Shun Hing Institute of Advanced Engineering
信興高等工程研究所

Report and Research Highlights
2024 - 2025

September 2025

香 港 中 文 大 學
The Chinese University of Hong Kong

Printed in September 2025, this report contains information known as at June 2025.

For further information, please visit our website: http://www.shiae.cuhk.edu.hk

Contents

Introduction of SHIAE

Mission of SHIAE

The MISSION of the Institute is to spearhead, conduct, promote and co-ordinate research in advanced engineering. There is no end to the list of areas to be explored and the plan is to give priority to research topics that are both exciting and innovative. The Institute also aspires to transferring its research results to industry for practical application and to put across to the community at large the role of engineering as a driving force for human development through educational activities.

As a pioneering institute exploring the forefront of the engineering science, The Shun Hing Institute of Advanced Engineering will

- spearhead state-of-the-art advanced engineering research
- create and sustain synergy with world-class researchers
- develop with and transfer to industries cutting edge technologies
- promote appreciation of engineering in society through educational programmes

The Shun Hing Education and Charity Fund was founded the late by Dr. William Mong Man Wai with the aim of enhancing educational opportunities for the younger generations. The Fund has already sponsored numerous educational and research programmes in Hong Kong, the Mainland, and overseas educational institutions. Himself an engineer and a firm believer in advancing the quality of life through the development of science and technology, Dr. Mong had been there to support the establishment and growth of this Institute from the beginning.

Centre of Excellence at CUHK

The Chinese University of Hong Kong (CUHK) is an internationally renowned institution of higher learning devoted to quality teaching and both academic and applied research. The University has established 29 research institutes and a number of research centres with a view to pursuing up-front research endeavours with focused goals and objectives. The Shun Hing Institute of Advanced Engineering was established in 2004 with the generous support of the late Dr. Willian Mong. It plays a crucial part in the research infrastructure of CUHK, particularly for the Faculty of Engineering, which is committed to the development of the state-of-the-art technologies in various advanced engineering areas. The Institute is now progressing to the

third decade of its milestone, and we are extremely grateful to have received continual staunch support and guidance from Mr. David Mong Tak Yeung, Chairman and CEO of the Shun Hing Group and the Shun Hing Education and Charity Fund.

As a strategic centre of excellence at CUHK, the Institute supports greater regional and international research collaborations, and strives to attract talent from the world over to achieve greater internationalization, a vision strongly advocated by every member of the University.

Commitment of the Faculty of Engineering

The Faculty of Engineering was founded in 1991 and was built upon existing strengths with added talent from all over the world. The Faculty has been able to attract some of the best minds. Many received their training in leading universities in North America, Great Britain and Australia. Most of them have extensive experience in industry and many are leaders in their fields. This team of top-notch talent is gathered to nurture local talent through educational programmes, and break new frontiers in research through innovative and exciting research endeavours.

The positioning of The Shun Hing Institute of Advanced Engineering in the William M.W. Mong Engineering Building is deliberate as a key nucleating point to integrate research endeavours in the Engineering Faculty and its neighbours. Our members join hands with their counterparts from the Faculties of Science and Medicine in many interesting research collaborations. It is the ambitious goal of the Faculty of Engineering that the Institute should become a lighthouse for the local technology landscape to herald the migration towards high value-added technology and an information economy.

The mission of the Institute is to spearhead, conduct, promote and co-ordinate research in advanced engineering. There is no end to the list of areas to be explored and the plan is to give priority to research topics that are both exciting and innovative. The Institute also aspires to transferring its research results to industry for practical application and to put across to the community at large the role of engineering as a driving force for human development through educational activities.

Building on Strength and The Way Ahead

Many of the Institute's research projects are built upon areas in which the Faculty has already achieved outstanding performance. These are areas that have great potential for further technological advancement and in line with industrial development in Hong Kong. The Institute provides a vibrant R&D environment to spur new discoveries and speed up their translation into applications. Since 2012, we have expanded our scope to cover new frontiers in Renewable Energy striving to answer tomorrow's energy challenges. In year 2017, we further expand the research scope in Multimedia Technologies to include Artificial Intelligence, Big Data Analytics and Deep Learning as well. In the past few years, the Engineering Faculty has recruited many young and talented researchers, and the Institute has given priority to provide them with the needed research support as far as possible.

Technology Transfer

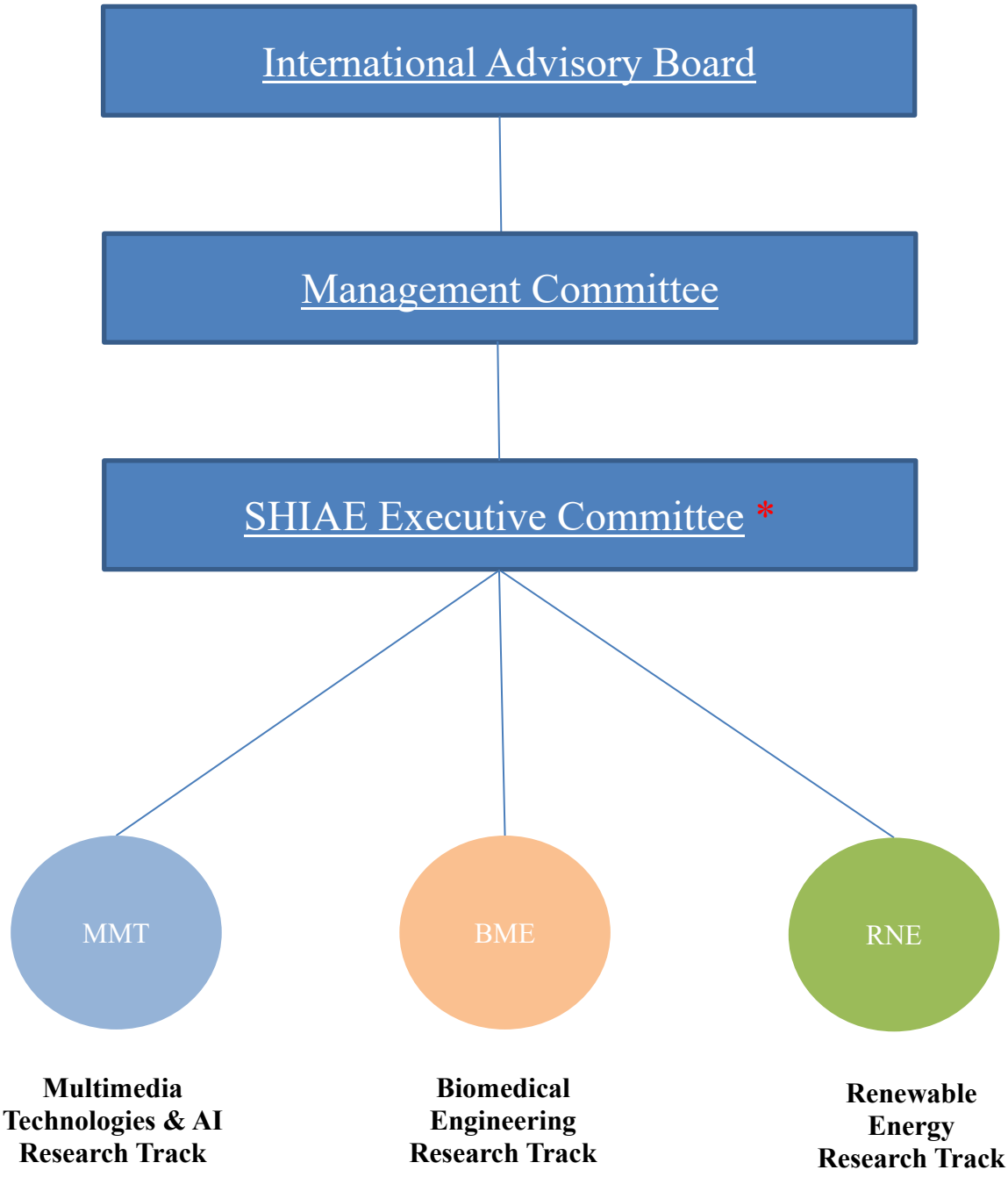
Synergy with industry is the ultimate goal of research and development in Hong Kong. External experts have been brought in to the Institute to lead research projects that could benefit the industrial sector.

The technology transfer arm of the Faculty of Engineering plays an important role in the traffic between the Institute and industry. The Institute houses an array of top-notch research and development activities encompassing contract research, spin-off companies, and consultancies.

Contribution to Society

The Institute has been making contributions to the progress of Hong Kong through a wide range of educational activities like training courses, seminars, symposiums which disseminate the latest technologies to promote appreciation of engineering in society and arouse interest of the younger generations in engineering.

Organization



We also provide support and sponsorship to the Faculty of Engineering in organizing prestigious academic conferences in Hong Kong so as to raise our international profile.

* In compliance with CUHK's guidelines in strengthening the governance of research units, an Executive Committee was formed to oversee the daily operation of the Institute in April 2020, headed by the Director, while the Dean of Engineering served as the Chairman of the Management Committee.

Composition of International Advisory Board

Chairman:

Dr. David T.Y. MONG 蒙德揚先生

Chairman & Group CEO
Shun Hing Electronic Holdings Limited
Hong Kong

Members:

Professor Victor ZUE

Delta Electronics Professor of Electrical Engineering and Computer Science Massachusetts Institute of Technology
U.S.A

Professor Zhi DING

Distinguished Professor
Department of Electronical and Computer Engineering
University of California
U.S.A.

Professor C.C. Jay KUO

Professor of Electrical Engineering and Computer Science
Viterbi School of Engineering University of Southern California
U.S.A.

Professor Kin LEUNG

Tanaka Chair in Internet Technology
Department of Electrical and Electronic Engineering
Imperial College London
United Kingdom

Professor Wai-yee CHAN 陳偉儀教授

Research Professor
School of Biomedical Sciences
The Chinese University of Hong Kong
Hong Kong

Professor Wing-shing WONG 黃永成教授

Emeritus Professor
Department of Information Engineering
The Chinese University of Hong Kong
Hong Kong

Professor TSANG, Hon Ki 曾漢奇教授

Dean of Engineering
The Chinese University of Hong Kong
Hong Kong

Professor Pak Chung CHING 程伯中教授

Director of Shun Hing Institute of Advanced Engineering
Research Professor of Electronic Engineering
The Chinese University of Hong Kong
Hong Kong

Composition of Management Committee

Chairman: **Professor TSANG, Hon Ki**
Dean of Engineering

Deputy Chairman: **Professor Pak Chung CHING**
Director, Shun Hing Institute of Advanced Engineering (ex-officio)

Members:

Mr. Gary NG
Managing Director of Shun Hing Technology Co., Limited

Professor Tan LEE
Department of Electronic Engineering

Professor Wei-Hsin LIAO
Department of Mechanical and Automation Engineering

Professor Soung-chang LIEW
Department of Information Engineering

Professor Anthony Man-cho SO
Department of Systems Engineering and Engineering Management

Professor Raymond Kai-yu TONG
Department of Biomedical Engineering

Professor Benny C.Y. ZEE
Director, Office of Research and Knowledge Transfer Services

Member and Secretary: **Professor John C.S. LUI**
Department of Computer Science and Engineering

Composition of Executive Committee

(*with effect from March 20, 2025*)

Chairman: **Director, Shun Hing Institute of Advanced Engineering** ***(ex officio)***
Professor Pak Chung CHING

Members: **Professor Jonathan Chung Hang CHOI**
Department of Biomedical Engineering

Prof. Alex Ka Nang LEUNG
Department of Electronic Engineering

Professor John C.S. LUI
Department of Computer Science and Engineering

Professor Li ZHANG
Department of Mechanical and Automation Engineering

Secretary: **Miss Kaia LI**
Shun Hing Institute of Advanced Engineering

Shun Hing Fellows and Research Associate

The Institute has launched a Shun Hing Distinguished Scholar Program with an aim to attract distinguished scholars to pursue research collaboration with our faculty and to strengthen our research profile.

Dr. Weidong LI 2025
The Chinese University of Hong Kong, Hong Kong SAR

Dr. Han HUANG 2025
Beihang University, Beijing

Dr. Qi QIAN 2024-2025
National University of Singapore, Singapore

Dr. Junning QIAN 2023-2025
The Chinese University of Hong Kong, Hong Kong SAR

Dr. Rui XIE 2023-2025
Tsinghua University, China

Dr. Haoran LIU 2024
Xi'an Jiaotong University, China

Dr. Wen LYU 2024
Tsinghua University, China

Dr. Haoming LIU 2023-2024
Peking University, China

Dr. Xiangguo SUN 2023-2024
The Chinese University of Hong Kong, Hong Kong

Dr. Rui ZHANG 2023-2024
The Chinese University of Hong Kong, Hong Kong

Mr. Bo WANG 2023-2024
Beijing University of Posts and Telecommunications

Financial Status of SHIAE

INCOME AND EXPENDITURE STATEMENT 2024-2025

(Fiscal Year: April 1, 2024 – March 31, 2025)

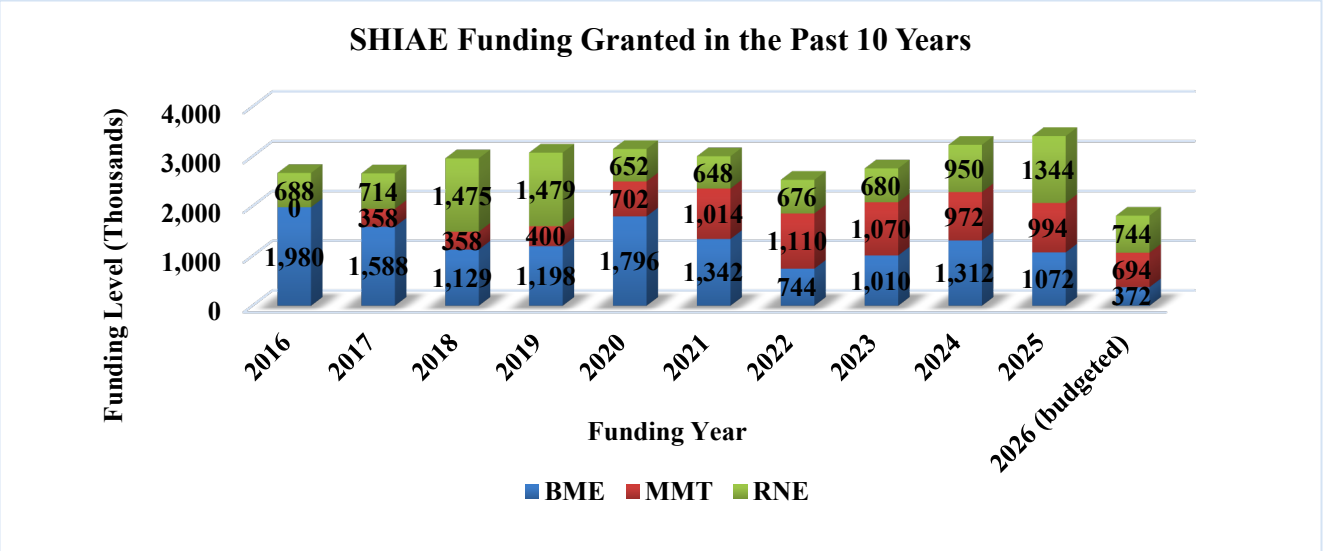
		Notes	As at 31 March 2025	As at March 31, 2024
INCOME				
Funding Source				61,000,000
Accumulated fund brought forward			9,409,963	-
Interest and investment income			401,382	8,317,382
	Sub-total:		9,811,345	69,317,382
EXPENDITURE				
Research Funding		(1)	3,234,000	58,132,200
Remaining fund from completed projects			(303,670)	(4,975,091)
Operating cost			89,572	6,750,310
	Sub-total:		3,019,902	59,907,419
BALANCE as at 31 March 2025			**6,791,443**	**9,409,963**

APPROVED BUDGET 2025-2026

(Fiscal Year: April 1, 2025 – March 31, 2026)

		Notes	
INCOME			
Accumulated fund brought forward			6,791,443
Projected interest and investment income			290,000
New donation - 2nd biannual installment on April 1, 2025			6,500,000
	Sub-total:		13,581,443
EXPENDITURE			
Research Funding			
On-going projects (Year 2024 batch)		(2)	1,600,000
Newly funded projects (Year 2025 batch)		(2)	1,810,000
Operating cost			
Staff and Admin. Cost			73,000
Office Expenses			10,000
Distinguished lectures			10,000
Activities Sponsorship			100,000
	Sub-total:		3,603,000
Projected Balance in March 2026			**9,978,443**

Note (1) Annualized Research Funding to each research areas granted in the past ten years.



This figure shows the distribution of the SHIAE funding granted to each track of research projects, namely Biomedical Engineering (BME), Multimedia Technology & AI (MMT) and Renewable Energy (RNE) annually.

Note (2) Total funding for each batch of projects (in HK$ '000)

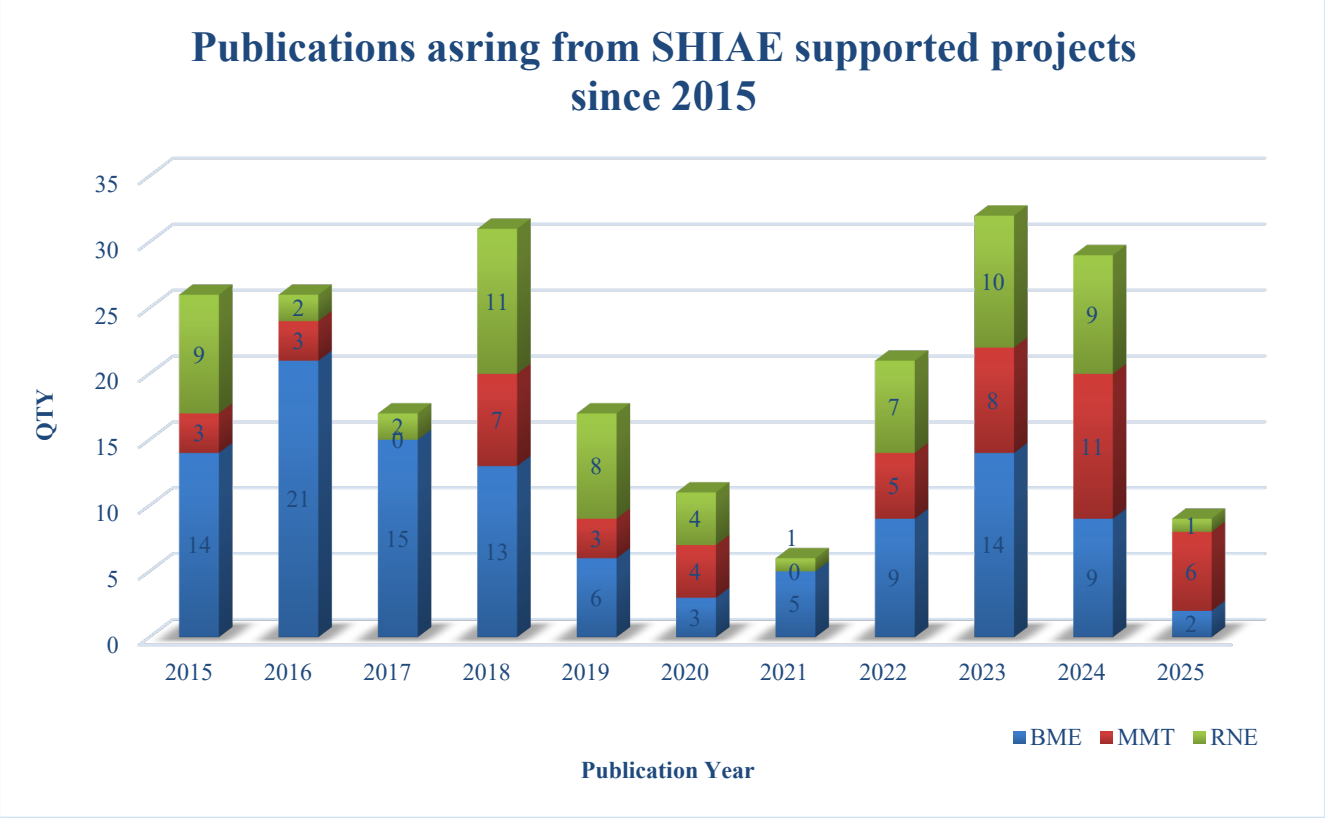
Funding Year / (No. of projects)	2026 (committed)	2025	2024	2023	2022	2021	2005 - 2020
Year 2005 / (6)	-	-	-	-	-	-	6,108
Year 2006 / (5)	-	-	-	-	-	-	3,175
Year 2007 / (7)	-	-	-	-	-	-	4,146
Year 2008 / (4)	-	-	-	-	-	-	3,976
Year 2009 / (5)	-	-	-	-	-	-	3,306
Year 2010 / (5)	-	-	-	-	-	-	2,789.2
Year 2011 / (4)	-	-	-	-	-	-	2,476
Year 2012 / (5)	-	-	-	-	-	-	3,040
Year 2013 / (4)	-	-	-	-	-	-	2,948
Year 2014 / (3)	-	-	-	-	-	-	2,004
Year 2015 / (4)	-	-	-	-	-	-	2,656
Year 2016 / (4)	-	-	-	-	-	-	1,340
Year 2017 / (4)	-	-	-	-	-	-	2,660
Year 2018 / (4)	-	-	-	-	-	-	2,962
Year 2019 / (4)	-	-	-	-	-	-	3,077
Year 2020 / (5)	-	-	-	-	-	1,600	3,150
Year 2021 / (5)	-	-	-	-	1,404	1,404	
Year 2022 / (3)	-	-	-	1,126	1,126	-	
Year 2023 / (5)	-	-	1,634	1,634	-	-	
Year 2024 / (5)	-	1,600	1,600	-	-	-	
Year 2025 / (5)	1,810	1,810					
WOSP2007	-	-	-	-	-	-	25
	1,810	3,410	3,234	2,760	2,530	3,004	49,838.2

Accumulated Total: **HK$66,586.200**

Research Reports and Highlights

Academic Publications

So far **79 projects** have been successfully completed and **541 articles** arising from the results of these research projects have been published in international conference proceedings and journals. Many of their academic outcomes have received Best Paper Awards from prestigious journals and top-tier conference with international recognition. The other **10 on-going projects** are also progressing well with encouraging results produced. All publications generated by each individual project are kept in the archive of SHIAE office. The chart below shows the number of academic publications produced from 2014 onward.



