解析によって示した．また，熱・運動量輸送の非相似性は，ケルビンヘルムホルツ不安定波誘起の乱流に起因していることを示した．しかし，リブの幾何的な形状やレイノルズ数が熱・運動量輸送の非相似性に与える影響は十分に理解されていない．そこで，本研究では，リブ間隔とレイノルズ数が熱伝達率や摩擦係数に与える影響を直接数値解析によって調査を行った．その結果，ケルビンヘルムホルツ不安定波誘起の大規模乱流構造が顕著になるケースで熱・運動量輸送の非相似性が最大となり，非相似性の強さは，バルク平均速度とリブ頂点の面平均速度の比を用いてスケーリングすることができることが明らかとなった．

概要

ケルビンヘルムホルツ不安定波に起因する大規模渦を誘起することで，流動抵抗の増加を押さえつつ高い伝熱促進効果をもたらす非相似伝熱制御の検討を行う．本研究では，ケルビンヘルムホルツ不安定性に起因する大規模渦を誘起するために，主流方向に並ぶリブ列を用いる．数値解析手法には密度分布関数とエネルギ分布関数の2つの分布関数を解く格子ボルツマン法を用いる．リブ列の間隔を変えた複数のケースに対して，直接数値解析を実施することで，大規模渦の

特性や，運動量・熱輸送の非相似的な挙動について議論を深める．

対象とした流れ場は図 1 に示すリブ付きオープンチャネルである．上面は断熱のすべり壁とし，下壁には主流方向に並ぶ縦リブ列を考える．リブ列上部の流体領域の高さH_cに対して，リブの高さ$h = 0.67H_c$，リブ幅$w = 0.05H_c$とし，リブ間隔$s/w = 1, 1.5, 4, 9, 24$の 5 種類のリブ列を対象とした．リブ列頂点の位置における摩擦速度によって定義された摩擦レイノルズ数を 180, 500 とした．プラントル数は 1 として，浮力や温度変化に伴う物性値の変化を無視した非圧縮の流体を考えた．リブ表面と下壁は等温条件，流体には一様発熱を与えた．主流方向・スパン方向には周期境界条件を用い，主流方向・スパン方向の領域サイズはそれぞれ$8H_c$，$6H_c$として十分大きな計算領域を確保した．

数値解析手法に関しては，速度場は27方向速度多緩和時間格子ボルツマン[2]を用い，温度場には19方向速度正規化格子ボルツマン[3]を用いた．計算格子は立法格子を用い，リブ列近傍の格子幅は乱流渦の最小スケールであるコロモゴロフスケールと同程度となるように設定した．リブ列から離れた領域では，不均衡修正局所細密格子法[4]を用いて，粗い計算格子を設定した．総格子点数は摩擦レイノルズ数 500 のケースで約 7.6 億点となった．並列化には領域分割法を採用し，領域ごとに割り当てられた GPU によって解析を行った．

結果および考察

乱流構造について議論する為に，図2にリブ頂点近傍の主流方向の乱流変動速度を示す．リブ列上では，主流方向に伸びる高速・低速ストリーク構造が明確にみられないが，リブ間隔が最も狭い時には，主流方向に伸びた微細な乱流変動速度が観察される．リブ間隔が広がるにつれてスパン方向につながった乱流構造が出現し，主流方向には低速・高速領域が交互に現れる．

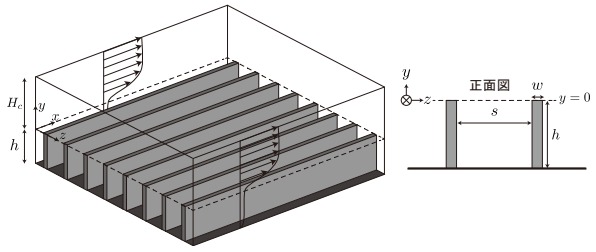
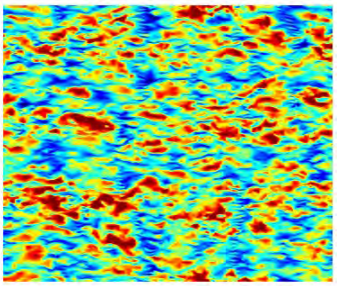


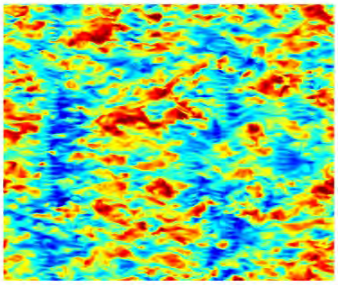
図1　リブ列を有するオープンチャネル．

このような乱流構造は，多孔質体や植生流れなどで見られるケルビンヘルムホルツ不安定波に起因した大規模構造であると考えらえる．この構造は，リブ間隔が$s/w = 4$となる時に最も顕著に表れる．さらにリブ間隔

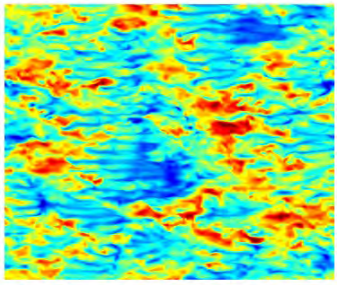
$s/w = 1$



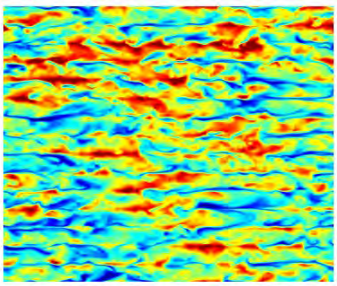
$s/w = 1.5$



$s/w = 4$



$s/w = 9$



$s/w = 24$

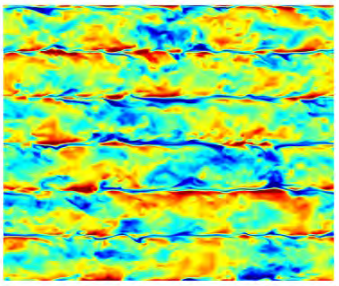


図2 リブ頂点付近の x−z 面の主流乱流変動速度．

が広がると，スパン方向の繋がりは消滅し，リブ列に沿って主流方向に伸びた乱流変動速度が観察される．リブ列の間隔が最も広いケースにおいては，この傾向が特に顕著であり，リブ壁面で発達した乱流境界層に伴う乱流変動が観察される．

次に，熱・運動量輸送の非相似性について議論する為に，スタントン数と摩擦係数の比で表されるレイノルズアナロジファクタを調べる．レイノルズアナロジファクタは，熱・運動量輸送の相似性が成立する平滑面乱流では 1 となり，熱輸送が運動量輸送を上回ると 1 を超え，下回ると 1 より小さな値を取る．図 3 にレイノルズアナロジファクタRAを縦軸に，滑り速度パラメータF_sを横軸としてプロットする．F_sは，バルク速度U_bとリブと流体領域の界面速度U_sとの比を用いて$F_s = U_s/U_b$として定義する．また，参照データとして摩擦レイノルズ 300 の文献データ[1]もプロットしている．図より，レイノルズアナロジファクタはすべてのレイノルズ数・リブ間隔において，1を超えており，熱輸送が運動量輸送を上回る理想的な非相似状態が実現できていることが分かる．また，レイノルズアナロジファクタの大きさはリブ間隔やレイノルズ数に強く依存するが，滑り速度パラメータを用いておおむねスケーリング可能であることが分かる．滑り速度パラメータが 0 の値から上昇するにつれて，アナロジファクタは増大し，熱・運動量輸送の非相似性が強くなる．アナロジファクタは，滑り速度パラメータが 0.3 付近で最大値を取り，滑り速度パラメータがさらに上昇すると，アナロジファクタは減少する．リブ間隔が最も大きいケースでは，滑り速度パラメータも増大し，アナロジファクタは再び1程度の値となる．これらの結果より，滑り速度パラメータが 0.3 付近で，熱輸送が運動量輸送を上回る最も理想的な伝熱制御が可能であることが分かった．このケースはリブ間隔が$s/w = 4$となるケースに対応しており，ケルビンヘルムホルツ不安定波に由来する大規模渦が最も顕著にみられるケースと対応する．よって，輸送が運動量輸送を上回る非相似状態は，ケルビンヘルムホルツ不安定波に由来する大規模渦に起因して現れることが示唆された．

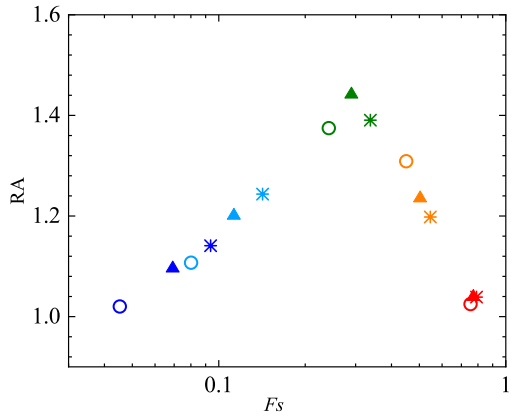


図 3 レイノルズアナロジファクタ．

まとめ、今後の課題

ケルビンヘルムホルツ不安定波誘起の大規模渦を利用した伝熱促進に関する検討を直接数値解析によって行った．その結果，リブ列の上部には，ケルビンヘルムホルツ不安定波に起因した大規模渦が発達し，摩擦係数とスタントン数の比で現れるレイノルズアナロジファクタは1を大きく超えて，理想的な非相似伝熱状態が実現できることを確認した．非相似伝熱は，大規模乱流構造によって引き起こされることが示唆され，非相似伝熱の強さを表すレイノルズアナロジファクタはバルク平均速度とリブ頂点の面平均速度の比を用いてスケーリングすることができることが明らかとなった．

参考文献

(1) Y Kuwata , “Dissimilar turbulent heat transfer enhancement by Kelvin–Helmholtz rollers over high-aspect-ratio longitudinal ribs”, *Journal of Fluid Mechanics* **952**, A21 (2022).

(2) K Suga, Y Kuwata, K Takashima, R Chikasue, “A D3Q27 multiple-relaxation-time lattice Boltzmann method for turbulent flows”, *Computers & Mathematics with Applications* **69** (6), 518-529 (2015).

(3) K Suga, R Chikasue, Y Kuwata, “Modelling turbulent and dispersion heat fluxes in turbulent porous medium flows using the resolved LES data”, *International Journal of Heat and Fluid Flow* **68**, 225-236 (2017).

(4) Y Kuwata, K Suga , “Imbalance-correction grid-refinement method for lattice Boltzmann flow simulations”, *Journal of Computational Physics* **311**, 348-362 (2016).

TSUBAME 共同利用　令和 6 年度　学術利用　成果報告書

利用課題名　無機有機ハイブリッド結晶の格子欠陥計算に関する研究
英文：Theoretical computational studies of lattice defects in inorganic-organic hybrid crystals

利用課題責任者
齋藤　典生

所属
山梨大学大学院総合研究部附属クリスタル科学研究センター
n-saito@yamanashi.ac.jp

邦文抄録（300 字程度）
本課題は、ギ酸分子アニオンと Al^{3+}イオンが 3 次元格子状に配列した化合物 $Al(HCOO)_3$ の結晶中に存在する不純物や格子欠陥について、理論計算と実験データと組み合わせて、その構造や材料特性との関係を明らかにすることを目的とした。$Al(HCOO)_3$ の結晶格子に残存する不純物分子（H_2O や CO_2）の構造や、$Al(HCOO)_3$ との相互作用を明らかにするため、$Al(HCOO)_3$ の 2×2×2 超格子の格子空隙に、位置や配向を変化させたゲスト分子を配置し、エンタルピー変化を計算した。その結果、特定の格子間隙において、ゲスト分子が $Al(HCOO)_3$ と水素結合を形成し、安定化することがわかった。

英文抄録（100 words 程度）
This study aimed to elucidate the structure and properties of impurities and lattice defects in $Al(HCOO)_3$. Impurity molecules such as H_2O and CO_2 were introduced into a 2×2×2 superlattice of $Al(HCOO)_3$, and their enthalpies were evaluated using density functional theory (DFT) calculations. The results revealed that hydrogen bonding between the impurity molecules and $Al(HCOO)_3$ stabilizes these impurities within specific interstitial sites.