Renewable Energy Track

Research Reports (2024-2025) In Renewable Energy

Newly Funded Projects (2025-2027)	* Inverse Design of Battery Cathodes with Advanced Characterization and Generative AI
	* Advancing Flexible Indoor photovoltaics for emerging autonomous sensor and consumer electronic applications (ISense)
Continuing Projects (2024-2026)	* Segmented High-entropy Thermoelectric Materials for Geothermal Heat Harvesting
	* Development of Carbon-14 Detection System at Part-per-Quadrillion Level Based On Doubly Resonant Photoacoustic Spectroscopy
(2023-2025)	* A Hierarchical Carbon-Centric Management System for Energy Storage-Assisted Data Center

Completed Projects (2022-2024) *(Funded Year)*	* Metasurface Based All-optical CNN for Real-time and Power - Efficient Machine Vision

INVERSE DESIGN OF BATTERY CATHODES WITH ADVANCED CHARACTERIZATION AND GENERATIVE AI

Principal Investigator: Professor Jizhou LI
Department of Electronic Engineering CUHK

Project Start Date: 1 July 2025

ABSTRACT

Lithium-ion batteries power modern life through applications ranging from portable electronics to electric vehicles. However, optimizing their performance involves navigating critical trade-offs among energy density, safety, and lifespan. These competing priorities arise from complex relationships between manufacturing parameters, cathode microstructure and chemistry, and overall battery performance. This project proposes to integrate multiscale battery characterization with generative AI to facilitate cathode materials design. We will develop interpretable AI models to reveal how manufacturing parameters, like material mixtures or microstructural features, shape a cathode's structure and chemistry. By training these models on experimental data from synchrotron X-rays, ultrasound imaging, and electron microscopy, the model will predict optimal manufacturing conditions to maximize the desired performance. For example, the AI could suggest how to adjust electrode fabrication to create ideal pore structures for faster charging or stable chemical interfaces for longer life. This "inverse design" approach, starting from desired performance metrics and working backward to identify manufacturing parameters, aims to accelerate the development of high-performance batteries while reducing trial-and-error experimentation. By bridging the gap between materials science and AI, this project could pave the way for next-generation batteries that are safer, more powerful, and longer-lasting, benefiting consumers and industries worldwide.

INNOVATION AND PRACTICAL SIGNIFICANCE:

Traditional forward cathode design, while historically successful, relies heavily on iterative and trial-and-error experimentation. This is a slow and resource-intensive process that struggles to unravel the interdependencies between material morphology, chemical stability, and electrochemical performance in lithium-ion batteries. This conventional approach often fails to efficiently optimize cathode materials, limiting the pace of innovation in energy storage technologies. For instance, it took over three decades to achieve approximately a threefold increase in energy density between 1990 and 2024. Our approach disrupts this paradigm by merging generative AI with advanced multiscale characterization techniques to establish an inverse design framework.

Unlike conventional methods, which incrementally adjust manufacturing variables to observe outcomes, our AI models learn the hidden relationships between production parameters, such as sintering conditions, binder formulations, and particle geometries, and their ultimate impact on battery performance. This enables the prediction and design of tailored microstructures that minimize degradation during charge cycles, improve energy density, and enhance overall battery longevity. By leveraging comprehensive imaging datasets and generative AI models, this framework provides a systematic and data-driven pathway to optimize cathode design.

Practically, this approach empowers manufacturers to rapidly prototype cathodes that meet specific performance goals, such as balancing energy density with safety or enabling fast charging without compromising cycle life. These capabilities address key barriers to widespread electrification, particularly in industries like electric vehicles and grid storage, where cost, performance, and reliability are critical. By

significantly reducing the time and cost associated with traditional experimental methods, this project is expected to lower battery production costs while enhancing energy density and safety, directly supporting global decarbonization goals.

By bridging AI and materials science, much like the successful integration of AI and biology exemplified by AlphaFold, our work not only can potentially be incorporated into industrial manufacturing but also brings insightful considerations for the establishment of a scalable blueprint for sustainable energy technologies, aligning with global priorities for climate resilience and equitable energy access.

PROJECT OBJECTIVES:

1) Build comprehensive multiscale characterization dataset for solid-state batteries: This objective aims to systematically gather high-resolution structural, chemical, and electrochemical data related to various manufacturing parameters, cathode microstructure, and chemistry across multiple scales, from atomic to macroscopic levels, which is not available in the literature. Utilizing advanced techniques such as synchrotron X-ray imaging, electron microscopy, and ultrasound, we seek to create a unified dataset that captures detailed information about solid-state battery materials and interfaces. By integrating diverse data types, we can more accurately capture the dynamics and responses during battery operation, thereby providing a robust foundation for training AI models and enhancing the understanding of solid-state battery performance and reliability.

2) Develop interpretable AI models to establish inverse design framework: This objective focuses on developing generative AI models that learn the causal relationships between manufacturing parameters (e.g., particle morphology, size distributions), heterogeneous microstructure (e.g. connectivity, interfaces) and cell performance (e.g., cycle life, energy density). The goal is to enable the inverse prediction of optimal manufacturing protocols tailored to specific performance metrics. The significance of this work lies in pioneering interpretable AI in battery materials, fostering collaboration between AI experts and materials engineers. The value includes reducing R&D cycles by predicting viable cathode designs without costly trial-and-error, ultimately lowering production costs and time-to-market.

3) Validate and optimize AI predictions experimentally: This objective aims to synthesize and test AIproposed cathode designs under real-world conditions, refining model accuracy and ensuring scalability for industrial adoption. We will establish a closed-loop feedback system where experimental data continuously informs and improves the AI models. This approach ensures that the inverse design framework is both scientifically robust and practically applicable in industrial settings, driving adoption by battery manufacturers to meet decarbonization targets effectively.

ADVANCING FLEXIBLE INDOOR PHOTOVOLTAICS FOR EMERGING AUTONOMOUS SENSOR AND CONSUMER ELECTRONIC APPLICATIONS (ISENSE)

Principal Investigator: Professor Martin STOLTERFOHT
Department of Electronic Engineering CUHK

Co-investigator(s): Professor Ni ZHAO [1]

[1] Department of Electronic Engineering CUHK

Project Start Date: 1 July 2025

ABSTRACT

Humanity is entering an era of the Internet of Things (IoT) - a vast ecosystem of interconnected devices that gather and exchange data collectively. For now, most IoT end consumer electronic devices rely on batteries, which increases maintenance costs and disrupts operations. Indoor photovoltaics (iPVs), however, offer a promising solution for reliable device operation while enabling seamless device integration. Perovskites already demonstrate considerable advantages over commonly used a-Si iPV devices, however, their photovoltaic performance (~35%) is currently far below the maximum potential efficiency (~57%) due to losses in wide-bandgap (WBG) solar cells (1.7–2 eV). Moreover, these devices suffer from relatively poor device stability, often related to halide segregation, necessitating novel approaches to prolong their lifetime. ISense will advance the development of phase-stable WBG perovskite semiconductors through fundamental research of recombination and degradation losses under low-illumination conditions and implementation of state-of-the-art charge transport materials tailored for integration into flexible large area devices. Moreover, we explore other photoactive layers, including organic semiconductors, and optimize the devices for efficient light harvesting under both indoor and outdoor conditions using 4-terminal tandem solar cells, thereby broadening the application potential of our devices. Finally, we will demonstrate their integration into state-of-the-art autonomous health sensors.

INNOVATION AND PRACTICAL SIGNIFICANCE:

While deeply embedded in the consumer electronics sector, IoT reaches well beyond smartphones and household gadgets. In fact, IoT is already driving advancements in areas such as e-health (e.g., wearable and implantable smart devices), e-energy smart buildings (e.g., smart windows), smart cities, transport, manufacturing and more. There are virtually endless possibilities for IoT applications, for private and corporate markets. Many of these devices would greatly benefit from solutions offering minimal maintenance. Thus, the development of indoor PV devices is of great practical significance.

Perovskite photovoltaic cells provide many advantages over amorphous silicon (a-Si) cells that are typically used for IoT applications, including higher efficiencies under indoor light, tunable bandgaps enabling a wide range of applications, exceptional power-to-weight ratios (>40 W/g) due to ultrathin active layers enabling lightweight sensor applications, and high open-circuit voltages that are beneficial for powering low-energy devices. In this project, we will leverage our expertise in development and characterization of recombination and degradation losses in high performance (WBG) perovskite devices. We will do so through fundamental research of recombination and ion-induced degradation pathways aiming for longer lifetimes. Furthermore,

ongoing collaborations with leading chemists in the development of next generation transport materials (Prof. Wu, ECUST) will allow us to overcome critical challenges with regard to the integration of perovskite semiconductors onto flexible substrates. ISense also aims to advance devices that are currently used in autonomous consumer electronics via implementation of tandem cells that allow not only to harvest outdoor light more efficiently but also maintaining a much better performance under indoor illumination conditions. Finally, internal collaboration with Prof. Zhao will facilitate the seamless integration of iPV cells into state-of-the-art health sensors. Lastly, the exploration of a wider class of materials will provide a platform for the development of next generation solar cells for emerging sensor applications.

PROJECT OBJECTIVES:

1. Optimize the performance and stability of wide-bandgap (WBG) perovskites under 1000-lux standardized white LED illumination with a power conversion efficiency (PCE) of 38%.

2. Demonstrate devices on flexible substrates with small performance losses (<20% relative).

3. Demonstrate semitransparent WBG perovskite cells and 4-terminal tandem solar cells (4T-TSCs) with Si with a PCE >30% under 1-sun conditions. Evaluate the performance of the 4T-TSCs under indoor and outdoor conditions and compare the results to the single-junction cells.

4. Compare and evaluate different technologies based on perovskite, organic semiconductors and dye-sensitized materials aiming to reach an indoor PCE >30% for all technologies

5. Demonstrate the application of flexible iPV cells in autonomous IoT and state-of-the-art health sensors.

WORK IN PROGRESS

SEGMENTED HIGH-ENTROPY THERMOELECTRIC MATERIALS FOR GEOTHERMAL HEAT HARVESTING

Principal Investigator: Professor Ady SUWARDI
Department of Electronic Engineering and Shun Hing Institute of Advanced Engineering, CUHK

Research Team Members:
Dr. Qi QIAN [1]
[1] Department of Electronic Engineering and Shun Hing Institute of Advanced Engineering, CUHK

Reporting Period: 1 July 2024 – 30 April 2025
(to be completed in June 2026)

INNOVATION AND PRACTICAL SIGNIFICANCE:

Engineering an efficient thermoelectric device starts with innovative material. In this project, we intend to study and develop a "winning" material by combining materials segmentation approach together with high-entropy strategy. This strategy is expected to not only improve the properties of a single material, but also result in higher overall efficiency of the combined segments. In terms of practical applications, the use of thermoelectrics for low-grade (up to 200 °C) heat harvesting can help to increase energy efficiency and alleviate the high carbon emissions. In addition, the nature of the proposed project is highly relevant to the abundance of geothermal resources across greater China and the rest of the regions. Therefore, more research efforts should be invested to fully utilize them. In terms of technology adoption, the development of optimized high-entropy material compositions and segmentation conditions does not involve significant change in processing conditions. Therefore, the strategy can easily be adopted by thermoelectric material manufacturers. On a more immediate term, the findings from this project will serve as a seed for a more downstream project for prototype development, for example via innovation and technology fund.

ABSTRACT

Thermoelectric material is one of the most efficient technologies to harvest electricity from low grade (<200°C) heat sources, such as geothermal and manufacturing plants. To achieve high efficiency, high performance materials are desirable, especially between room temperature to 200°C. Specifically, with an average figure of merit (zT) of 1.5, efficiency > 10% can be realized. While the current stateof-the-art room temperature (25°C) zT is around 1.5, it drastically decreases as temperature approaches 200°C. On the other hand, while there are materials with zT > 1.5 at 200°C, they have low zT at room temperature. One intuitive approach is to combine materials with high performance at the respective temperatures into segmented legs. However, the fundamental transport properties are far from straightforward. To overcome this, we propose to study the fundamental electronic and thermal transports of segmented thermoelectrics. We intend to approach this via high-entropy alloying using Bi2Te3 and GeTe-based alloys as the candidate materials due to their intrinsically high performance at room temperature and also up to 200°C. We hypothesize that such segmentation approach will potentially enable average zT of 1.5 and beyond. The insights derived from this study will bring the community a step closer towards realizing efficient thermoelectric devices.

1. OBJECTIVES AND SIGNIFICANCE

Engineering an efficient thermoelectric device starts with innovative material. In this project, we intend to study and develop a "winning" material by combining materials segmentation approach together with high-entropy strategy. This strategy is expected to not only improve the properties of a single material, but also result in higher overall efficiency of the combined segments. In terms of practical applications, the use of

thermoelectrics for low-grade (up to 200 °C) heat harvesting can help to increase energy efficiency and alleviate the high carbon emissions. In addition, the nature of the proposed project is highly relevant to the abundance of geothermal resources across greater China and the rest of the regions. Therefore, more research efforts should be invested to fully utilize them. In terms of technology adoption, the development of optimized high-entropy material compositions and segmentation conditions does not involve significant change in processing conditions. Thus, the strategy can easily be adopted by thermoelectric material manufacturers. On a more immediate term, the findings from this project will serve as a seed for a more downstream project for prototype development, for example via innovation and technology fund.

2. RESEARCH METHODOLOGY

2.1. Phase 1: Synthesis and properties investigation of high-entropy material composition

Excessive atomic substitution of common TE materials to form high entropy alloys have gained increasing interests in the TE research community.[1, 2] Specifically, such entropy-driven alloying strategies have transformed lower symmetry TE materials such as SnSe, GeTe, and MnTe into high symmetry cubic structures,[2-5] where the enhanced long-range order is beneficial to their electronic band structures and electrical transport properties. In addition, detrimental phase transitions and secondary phase impurities are often eliminated in the process, leading to an overall higher material quality. Furthermore, the resulting short-range disorder effectively reduces the intrinsic lattice thermal conductivity, which can potentially allow the material to still achieve a low thermal conductivity in its single-crystalline or large-grained form (i.e. from processing by melting techniques), without relying on other microstructural defects that are usually responsible for the low thermal conductivities and higher performances in polycrystalline TE materials.[6, 7]

Therefore, for Phase 1 of the project, we aim to investigate the material composition through small batch melting methods to obtain a high entropy composite with a low thermal conductivity of < 1 W/m·K at room temperature that is comparable to those of state-of-the-art high entropy polycrystalline TE materials processed by powderization-sintering (**Figure** 1a-b) A low thermal conductivity of the TE material is crucial to maintaining the temperature difference for maximum device efficiency.

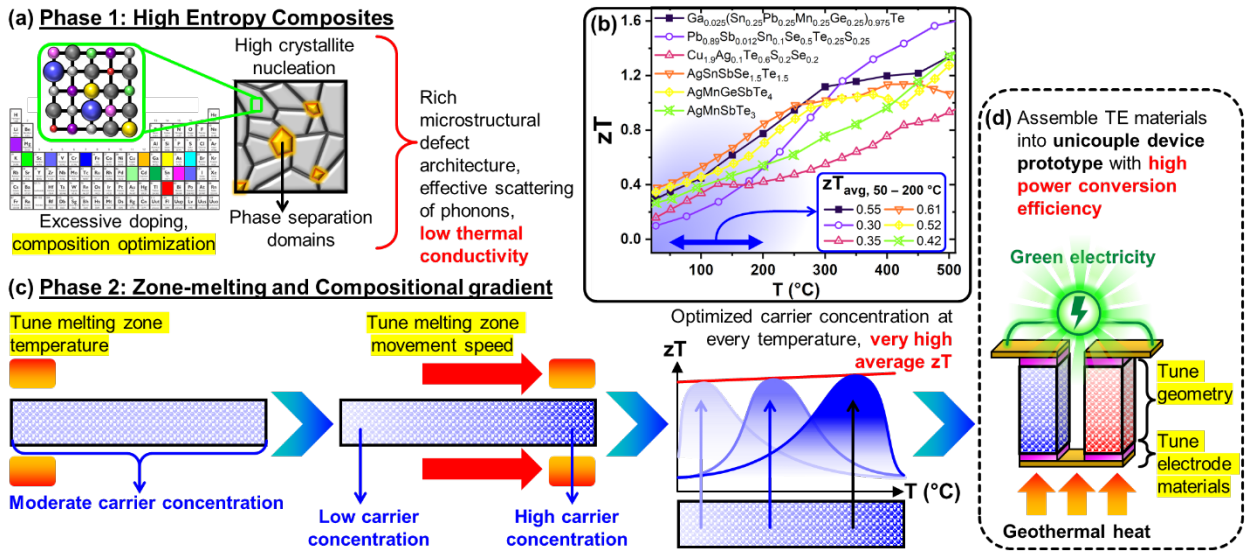


Figure 1: (a) High-entropy composites with low thermal conductivities (Phase 1). **(b)** Reported zT's and average zT's across the 50 – 200 °C range.[20, 25-29] **(c)** Tuning of zone-melting conditions to produce segmented high-entropy composites with high average zT across the 50 – 200 °C range (Phase 2).**(d)** device assembly.

2.2. Phase 2: Investigation of segmented thermoelectrics for high-efficiency prototype

Currently, most reported high entropy alloys processed by powderization-sintering can easily achieve low thermal conductivities of <1 W/mK at room temperature, which continues to decrease with temperature. However, their zT's generally have an increasing trend with temperature, with high zT's at T > 500 °C, but average zT's that are < 0.7 at the typical temperatures of interest for geothermal heat (50 − 200 °C). To address

the other issue of generally low average zT's over the entire temperature range due to the unoptimized carrier concentration at each temperature zone, we also propose an unorthodox gradient doping to obtain segmented high-entropy zone-melted ingots, where the carrier concentration varies across the length of the material. When a temperature gradient is applied across the length, each temperature segment will have a carrier concentration that is optimized for that particular temperature, thus improving the average zT over the entire temperature range.

Most reports on the zone-melting process focus on concentrating impurity atoms to the ends of the zone melted ingot for removal later, usually by slowing down the melting zone movement or increasing the number of zone melting passes, which will lead to slower processing times.[8] However, if the goal is to obtain a gradient distribution of dopants instead, it is possible to achieve this by doing the opposite (less zone passes and faster zone movement), potentially leading to the development of a process that is not only faster but will also produce a material with higher performance. Generally, it was also found that faster zone movement speeds and lower zone heating temperatures can lead to smaller grain sizes.[9] For alloys containing multiple elements, it is also possible to vary the heat treatment conditions to control the phase separation domains, leading to the formation of composites with different regions of stoichiometric variations, which can further improve zT.[10, 11]

Therefore, for Phase 2 of the project, we aim to develop an optimized processing conditions to achieve a high average zT of >1.5 over a temperature range of 50 – 200 °C (**Figure** 1c-d). This phase of the project will build upon the previous phase by directly taking the optimized small batch melted sample and further process it by zone-melting. The successful completion of the project will pave the way for the production of high performance segmented high-entropy materials that can be obtained with minor adjustments to commercial manufacturing processes. As a proof of concept, we also aim to optimize the assembly of the materials to make a unicouple TE device prototype with a power conversion efficiency of >10%. This will involve tuning the unicouple geometry and electrode materials to maximize heat transfer and minimize contact resistance.

3. RESULTS ACHIEVED SO FAR

The original project objectives, as outlined in the proposal, are twofold:

1. Work Package 1 (WP1): Develop a high-entropy thermoelectric composite with a thermal conductivity < 1 W/m·K at room temperature, comparable to state-of-the-art polycrystalline materials processed via powderization-sintering, targeted for completion by Year 2, Q2 (mid-2025).

2. Work Package 2 (WP2): Achieve a zone-melted functionally-graded high-entropy thermoelectric composite (ZFHETEC) with an average figure of merit (zT) > 1.5 over 50–200°C, and fabricate a prototype with >10% energy conversion efficiency by Year 3, Q4 (end-2025).

As of mid-2025 (Year 2, Q2), we have made significant progress toward WP1's deliverable. We successfully optimized the composition of an Mg3Sb2-based high-entropy alloy, achieving a thermal conductivity of 0.85 W/m·K at room temperature, meeting the target of < 1 W/m·K. This milestone aligns with the timeline in Section 6 of the proposal. The material was synthesized using small-batch melting techniques, followed by detailed characterization of its electronic and thermal transport properties. Preliminary results indicate a zT of 1.2 at room temperature, which is promising but requires further enhancement to meet the ultimate goal of an average zT > 1.5 across the target temperature range.

For WP2, we have initiated optimization of zone-melting conditions using the WP1-optimized composition. Early trials have produced ingots with a gradient in carrier concentration, and we have achieved an average zT of 1.1 over 50–200°C in our best sample to date. While this falls short of the final target (zT > 1.5), it demonstrates proof-of-concept for the functionally-graded approach. Prototype development (WP2.2) has not yet commenced, as it is scheduled for Year 3, but groundwork for unicouple design is underway. Overall, we are on track with WP1 and slightly ahead of schedule for initiating WP2, though further refinements are needed to meet the ultimate zT and efficiency goals.

3.1. WP1: Optimization of High-Entropy Material Composition

In terms of material composition, we explored Mg_3Sb_2-based alloys with multi-element doping (e.g., Bi, Te, Zn) to maximize configurational entropy. The optimized composition, $Mg_{2.8}Bi_{0.1}Sb_{1.8}Te_{0.1}Zn_{0.1}$, was synthesized via small-batch melting. The as-synthesized samples subsequently were measured for thermal conductivity using a laser flash diffusivity system, with the **lattice thermal conductivity at 25°C found to be as low as 0.85 W/m·K, decreasing to 0.75 W/m·K at 200°C**, as shown in Figure 2a. This is attributed to enhanced phonon scattering from lattice strain and short-range disorder. Electrical conductivity was determined using ZEM-3, with values as high as 450 S/cm at 25°C, with a Seebeck coefficient of 180 µV/K, yielding a power factor of 14.6 µW/cm·K² and a **zT of 1.1 close to room temperature (Figure 2a)**. At 200°C, zT drops below 1.0 due to carrier concentration mismatch. In terms of microstructure, X-ray diffraction (XRD) and scanning electron microscopy (SEM) confirm a single-phase cubic structure with grain sizes of 10–50 µm (Figure 2b-d), smaller than typical zone-melted ingots, supporting our hypothesis of dopant-induced nucleation sites.

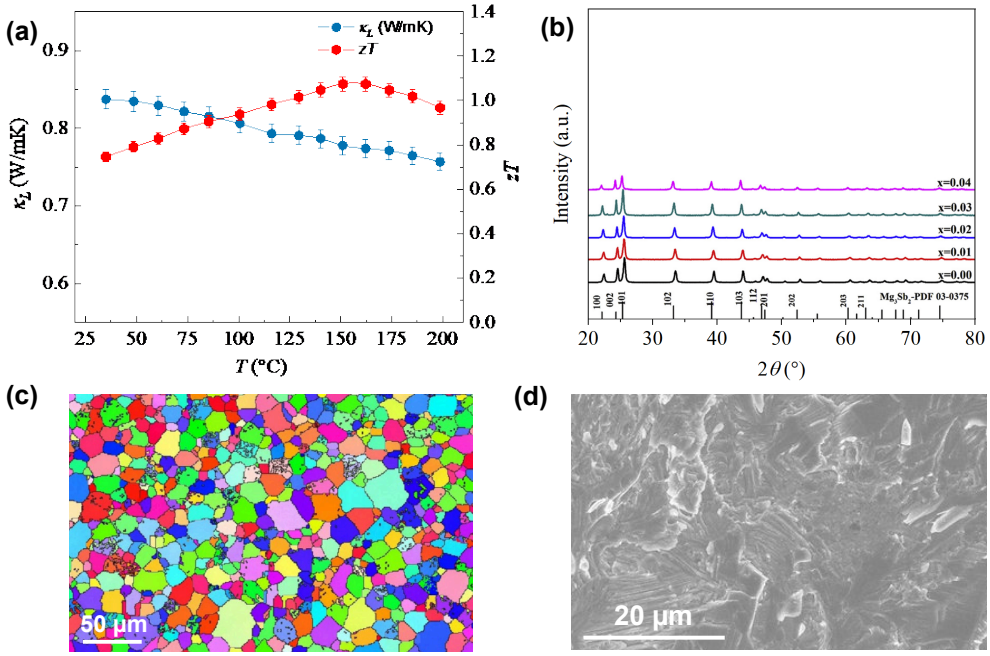


Figure 2. (a) lattice thermal conductivity and zT of the optimized sample so far. (b) XRD (x-ray diffraction) showing single phase of $Mg_{2.8}Bi_{0.1}Sb_{1.8}Te_{0.1}Zn_{0.1}$, with various amount of Se doping. (c) EBSD (electron backscatter diffraction) showing average grain size of around 20 µm. (d) SEM (scanning electron microscopy) image showing homogeneous surface and absence of secondary precipitates, consistent with the single phase shown in the XRD plot. *Unpublished data*

3.2. WP2: Zone-Melting Optimization (Preliminary)

Zone-melting was conducted at a zone speed of 10 mm/h and a temperature of 650°C, producing a 10 cm ingot with a doping gradient (Bi concentration varying from 0.05 to 0.15 along the length). The average zT across 50–200°C is 1.1, with peak zT of 1.3 at 100°C. Thermal conductivity remains low (0.8 W/m·K average), but electrical conductivity varies along the gradient, optimizing zT locally. However, a major technical challenge so far lies in the fact that dopant diffusion during zone-melting is slower than anticipated, limiting the gradient steepness. Adjustments to zone speed and multi-pass strategies are under investigation. These results validate the potential of high-entropy alloys and gradient doping, though further tuning is required to achieve the target zT and efficiency.

4. PUBLICATION AND AWARDS

4.1. Publications

[1] Z. Gong, A. Suwardi, and J. Cao, "Electricity Generation From Ambient Water Evaporation in the Absence of Sunlight via PVA-Based Porous Hydrogels", *Advanced Functional Materials*, Wiley, Germany, published, 2025 (DOI: 10.1002/adfm.202423371).
[2] S. F. D. Solco, K. Saglik, D. Zhang, X. Y. Tan, Q. Zhu, H. Liu, A. Suwardi, and J. Cao, "Thermoelectric performance enhancement of Mg2Si-based silicides synthesized in nitrogen atmosphere", *Materials Research Express*, IOP, United Kingdom, pp. 125505, 2024. (DOI: 10.1088/2053-1591/ad9cf1)

REFERENCES

1. Jiang, B.; Yu, Y.; Cui, J.; Liu, X.; Xie, L.; Liao, J.; Zhang, Q.; Huang, Y.; Ning, S.; Jia, B.; Zhu, B.; Bai, S.; Chen, L.; Pennycook, S. J.; He, J., High-entropy-stabilized chalcogenides with high thermoelectric performance. *Science* **2021,** *371* (6531), 830-834.
2. Jiang, B.; Wang, W.; Liu, S.; Wang, Y.; Wang, C.; Chen, Y.; Xie, L.; Huang, M.; He, J., High figure-of-merit and power generation in high-entropy GeTe-based thermoelectrics. *Science* **2022,** *377* (6602), 208-213.
3. Luo, Y.; Hao, S.; Cai, S.; Slade, T. J.; Luo, Z. Z.; Dravid, V. P.; Wolverton, C.; Yan, Q.; Kanatzidis, M. G., High Thermoelectric Performance in the New Cubic Semiconductor AgSnSbSe3 by High-Entropy Engineering. *Journal of the American Chemical Society* **2020,** *142* (35), 15187-15198.
4. Ma, Z.; Xu, T.; Li, W.; Cheng, Y.; Li, J.; Zhang, D.; Jiang, Q.; Luo, Y.; Yang, J., High Entropy Semiconductor $AgMnGeSbTe_4$ with Desirable Thermoelectric Performance. *Advanced Functional Materials* **2021,** *31* (30), 2103197.
5. Luo, Y.; Xu, T.; Ma, Z.; Zhang, D.; Guo, Z.; Jiang, Q.; Yang, J.; Yan, Q.; Kanatzidis, M. G., Cubic $AgMnSbTe_3$ Semiconductor with a High Thermoelectric Performance. *Journal of the American Chemical Society* **2021,** *143* (34), 13990-13998.
6. Yang, A.; Lin, H.; Chen, D.; Yu, Y.; Wang, G.; Wang, Y., Crystallization mechanism and optical properties of Nd3+ doped chalcohalide glass ceramics. *Materials Research Bulletin* **2012,** *47* (11), 3078-3082.
7. Goncharuk, V.; Mamaev, A.; Silant'ev, V.; Starodubtsev, P.; Maslennikova, I., Nucleation and crystallization behavior of RE - doped tellurite glasses. *IOP Conference Series: Materials Science and Engineering* **2016,** *112* (1), 012023.
8. Wan, H.; Xu, B.; Zhao, J.; Yang, B.; Dai, Y. In *Analysis of the High-Purity Aluminum Purification Process Using Zone-Refining Technique*, TMS 2019 148th Annual Meeting & Exhibition Supplemental Proceedings, Cham, 2019//; Springer International Publishing: Cham, 2019; pp 1697-1706.
9. Chen, Y.-R.; Hwang, W.-S.; Hsieh, H.-L.; Huang, J.-Y.; Huang, T.-K.; Hwang, J.-D., Thermal and microstructure simulation of thermoelectric material Bi2Te3 grown by zone-melting technique. *Journal of Crystal Growth* **2014,** *402*, 273-284.
10. Gelbstein, Y., Phase morphology effects on the thermoelectric properties of Pb0.25Sn0.25Ge0.5Te. *Acta Materialia* **2013,** *61* (5), 1499-1507.
11. Tsai, Y.-F.; Wei, P.-C.; Chang, L.; Wang, K.-K.; Yang, C.-C.; Lai, Y.-C.; Hsing, C.-R.; Wei, C.-M.; He, J.; Snyder, G. J.; Wu, H.-J., Compositional Fluctuations Locked by Athermal Transformation Yielding High Thermoelectric Performance in GeTe. *Advanced Materials* **2021,** *33* (1), 2005612.

WORK IN PROGRESS

DEVELOPMENT OF CARBON-14 DETECTION SYSTEM AT PART-PER-QUADRILLION LEVEL BASED ON DOUBLY RESONANT PHOTOACOUSTIC SPECTROSCOPY

Principal Investigator: Professor Zhen WANG
Department of Mechanical & Automation Engineering
CUHK

Co-Investigator: Professor Wei REN [1]

Research Team Members:
Wenkai LAI, Research Assistant [1]

[1] Dept. of Mechanical and Automation Engineering

Reporting Period: 01 July 2024 – 30 April 2025
(to be completed in June 2026)

INNOVATION AND PRACTICAL SIGNIFICANCE:

Carbon-14 (14C) is a very rare and important carbon isotope with a natural abundance about 1 part-per-trillion (ppt, 0.0000000001%). It is radioactive, decaying with a half-life of about 5730 years which makes it a significant radiolabel in archaeology, oceanography, climate sciences, nuclear plant monitoring and medical science [1-5].

Hong Kong's Climate Action Plan 2050 outlines the strategies and targets for combating climate change and achieving carbon neutrality [6]. Carbon sources apportion is the premise of carbon neutrality and future carbon trading market construction in Greater Bay Area. Quantifying 14CO2 has been proven to be a promising method to measure the CO2 from fossil fuels [7]. The basic principle is that the fossil fuel is totally depleted of 14C because of the much longer existence time. It is a marker to differentiate the fossil fuel emissions from other renewable power sources. A widespread use of biofuels is a key recommendation of the Paris Agreement to tackle global warming issues induced by the greenhouse-gas emissions from fossil fuels. Considering the low concentration of 14CO2 (1.2 ppt) in pure CO2, we need instruments with ultra-high sensitivity to trace and elucidate carbon cycle. To construct a regional carbon emission network, the instrument should have small footprint and affordable price for distributed monitoring.

Additionally, for the purpose of carbon reduction, 10 nuclear power plants in total will be in Guangdong province with distances about 50-400 km away from Hong Kong in the future [8]. 14CO2 artificially generated in nuclear reactors may pose a risk to living organisms if not properly confined. The precise and real-time monitoring of 14CO2 in radioactive waste streams generated during nuclear decommissioning is important for establishing the best-suited nuclear waste management. However, the concentration of 14CO2 may change by orders of magnitude at different locations. For example, the biodegradation of radioactive waste leads to 14CO2 emissions with an activity concentration of 10 ppb to 1 ppm. In this scenario, we need instruments with a wide dynamic range and in-situ measurement.

The radiocarbon dating is a well-established technique in archaeology. In specialized corporations of Hong Kong, radiocarbon dating is one of the key scientific methods for authenticating antiques [9]. The basic principle is the amount of 14C in organisms keeps an equilibrium with living environment and subsequently decreases as a result of radioactive decay after death. The date of death can be determined by measuring the remaining amount of the isotope with extremely low concentration. This radiocarbon dating technique can determine the age of carbon-containing samples up to about 50,000 years old. However, due to the extremely

high price and room size of the professional instrument, called Accelerator Mass Spectrometry (AMS), the dating work needs to be only conducted in a few professional labs outside Hong Kong.
The project aims to develop a 14CO2 sensing instrument which can provide ultra-high sensitivity and wide dynamic range with portable size and low cost. The project will provide the third type of cutting-edge technique world-wide for 14CO2 detection and make breakthroughs in resolving bottlenecks.

[1] Bronk Ramsey, Christopher. "Radiocarbon dating: revolutions in understanding." Archaeometry 50.2 (2008): 249-275.
[2] Shepard, F. P., and J. R. Curray. "Carbon-14 determination of sea level changes in stable areas." Progress in oceanography 4 (1965): 283-291.
[3] Heaton, Timothy J., et al. "Radiocarbon: A key tracer for studying Earth's dynamo, climate system, carbon cycle, and Sun." Science 374.6568 (2021): eabd7096.
[4] Povinec, P. P., et al. "Forty years of atmospheric radiocarbon monitoring around Bohunice nuclear power plant, Slovakia." Journal of Environmental Radioactivity 100.2 (2009): 125-130.
[5] Kratochwil, Nicole A., et al. "Nanotracing and cavity-ring down spectroscopy: A new ultrasensitive approach in large molecule drug disposition studies." PloS one 13.10 (2018): e0205435.
[6] https://www.info.gov.hk/gia/general/202110/08/P2021100800588.htm?fontSize=1
[7] Basu, Sourish, et al. "Estimating US fossil fuel CO2 emissions from measurements of 14C in atmospheric CO2." Proceedings of the National Academy of Sciences 117.24 (2020): 13300-13307.
[8] https://www.rfa.org/cantonese/news/nuke-09152022093318.html
[9] http://www.antiqueauthentication.com/Radiocarbon/

ABSTRACT

As many governments have announced the carbon neutrality goal, carbon sources apportion requires precision instruments to detect radiocarbon, which is the marker to differentiate the fossil fuel and renewable energy source emissions. However, the present commercial instruments based on accelerator mass spectrometry and cavity ring-down spectroscopy have issues of long measurement time, extremely high price and large footprint. To address this issue and achieve on-site and on-line radiocarbon detection, we propose to develop a carbon-14 detection instrument based on the novel doubly resonant photoacoustic spectroscopy. It is the one and only photoacoustic technique which has enough sensitivity for atmospheric radiocarbon detection, and inherits the features of low cost and compact sensing system. With the completion of the project, we expect to deliver a new type of radiocarbon instrument which can fulfill the in-situ distributed measurement requirement of carbon sources apportion, and provide technical support for Hong Kong to become an international city with accurate carbon traceability and trading.

1. OBJECTIVES AND SIGNIFICANCE

Hong Kong's Climate Action Plan 2050 outlines the strategies and targets for combating climate change and achieving carbon neutrality. Carbon sources apportion is the premise of carbon neutrality and future carbon trading market construction in Greater Bay Area. Quantifying 14CO2 has been proven to be a promising method to measure the CO2 from fossil fuels. The basic principle is that the fossil fuel is totally depleted of 14C because of the much longer existence time. It is a marker to differentiate the fossil fuel emissions from other renewable power sources. Considering the low concentration of 14CO2 (1.2 ppt) in pure CO2, we need instruments with ultra-high sensitivity to trace and elucidate carbon cycle. Additionally, for carbon reduction, 10 nuclear power plants in total will be in Guangdong province with distances about 50-400 km away from Hong Kong in the future. 14CO2 artificially generated in nuclear reactors may pose a risk to living organisms if not properly confined. The precise and real-time monitoring of 14CO2 in radioactive waste streams generated during nuclear decommissioning is important for establishing the best-suited nuclear waste management. However, the concentration of 14CO2 may change by orders of magnitude at different locations. In this scenario, we need instruments with a wide dynamic range and in-situ measurement. However, due to the extremely high price and room size of the professional instrument, called Accelerator Mass Spectrometry (AMS), the dating work needs to be only conducted in a few professional labs outside Hong Kong. The project aims to develop a 14CO2 sensing instrument which can provide ultra-high sensitivity and wide dynamic range

with portable size and low cost. The project will provide the third type of cutting-edge technique world-wide for 14CO2 detection and make breakthroughs in resolving bottlenecks.
The objectives are as follows:
1. The acoustic resonator designed for silicon-based cantilever is lacking in existing photoacoustic sensors, which limits the sensitivity. We will design a high Q-factor acoustic resonator to amplify the acoustic waves by standing wave effect. The resonator with the cantilever will enhance the sensitivity of photoacoustic gas sensors by at least one order of magnitude.
2. The photoacoustic signal is proportional with laser power. The sensitivity of gas sensors using mid-infrared lasers is limited by lack of mechanism which can enhance the laser power. We plan to develop a high efficiency locking technique between mid-infrared laser and high finesse optical resonator. The intracavity constructive interference will enhance laser power, thus the photoacoustic signal by 2-3 orders of magnitude.
3. The first generation of 14CO2 instrument based on photoacoustic spectroscopy will be produced by combining the acoustic resonator and optical resonator. The instrument should have parts-per-quadrillion sensitivity and compact size. To validate the performance and feasibility, we will design experiments to compare with commercial instruments.

2. RESEARCH METHODOLOGY

Design and fabrication of a high Q-factor acoustic resonator.
A silicon cantilever will be used as the acoustic transducer because the cantilever-based interferometric readout shows an unprecedented sensitivity compared to conventional acoustic transducers. The molecules at ground state absorb photons and excited to a higher energy level. After colliding with other molecules, excited molecules will relax back to ground state. If a pump laser (i.e., QCL in the project) is modulated in intensity, the process will induce acoustic waves. The generated acoustic wave forms standing wave in the acoustic resonator. As the resonant frequency of cantilevers normally features a narrow bandwidth, the designed acoustic resonator must match well with the cantilever resonance. The device consists of a cylindrical resonator and two buffering volumes on both sides. The cantilever is mounted at the centre of the resonator, corresponding to the antinode position of the acoustic wave. The acoustic wave induces the displacement of cantilever which is detected by an interferometer. The interference signal is then demodulated by a lock-in amplifier to retrieve the PAS signal which is proportional with $^{14}CO_2$ concentration.

Development of $^{14}CO_2$ sensing instrument based on the doubly resonant PAS technique.
We will lock the QCL frequency to a miniaturized high-finesse optical resonator to enhance the intracavity power. Lasers can be locked with optical resonator based on Pound-Drever-Hall technique. One technical challenge is to maintain the high efficiency locking condition between laser and optical resonator. The coupling efficiency is based on the design of mode matching lens, the PID parameters and noise level of laser current driver. We expect to increase the intracavity power by 2500 times considering a coupling efficiency of 50%. In the doubly resonant PAS, the acoustic resonator will be inserted into the optical resonator to take advantage of the high intracavity power. In this way, both acoustic and optical standing waves contribute to a stronger PAS signal. Here, a 4.5 μm QCL with a emission power exceeding 100 mW will be used to access the P(20) line of $^{14}CO_2$. Due to the short optical resonator (i.e., 6 cm in this project), the PAS sensing part of the instrument is very compact. The developed doubly resonant PAS instrument will be tested and calibrated by standard $^{14}CO_2$ samples. The performance will be compared with commercial instruments.

3. RESULTS ACHIEVED SO FAR

(1) Cantilever-cavity with acoustic resonator

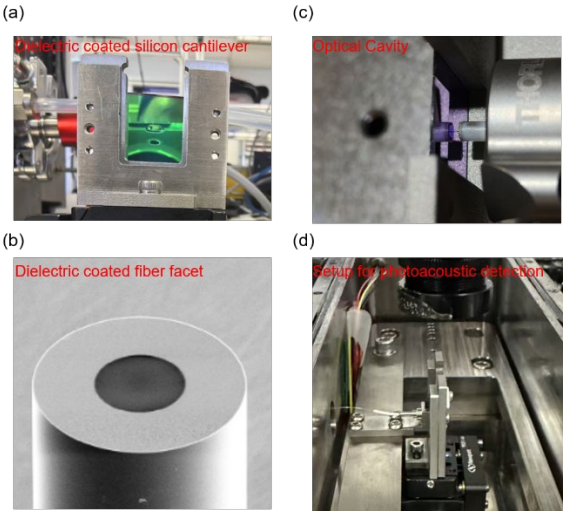


Fig.1 The cantilever cavity design

A silicon cantilever with drum shape will be used as the acoustic transducer. To enhance the sensitivity, the cantilever is high reflectivity coated at 1550 nm, as shown in Fig.1(a). The facet of single mode fiber is fabricated with a spherical surface with a surface diameter of 50 μm, and radius of curvature of 910 μm, shown in Fig.1(b). The spherical surface is high reflectivity coated in a same batch with the cantilever. The cantilever and fiber tip are aligned by a multi-axis translator to form an optical cavity, which has a finesse beyond 500 and a length of hundreds of μm, as shown in Fig.1(c). As the acoustic transducer, the acoustic wave induces the displacement of cantilever which will be amplified by the cavity, leading sensitivity enhancement. As shown in Fig.1(d), the cavity-based acoustic transducer is placed in a gas chamber for photoacoustic detection.