Keywords: MOF, Formate, Aluminium, Impurity, H_2O, CO_2, Inorganic-organic hybrid

背景と目的
無機有機ハイブリッド結晶の一種である多孔性金属錯体（MOF）は、金属イオンと配位子からなる錯体ユニットが連続的に結合することで、2 次元や 3 次元の高次構造を形成する。均一な微細孔をもつことが特徴で、分離・貯蔵・吸着・変換など様々な機能を示す。MOF の結晶格子には、全無機化合物と同様に原子の配列乱れや欠陥の存在が示唆されているが、どのような構造の欠陥が存在するかや、格子欠陥が材料物性に及ぼす影響はよくわかっていない。また、MOF の結晶格子中には水などの不純物が容易に侵入し、吸着などの物性に影響を及ぼす。

そこで本課題は、ギ酸分子アニオンと Al^{3+}イオンが三次元格子状に配列した $Al(HCOO)_3$ について、格子欠陥や不純物を含む結晶構造を、密度汎関数法（DFT）計算を用いてモデル化し、実験と連携してその構造や物性に及ぼす影響を明らかにすることを目的した。

概要
$Al(OH)_3$ とギ酸を混合した溶液を 100 °C で 24 h 還流することで $Al(HCOO)_3$ を合成した。作製したサンプルを、赤外分光法や粉末 X 線回折（XRD）などを用いて分析したところ、結晶構造中に H_2O や CO_2 を不純物として含有することを確認した。結晶格子に残存する不純物分子の構造や $Al(HCOO)_3$ との相互作用を明らかにするため、$Al(HCOO)_3$ の 2×2×2 超格子を作成し、格子空隙中に位置や配向を変化させたゲスト分子を配置し、構造緩和計算を検討した。超格子と孤立分子モデルのエンタルピーの差から、インターカレーションによるエンタルピー変化（ΔH）を求めた。

$Al(HCOO)_3$ の結晶構造に存在する空隙は、ギ酸の H 原子が空隙の内側へ向いたもの（Void_inH）と、外側へ向いたもの（Void_outH）の 2 種類が存在する。そこで、それぞれの空隙にゲスト分子を配置し、ΔH を比較したところ、H_2O・CO_2 分子ともに Void_inH の方が 160−190 meV 安定であることがわかった。Void_inH へのインターカレーションが安定化する原因は、CO_2 の

O 原子とギ酸の H 原子、水分子の H 原子とギ酸の O 原子で水素結合を形成するためである。一方、Void_outH ではいずれのモデルも水素結合の形成は確認されなかった。各モデル間の ΔH の差は約 60 meV であるため、格子間隙中にゲスト分子の安定サイトが幾つか存在する可能性が考えられる。$Al(HCOO)_3$ 結晶格子における H_2O や CO_2 分子の最安定構造を明らかにするため、ゲスト分子の占有位置や配向を変化させた様々なモデルで ΔH を計算中である。

結果および考察

計算に用いた $Al(HCOO)_3$ の超格子モデルを図 1a に示す。この超格子は $Z = 64$ であり、格子中に存在する 1 つの格子間隙に H_2O または CO_2 を 1 分子だけ挿入し、格子定数と原子位置の最適化計算を行った。ゲスト分子の位置とその配向は、図 1b に示す 5 パターンを検討し、構造緩和計算で得られたそれぞれのエンタルピーを表 1 に示す。

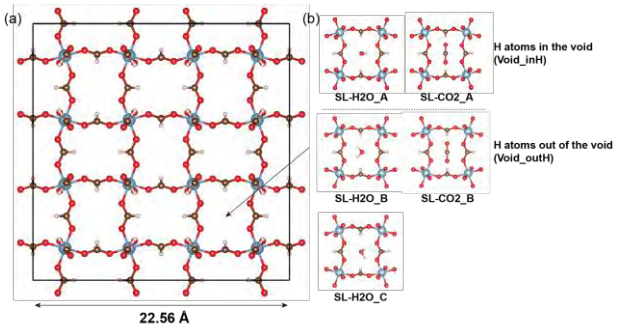


図. 1 (a) $Al(HCOO)_3$ 2×2×2 超格子の結晶構造、(b) H_2O および CO_2 分子を格子間隙に挿入した構造モデル

表 1. $Al(HCOO)_3$ 超格子モデルの構造緩和計算結果

Model	Void type	Enthalpy / eV	ΔH / meV
SL-Orig	–	−5871.5524	–
SL-H2O-A	Void_inH	−5886.3395	−488.7
SL-H2O-B	Void_outH	−5886.1142	−263.4
SL-H2O_C	Void_outH	−5886.1760	−325.2
SL-CO2-A	Void_inH	−5895.7402	−524.6
SL-CO2_B	Void_outH	−5895.5742	−358.6

インターカレーションによるエンタルピー変化（ΔH）は下記の式によって求めた。

$$\Delta H = H(\text{SL-Molecule}) - (H(\text{SL-orig}) + H(\text{Molecule})) \quad \text{(eq.1)}$$

ここで、H(SL-Molecule)はゲスト分子を挿入した超格子、H(SL-Orig) はゲスト分子なしの超格子、H(Molecule)は孤立したゲスト分子のエンタルピーを表す。

eq.1 を用いて、各モデルの ΔH を求めたところ、全ての ΔH が負となった。したがって、0 K において H_2O および CO_2 分子のインターカレーションが自律的に進行すると予測される。ΔH の絶対値を比較すると、Void_inH にインターカレーションさせたモデルの方が Void_outH よりも 160−190 meV 程度値が大きく、Void_inH が吸着の安定サイトであることがわかる。Void_inH において、CO_2 の O 原子とギ酸の H 原子の原子間距離が 2.33 Å であり、両者が水素結合を形成することが示唆された。また、H_2O についても H 原子とギ酸の O 原子の原子間距離が 2.43 Å であり、両者が水素結合を形成すると推察される。

Void_outH では、いずれのモデルでも水素結合の形成は確認されず、各モデル間の ΔH の差は約 60 meV であった。粉末 XRD のリートベルト解析において、Void_outH にインターカレーションした H_2O や CO_2 分子がディスオーダーによって格子空隙中に広く分布する結果を得ており、上述の計算結果は実験結果を支持している。

まとめ、今後の課題

本課題は、$Al(HCOO)_3$ の超格子に H_2O および CO_2 分子を挿入し、そのエンタルピーからインターカレーションの構造を予測した。$Al(HCOO)_3$ における H_2O や CO_2 分子の最安定構造を明らかにするため、分子の占有位置や配向を変化させた様々なモデルで ΔH を計算中であり、次年度も引き続き検討を行う予定である。また、一部のギ酸配位子が欠損した欠陥モデルについても検討を行っており、実験結果との連携を計画している。

TSUBAME 共同利用　令和6年度　学術利用　成果報告書

利用課題名　多言語マルチモーダル大規模言語モデルに関する研究

利用課題責任者　小町守

所属　一橋大学大学院　ソーシャル・データサイエンス研究科
URL https://hit.komachi.live

邦文抄録（300 字程度）
多言語マルチモーダル大規模言語モデルの研究について、視覚言語モデルにどれくらい画像の認識能力があるのかの検証を行いました。視覚言語モデルを用いて画像特徴量を得て、テキストに関連した画像を検索するような予備的なタスクの設計と、画像の時間情報を考慮する画像質問応答に関する新しいアーキテクチャの検討を行いました。画像質問応答に関しては、大規模言語モデルを用いて自動的に重要度を付与したデータセットを作成し、重要度を自動的に推定するアーキテクチャを考案し、実験したところ、提案手法はベースラインの手法と比較し、画像の時間情報を考慮した質問応答システム性能の向上に寄与することが示唆されました。

背景と目的

画像やテキストを統合的に利活用するマルチモーダルな大規模多言語言語モデル（視覚言語モデル）に関する研究が盛んに行われています。視覚言語モデルがどれくらい言語と視覚のインタラクションを扱うことができるかは未知数であるため、視覚言語モデルの認識能力がどれくらいあるのかを検証しました。また、画像質問応答において、関連性が低い画像を用いて質問応答を行う可能性があるため、重要度の高い画像を推定しつつ質問応答ができるアーキテクチャの考案を行いました。

概要

画像やテキストを統合的に利活用するマルチモーダルな大規模多言語言語モデルに関する研究を行いました。ViCLIP を用いて画像に対する画像特徴量を得てテキストに関連した画像を検索するような予備的なタスクの設計と、それを用いた画像質問応答のアーキテクチャに関する検討を行いました。データセットは大規模言語モデル（GPT-4o, Gemini）によって評価および説明の生成を行い、質の高いデータセットとなるようにフィルタリングを行いました。また、画像質問応答については画像に付与された前後の時間情報を考慮したアーキテクチャにすることで、画像の質問に対する関連度（重要性）を適切に推定することを狙いました。

結果および考察

画像とテキストを CLIP によって統合的に扱うことが可能である、ということが検証できました。画像質問応答に関しては検索された画像の重要度を推論するアーキテクチャを考案し、重要度の判断を自動付与したデータセットを作成してファインチューニングすることで、重要度を自動的に推定できることも示しました。直接的評価、間接的評価の両方でベースラインと比較して性能の向上を達成することができました。一方、前後の時間情報を両方考慮するようなアーキテクチャには必ずしも大きな性能の向上が得られないことも分かりました。

まとめ、今後の課題

マルチモーダル大規模言語モデルの利活用に関する研究に取り組みました。具体的には画像質問応答において、画像とテキストのインタラクションを用いて検索する新しいタスクを提案し、時間情報を考慮しつつ関連度の高い画像を用いて質問応答する新しいアーキテクチャを考案し、実験的に有効性を検証しました。

ダウンストリームのタスクでハイパーパラメータの探索をすることに時間がかかることが分かり、どのように TSUBAME のプラットフォームを使いこなしていくのかについて研究グループ内で知見を溜めていく必要があることが分かりました。

TSUBAME 共同利用　令和 6 年度　学術利用　成果報告書

利用課題名　計算資源が限られた状況下における Low-Rank Adaptation Fine-Tuning に関する研究
英文：A Study on Low-Rank Adaptation Fine-Tuning under Limited Computational Resources

利用課題責任者
波多野 賢治
Kenji Hatano

同志社大学文化情報学部
Doshisha University, Faculty of Culture and Information Science
https://www-mil.cis.doshisha.ac.jp

本課題では，学習済みの LoRA アダプタが，重みのビット幅が異なる基盤モデル間で転移可能であるかどうかを検証した．具体的には，複数のタスクとモデルを用いて，低ビット幅へ量子化された基盤モデル上で学習された QLoRA アダプタが，より大きなビット幅の基盤モデル上でも有効に動作する事を明らかにした．特に，重みを 4 ビット幅や 3 ビット幅に量子化した基盤モデル上で学習したアダプタであっても，推論時に重みを 16 ビット幅に復元することで，基盤モデルを量子化しない通常の LoRA とほぼ同等の推論精度を達成できることを示した．本手法は，学習済み QLoRA モデルの性能向上や，量子化によって得られた余剰メモリを活用したバッチサイズの拡大による LoRA の学習速度の向上などに活用可能である．

In this study, we investigated whether trained LoRA adapters can be transferred across foundation models with different weight bit-widths. Specifically, we showed that LoRA adapters trained on foundation models quantized to low bit-widths can still function effectively when applied to foundation models with higher bit-widths, using multiple tasks and models. We demonstrated that even adapters trained on 4-bit or 3-bit quantized foundation models can achieve inference accuracy comparable to standard LoRA trained at 16-bit precision, by restoring the foundation model to 16-bit at inference time. This approach can be applied to enhance the performance of pre-trained QLoRA models and to accelerate LoRA training by expanding the batch size using memory savings gained through quantization.