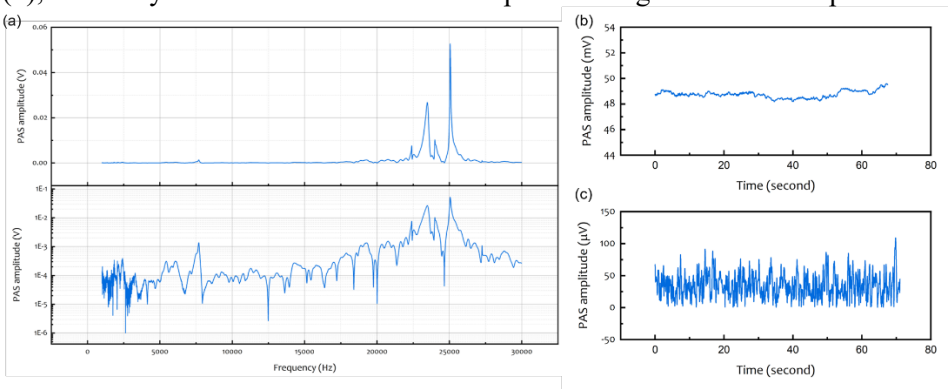


Fig.2 The cantilever characterization

To characterize the acoustics transducer, a pump laser with a wavelength of 1531.58 nm which access the absorption line of C_2H_2 is applied to generate acoustic wave. The pump bypasses the cantilever and intensity modulated at various frequencies. As shown in Fig.2(a), the frequency response of the cantilever is measured with the strongest resonant frequency of 25.056 kHz. With an incident pump power of 5.8 mW, the photoacoustic signal of 10000 ppm C_2H_2 is measured at a pressure of 40 mbar, as shown in Fig.2(b). The noise is measured by turn off the pump, as shown in Fig.2(c). The signal-to-noise ratio is evaluated to be about 2733, corresponding to a minimum detection limit of 3.6 ppm and normalized noise equivalent absorption coefficient of $1*10^{-8}$ level.

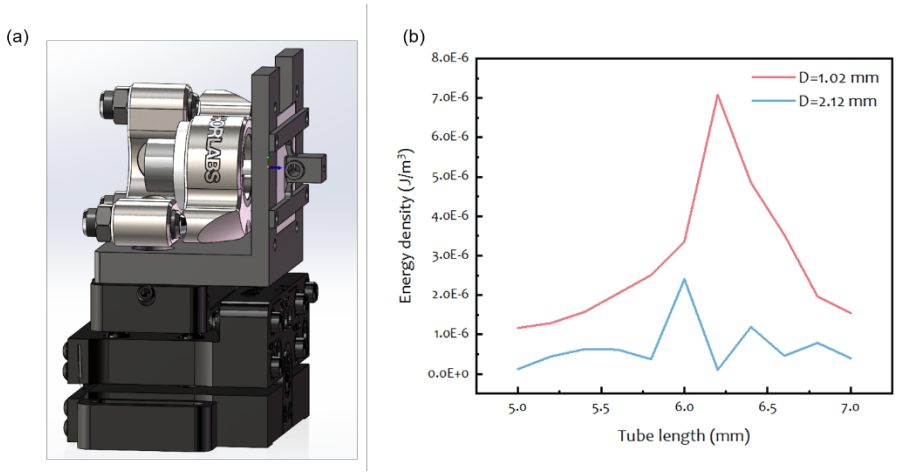


Fig.3 The acoustic resonator design

To further enhance the sensitivity, we start designing the acoustic resonators which can enhance the acoustic wave via standing wave effect. To achieve the highest enhancement effect, the inner diameter and length of the resonator should be carefully designed based on the resonant frequency of the cantilever. As shown in Fig.3(a), we fabricated the setup which can integrated the cantilever and acoustic resonator. The distance between them can be finely tuned. Based on Finite Element Analysis, we simulated the relationship between acoustic wave intensity and resonator dimensions including length and inner diameter, as shown in Fig.3(b). In this way, we start fabricating the acoustic resonator based on the simulation.

(2) Quantum cascade laser locking technique with optical cavity

We applied a quantum cascade laser from Hamamatsu for $^{14}CO_2$ detection. As shown in Fig.4, the laser wavelength and power are characterized, with a 62.5 mW at 2209.1 cm^{-1} which is the absorption wavelength of $^{14}CO_2$. The quantum cascade laser acts as pump for photoacoustic excitation.

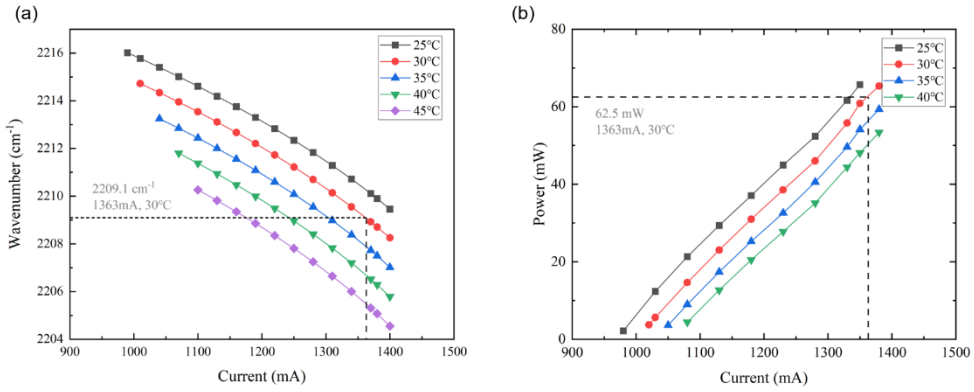


Fig.4 The quantum cascade laser characterization

We utilize an optical cavity with a finesse beyond 2000 for pump laser enhancement. Based on Pound-Drever-Hall method, the pump is locked with the cavity tightly, as shown in Fig.5(a). The noise while locking is due to the large laser linewidth which can be reduced by reducing the current noise of the laser driver. An acousto-optic modulator (AOM) is applied to modulate the laser intensity and generate the acoustic waves. The results in Fig.5(b) shows that the pump can achieve locking with the cavity for power enhancement and also be modulated without losing locking.

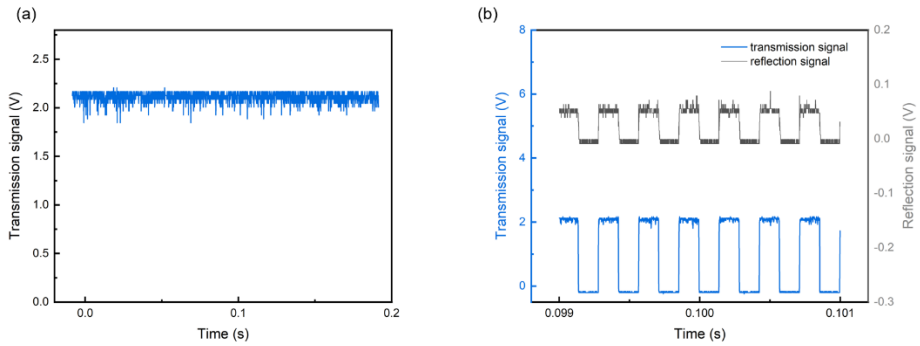


Fig.5 The laser-cavity locking and intensity modulation

4. PUBLICATION AND AWARDS

[1] X. Xiao, Z. Liu, H. Sun, Z. Wang, B. Duan and W. Ren, “Neural network-enhanced FMCW gas spectroscopy” *submitted*.

WORK IN PROGRESS

HIERARCHICAL CARBON-CENTRIC MANAGEMENT SYSTEM FOR ENERGY STORAGE-ASSISTED DATA CENTERS

Principal Investigator: Professor CHEN Yue
Department of Mechanical & Automation Engineering
CUHK

Research Team Members:
Rui XIE, SHIAE Research Associate (1)
Dongxiang YAN, Postdoctoral Fellow (1)
Wenyi ZHANG, Postdoctoral Fellow (1)
Shihan HUANG, PhD student (1)
Tao TAN, PhD student (1)

(1) Dept. of Mechanical and Automation Engineering

Reporting Period: 01 July 2023 – 1 May 2024
(to be completed in June 2025)

INNOVATION AND PRACTICAL SIGNIFICANCE:

With the explosion of data-driven workloads, the energy demand and carbon emissions of cloud/edge data centers grow sharply, threatening environmental sustainability. In the current practice, data centers balance their carbon emissions with carbon offsets such as signing power purchase agreements with large-scale renewable energy projects. This practice, however, has some potential problems: 1) Large-scale renewable generation sites are often located far from data centers. Huge transmission losses occur when their generated electricity is used. These losses need to be compensated by other carbon-intensive energy sources. 2) Even if data centers' net carbon emissions are offset to zero, they still emit massive amounts of carbon into the environment. In fact, reducing absolute carbon emissions is the only way to eventually achieve sustainability. This requires emphasizing carbon efficiency in data center operations. This project aims to develop a hierarchical carbon-centric data center management system. By adopting a bottom-up approach, we provide carbon-efficient solutions to individual, regional, and overall data center systems, respectively. If successful, it can benefit: 1) end-users, by allowing them to gain revenue through participation in carbon emission reduction; 2) data centers, by lowering their carbon emissions and operational costs while promoting the use of their co-located renewable generations; 3) the power grid, by alleviating the impact of volatile renewable generations on system reliability. The outcome of this project will be turned into a set of software packages. The PI will then reach out to our industry partners to address critical issues in engineering practices, such as compatibility with the existing management system. This may bring about follow-up funding and opportunities for implementation and commercialization.

ABSTRACT

The energy demand of data centers increases dramatically with the explosion of data-driven workloads. This positions data centers among the main contributors to global carbon emissions. In fact, the increasing carbon footprint is a more serious problem than merely the rising energy demand. Noticing that energy efficiency and carbon efficiency are not necessarily correlated, this project aims to develop a carbon-centric data center management system. Unlike previous research that focused on energy costs, it elevates carbon to a first-priority metric in the design. The main difficulties come from the diversity of computing tasks, the volatility of carbon-free energy sources, and the invisibility of the carbon footprint of grid power. This project adopts a bottom-up approach with the following tasks: (1) Enhancing carbon efficiency within an edge data center by co-optimizing computing task assignment and energy storage utilization considering diverse user requirements;

(2) Enhancing carbon efficiency regionally across edge data centers considering the tradeoff between reduced emission and lower latency; and (3) Enhancing carbon efficiency of the overall system by coordinating the operations of cloud and edge data centers through well-designed prices. A simulation platform will be built for demonstration and validation. This project will contribute to carbon emission reduction and sustainability.

1. OBJECTIVES AND SIGNIFICANCE

(1) For each edge data center, develop a computing task and energy storage co-optimization approach to enhance the use of co-located renewable generation.
(2) For regional edge data center clusters, develop a computing task and energy storage co-sharing approach to take advantage of the complementarity between their renewable generations.
(3) For the overall data center system, develop a carbon-integrated electricity price-based coordination mechanism to reduce the use of carbon-intensive grid power.
(4) Construct a simulation platform for demonstration and validation.

2. RESEARCH METHODOLOGY

First, we will focus on improving the carbon efficiency of an edge data center. Edge data centers typically have low power density, allowing them to be self-powered by their co-located renewable energy sources without relying on grid power. Hence, it is possible to have a zero-carbon footprint as long as the volatility of renewable energy sources is well-tackled. We will create appropriate models to characterize the diverse features of computing tasks from end-users. With these models, an online computing task assignment approach will be developed to match the energy consumption of the edge data center with the co-located renewable generation. Furthermore, we will co-optimize energy storage utilization with computing task assignments to promote the use of clean renewable energy.

Second, we will focus on improving the carbon efficiency of edge data center clusters based on the fact that the aggregate renewable power supply across a region is smoother and more predictable. Thus, shifting computing tasks across data centers would help when an individual edge data center struggles with its co-located volatile renewable generation. However, shifting computing tasks across a larger region will trade off low latency of computation for decreased aggregate renewable power volatility. We will characterize the latency and the associated disutility for end-users when their computing tasks are shifted. A computing task sharing approach considering the tradeoff above will be developed. Furthermore, as some computing tasks are non-shiftable, edge data centers may still need carbon-intensive grid power. A computing task and energy storage co-sharing approach will be proposed to further reduce reliance on grid power by lowering the energy supply-demand mismatch.

Third, we will focus on improving the carbon efficiency of the overall data center system via proper carbon-integrated electricity prices. The response of edge data centers to carbon-integrated electricity prices will be modeled through sensitivity analysis. Based on this, a price-based coordination mechanism will be developed to regulate the service demand of edge data centers. Furthermore, we will design incentives to encourage edge data center clusters to support low-carbon operation of cloud data centers via their individual/shared energy storage. With the proposed technologies, a simulation platform will be built for demonstration and validation.

3. RESULTS ACHIEVED SO FAR

We have developed a distributed online algorithm for combined computing workload and energy coordination of data centers. The proposed algorithm is prediction-free and easy to implement. It well addresses the two major challenges of data center operation: (1) The high uncertainty due to the unpredictable computing workload and renewable generation, and (2) the need for a distributed implementation framework to avoid the high computational burden and privacy leakage. As shown in Fig. 1, the proposed algorithm can achieve the lowest accumulated operation cost compared to other existing online algorithms. Moreover, the proposed accelerated alternating direction method of multipliers (ADMM) algorithm can achieve an operation cost close to the centralized benchmark in a distributed manner, as illustrated in Fig. 2. The small gap is the efficiency

loss due to the truncation for acceleration purposes. The proposed algorithm can improve the renewable power utilization of data centers, thereby improving their carbon efficiency.

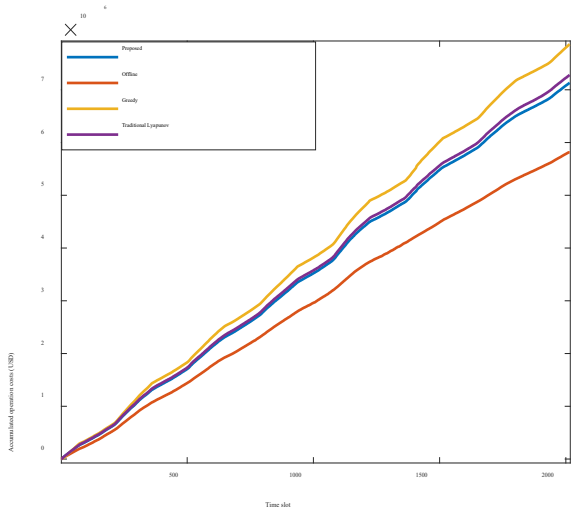


Fig. 1 Accumulated operation costs under the offline benchmark and three online algorithms

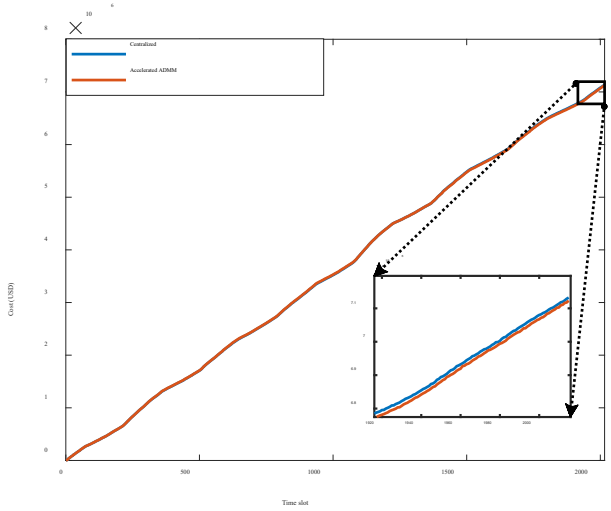


Fig. 2 Cost traces of the proposed algorithm and its centralized counterpart

We have developed an Aumann-Shapley price-based method to allocate carbon responsibility within a power network. Electricity consumers should be responsible for at least part of the carbon responsibility since they are the cause of electricity generation. In order to encourage the adoption of low-carbon practices in data centers, it is crucial to properly allocate and price the carbon responsibility of data centers at various buses within the power network. The proposed Aumann-Shapley price-based allocation method possesses several desirable properties, such as cost sharing, scale invariance, monotonicity, additivity, and consistency. These properties contribute to a fair and effective allocation of carbon responsibility. Based on this, we have developed a carbon-integrated electricity pricing method. The prices for a modified IEEE 30-bus system are shown in Fig. 3, which effectively captures the contributions of electricity demand from various locations to the total system emissions. We can observe from Table 1 that with the proposed pricing method and energy storage, we can achieve the lowest total emission and renewable power curtailment.

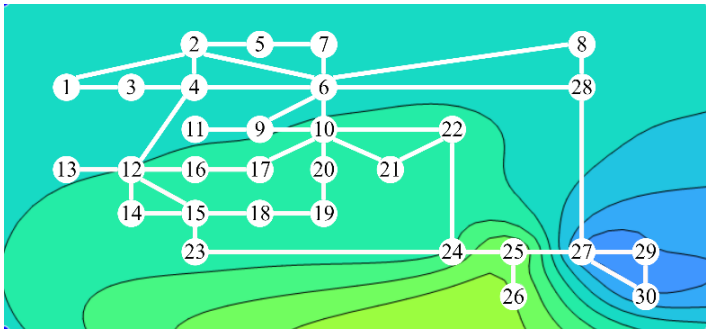


Fig. 3 Carbon-integrated electricity prices for a modified IEEE 30-bus system

Table 1 Results with/without energy storage and carbon responsibility allocation

Case	Proposed	A1	A2	A3
Energy storage	√	√	×	×
Carbon responsibility allocation	√	×	√	×
Total generation cost ($/h)	3387	3121	3443	3173
Total emission (kgCO_2/h)	30546	53701	31063	54457
Renewable curtailment rate	1.84%	1.84%	3.25%	3.25%

We have developed a prediction improvement approach to forecast renewable power and electricity demand more accurately by aggregating predictions from the system operator and distributed agents. Fig. 4 shows that with the improved predictor (Best Linear Predictor), the variation ranges of prediction errors are greatly narrowed. In the figure, the green area represents the variation range of the original prediction error, while the blue area represents the variation range of the improved prediction error. Furthermore, a robust optimization method has been proposed to leverage the improved prediction for better decision-making. The proposed method can lower operation costs and enhance carbon efficiency by reducing the uncertainty associated with renewable generation.

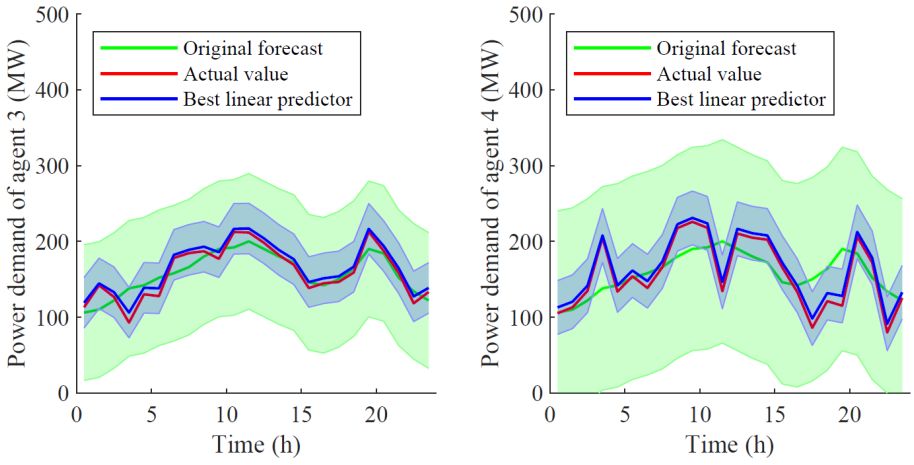


Fig. 4 Original and improved uncertainty sets of a wind farm (left) and an electricity demand (right)

4. PUBLICATION AND AWARDS

4.1. Publications

J[1] S. Huang, D. Yan, and Y. Chen, “An Online Algorithm for Combined Computing Workload and Energy Coordination within a Regional Data Center Cluster,” International Journal of Electrical Power & Energy Systems, vol. 158, pp. 109971, 2024.
J[2] R. Xie, and Y. Chen, “Real-time Bidding Strategy of Energy Storage in an Energy Market with Carbon Emission Allocation Based on Aumann-Shapley Prices,” IEEE Transactions on Energy Markets, Policy and Regulation, 2024, early access.
J[3] R. Xie, P. Pinson, Y. Xu, and Y. Chen, “Robust Generation Dispatch with Purchase of Renewable Power and Load Predictions,” IEEE Transactions on Sustainable Energy, 2024, early access.
J[4] D. Yan, S. Huang, and Y. Chen, “Real-time Feedback Based Online Aggregate EV Power Flexibility Characterization,” IEEE Transactions on Sustainable Energy, vol. 15, no. 1, pp. 658 – 673, 2024.
J[5] Y. Zhang, Y. Su, Y. Chen, and F. Liu, “Asynchronous Distributed Charging Protocol for Plug-in Electric Vehicles,” Journal of Economy and Technology, vol. 1, pp. 29-47, 2023.
C[1] R. Xie, and Y. Chen, “Privacy-Preserving Aggregated Load Forecasting Based on Vertical Federated Learning”, Nexus Forum, Hong Kong, China, pp. 1-6, May 9-10, 2024.
C[2] M. Yang, and Y. Chen, “Robust Operation of Distribution Systems with Uncertain Renewable Generation via Energy Sharing”, The 3rd Conference on Fully Actuated System Theory and Applications, Shenzhen, China, pp. 1-6, May 10-12, 2024.