Keywords: PEFT, LoRA, QLoRA, Quantization, Transferability

背景と目的

大規模言語モデルをはじめとする深層学習モデルは，重みの勾配や最適化状態を保持するため，推論時よりも訓練時に要求される GPU メモリ量が多いことが知られている．したがって，例えば同一の計算資源上で基盤モデルのファインチューニングと推論を行う際には，基盤モデルのサイズは訓練時の資源制約から決定されるため，推論時にはメモリの余剰が発生する．

その一方で，既存手法はこの余剰メモリを考慮しておらず活用できていない[1,2,3]．そこで本課題では，この推論時のメモリの余剰を活用して LoRA モデルの性能を向上させる新たな 量子化-LoRA フレームワークとして，Post LoRA Restoration（PLR）を提案する．PLR では，訓練時の資源制約から生まれる推論時のメモリの余剰を活用し，量子化された基盤モデルの重みのビット幅の復元を行う．本手法はあるビット幅の基盤モデル上で学習したアダプタが他のビット幅の基盤モデルの上でも効果的に機能できるという転移性を LoRA アダプタが持つという仮説に基づいている．

TSUBAME4.0 の計算資源を用いて評価実験を行った結果，PLR の適用による精度の向上を確認でき，アダプタの転移性および提案手法の有効性を確認できた．

概要

提案手法である PLR の概念図を図1に示す．モデルの重みを k ビットへ量子化した後に LoRA を実施し推論を行う QLoRA [2] に対し，PLR では推論時のメ

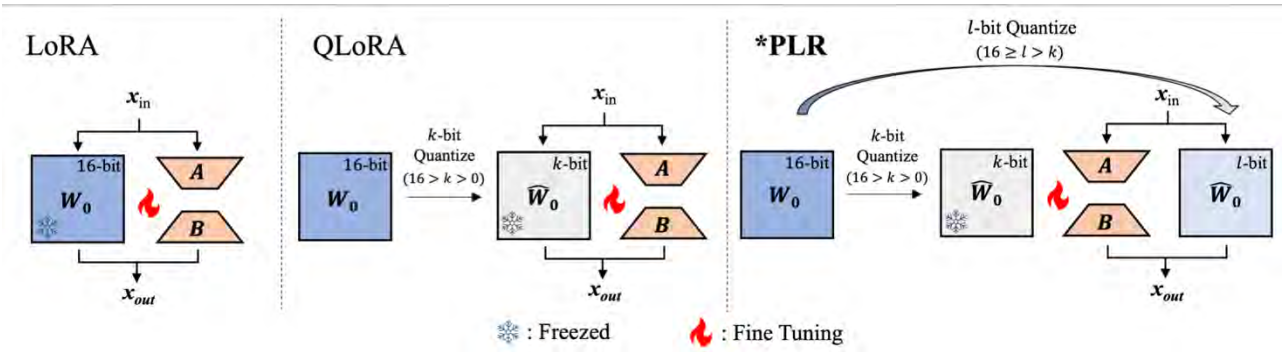


図1：LoRA, QLoRA, PLR の概念図

モリ余剰に着目して，量子化-LoRA モデルを学習したのちに 量子化-LoRA の基盤モデルの重みの精度をより高いビット幅へ復元する．

PLR の具体的な流れは，以下の通りである．基盤モデルの重みを $\boldsymbol{W}_0^{16\text{-bit}}$，LoRA アダプタの重みを $\Delta\boldsymbol{W}_{\text{LoRA}}$ とすると，LoRA 適用後の各層の重み $\boldsymbol{W}$ は，$\boldsymbol{W} = \boldsymbol{W}_0^{16\text{-bit}} + \Delta\boldsymbol{W}_{\text{LoRA}}$ で表される．QLoRA などの量子化-LoRA 手法では，$\boldsymbol{W}_0^{16\text{-bit}}$ を k ビット幅へ量子化し $\boldsymbol{W}_0^{k\text{-bit}}$ に置き換えた，$\boldsymbol{W} = \boldsymbol{W}_0^{k\text{-bit}} + \Delta\boldsymbol{W}_{\text{LoRA}}$ について，$\Delta\boldsymbol{W}_{\text{LoRA}}$ の訓練を行う．ここで，k は $16 > k > 0$ である．PLR では，QLoRA の訓練完了後，$\boldsymbol{W}_0^{k\text{-bit}}$ を $\boldsymbol{W}_0^{l\text{-bit}}$ の重みへ置き換えた，$\boldsymbol{W} = \boldsymbol{W}_0^{l\text{-bit}} + \Delta\boldsymbol{W}_{\text{LoRA}}$ によって推論が行われる．ここで，l は $l > k$ を満たし，推論を行う計算資源の余剰に合わせて選択される．

一般的に，LLM の量子化は低いビット幅であるほど近似値に丸め込まれた際に生じる量子化誤差が大きく，性能が低下することが知られている．PLR を適用することでよりビット幅の大きく劣化の少ない基盤モデルを用いることが可能となるため，量子化誤差から生じる基盤モデルの性能低下を回避し QLoRA モデルの性能を復元できることが期待される．

PLR の有効性を確かめるために，QLoRA モデルに対し PLR を適用した際に下流タスクの正解率が向上するかを検証する．使用するモデルは，Llama3 ファミリー[4]のうち，Llama 3.2-1B, Llama 3.2-3B, Llama 3.1-8B の三種類のモデルサイズを選定した．また，評価対象とするタスクには，Grade School Math 8K (GSM8K) [4] および SQL Create Context (SCC) を使用する．

次に，学習の設定について述べる．LoRA および QLoRA では，Adapter の Rank と は，すべてのモデルで 16 に設定し，LoRA を適用する LLM の層は Attention 層のみとした．学習率は 2e-4，バッチサイズは 32 とし，10 エポック学習した内，検証データに対する Loss が最も低いエポックの重みを評価に使用する，QLoRA および PLR で用いる基盤モデルの量子化手法には，GPTQ [7] を使用し，8,4,3,2 ビットの4パターンのビット幅で PLR の有効性を検証する．GPTQ は，重みの量子化後にキャリブレーションデータを用いて補正を行い，量子化誤差を最小限に抑える手法である．キャリブレーションデータには，訓練データから無作為抽出した 500 件を使用する．これらの実装は，Python 3.9.18，Cuda 12.8，PyTorch 2.5.0，Transformers 4.47.0，Peft 0.14.0 のソフトウェアおよびモジュールを用いて，TSUBAME4.0 の計算機上で行われた．

結果および考察

上記の設定で学習した LoRA，QLoRA，QLoRA+PLR モデルの各タスクに対する正解率を，表1に示す．なお，表中の PLR16 や PLR8 は，QLoRA の学習後にそれぞれ 16-bit，8-bit へ基盤モデルの重みを復元することを指す．

モデルとデータセット全体を見ると，ほぼすべてのケースにおいて PLR の適用によって正解率が向上しており，PLR の有効性が確認された。一方で，8-bit においては唯一 PLR の性能が低下している．ここで，16-bit LoRA に着目すると，16-bit LoRA よりも 8-bit QLoRA および 8-bit QLoRA + PLR16 の正解率が高い．このことから，基盤モデルとしての性

表1：LoRA，QLoRA および QLoRA+PLR の各モデルのそれぞれのタスクに対する正解率

Datasets	Models	16-bit LoRA	8-bit QLoRA		4-bit QLoRA			3-bit QLoRA				2-bit QLoRA				
			QLoRA	PLR16	QLoRA	PLR8	PLR16	QLoRA	PLR4	PLR8	PLR16	QLoRA	PLR3	PLR4	PLR8	PLR16
GSM8k	Llama 3.2-1B	21.60	**22.13**	21.83	18.19	**18.95**	**19.41**	11.22	**15.31**	**16.91**	**16.53**	**3.18**	3.11	3.03	2.81	2.43
	Llama 3.2-3B	42.91	**43.06**	42.07	37.75	**41.32**	**42.38**	29.56	**34.50**	**37.76**	**37.30**	4.17	**6.60**	**15.61**	**15.39**	**15.85**
	Llama 3.1-8B	59.66	58.75	**58.98**	56.56	**58.15**	**58.00**	50.34	**57.31**	**58.98**	**59.59**	5.16	**18.27**	**33.35**	**33.73**	**34.34**
	Avg.	41.39	**41.31**	40.96	37.50	**39.47**	**39.93**	30.37	**35.71**	**37.88**	**37.81**	4.17	**9.33**	**17.33**	**17.31**	**17.54**
SCC	Llama 3.2-1B	69.21	**67.29**	63.33	25.84	**30.11**	**31.60**	23.90	22.19	**37.69**	**37.63**	0.00	0.00	**0.22**	**0.22**	**0.19**
	Llama 3.2-3B	72.30	**83.02**	82.79	75.22	**80.49**	**80.41**	68.65	68.15	**75.54**	**75.44**	2.97	**32.93**	**39.51**	**35.83**	**35.87**
	Llama 3.1-8B	84.96	**84.30**	84.21	84.38	**84.47**	**84.46**	77.35	**79.46**	**80.04**	**79.94**	4.81	**36.73**	**32.20**	**36.13**	**36.20**
	Avg.	75.49	**78.20**	76.78	61.81	**65.02**	**65.49**	56.63	56.60	**64.42**	**64.34**	2.59	**23.22**	**23.98**	**24.06**	**24.09**
Avg.	Llama 3.2-1B	45.40	**44.71**	42.58	22.02	**24.53**	**25.51**	17.56	**18.75**	**27.30**	**27.08**	1.59	1.55	**1.63**	1.52	1.31
	Llama 3.2-3B	57.61	**63.04**	62.43	56.49	**60.91**	**61.40**	49.11	**51.33**	**56.65**	**56.37**	3.57	**19.77**	**27.56**	**25.61**	**25.86**
	Llama 3.1-8B	72.31	71.53	**71.60**	70.47	**71.31**	**71.23**	63.85	**68.39**	**69.51**	**69.77**	4.99	**27.50**	**32.78**	**34.93**	**35.27**
	Avg.	58.44	**59.76**	58.87	49.66	**52.25**	**52.71**	43.50	**46.15**	**51.15**	**51.07**	3.38	**16.27**	**20.65**	**20.69**	**20.81**

能が 16-bit のものよりも 8-bit の方が優れていることが推測でき，これが基盤モデルを 16-bit へ復元する PLR16 が 8-bit QLoRA に敗北した要因であると解釈できる．この 8-bit が 16-bit に打ち勝つ現象は，量子化による正則化効果により 16-bit よりも 8-bit の方がアダプタの学習が効率的に進んだためであると考えられる．

次に，モデルサイズごとに結果を比較すると，モデルサイズが大きく，性能が高いほど，PLR の有効性が高まっていることが読み取れる．特に，GSM8k の 3-bit QLoRA+PLR16 や SCC の 4-bit QLORA+PLR16 の精度は，いずれも 16-bit LoRA と同等の精度に達している，これは，4-bit や 3-bit 幅の量子化基盤モデル上で，通常の LoRA と同程度にアダプタがタスクの解法を学習できていることを示唆している．この結果から，LoRA と比較した際の QLoRA による性能低下は量子化誤差による基盤モデルの性能劣化の影響を受けたものであり，一定のビット幅まではアダプタ自体の学習は順調に進んでいることが明らかになった．

まとめと今後の課題

本課題では，低ビット幅の基盤モデル上で学習された LoRA アダプタが，より高いビット幅の基盤モデル上でも有効に動作するかを検証した．実験の結果，3 ビットまでの QLoRA アダプタは，通常の LoRA アダプタと同程度の性能を有することが確認できた．本手法を適用することにより，学習済み QLoRA モデルの性能を追加の学習なしに向上させることが可能である．

今後の課題としては，具体的にどれほどの余剰メモリがある場合に PLR が適用可能かについての調査を行う予定である．また，仮に余剰メモリがないケースでも，処理の一部を CPU に分散させるオフローディング技術などの推論時の最適化手法を用いることで，余剰メモリを創り出せる可能性がある．そうした最適化手法の併用も追加実験として行う必要がある．

参考文献

[1] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In International Conference on Learning Representations, 2021.