4.2. Awards

[1] 2023 Best Paper Award of Journal of Economy and Technology

COMPLETED

MANUFACTURING METASURFACE BASED ALL-OPTICAL CNN FOR REAL-TIME AND POWER -EFFICIENT MACHINE VISION

Principal Investigator: Professor Chaoran HUANG
Department of Electronic Engineering
CUHK

Research Team Members:
Mingcheng Luo[(1)],
Li Fan [(1)];

[(1)] Dept. of Electronic Engineering, CUHK

Project Start Date: 1st July 2022
Completion Date: 30th June 2024

INNOVATION AND PRACTICAL SIGNIFICANCE:

Device-level innovation*: This is the first time that the photonic metasurface is proposed for CNNs. Photonic metasurface was originally used for more conventional applications such as optical imaging [10]. Recently, there has been a surge of research on using photonic metasurface for computing [11]. However, those studies have only demonstrated simple arithmetic operations (e.g., differentiators [12], [13]). The photonic metasurface, according to our preliminary results, can perform convolution operations to the optical images encoded in visible light, and extract multiple image features from the entire image at one time. The proposed all-optical CNN is completely passive, thus can bring unprecedented low latency and power efficiency as compared to electronic accelerators.*

System-level innovation: *The optical convolutional accelerators that have been demonstrated so far need to manipulate how the image pixels enter the photonic devices, which as a result, need to preprocess the image at the digital domain first and then convert the processed signal back to the optical domain[5], [9]. The proposed all-optical CNN, in contrast, can process the whole image encoded on light directly at the photonic domain, without power-hungry domain crossings (e.g., optical-to-electrical-optical and analog-to-digital-to-analog conversions). As such, our all-optical CNN can significantly reduce the system overhead in terms of power consumption, latency, device footprint, and cost.*

Application-level innovation*: The proposed all-optical CNN can be used as a general-purpose AI accelerator, since it can compute the convolutional layer in only sub-picosecond. This means that the proposed all-optical CNN can achieve extremely large computing throughput as many AI accelerators target. However, we want to set a more ambitious goal from the application level – we want to fully exploit the unprecedented low latency and power efficiency that our all-optical CNN uniquely has, in order to benefit those extremely latency-, power-sensitive, and computation-intensive applications [3].*

In the following section, we will discuss some of the targeted applications and how our all-optical CNN can bring innovative and practical significance to them.

ABSTRACT

Conventional integrated circuits (ICs) struggle to meet the escalating demands of artificial intelligence (AI). OpenAI estimates that these demands have been growing 100 times every two years, outpacing Moore's Law by 50 times. To surpass Moore's Law, neuromorphic computing has emerged. Unlike traditional processors solving problems in a sequential manner, neuromorphic computing executes in a highly parallel manner, leading to substantial speed and energy efficiency improvement. However, scaling up components in

neuromorphic hardware remains a challenge. As a result, most neuromorphic hardware is limited to basic benchmark demonstrations, hindering its application to real-world AI challenges. Our research develops a practical pathway to realize large-scale neuromorphic computing systems, not only to surpass digital electronics in speed and energy efficiency, but also capable of handling a large number of parameters to close the performance gap with large-scale AI models. To achieve this, our research develops a 3D photonic-electronic neuromorphic computing system, leveraging a novel device optical metasurface and its numerous spatial modes. Optical metasurfaces, comprising sub-wavelength meta-atoms on a 2D plane, offer unmatched parallelism, processing tens of millions of weights in one operation with zero power consumption. Our chip integrating over 40 million meta-atoms on a metasurface chip showed a single-layer metasurface can provide matched performance of a 50-layer convolutional NN (ResNet 50), while reducing computing time and energy consumption by over 1000 times compared to GPU. Our system is delivering high-performance solutions to real-world AI challenges through its unprecedented scale. We demonstrate the practical application in acceleration the analysis of whole slide images (WSIs) for cancer detection and object detection for autonomous driving.

1. OBJECTIVES AND SIGNIFICANCE

1. Achieving unprecedented parallel computing using optical metasurfaces for computing over tens of millions of weights in a single metasurface layer, with over 100 times reduction in power consumption compared to digital electronics.
2. Innovating a new computing framework and architecture suitable for the optical metasurfaces-based neural network to scale to greater widths, depths, and complexity, while remain immune to physical errors.
3. Delivering high-performance solutions to real-world AI challenges through the unprecedented scale and efficiency of our system.

2. RESEARCH METHODOLOGY

The implementation of the proposed meta-ONN is illustrated in Fig. 1a. The meta-ONN is a dielectric metasurface made up of 41 million silicon cylindrical nanodisks in a 10 mm2 chip area. Each nanodisk controls the transmissive and reflective phase and amplitude of incident light at a subwavelength scale. These modulation values are determined by the interference between electric and magnetic dipole resonances within each nanodisk, which can be adjusted by varying the nanodisk's radius. The incident beam, carrying encoded input information such as images, is reflected by the metasurface. According to the Huygens–Fresnel principle, the wavefront, after being modulated by each meta-atom, acts as secondary spherical waves and spreads out in all directions. The resulting wavefront at the following plane is formed by the combination of all these secondary wavelets, with each wavelet contributing to the overall shape of the wavefront. Consequently, each meta-atom can be regarded as an optical neuron, which is fully connected to the input nodes of the spatial light modulator (SLM) plane and the output nodes of the receiver. The weights in such a fully-connected ONN is predetermined by each meta-atom’s radius. Subsequently, the reflected beam is collected by an optical lens and focused onto an image sensor array. The image sensor introduces an optoelectronic nonlinear activation function through square-law detection. The final decision is made by a highly compact NN implemented by a digital processor. Experimental setup details are provided in the Methods.

The overall system is a multilayer neural network. The first hidden layer made by optical metasurface is an exceptionally wide layer providing abundant parameters over 41 millions, which encodes the input into a hyperspace. Following this hidden layer is an optical lens that facilitates Fourier feature mapping within the optical domain. This feature mapping layer enhances the system's ability to learn high-frequency features. The parameters of the hidden layers are initialized by sampling from the Gaussian distribution, ensuring the system to converge to a global minimum. This is realized by engineer 41 million circular silicon posts with varying transmission coefficients and phases by adjusting the diameters varying from 100 nm to 400 nm at a fixed unit cell period of 500 nm. Each diameter is sampled independently from a Gaussian distribution. The resulted ONN is a large-scale kernel machine with Fourier feature mapping function, making it a generic pre-processor for different applications. Therefore, the same metasurface chip is used for all the demonstrated applications. Only the digital NN at the backend is trained for different applications.

Our method draws inspiration from optical computing systems that leverages random scattering and

shares similarities with reservoir computing, a computing architecture implemented in various optical systems. However, our system introduces two critical differences. First, the introduction of metasurface provides an almost infinite-wide layer to encode the input to extremely high-dimensional space, a scale never demonstrated in prior systems. Second, using metasurfaces provides full controllability of each entry (meta-atom) to optimize the projection matrices. This is in contrast to solely relying on the random nature of a physical system. By designing the geometry distribution of circular silicon posts in the metasurface, we can ensure that the neural network is initialized with Gaussian distribution which is prerequisite for reaching optimal performances. Third, we use an optical lens to provide Fourier transform which is critical for the system to learn high frequency functions. These distinctions enable a single-layer metasurface to rival cutting-edge neural network models with many nonlinear layers such as ResNet and Vision Transformer, a performance that has never been realized in other ONN systems.

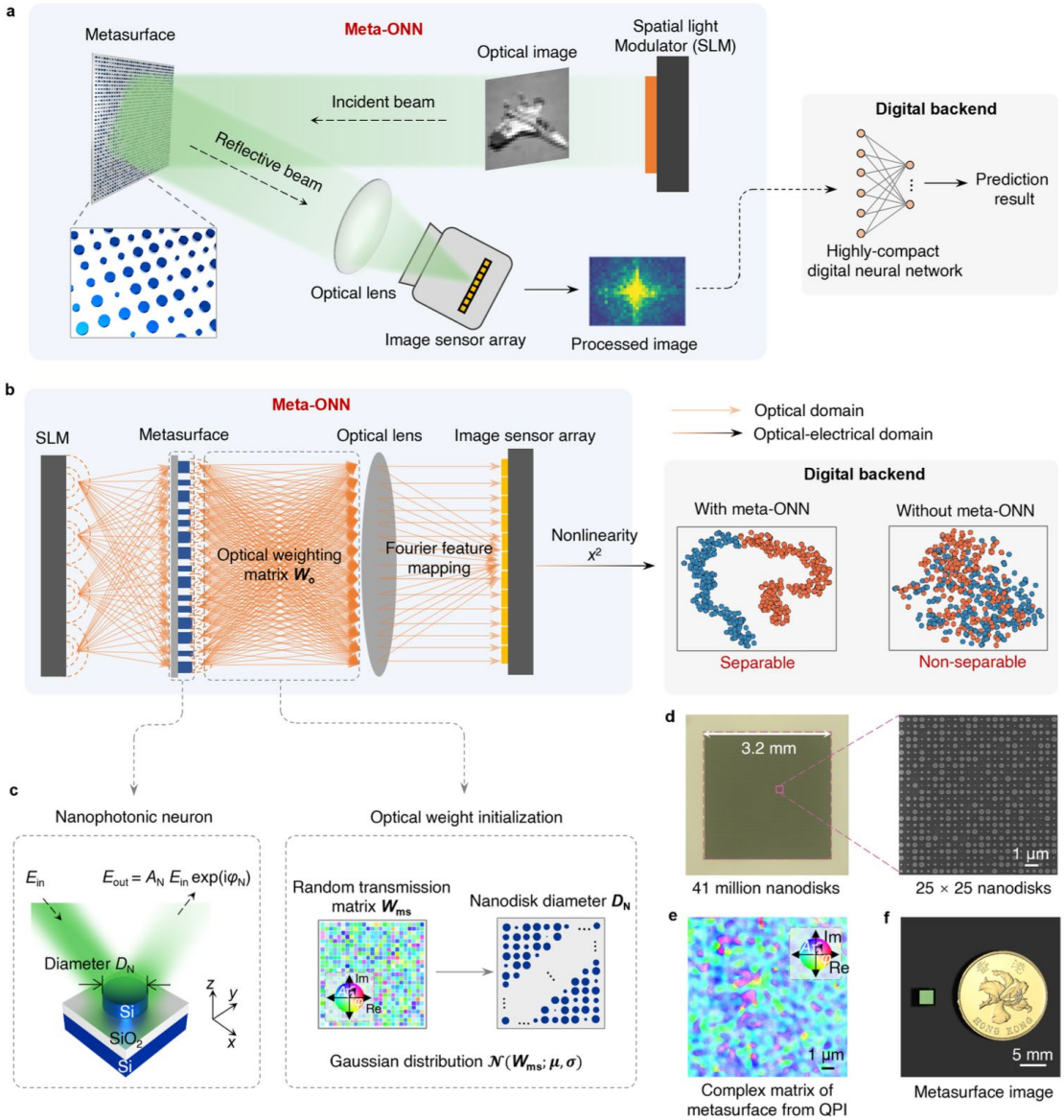


Fig. 1 The metasurface-based optical neural network (meta-ONN). (a) Schematic of the meta-ONN. The optical image generated by the SLM is projected onto a single-layer metasurface consisting of massive cylindrical silicon nanodisks. An optical lens then collects the reflected optical image by the metasurface. Following that, the optical field at the focusing plane is captured by an image sensor array. Lastly, the captured image is fed into a digital neural network to produce the prediction result. (b) Mathematics operations represented by meta-ONN. In the right graph of scattering points, the different colors represent different categories of the input dataset. (c) The illustration of the nanodisk representing a photonic neuron (left graph) and the design flow of optical weight initialization (right graph). (d) Optical microscope image of the fabricated metasurface chip with a compact area of 3.2 × 3.2 mm2 (left graph) and the scanning electron microscope

(SEM) image of a zoom-in metasurface region containing 25 × 25 nanodisks (right graph). (e) The measurement result of quantitative phase imaging (QPI) of a zoom-in region of the metasurface chip. In the color wheel, the color represents the phase modulation coefficient, and the brightness represents the amplitude modulation coefficient. (f) Image of the metasurface chip compared with a Hong Kong dollar coin.

3. RESULTS ACHIEVED

We first demonstrate our metasurface-based ONN can rival deep NNs in machine vision tasks. For each task, the input images are generated using a SLM. These images are processed by the metasurface and then collected by an optical lens before being detected by a CMOS digital camera. The detected digital image is downsampled, with the downsampling ratios being task-dependent. The same metasurface chip is used for all the tasks. Only the digital neural network at the backend is trained for different applications. For benchmarking, we compared the performance of our meta-ONN with three benchmarking large-scale deep learning models: ResNet-50, a classical 50-layered CNN with approximately 23.5 million parameters; the Segment Anything model (SAM), a cutting-edge large promotable segmentation model with 93.7 million parameters; and Vision Transformer (ViT), a transformer encoder model for image classification with more than 85.8 million parameters. The results are shown in Fig. 2.

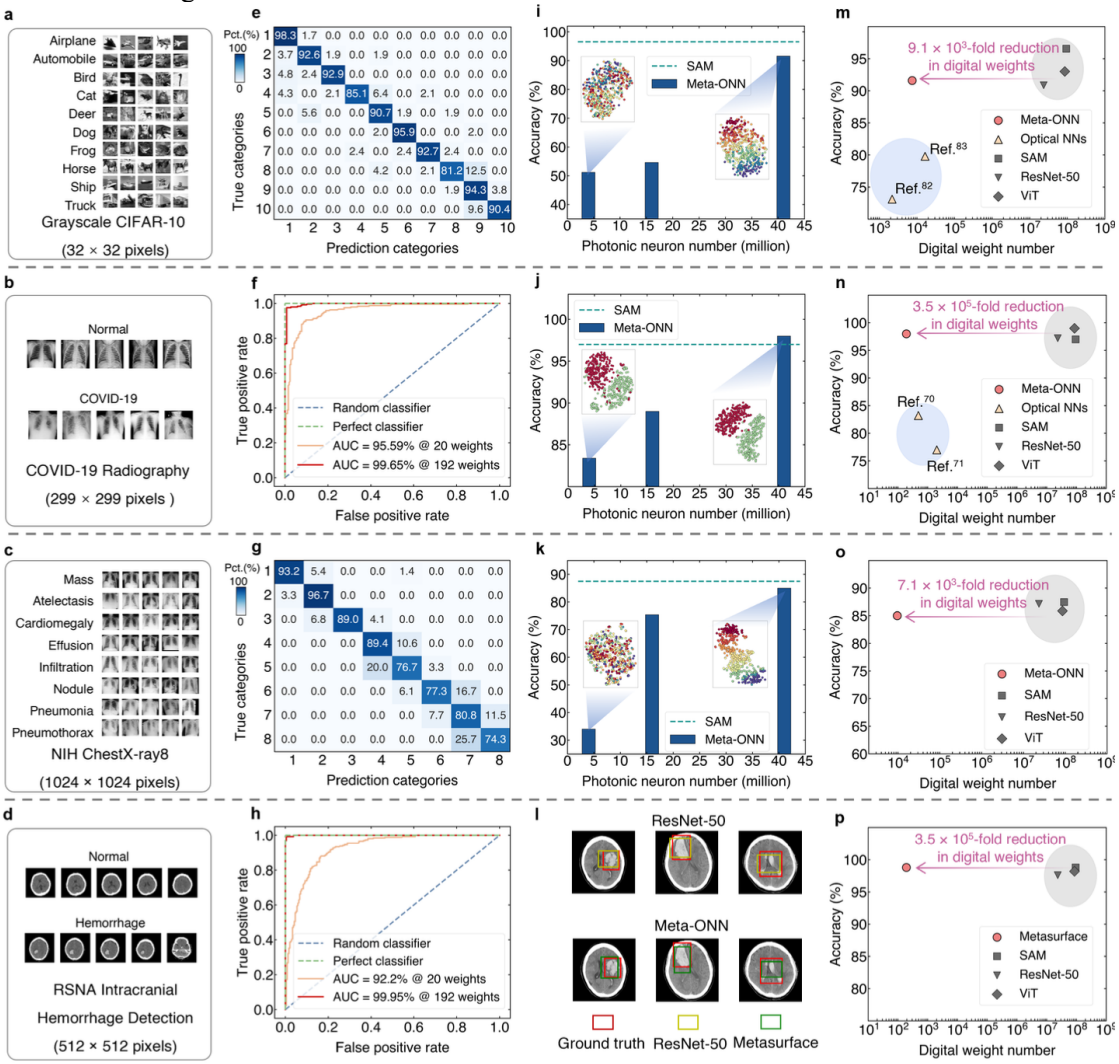


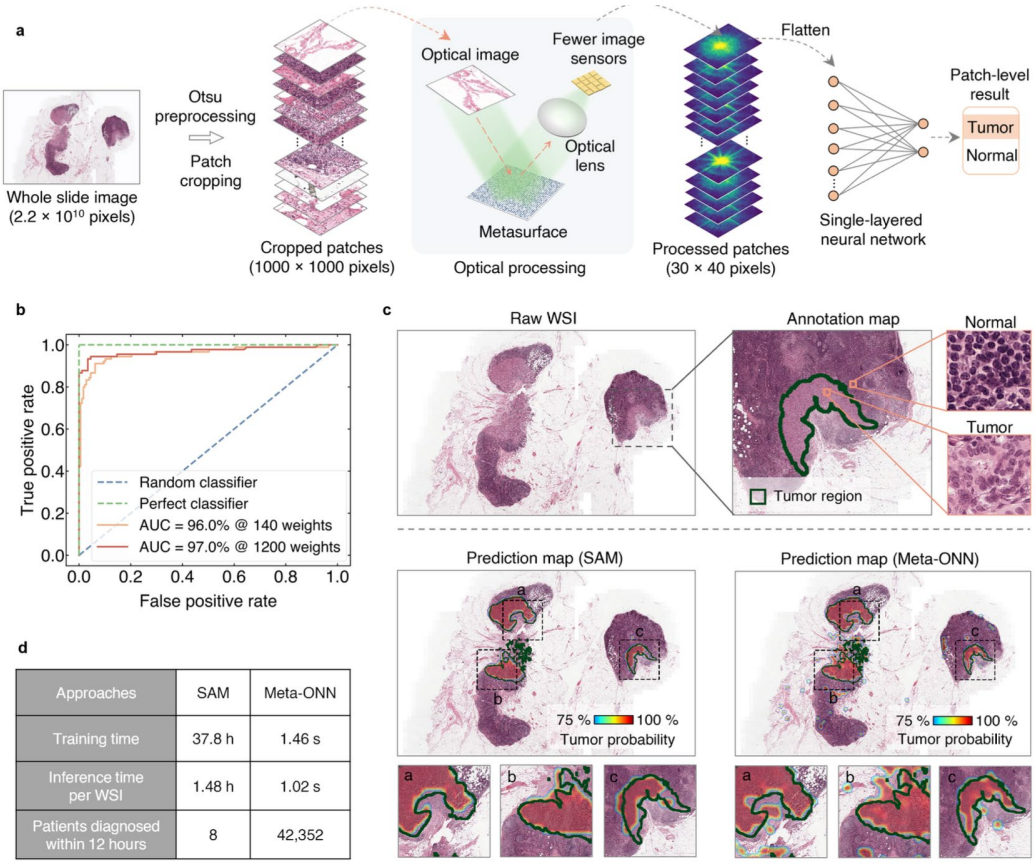
Fig. 2 Experimental results of the meta-ONN for four benchmark tasks. (a)–(d) Illustration of the dataset images of the CIFAR-10 (a), the COVID-19 Radiography (b), the NIH ChestX-ray8 (c), and RSNA Intracranial Hemorrhage Detection tasks (d). (e) and (g) The confusion matrix of the predic- tion results for CIFAR-10 and NIH ChestX-ray8. (f) and (h) Receiver operating characteristic (ROC) curve of the prediction results for COVID-19 Radiography and the RSNA Intracranial Hemorrhage Detection. (i)–(k) The accuracy versus optical neuron number of the meta-ONN for CIFAR-10 (i), COVID-19 Radiography (j), and NIH ChestX-ray8

(k). The dashed line represents the accuracy of SAM as a comparison. The inset graphs with colorful scatters represent the results of the t-SNE anal- ysis. (l) Bleeding regions inside the brain predicted by the meta-ONN (bottom graph) and digital ResNet-50 (upper graph). (m)–(p) Comparison of accuracy and electronic weights of the meta-ONN with other optical approaches and digital models for the CIFAR-10 (m), the COVID-19 Radiography (n), the NIH ChestX-ray8 (o), and RSNA Intracranial Hemorrhage Detection tasks (p).

Leveraging on these advantages, we finally demonstrate the application of our metaONN in a real-world challenge, showcasing its distinct benefits in computationally intensive applications. For many diseases, particularly cancers, pathological diagnosis is the gold standard in clinical practice. The introduction of Whole Slide Imaging (WSI) scanners, which generate digitized pathology microscopic images, has revolutionized pathology image analysis. This technology enables computer-aided diagnostics, leveraging advanced deep-learning techniques to reduce the workload of pathologists and optimize the regional distribution of medical resources. However, WSIs present a challenge for deep learning due to their extremely large size. With single slides containing multi-gigapixel images, efficient processing of these images is crucial for automated diagnosis.

In our work, we apply a meta-ONN to detect and localize breast cancer that has metastasized to nearby lymph nodes, a task of significant clinical importance but requiring substantial reading time from pathologists. Pathologists typically need to review thousands of megapixel photos from a single WSI exceeding 10 gigapixels. We use the CAMELYON16 dataset for our study. To address the processing of extensive WSIs, each with dimensions of more than 2 billion pixels, we employ a patch-based framework. Initially, a preprocessing algorithm, Otsu algorithm, is used to separate the raw whole slide image into the useful foreground and the nontissue background, resulting in a reduced total pixel of the whole slide image to be processed, as shown in Fig. 4a. We then divide the large WSIs into smaller patches, each containing 1,000 x 1,000 pixels. These patches consist of 1,775 normal patches and 887 tumor patches. During the training phase, we convert the patches into optical images using the SLM. The modulated patch samples are processed by our meta-ONN and detected by the sensor array. Subsequently, the output from the sensor array is downsampled. The final step is to train a single-layered neural network with the patch samples to create a classifier capable of distinguishing between normal and tumor classes. The mean AUC is 96.0% when the trained weight number is 140 and increases to \$97.0 \%\$ at the weight number of 1,200 , as shown in Fig. 4b. The training process takes only 1.46 s , achieving a training accuracy of 95.1%, as shown in Fig. 4 d .