[2] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: efficient finetuning of quantized llms. In Proceedings of the 37th International Conference on Neural Information Processing Systems, pp. 10088–10115, 2023.

[3] Yuhui Xu, Lingxi Xie, Xiaotao Gu, Xin Chen, Heng Chang, Hengheng Zhang, Zhengsu Chen, XIAOPENG ZHANG, and Qi Tian. Qa-lora: Quantization-aware low-rank adaptation of large language models. In The Twelfth International Conference on Learning Representations, 2024.

[4] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. arXiv preprint arXiv:2407.21783, 2024.

[5] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. arXiv preprint

arXiv:2110.14168, 2021.
[6] b mc2. sql-create-context dataset, 2023. URL https://huggingface.co/datasets/b-mc2/sql-create-context.
[7] Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers. In The Eleventh International Conference on Learning Representations, 2023.

TSUBAME 共同利用　令和6年度　学術利用　成果報告書

MD シミュレーションを用いた医薬品等の分子設計
Molecular Design Using MD Simulations

宮村　尚明
Naoaki Miyamura

星薬科大学
Hoshi University
https://www.hoshi.ac.jp/

本研究では、短期の TSUBAME4.0 利用を通じて既存の VHH 抗体—卵白リゾチーム複合体の MD シミュレーションデータを再解析し、MM-PBSA 法を適用することで結合自由エネルギーを定量的に評価した。解析の結果、変異導入の有無や変異部位の違いが電荷相補性や水素結合ネットワークに大きく影響し、親和性向上を可能にするアミノ酸置換の有効性が示唆された。また、自由エネルギーの寄与分解により、結合安定化の要因を分子レベルで把握できた。今後は実験データとの照合や多様な変異体への適用を進め、より精密な抗体設計指針の確立を目指す。本成果は抗体工学の一層の精密化や医薬品開発支援に寄与することが期待される。

In this study, we reanalyzed existing MD simulation data for a VHH antibody–hen egg white lysozyme complex using TSUBAME 4.0 over a short period and applied the MM-PBSA method to quantitatively evaluate binding free energy. The results indicated that the presence or absence of mutations, as well as differences in mutation sites, significantly impacted electrostatic complementarity and hydrogen-bond networks, suggesting the effectiveness of amino acid substitutions that enhance affinity. Furthermore, free energy decomposition clarified the factors contributing to complex stabilization at the molecular level. We plan to compare these findings with experimental data and extend their application to a broader range of variants to establish more precise guidelines for antibody design. These outcomes are expected to further refine antibody engineering and contribute to drug development support.

Keywords: MD simulation, MM-PBSA, Antigen-antibody interaction, mutation

背景と目的

先行研究では、スーパーコンピュータ TSUBAME の高度な計算能力を活用し、VHH 抗体と卵白リゾチームの複合体に関する MD シミュレーションを実施し、特定変異による親和性変化を検討してきました。これまでの解析では、変異導入に伴うエネルギー収支をより定量的に評価する必要性が指摘されており、そのための自由エネルギー解析手法が求められていました。そこで本追加プロジェクトでは、短期（約 3 か月）の TSUBAME4.0 利用を通じ、既存の VHH 抗体—卵白リゾチーム複合体の MD トラジェクトリを用いて、MM-PBSA 計算によるエネルギー評価を行うことを目的としました。

概要

本プロジェクトの主目的は、既存の VHH 抗体−卵白リゾチーム複合体系の MD シミュレーションデータを再解析し、MM-PBSA 法を用いて結合自由エネルギーを算出することです。短期間ながら TSUBAME4.0 の計算資源を集中的に用いることで、多数のスナップショットから統計的に安定したエネルギー評価を得ることを目指しました。また、得られたエネルギー分解結果を、これまでの解析結果と照合することで、抗体工学におけるさらなる精密化に資する情報を集積します。

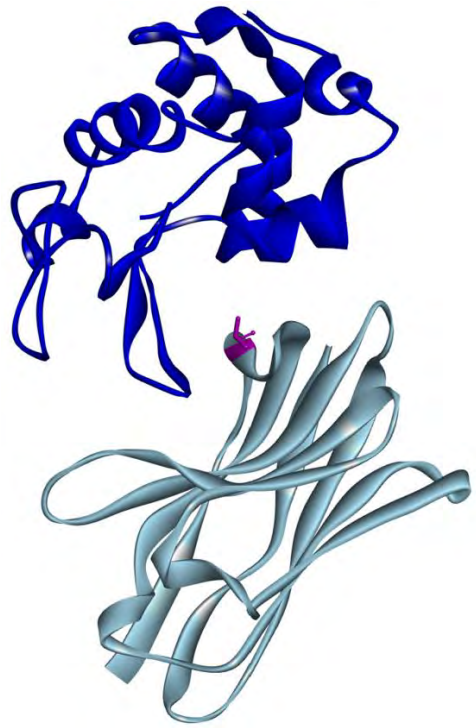


図:VHH 抗体(シアン)と卵白リゾチーム(ブルー)の複合体

結果および考察

VHH 抗体—卵白リゾチーム複合体の MM-PBSA 解析では、変異導入の有無や変異部位の違いによって結合自由エネルギーがどのように変化するかを定量的に評価しました。解析の結果、電荷相補性や疎水性相互作用に加えて、水素結合ネットワークの最適化が結合安定性に大きく寄与することが示唆されました。高い結合親和性を示す変異では、特定のアミノ酸置換が相互作用面の電荷分布や極性相互作用を有利に再編成していることが確認されました。さらに、変異による結合自由エネルギーの寄与分解を通じ、アミノ酸置換の有効性や逆効果となる要因が定量的に示され、実験的アプローチと併せて親和性向上に有望な変異候補を絞り込むための基礎データが得られました。