Another WSI consisting of 2,030 unlabeled patches is used to test the performance of tumor tissue segmentation. Initially, these unlabeled patches are first processed using our meta-ONN. Subsequently, the processed patches are inferred with the trained single-layered NN to produce the tumor-positive probability. Finally, the probability heat map is generated by mapping the predicted probability of the patches to the raw WSI. As shown in Fig. 4c, the resulting heat map demonstrates that our metaONN achieves accurate segmentation of three different tumor tissue regions from the billion-pixel-scale WSI, achieving an IOU of 0.60 , which is comparable to that of SAM (0.63). More importantly, our meta-ONN exhibits an impressively fast inference time of 1.02 s per whole slide image (WSI), representing a significant reduction compared to SAM which requires 1.48 hours to analyze one WSI. This remarkable reduction in inference time allows our meta-ONN to diagnose more than 42,352 patients within a single 12-hour working day, while only 8 patients can be diagnosed using SAM in the same timeframe.



Approaches	SAM	Meta-ONN
Training time	37.8 h	1.46 s
Inference time per WSI	1.48 h	1.02 s
Patients diagnosed within 12 hours	8	42,352

Fig. 4 Experimental results of the meta-ONN for cancer diagnosis based on the whole slide image. (a) The working flow of the cancer diagnosis based on the whole slide image. The input WSI with over 2.2 x10^{10} pixels is preprocessed with the Ostu method and cropped into a series of patch images with 1000x1000 pixels. The patch images are processed by the meta-ONN in a single shot and compressed into 30x40 pixels. Finally, these processed images are fed to a highly compact digital neural network to generate the final prediction result of whether the tumor cell is detected in the image. (b) ROC curve of the patch-level prediction results during the training phase. (c) The heat map of the prediction probabilities of the meta-ONN and the state-of-the-art segmentation model (SAM). The inset graphs a, b, and c represent three different zoom-in regions of the WSI, respectively. (d) Comparison of the training time and inference time of the meta-ONN with SAM. The time consumption of the meta-ONN is the predicted one based on current experimental results (see the details in Supplementary Materials section 4.1).

The research work has applied for US patent with a Serial No. 63/643,973

4. PUBLICATION AND AWARDS

J[1] M. Luo, T. Xu, S. Xiao, H. K. Tsang, C. Shu, and C. Huang, “Meta-optics based parallel convolutional processing for neural network accelerator,” Laser & Photonics Reviews, p. 2300984, 2024.
J[2] T. Xu, W. Zhang, J. Zhang, Z. Luo, Q. Xiao, B. Wang, X. Xu, M. Luo, B. Shastri, P. Prucnal, C. Huang, “Control-free and efficient integrated photonic neural networks via hardware- aware training and pruning”, Optica, Vol. 11 No. 7, 2024.
C[1] M. Luo, S. Xiao, T. Xu, H. K. Tsang, C. Shu, and C. Huang, “Ultra-compact optical convolutional accelerators based on polarization-independent metasurfaces,” in 2022 Conference on Lasers and Electro- Optics (CLEO), 2023, pp. 1–2.

The conference paper C[1] is selected as the highlight conference paper at CLEO 2023 (top 2%).
(Part of the research outcome are reported in a paper under review by Nature Communications)

Research Reports (2024-2025) In Biomedical Engineering

Newly Funded Projects (2025-2027)	* Low-temperature Liquid Crystal Elastomer Robotic Textiles and their Biomedical Applications
Continuing Projects (2024-2026)	* Cost-efficient Highly Potent Antimicrobial Peptide Discovery for Livestock Farming Antibiotic Alternatives with Protein Language Model-Powered AI Methods
	* Development of Mitochondria-Targeting, Single-atom Nanozyme for Accelerated Bone Regeneration
(2023-2025)	* Design, Optimization, And Experimental Validation of a Handheld Variable-curvature Hybrid-Structure Robotic Instrument (HVHRI) for Maxillary Sinus Surgery
	* Smart Bandage with Integrated Organic Electronic Sensor and Iontronic Drug Delivery Platform for Advanced Chronic Wound Care

Completed Projects (2022-2024)	* Coupling Mos2 Field-effect Biosensors with Hybridization Chain Reaction Self-assembly Amplification for Highly Sensitive and Label-free Nucleic Acid Detection

(Funded Year)

LOW-TEMPERATURE LIQUID CRYSTAL ELASTOMER ROBOTIC TEXTILES AND THEIR BIOMEDICAL APPLICATIONS

Principal Investigator: Professor Qiguang HE
Department of Mechanical and Automation Engineering, CUHK

Project Start Date: 1 July 2025

ABSTRACT

Robotic textiles have received significant interest for their flexibility, lightweight nature, and biocompatibility, making them highly promising for biomedical applications. While various actuation materials have been explored for textile fabrication, liquid crystal elastomers (LCEs) stand out due to their exceptional actuation strain and energy density. However, the application of the-state-of-the-art LCE textiles in biomedical scenarios has yet to be realized due to two critical challenges: the high phase transition temperature makes them incompatible with human skin, posing potential thermal damage risks, while the lack of real-time control and monitoring capabilities limits their ability to meet the demanding requirements of precision medical applications. This project aims to develop low-temperature LCE robotic textiles integrated with sensing, control, and actuation functions. To achieve this, the molecular structure and composition of LCEs will be tailored to achieve actuation at a temperature suitable for human skin. Liquid metal will be integrated into textile designs, along with flexible circuits, to create a fully integrated system capable of dynamically sensing and responding to user input and environmental stimuli. The textiles will be applied to various biomedical scenarios, including locomotion assistance, thermoregulation, and therapeutic compression. This project establishes an innovative "materials-devices-systems-applications" framework for advancing smart medical textiles.

INNOVATION AND PRACTICAL SIGNIFICANCE:

This project aims to transform LCEs from a mere abstraction to a fully integrated medical device by developing the first biocompatible LCE robotic textile system capable of sensing, controlling, and low-temperature actuation. Through molecular engineering, the proposed system addresses the critical safety limitations of conventional LCEs while enabling closed-loop feedback control that dynamically senses and responds to users' inputs and environmental stimuli. Unlike conventional textiles based on rigid smart materials or bulky deformable structures, the LCE robotic textile developed in this project offers a highly flexible, lightweight, and compact design with untethered, wireless functionality. This innovation provides a clinically viable solution for biomedical applications, including proprioceptive motion assistance, adaptive thermoregulation, and precision compression therapy. Through collaboration with Beijing Tempaware Materials Technology Co., Ltd., this project aims to translate low-temperature LCE actuation technology into customizable medical textiles suitable for large-scale production. Aiming to address critical health challenges such as muscle degeneration and venous diseases associated with an aging population, the proposed lightweight, low-cost, and home-use wearable devices offer an accessible and efficient approach to motion augmentation and vascular symptom management. For instance, adaptive compression stockings made from LCE robotic textiles can autonomously regulate compression intensity in response to real-time swelling, while smart knee pads can provide precise locomotion assistance for individuals with movement disorders. This home-based solution reduces reliance on public healthcare resources and lowers family caregiving costs, providing excellent synergy with the national Healthy China 2030 initiative to advance inclusive smart medical technologies.

PROJECT OBJECTIVES AND LONG-TERM IMPACT:

1. Develop low-temperature liquid crystal elastomer (LCE) materials.
Conventional LCEs are unsuitable for direct contact with human skin due to their high phase transition temperature (>70°C). This project aims to *lower the phase transition temperature of LCE (< 43°C) by adjusting its molecular structure and composition while maintaining exceptional actuation performance and eliminating the risk of thermal damage*. These advancements establish a foundation for the safe and effective biomedical application of LCE materials.

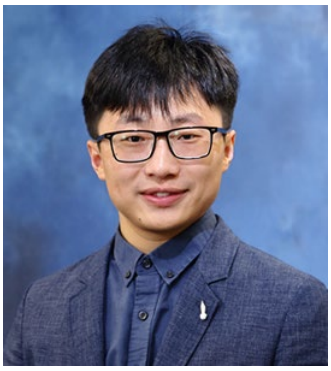
2. Develop LCE robotic textiles integrated with sensing, control, and actuation functions.
Conventional LCE textiles rely on passive activation, lacking real-time, high-precision, and user-specific control. This project *integrates liquid metal with flexible circuits to establish a closed-loop system for sensing, control, and actuation*. This advancement enables LCE robotic textiles to dynamically sense and respond to user inputs and environmental stimuli, achieving seamless and adaptive operation to accommodate individual needs.

3. Demonstrate biomedical applications of LCE robotic textiles.
The practical biomedical applications of conventional LCE textiles remain largely unexplored beyond proof-of-concept demonstrations. This project advances LCE robotic textiles for diverse biomedical applications, including locomotion assistance, autonomous thermoregulation, and therapeutic compression. These innovations provide an effective and accessible solution for muscle augmentation and vascular condition management, addressing critical clinical needs.

WORK IN PROGRESS

COST-EFFICIENT HIGHLY POTENT ANTIMICROBIAL PEPTIDE DISCOVERY FOR LIVESTOCK FARMING ANTIBIOTIC ALTERNATIVES WITH PROTEIN LANGUAGE MODEL-POWERED AI METHODS

Principal Investigator: Professor LI Yu
Department of Computer Science & Engineering, CUHK



Co-investigator(s):
Dr. Irwin King [(1)],

Research Team Members:
Dr. Christopher K C LAI [(2)],
Dr. Lei Dai [(3),]

[(1)] Dept. of Computer Science & Engineering, Faculty of Engineering
[(2)] Dept. of Microbiology, Faculty of Medicine
[(3)] SIAT, Shenzhen, China

Reporting Period: 1st July 2024 – 30th April 2025

INNOVATION AND PRACTICAL SIGNIFICANCE:

Innovation:

1. We develop the first method to using protein language model for AMP discovery. Our preliminary results show that it's very efficient and sensitive and can discover AMPs omitted by the previous method because of its awareness of structural information.

2. We are the first team to discover AMPs from both the bacterial genomes and host genome, dedicated to animal husbandry. The interaction between host and microbiome makes us easier to discover novel AMPs.

3. Based on our preliminary study, we have experimentally validated 62 candidates from our method. Impressively, 84% of our candidates are potent AMPs, better than all previous methods, and 8 of them even surpass medicinal AMP polymyxin B and swine-gut AMPs like PR-39, Cecropin P1, and Porcine beta-defensin 2. We will discover more AMPs on other animal genomes and further refine the comprehensive evaluation in this project.

Practical significance:

1. We propose a paradigm for cost-efficient and effective ways to discover and develop peptide drugs. Such a platform can give rise to a new drug discovery start-up.

2. We discovered novel highly potent AMPs and will discover more. They can be potentially used in the actual animal husbandry industry.

ABSTRACT

From 2020, the addition of antibiotics has been banned in China's feed, aiming to reduce the harm caused by abusing antibiotics. To maintain normal development of animal agriculture, it is essential to find antibiotic alternatives, which are needed for animal normal growth. Antimicrobial peptides (AMPs), as natural antibacterial drugs, are less vulnerable to drug resistance because of their unique mechanism of action and are ideal substitutes for antibiotics. However, only **7 antimicrobial peptides are currently approved by FDA**, so it is important to discover more useful antimicrobial peptides. Previous AMP mining methods omitted the tertiary structure of proteins, failed to efficiently find AMPs with strong bactericidal activity, and were not tested on animal pathogen strains. This project uses evolutionary and tertiary structural information based on protein language models to discover new potent AMPs from animal data. Our pipeline is very cost-efficient and 84% of the candidates identified by our method are potent AMPs. Results have shown **8 experimentally validated highly potent novel AMPs** from the organism and their microbes' genome data, even **more potent than medicinal AMP polymyxin B and swine-gut AMPs like PR-39, Cecropin P1, and Porcine beta-defensin 2**. In addition, none of them show obvious cytotoxicity. Our

work has been sent out to **external peer review** **by** ***Nature Biomedical Engineering***. Our discovered 8 highly potent AMPs have great potential, we will further investigate them, such as conducting animal test, to see if they can be applied as feed additives in animal husbandry and find ways for transformational purposes.

1. OBJECTIVES AND SIGNIFICANCE

1. **To develop an AMP prediction model based on protein language models.** Protein language models are cutting-edge approaches that combine machine learning and bioinformatics. This ensures the model is both fast and accurate, enabling the discovery of novel AMPs at an unprecedented scale. Protein language models also encode significant evolutionary and functional information about proteins, enabling the discovery of evolutionary remote AMPs with novel antibacterial mechanisms that previous methods cannot detect.
2. **To discover highly potent AMPs for animal husbandry with the new pipeline from the large-scale animal metagenomic data.** Animal husbandry is a major contributor to antimicrobial resistance due to the overuse of antibiotics. The discovery of potent, targeted AMPs offers an alternative to antibiotics in livestock, reducing resistance pressure while maintaining animal health. By focusing on large-scale animal metagenomic data, our pipeline has the potential to discover AMPs specifically tailored to the microbiomes of livestock. This approach ensures greater efficacy and specificity, addressing the unique challenges in animal agriculture.
3. **To biosynthesize and validate the above discovered AMP candidates to see if these AMPs are suitable to be used as feed additives in animal husbandry.** Successful biosynthesis and validation of AMPs will provide proof-of-concept for our pipeline, demonstrating its potential to deliver tangible solutions to improve animal health, reduce antibiotic use, and promote sustainable agriculture.
   Our project proposed a holistic approach addresses urgent global challenges, including antibiotic resistance, sustainable livestock production, and food security. By applying AMPs as antibiotic alternatives in animal husbandry industry, we seek to create real-life impact and revolutionize animal husbandry.

2. RESEARCH METHODOLOGY

1. **Collect and integrate the protein sequences with the billion-scale metagenomics data.**
   The project utilizes a high-precision AMP prediction model to scan large-scale databases, enabling the discovery of novel antimicrobial peptides. We Integrated protein sequences and metagenomic databases, translating genomic data to proteins via PLASS. Unified datasets were deduplicated using CD-HIT, generating a billion-scale non-redundant database. Additionally, we classify genetic data at the in vivo biological environment level, which allows us to design AMPs tailored to specific animals.
2. **Train a protein language model dedicated to metagenomic data.**
   We trained a BERT-based protein language model on hundreds of millions of collected deduplicated sequences (CD-HIT processed) using high-performance computing. Unsupervised learning captured structural, evolutionary, and functional protein features for downstream prediction.
3. **Build an AI method to predict antimicrobial peptide, based on the annotated data, protein language model, and AlphaFold2.**
   Combined annotated AMP datasets (cleaned/standardized) with sequence-similar non-AMP data for model training. Integrated protein language model embeddings with AlphaFold2-derived structural features and traditional bioinformatics metrics (e.g., amino acid composition). The developed deep learning model was benchmarked using Accuracy, Recall, Precision, F1-score, ROC metrics.
4. **Use the prediction model to screen AMP candidates in the collected dataset**
   We applied the prediction model to large-scale metagenomic data, incorporating host-microbiome interactions and habitat-specific microbial community features. Automated filters included sequence redundancy removal, model score thresholds, and metaproteomic alignment for candidate prioritization.
5. **Biosynthesize antimicrobial peptide candidates and wet lab functional validation.**
   Candidate AMPs are synthesized using chemical synthesis techniques, and their antimicrobial activities are evaluated through biological experiments against common animal pathogens, including *Staphylococcus aureus*, *Salmonella spp.*, *pathogenic Escherichia coli*, *Streptococcus suis*, *Pseudomonas aeruginosa*, and *Klebsiella pneumoniae*. Additionally, experiments such as cytotoxicity assays are conducted to validate the biological safety of the AMPs.

3. RESULTS ACHIEVED SO FAR

Objective 1

Investigation scope: We checked the metagenome databases and their usage for large language model pre-training, specifically focusing on the improvement of models in protein/peptide-related tasks.

Results achieved: In terms of data collection, we conducted a comprehensive literature review and performed extensive searches of the PubMed database, prioritizing core publications related to genomic catalogs. This process yielded approximately 60,000 metagenome-assembled genomes (MAGs). We subsequently focused on acquiring samples derived from specific mammalian hosts, targeting gastrointestinal tract sites (including fecal, intestinal, and ruminal specimens). These data were integrated to construct the Mammalian Digestive Tract Microbial Genome Database (MDTMGD), a resource designed to facilitate antimicrobial peptide discovery across diverse animal species.

For protein language model construction, we employed the ESM-2 model architecture, and retrained it using metagenomic and AMP data to enhance its suitability for AMP-related tasks. Building upon this model, we developed an end-to-end hierarchical multi-label deep forest framework, HMD-AMP, to enable accurate AMP prediction and annotation of antimicrobial spectra. The current framework achieves SOTA performance in predicting evolutionary remote AMPs and AMP targets.

Problems encountered: Key challenge involves handling the MAGs for model training and balancing the training cost and model performance. High-quality data and model architecture is essential for the language model. Through comprehensive pipeline with rigorous data quality control and deeply exploration for the available protein language models, our developed models demonstrate significant improvement in AMP-related tasks. The curated database serves as a valuable and detailed resource for mining AMPs from animals.

Objective 2

Investigation scope: We explored the functional proteins discovery pipelines, specifically focusing on processing the large genome data into short peptides.

Results achieved: We established a comprehensive AMP mining pipeline, focusing on the porcine and seven mammalian species: yellow baboon, mouse, cattle, water deer, Siberian roe deer, horse, and porcine. The pipeline processes metagenomic data as input, beginning with open reading frame (ORF) prediction followed by translation of nucleotide sequences into polypeptide sequences. By applying our developed AMP prediction model, HMD-AMP, to these sequences, we identified more than 20 million candidate AMP sequences. Then, considering there is a pressing need for new therapeutic agents in swine production due to the increasing resistance to existing antibiotics, we prioritized the swine and its gut microbe genomes to uncover AMPs that can potentially serve as viable alternatives to alleviate antibiotic resistance.

We cross-referenced metaproteomic data from pigs and their gut microbiota with our candidate sequences, merging the datasets to retain only candidate sequences experimentally validated as expressed polypeptides. A total of 187 candidate AMPs were identified from the porcine genome and 7,460 candidate AMPs from the porcine gut microbial genomes.

Problems encountered: Our model identified millions of candidate AMPs, synthesize and validate all of them is impossible. Narrow down the candidates significantly to obtain more promising ones to validate is needed. At the same time, this operation should not be time-consuming and labor-intensive, to ensure the efficiency of the pipeline. Our cross-reference from metaproteome largely and efficiently reduced the candidates' number while maintaining ones are more likely to be AMPs.

Objective 3

Investigation scope: We investigated the experimental subjects to verify the antibacterial effect of AMPs. To discover feed additives, we focused on the Gram-positive and Gram-negative bacteria.

Results achieved: 2 porcine-derived and 60 microbiota-derived candidate AMPs with prediction scores ≥0.9 were chemically synthesized and subjected to antimicrobial activity assays against six common porcine pathogens. All 62 tested peptides exhibited antimicrobial activity, with 52 demonstrating potent bactericidal effects. Notably, 8 AMPs displayed exceptional broad-spectrum activity, inhibiting all six pathogens with efficacy comparable to polymyxin B and vancomycin. For these eight AMPs, we established a comprehensive evaluation framework, including minimum inhibitory concentration (MIC) determination and cytotoxicity assays (Figure 1), to further assess their antimicrobial potency and safety profiles. Cytotoxicity assays confirmed their low toxicity, with partial

adverse effects observed only at concentrations far exceeding their MICs (Figure 1). These AMP shows the potential to serve as antibiotic alternatives in animal husbandry.
Our work for the above-mentioned results is sent out for peer review by the top international journal Nature Biomedical Engineering.
Problems encountered: We have thousands of candidate AMPs, however, considering the high cost of wet-lab experiments, we sampled 62 peptides for validation. Such amount may lead to the inability to find AMPs that have great potential as feed additives. Despite the limitation, as many as 8 highly potent AMPs were discovered, indicating the effectiveness of our pipeline.

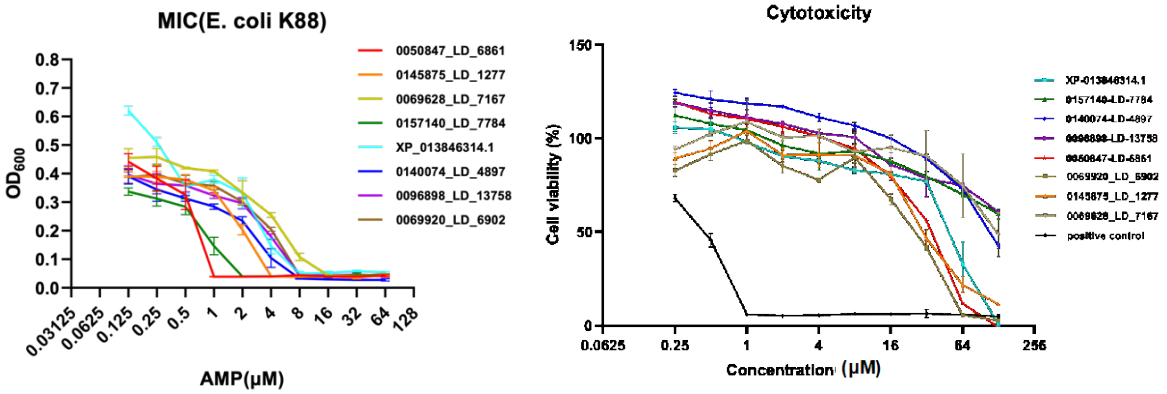


Figure 1 **Left:** The MIC of our discovered most potent 8 AMPs on *E. coli K88.*
Right: The cytotoxicity test results of our discovered highly potent AMPs

Table 1 Summary of objectives addressed to date

Objectives	Percentage achieved (estimated)
To develop an AMP prediction model based on protein language models.	100%
To discover highly potent AMPs for animal husbandry with the new pipeline from the large-scale animal metagenomic data.	80%
To biosynthesize and validate the above discovered AMP candidates to see if these AMPs are suitable to be used as feed additives in animal husbandry.	50%

Results expected in the next reporting period
Objective 2
We have obtained thousands of candidate AMPs from swine with very high confidence. Since our goal is to discover feed additive AMPs for animal husbandry, based on the 20 million candidate AMP, we will extend the process of cross-reference metaproteomes to animals other than swine to obtain a diversity of candidate AMP sequences from various resources.
Objective 3
Besides cytotoxicity, we will test the hemolysis effect of discovered highly potent AMPs to comprehensively evaluate the toxicity of AMPs. Then, we will select highly effective and low-toxicity AMPs for animal experiments to evaluate their effects on promoting animal growth, regulating gut microbiota, and ensuring safety.
To improve the stability of AMPs *in vivo* and reduce their susceptibility to proteolytic degradation, chemical modifications such as cyclization and PEGylation will be employed. These modifications will also optimize the solubility and bioavailability of AMPs, making them more suitable for use as feed additives.

Patent
We applied PCT patent for our AMP discovery pipeline and our discovered AMPs. The patent's name is "MACHINE LEARNING PIPELINE FOR DISCOVERING NOVEL ANTIMICROBIAL PEPTIDES," and it was published with Publication no. WO 2024/169915 A1.

4. PUBLICATION AND AWARDS

J[1] Shen, T., Hu, Z., Sun, S. et al., "Accurate RNA 3D structure prediction using a language model-based deep learning approach," *Journal*, Nature Methods, 2024: 1-12.
J[2] Wang J, Fan Y, Hong L, et al., "Deep learning for RNA structure prediction," *Journal*, Current Opinion in Structural Biology, 2025, 91: 102991.

WORK IN PROGRESS

DEVELOPMENT OF MITOCHONDRIA-TARGETING, SINGLE-ATOM NANOZYME FOR ACCELERATED BONE REGENERATION

Principal Investigator: Professor LI Zhong Alan
Department of Biomedical Engineering, CUHK

Co-investigator(s):
Yuwen Wang, PhD Candidate [(1)],
Qiongjiao Zeng, Research Assistant [(1, 2)],
Xian Chen, Research Assistant [(1,2)]

(1) Department of Biomedical Engineering, Faculty of Engineering, CUHK
(2) Shun Hing Institute of Advanced Engineering, CUHK
(3) Center for Neuromusculoskeletal Restorative Medicine
(4) Institute for Tissue Engineering and Regenerative Medicine, Faculty of Medicine, CUHK

Reporting Period: 1st July 2024 – 30th April 2025

INNOVATION AND PRACTICAL SIGNIFICANCE:

The major innovation of the current work lies in the design of a biomaterial based on cell-derived ECM, which potentially enables the delivery and long-term stabilization of cell-derived therapeutic signaling factors. We have already filed a provisional patent application in May 2019 (Anna Blocki and Marisa Assuncao, Chinese University of Hong Kong (2019) "Process and material for tissue healing" US provisional patent application: 62/848,971), protecting this material design.

By utilizing this biomaterial, we are able to address major limitations of cell-based therapies, such as limited engraftment and survival of transplanted cells, limited therapeutic efficacy of conditioned media, immunological concerns, when allogeneic cell sources are utilized (ECM is highly conserved and thus not evoke an immune response), etc. At the same time, the engineered biomaterial exhibits the necessary complex bioactivity to guide complex tissue healing processes, in contrast to selected biologics or simple scaffolds

The unique techniques utilized, enable the synthesis of larger amounts of biomaterial, thereby ensuring a stable/reproducible bioactivity and are necessary for future scale-up production.
The bioactive material can be stored and thus utilized off-the-shelf. It can be processed and incorporated in all types of materials including tissue scaffolds, implants, wound dressings and (injectable) hydrogels. Hence, just by itself or incorporated into other materials, it can be applied to tissue areas with chronically inflamed and dysregulated microenvironments, thereby modulating and turning the diseased environment into a pre-healing one. This will advance the healing and regeneration process in non-healing and non-regenerative tissues such as osteoarthritis and beyond.

ABSTRACT
Critical-sized bone defects do not heal spontaneously throughout a patient's lifetime, posing a global challenge to musculoskeletal health. Resident stem cells in bone, which are indispensable in skeletal development and regeneration, undergo enhanced mitochondrial activities during osteogenic differentiation. However, accumulation of excessive reactive oxygen species produced by injured bone tissues can lead to oxidative stress and mitochondrial damage, which negatively affects the osteogenic differentiation of stem cells and tissue repair. In such an environment, it is crucial to target stem cells mitochondria for reactive oxygen species removal and restore mitochondrial homeostasis for optimal osteogenic differentiation. Single-atom nanozymes exhibit ultra-high atom utilization efficiency, making them highly promising materials for modulating mitochondrial energy metabolism. Herein, we developed a dendritic mesoporous silica nanoparticle (DMSN)-based nanozyme, named TPP-DMSN-Fe/Cu, loaded with Fe and Cu single atoms and modified with mitochondrion-targeting triphenylphosphonium (TPP). In vitro, TPP-DMSN-Fe/Cu was found to upregulate stem cells osteogenesis by scavenging reactive oxygen species, enhancing mitochondrial function, and promoting autophagy of divided and abnormal mitochondria.

Furthermore, in vivo experiments demonstrated that TPP-DMSN-Fe/Cu nanozymes significantly enhanced mitochondrial biogenesis, promoted bone regeneration, and increased bone volume and bone mineral density. Therefore, the multifunctional, mitochondria-targeting nanosystem of TPP-DMSN-Fe/Cu holds enormous potential in accelerating bone regeneration by regulating cellular energy metabolism.

1. OBJECTIVES AND SIGNIFICANCE

This study aims to develop mitochondria-targeted nanozymes (TPP-DMSN-Fe/Cu) that mimic NADH oxidase and cytochrome c oxidase to enhance the mitochondrial electron transport chain (ETC), leveraging TPP for precise mitochondrial delivery, thereby scavenging reactive oxygen species (ROS), restoring ATP synthesis via OXPHOS/TCA cycle reactivation, and reducing oxidative damage. Concurrently, it seeks to elucidate metabolic reprogramming mechanisms by analyzing the nanozyme's role in mitochondrial biogenesis, autophagy-mediated turnover, and metabolic shifts in stem cells—including glycolytic downregulation, lipid β-oxidation upregulation, and glutathione (GSH) elevation to stabilize redox homeostasis. Finally, the therapeutic efficacy was evaluated in critical-sized bone defect models using micro-CT, histology, and immunostaining to quantify bone regeneration. Notably, we compared outcomes to existing therapies (e.g., growth factors) and assessed systemic safety (biodistribution, inflammation) to validate the translational potential of our nanozymes.

This project offers dual therapeutic actions, namely mitochondrial ROS scavenging and OXPHOS restoration, to address oxidative stress and metabolic stagnation in bone defects. Its precision mitochondria-targeting ability ensures localized mitochondrial repair without systemic toxicity, and cost-effective synthesis enables scalable production. Preclinical results show higher bone density and faster regeneration with nanozyme treatment. The GelMA scaffold enables localized delivery, improving bioavailability and minimizing off-target effects. Unlike single therapeutic mechanism-based approaches (e.g., transient antioxidants), this strategy provides a multifunctional strategy for efficient bone regeneration with promising potential for clinical translation. Additionally, it offers a potential solution to treat various other bone conditions.

2. RESEARCH METHODOLOGY

Animal experimental protocols conducted in this study were approved by the Animal Experimentation Ethics Committee of the Chinese University of Hong Kong (Approval No. (24-427) in DH/HT&A/8/2/1 Pt.63). **Synthesis of TPP-Fe/Cu-DMSN (Fig. 1):** (1) dendritic mesoporous silicon (DMSN) synthesis, (2) DMSN-NH2 synthesis, (3) DMSN-Fe/Cu synthesis, and (4) TPP-Fe/Cu-DMSN synthesis.

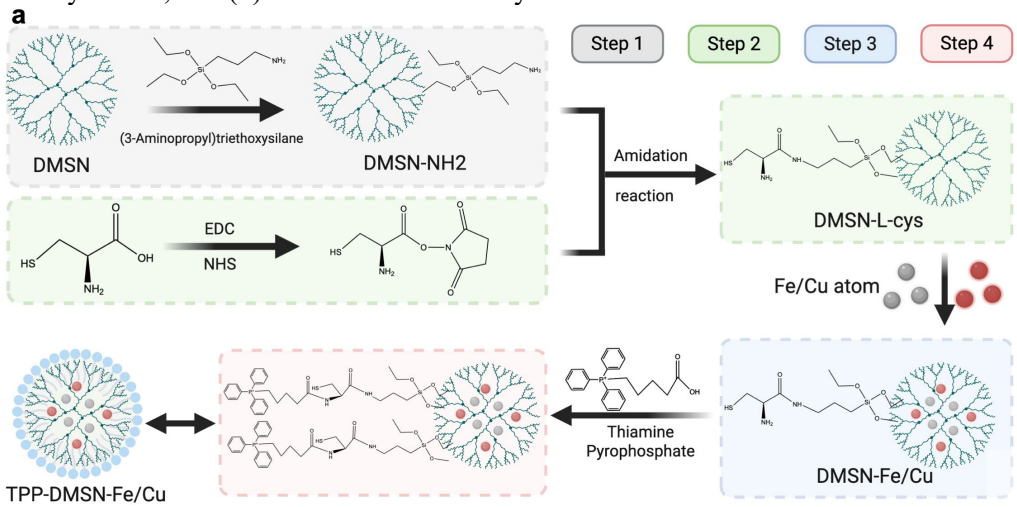


Fig. 1 Schematic of the synthesis process of TPP-DMSN-Fe/Cu nanozymes.

Characterization of NPs: The NPs were characterized by ultraviolet–visible (UV–vis) transmittance spectroscopy, transmission electron microscopy (TEM), scanning transmission electron microscope (STEM, for elemental analyses), Zetasizer (Malvern), and FT-IR (Cary 630). **Cell culture**: C3H/10T1/2, Clone 8 stem cells were procured from Oricell (Guangzhou, China). Low-glucose Dulbecco's modified Eagle's medium (DMEM; Gibco) supplemented with 10% fetal bovine serum (FBS) and 1% penicillin-streptomycin was used as cell culture medium. For osteogenic differentiation, cells were transitioned to high-glucose DMEM containing 100 nM dexamethasone,

10 mM β-glycerophosphate, and 50 μg/mL L-ascorbic acid (Sigma Aldrich). The differentiation medium was refreshed every 48–72 hours to ensure consistent nutrient availability. All culture reagents, unless specified otherwise, were sourced from Gibco. **Cell viability**: Cell Counting Kit-8 (CCK-8; APExBIO, USA) was used to determine the viability and proliferation of stem cells. The optical density was measured at 450 nm using a SpectraMax iD3 microplate spectrophotometer (Molecular Devices, USA). **Cellular uptake**: NPs were internalized by stem cells were observed using TEM (Hitachi H-7650; Hitachi, Tokyo, Japan) and confocal microscopy (Leica SP8). For the latter, the cells and NPs were stained with Mito-tracker green and Rhodamine B isothiocyanate (RBITC), respectively. **Alkaline phosphatase (ALP) activity and staining**: ALP activity of the cells was detected using the ALP Assay Kit (Shanghai Beyotime Biotechnology Institute). The BCIP/NBT ALP Staining Kit was used according to the manufacturer's instructions. The ALP staining results were observed using an inverted light microscope (Nikon, ECLIPSE Ts2). **Alizarin red staining and quantification**: ARS staining was conducted with 2% ARS staining solution (Sigma-Aldrich). Quantitative analysis was carried out by measuring the absorbance at 562 nm after desorption of the stained mineral deposits with 10% (w/v) cetylpyridinium chloride (Aladdin). **RNA sequencing**: Total RNA was isolated from stem cells using TRIzol reagent. Sequencing was performed on an Illumina Novaseq 6000 platform by LC-Bio Technologies (Hangzhou, China). **Real-time RT-PCR**: Total RNA was extracted using the RNA Extraction Kit (ZYMO). The primers were synthesized by BGI Genomics. The transcript levels of the target genes were calculated using the ΔΔCt method. **Western blotting**: Proteins were extracted using RIPA lysis buffer (Thermo Fisher). After BCA (Thermo Fisher) quantification, proteins were loaded into sodium dodecyl sulphate polyacrylamide gels and transferred to Trans-Blot Turbo Midi 0.2 μm PVDF Transfer Packs (Bio-Rad). **Calcium^{2+} influx**: Intracellular calcium ion influx imaging was performed using a Leica Thunder imager to record the intensity of the intracellular calcium dye (Fluo-4 AM, Beyotime). The calcium imaging data obtained were analyzed and visualized using ImageJ. **Seahorse**: Seahorse XFe96 analyzed oxygen consumption rate (OCR) and extracellular acidification rate (ECAR) to assess OXPHOS, glycolysis, and lipid metabolism. **H_2O_2 Catalytic Elimination Assay**: 100 μL of NP-treated H_2O_2 was added to 100 μL of Ti $(SO_4)_2$ solution, and the absorbance value was recorded every 1 h until 6 h. **DPPH Scavenging Assay**: 1ml DPPH radical working solution (0.1mM) was mixed with 1ml NPs solution and kept in the dark for 60 min at 25 °C. The absorbance was measured by a UV-spectrophotometer at 517 nm. **•OH Scavenging Assay**: 800 μL deionized water, 20 μL H_2O_2 (3%), 100 μL TMB (2 mg/mL) and 20 μL CuCl2 (1 mg/mL) were mixed with 20μL NPs solution, the mixture absorbance was then measured by UV-spectrophotometer at 652nm. **Cyt c oxidase mimicking activity**: 100 μL Cyt c solution (1 mg mL-1, PBS buffer) and 100 μL as-prepared nanozyme (1 mg mL-1) were added sequentially into a 5 mL ep tube containing 1 mL of solution (pH 7.4). UV-vis absorption measurements were performed after 60 min. **NAD$^+$ contents detection**: NAD^+ levels in cells treated with/without nanozymes were measured using a commercial NAD^+/NADH assay kit (Beyotime) via the WST-8 method. Absorbance at 450 nm was measured against a standard curve to calculate concentrations. **Therapeutic efficacy in a rat model of critical-size bone defects**:
50 rats (8-12 weeks, male, 200-220 g) were randomly divided into five groups: empty defect (control), GelMA scaffolds, DMSN + GelMA scaffolds, DMSN-Fe/Cu + SF/GelMA scaffolds, and TPP-DMSN-Fe/Cu + GelMA scaffolds. Bone defects with 3-mm diameter and 3-mm depth were created for 4- or 8-week treatments. **Histology, Immunohistochemistry, and Immunofluorescence Staining**: Sections underwent hematoxylin and eosin (H&E), Masson's trichrome using commercial kits. Images intensity was semi-quantitatively analyzed using ImageJ software (NIH, USA). **Micro-CT Analysis**: Mandibular defects were analyzed at 4- and 8-weeks post-treatment using a Quantum GX2 micro-CT system (Revvity). **Statistical analysis**: Comparisons between two groups were made using Student's t-test, and comparisons between groups were made using one-way ANOVA followed by Tukey's post hoc test using GraphPad Prism 8. The level of statistical significance was set at $P < 0.05$. At least three independent replications were performed for each experiment.

3. RESULTS ACHIEVED SO FAR

Scientific achievements: We developed a mitochondria-targeted nanozyme platform that mimics the enzymatic activity of NADH oxidase (Complex I) and cytochrome c oxidase (Complex IV), pivotal components of the mitochondrial respiratory chain. By enhancing FAO in stem cells, the TPP-DMSN-Fe/Cu nanozymes amplify electron transport in OXPHOS and modulate the TCA cycle, restoring redox balance and boosting ATP synthesis. The nanozymes have been shown to selectively localize to the mitochondria, where they function to scavenge ROS, elevate glutathione to fortify antioxidant defenses, and enhance GTPase activity to regulate mitochondrial dynamics. This process promotes fission and the clearance of damaged mitochondria by autophagy. Concurrently, the nanozyme activates the CaMKK/AMPK pathway, thereby stimulating mitochondrial biogenesis and creating a

microenvironment conducive to osteogenesis (**Fig. 2**). In animal models, the system demonstrated robust bone regeneration, marked by accelerated defect repair and increased bone mineral density (**Fig. 3**). This suggests a direct link between mitochondrial functional recovery and functional bone regeneration.

Potential for clinical translation, commercialization & technology transfer: Our bone regeneration technology offers dual therapeutic actions, namely mitochondrial ROS scavenging and OXPHOS restoration, to address oxidative stress and dysregulated metabolism in bone defects. Its precision mitochondria-targeting ability ensures localized mitochondrial repair without systemic toxicity. Unlike single therapeutic mechanism-based approaches (e.g., transient antioxidants), this strategy provides a multifunctional strategy for efficient bone regeneration with promising potential for **clinical translation**. Additionally, it offers a potential solution to treat various other bone conditions such as osteoporosis. Notably, we employed a cost-effective synthesis protocol, enabling **scalable production** for commercialization. To prepare for future **technology transfer**, we are currently drafting a patent application. Additionally, we are actively seeking grants to support large animal studies to further verify the efficacy of our nanozymes for bone regeneration.

4. PUBLICATION AND AWARDS

J[1] Y. Wang, X. Zhang, D. Xie, C. Chen, Z. Huang, and Z.A. Li, "Chiral Engineered Biomaterials: New Frontiers in Cellular Fate Regulation for Regenerative Medicine," *Advanced Functional Materials* (2024) 2419610.
J[2] Y. Wang, H. Jan, Z. Zhong, L. Zhou, K. Teng, Y. Chen, J. Xu, D. Xie, D. Chen, J. Xu, L. Qin, R.S. Tuan, and Z.A. Li, "Multiscale metal-based nanocomposites for bone and joint disease therapies", *Materials Today Bio* 32 (2025) 101773.

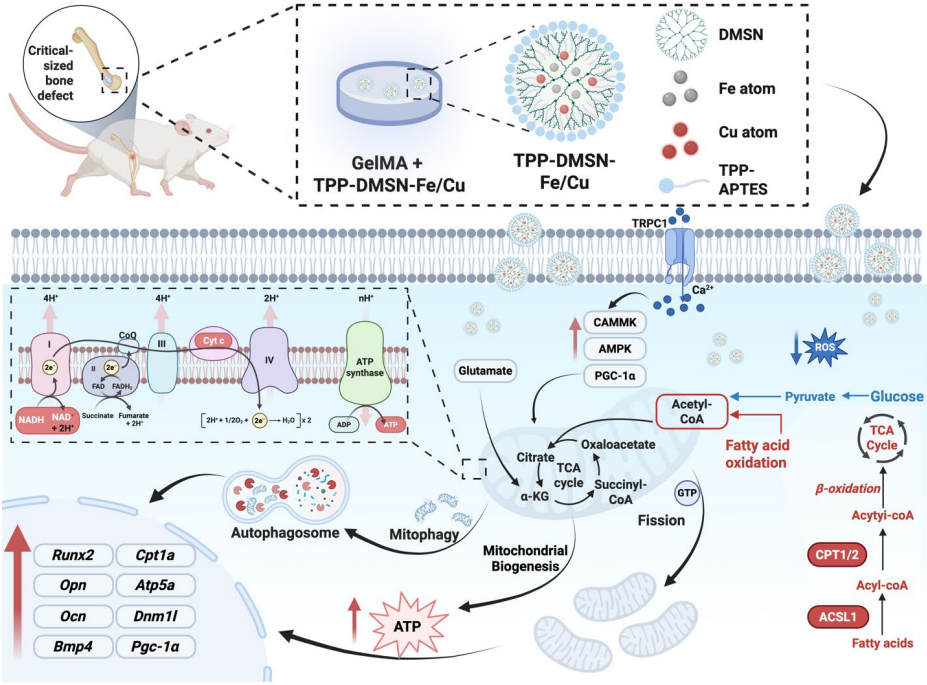


Fig. 2 Schematic summary of major findings thus far. TPP-DMSN-Fe/Cu nanozymes can reduce oxidative stress, improve mitochondrial function, and promote osteogenic differentiation of stem cells.