まとめ、今後の課題

本短期プロジェクトでは、VHH 抗体—卵白リゾチーム複合体の MD シミュレーションデータを再解析し、MM-PBSA 計算による結合自由エネルギーを定量的に評価しました。変異効果の要因を分子レベルで把握するとともに、抗体設計に役立つ具体的な指針を得ることができましたが、実験データとの比較検証を通じたパラメータ最適化や多様な変異体への適用をさらに拡大していくことが今後の課題です。また、量子化学計算を基盤にした分子モデルの改良も重要な課題になります。これらの取り組みにより、計算科学と実験との連携を深めながら、より精密な抗体設計指針を確立し、最適な親和性・選択性を実現するアミノ酸置換を提案できる可能性が期待されます。

TSUBAME 共同利用　令和 5 年度　学術利用　成果報告書

利用課題名　機能性ペプチド提示エクソソームの創製
英文：Development of exosomes with displaying functional peptides

利用課題責任者
Yoshimasa Kawaguchi

所属
Kyoto University, Institute for Chemical Research
URL: https://www.kuicr.kyoto-u.ac.jp/sites/

邦文抄録（300 字程度）
近年、エクソソームは薬物送達キャリアとして注目を浴びている。一方で、その特異性や膜融合効率については課題が残されている。そこで、本研究では、エクソソームに提示されている CD63 において機能性ペプチドをグラフトすることで、細胞への取り込み促進や標的化することを目的に検討を行った。まず、機能性ペプチドをグラフトするのに適したループかどうか CD63 の細胞外ループにおいて揺らぎの計算を実施した。その結果、CD63 の細胞外ループは比較的揺らぎが小さいことがわかった。それらのループに対してモデルペプチドをグラフトしたところ、野生型の CD63 と比較して発現が変化する領域があることがわかった。よって変化の小さい領域に対して、標的化ペプチドを搭載した。一方で、標的タンパク質への結合は確認できず、構造的揺らぎが標的タンパク質結合へ重要であることが示唆された。

英文抄録（100 words 程度）
Exosomes have gained attention as a drug delivery carrier, but their specificity and membrane fusion efficiency are still challenging. This study aims to promote cell uptake and targeting by grafting functional peptides onto CD63, which is presented on exosomes. The extracellular loops of CD63 were found to have relatively low fluctuations and were suitable for grafting. Grafting model peptides onto these loops showed regions where expression changed compared to wild-type CD63. Therefore, we equipped targeting peptides onto regions with minimal changes. However, binding to the target protein was not confirmed, suggesting that structural fluctuations may be crucial for target protein binding.

Keywords: 5つ程度
exosome, graft, peptide, CD63, fluctuation,

背景と目的

エクソソームは、細胞外小胞体であり、細胞から分泌され、異なる細胞間の情報伝達や、疾患の診断や治療のバイオマーカーとして注目されている。しかし、エクソソームを治療や診断に応用するには、標的細胞への送達や取り込みを促進させる必要がある。また、エクソソームに薬物等を内封して薬物送達に利用する場合、標的細胞に取り込まれた後に膜融合による薬物のサイトゾルへの放出が必要であり、効率的な膜融合の促進も課題の一つである。これらの問題を解決することで、エクソソームがより一層有用性の高い drug delivery system に発展させられると考えられる。そこで本プロジェクトでは、機能性ペプチドをエクソソームのマーカータンパク質である CD63 にグラフトし、標的細胞への送達や取り込み促進を達成することを目的とした。

概要

エクソソームは、細胞から放出される 50nm から 150nm の細胞外小胞の一種である。エクソソームは、マイクロ RNA やタンパク質などを内包しており、細胞間コミュニケーションツールとしての役割を果たしており、がん診断や薬物送達キャリアとしても注目されている。一方で、薬物キャリアとして活用するには、エクソソームを標的細胞に特異的に送達する必要がある。そこで、エクソソームの表面に機能性ペプチドを導入することで、新たな機能性エクソソームの創製を目指した。各種機能性ペプチドを CD63 の細胞外ループにグラフトすることで、標的化や取り込み促進が可能かどうか検討した。

結果および考察

CD63 の細胞外ループを 157-164、165-172、173-180 の 3 つの領域に分割して、それぞれの配列と

機能性ペプチドのモデル配列として FLAG tag を入れ替え、AMBER によって構造の揺らぎ計算を実施した（図 1）。その結果、CD63 の細胞外ループ領域はどのペプチド配列でも揺らぎはほとんどないと算出された。次に、FLAG をグラフトした CD63 を細胞に発現させた。その結果、165-172 にグラフトした際に、発現パターンが野生型の CD63 と比較して大きく変化することが明らかとなった。この原因として、ループ中にシステインが含まれており、ジスルフィド結合に大きな影響を与えることが原因であると考えられた。よって、ループ 157-164 および 173-180 の 2 箇所について、それぞれ標的化ペプチドをグラフトして標的タンパク質への結合を評価した。その結果、標的タンパク質への結合は確認できなかった。

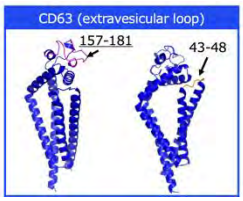


図 1. CD63 の細胞外ループ領域.

まとめ、今後の課題

CD63 に機能性ペプチドをグラフトし、構造の揺らぎを計算したところ変化はなかったことから、エントロピーロスなく機能性ペプチドを提示できる可能性が示唆された。揺らぎが小さく、発現パターンが維持される領域に対して標的化ペプチドをグラフトしたところ、標的タンパク質への結合は確認できなかったため、ある程度の揺らぎがあることが適切な結合モードを取ることに重要である可能性が考えられた。今後は、このエントロピーロスのない状態で結合可能な配列の探索を display screening を活用して行い、標的化を実現する。

東京科学大学 TSUBAME 共同利用
令和６年度利用終了課題 利用成果報告書集

発行 ： 令和８年 ７月
国立大学法人 東京科学大学 情報基盤センター 共同利用支援室
住所 ： 〒152-8550 東京都目黒区大岡山 2-12-1 I8-13
E-mail: tsubame-kyodo@cii.isct.ac.jp
URL: https://www.t4.cii.isct.ac.jp/tsubame-kyodo