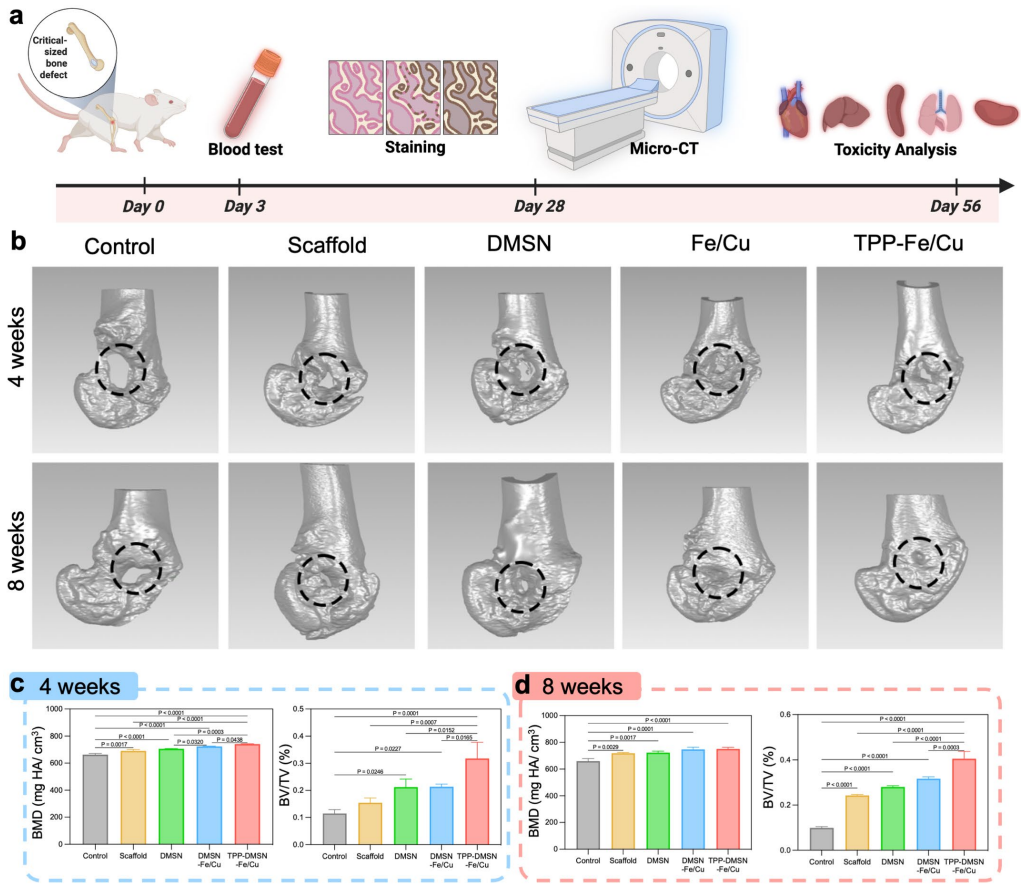


Fig. 3 In vivo tests of TPP-DMSN-Fe/Cu nanozymes in rats. a Schematic diagram of the treatment protocol. **b** 3D reconstructed micro-CT scan images. **c, d** Quantitative analysis of micro-CT results (n = 3).

WORK IN PROGRESS

DESIGN, OPTIMIZATION, AND EXPERIMENTAL VALIDATION OF A HANDHELD VARIABLE-CURVATURE HYBRID-STRUCTURE ROBOTIC INSTRUMENT (HVHRI) FOR MAXILLARY SINUS SURGERY

Principal Investigator: Professor MA Xin[1,4]
Department of Mechanical and Automation Engineering, CUHK

Co-Investigator:
Weibin Li[2], Xianfeng Xia[3], Samuel Au[1,4], Zheng Li[3,4]

Research Team Members: Yi Yang[4],, Xuchen Wang[1,4], Puchen Zhu[1,4]

(1) Dept. of Mechanical and Automation Engineering, The Chinese University of Hong Kong.
(2) School of Computer Science, Sun Yatsen University.
(3) Dept. of Surgery, The Chinese University of Hong Kong.
(4) Multi-Scale Medical Robotics Center, Hong Kong.

Reporting Period: 1st July 2023 – 31st May 2024

INNOVATION AND PRACTICAL SIGNIFICANCE:

To include a paragraph to highlight specifically the innovation and practical significance of your work. Both VC and the donor would like to see more research endeavors be directed to innovation and technology transfer for the betterment of mankind.

We will develop a dexterous, compact HVHRI with high bending and torsional stiffness for sinus surgery. Compared with existing robotic flexible instruments, the HVHRI has larger reachability and dexterity, which will reduce the trauma to the patients in surgery and the possibility of complications after surgery.

ABSTRACT

Existing robotic flexible medical tools for maxillary sinus surgery are still low in dexterity and big in diameter, which results in big damage to the patients. We will develop a novel 4-DOF handheld hybrid-structure, variable-curvature robotic instrument (HVHRI) (including a 1-DOF variable-curvature flexible bending section and a 3-DOF distal grasper), which is compact in size (diameter is 3.5 mm; actuation system weighs < 800g) and can provide sufficient bending (14 $N{\cdot}mm$) and torsional stiffness (0.5°/$N{\cdot}mm$). To our knowledge, it is the thinnest flexible instrument for sinus surgery. And it is capable of sharply bending at the distal end (bending radius is 1.5 mm). To further enlarge the reachable space and dexterity of the HVHRI inside the maxillary sinus, a novel structure parameter optimization framework (maximizing the reachable space and dexterity of the HVHRI) will be studied. The dexterity and reachability of the existing two-segment flexible tool, the hybrid-structure flexible tool, and the HVHRI will be compared by simulations and experiments. Besides, we will propose two intuitive control methods for the HVHRI (handheld and integrated with robot arm). Several phantom and cadaver experiments will be conducted to validate the feasibility, dexterity, reachability, bending stiffness and torsional stiffness of the HVHRI.

1. OBJECTIVES AND SIGNIFICANCE

1. Existing robotic medical tools for maxillary sinus surgery are still straight, low in dexterity and big in diameter (> 4 mm), which results in big damage to the patients (a big incision is often needed in the face). We will develop a novel 4-DOF HVHRI that has a small diameter (3.5 mm), high reachability and dexterity. With the HVHRI, surgeons can dramatically reduce the trauma caused to the patients in surgery.
2. Most of the existing medical instruments are operated manually. However, the enhanced dexterity adds to

the difficulty of manual operation, which in turn imposes upon surgeons longer learning curve and more workload. To facilitate sterilization, reduce surgeons' learning time and avoid surgeons' fatigue, we will develop a compact actuation system for the HVHRI (with weight < 800g) which can be easily handled by hand or integrated with robot arm. This actuation system enables convenient sterilization and easy change of instruments for the surgeons.

3. A novel structure optimization framework will be developed for maximizing dexterity and reachability of HVHRI inside the maxillary sinus area.

4. To validate the feasibility, operability, and workload of the HVHRI, we will conduct several simulations and experiments. Simulations will be conducted to compare the dexterity and reachability between existing medical instruments and the HVHRI. Several phantom and cadaver experiments will be conducted to validate the feasibility, dexterity, reachability, bending stiffness and torsional stiffness of the HVHRI. Results of the simulations and experiments will provide surgeons with detailed data when they choose instruments.

2. RESEARCH METHODOLOGY

2.1. Development of a novel variable-curvature, hybrid-structure manipulator (VHM) that has a small diameter (3.5 mm) and high reachability and dexterity in the maxillary sinus area.

As is shown in Fig. 1(a), we will propose a novel VHM. (1) In this VHM, we will develop a novel 3-DOF gripper (Fig. 1(b)), which can be driven by four cables. The gripper (the diameter is 3.5mm; the length is 9 mm) is more compact than existing ones and can bend 90° in two directions. (2) Besides, we will develop a 1-DOF variable-curvature flexible bending section (Fig. 1(c)). The joint is produced on metal tubes by laser cutting. And guide rings can be easily made by sheet metal forming technology to constrain the 6 actuating cables. Different maximum bending angles (α_i) will be set at each joint of the flexible bending section. In operation, due to the different joint limit of each joint (α_i), the curvatures of the flexible bending section can be manipulated to increase the reachability of the VHM. (3) Besides, we will design a novel metal-woven mesh to further increase the overall torsional stiffness of the manipulator. And the FEA simulation results (Fig. 1(d)) show that the torsional stiffness of the metal-woven mesh is twice better than the existing nitinol flexible bending section. All in all, the VHM contains a 3-DOF gripper, a 1-DOF flexible bending section, and a novel metal-woven mesh.

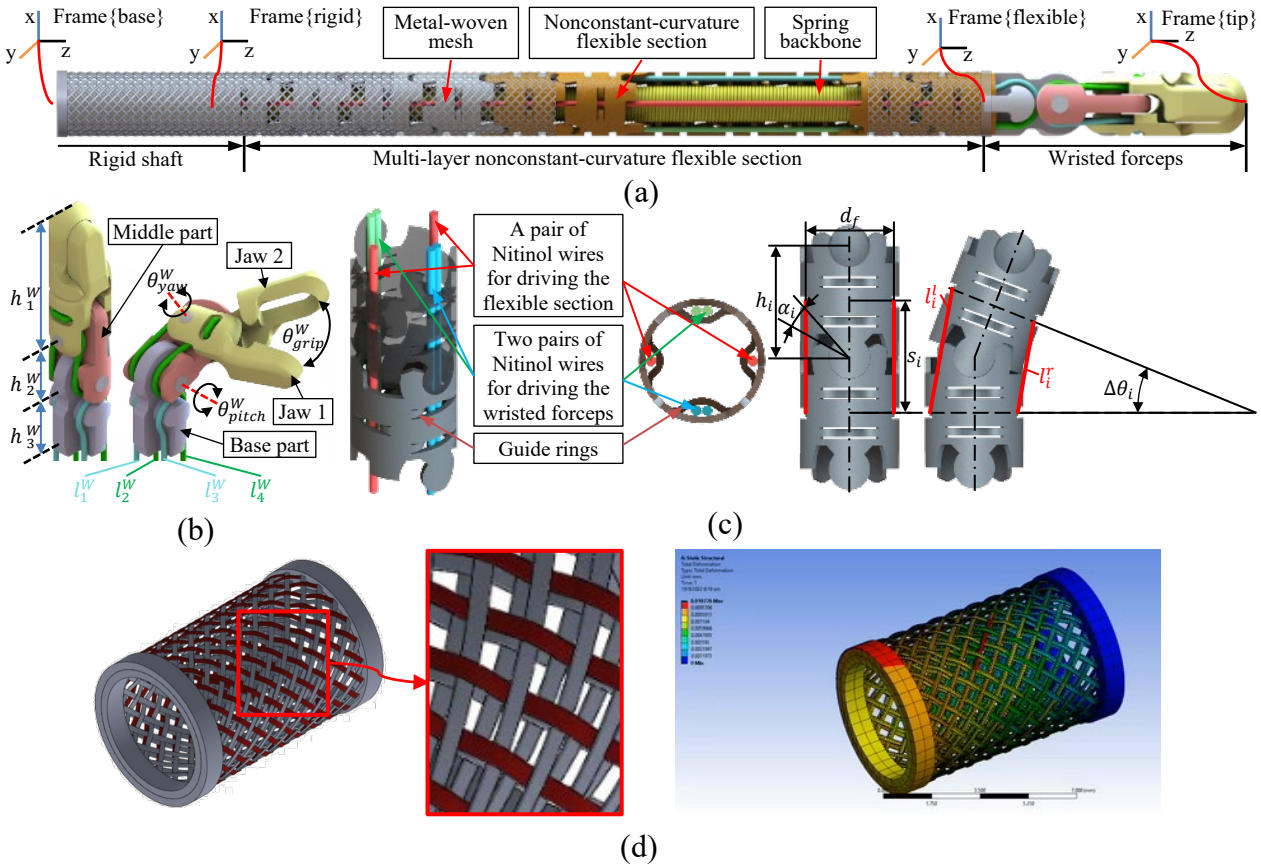
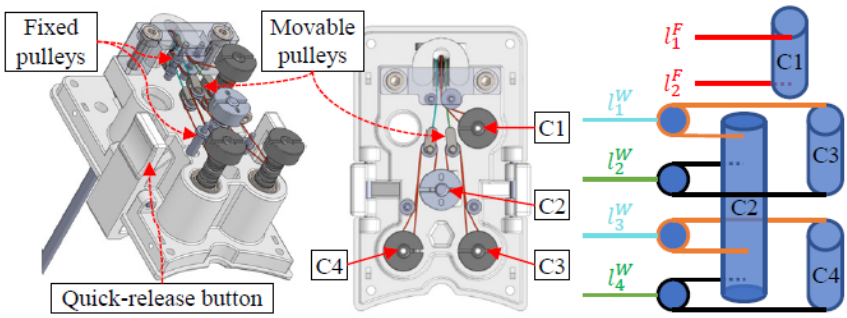


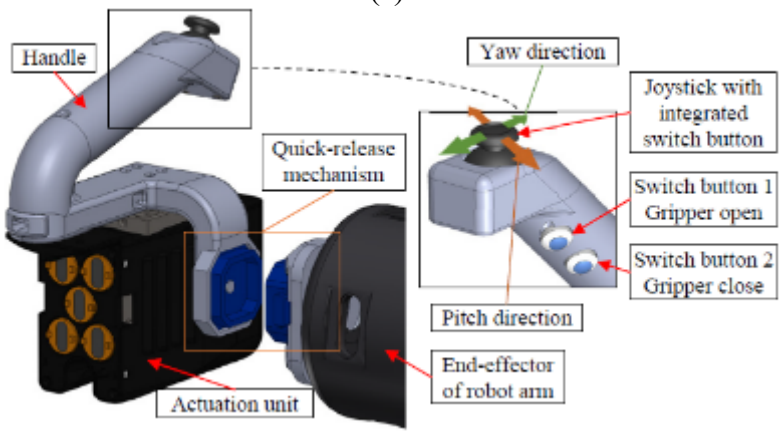
Fig. 1. (a) Overview of the variable-curvature, hybrid-structure manipulator. (b) The structure of the 3-DOF gripper. (c) The structure of a single joint of the 1-DOF variable-curvature flexible bending section. (d) The metal-woven mesh for further increasing the overall torsional stiffness of the manipulator and its FEA simulation.

2.2. Development of a compact actuation system (with weight < 800g) that can be easily handled by hand or integrated with robot arm.

First, we will develop a backend with a novel decoupling mechanism (see Fig. 2(a)) to connect HVM and the actuation system. With this backend, we can use three motors to separately control the motion of the 3-DOF gripper (pitch, yaw and open/close) and use one motor to control the motion of the flexible bending section so that the coupling motions of the gripper and the flexible section can be reduced. In addition, as shown in Fig. 2(b), we will develop a novel interface to operate the HVHRI. The interface will contain a joystick and two buttons. The joystick is used to control the manipulator and the two buttons are used to switch the gripper between “open” and “close”. The actuation system contains four motors. The manipulator can be detached via a quick-release structure, which is convenient for sterilization.



(a)



(b)

Fig. 2. (a) Backend with decoupling mechanism. (b) Interface for operating the HVHRI.

2.3. Structure optimization for the 4-DOF HVHRI by maximizing the dexterity and reachability inside the maxillary sinus area.

As is shown in Fig. 3, we will propose a novel structure optimization framework for the VHM. In this framework, we will further consider both the dexterity and reachability of the instruments in the maxillary sinus area. To my knowledge, this is the first time to optimize the structure of surgical instruments considering both the reachability and dexterity for the sinus surgery. Besides, more structure parameters will be optimized (including the different maximum bending angles for the joints of the flexible bending section) in the framework.

Algorithm 1 Structure Parameters Optimization Framework

Input: Positions of environment cloud points (P_{envir}), positions of target space's cloud points (P_{target}), and positions of waypoints (P_{way}).

Output: Number of sections (N_{seg}), length of each section (L_i, $i = 1, 2, \cdots, N_{seg}$), joint limit of each section ($\theta_{limit-i}, i = 1, 2, \cdots, N_{seg}$).

for $N_{seg} = 1 : N_{seg}^{max}$ **do**
  for $i = 1 : N_{seg}$ **do**
    for $L_i = L_i^{min} : L_i^{max}$ **do**
      for $\theta_{limit-i} = \theta_{limit-i}^{min} : \theta_{limit-i}^{max}$ **do**
        $p_i = [L_i, \theta_{limit-i}]$
      end for
    end for
  end for
  $\mathrm{Dex}_{global}^{max}(p) = MaxOpt(\mathrm{Dex}_{global}(p), P_{envir}, P_{target}, P_{way})$
end for
$[L_1, \theta_{limit-1}, \ldots, L_{N_{seg}}, \theta_{limit-N_{seg}}]_{optimal} = Argmax(\mathrm{Dex}_{global}^{max}(p))$

Fig. 3. Structure optimization framework for HVHRI

2.4. Experimental validation for the HVHRI.

To evaluate the performance of the HVHRI, we will conduct several experiments in six aspects: 1) tracking the trajectories of the HVHRI's motions to verify the proposed kinematics and decoupling mechanisms; 2) a bending stiffness test on the HVHRI to show how much force the hybrid structure endures; 3) a torsional stiffness test on the HVHRI to show how many twists it overcomes; 4) comparing the reachability and dexterity of the HVHRI and other flexible instruments; 5) a feasibility study in a 3D maxillary sinus phantom to show the reachability and dexterity of the HVHRI; and 6) cadaver experiments for validating the feasibility of the HVHRI.

3. RESULTS ACHIEVED SO FAR

3.1. Development of a novel variable-curvature, hybrid-structure manipulator (VHM) that has a small diameter (3.5 mm) and high reachability and dexterity in the maxillary sinus area.

In this project, we prototyped HVHRI (see Fig. 4(a)) comprises three major components: a detachable instrument unit, an actuation unit, and a handheld user interface. **The detachable instrument unit** consists of a VHM and a backend transmission mechanism (BTM). The VHM includes a pair of 3-DOF wristed forceps and a 1-DOF multi-layer variable-curvature flexible section. **The 3-DOF wristed forceps:** In order to address the challenges associated with reducing the diameter of surgical instruments for MSS, a specialized design of the 3-DOF wristed forceps with a 3.5 mm diameter was developed. There will be challenges to directly scale down the diameter of a conventional Φ8 mm instrument to Φ3.5 mm: (1) accommodating six driving cables within limited space, (2) ensuring high stiffness of thinner driving cables (Φ0.5 mm to Φ0.2 mm), and (3) avoiding sharp bending angles

for driving cables with small pulley diameters (Φ3.8 mm to Φ1.6 mm). To overcome these challenges, a novel approach is employed in the design. The 3-DOF wristed forceps incorporate a four-cable wrist mechanism, utilizing symmetrically fixed cables with high friction paths on the jaws to control pitch ($\boldsymbol{\theta}^{\boldsymbol{W}}_{\boldsymbol{pitch}}$), yaw ($\boldsymbol{\theta}^{\boldsymbol{W}}_{\boldsymbol{yaw}}$), and grip ($\boldsymbol{\theta}^{\boldsymbol{W}}_{\boldsymbol{grip}}$) motions. By using only four driving cables, more space is made available to accommodate thicker cables, thus effectively enhancing stiffness. Nitinol wires with a diameter of 0.3 mm are selected as the

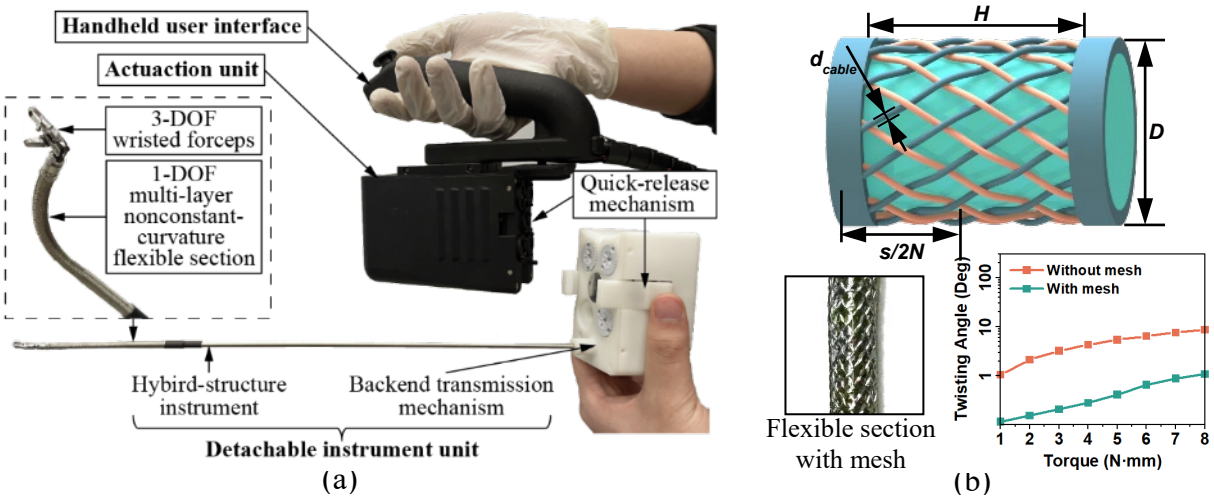


Fig. 4. (a) Overview of the HVHRI prototype. (b) Model, prototype, and torsional stiffness test of the metal-woven mesh.

driving cable, which provides up to 50 N tension. To avoid bending the driving cables with sharp angles, the four driving cables are designed to be in contact with the smooth convex sliding surfaces of the parts to eliminate the need of pulleys with smaller diameters. As well, the streamlined design of the wristed forceps assembly comprised only six parts, facilitating both manufacturability and assemble ability. **The 1-DOF multi-layer variable-curvature flexible section with high stiffness:** Flexible medical tools face several challenges that hinder their effectiveness in various procedures. Firstly, the use of constant curvature limits flexible instruments' dexterity and reachability, particularly in complex anatomical structures such as the nasal and maxillary sinus cavities. Secondly, existing robotic flexible instruments are often low in bending and torsional stiffness, which impacts their stability during tissue manipulations. This compromises the surgeon's ability to perform delicate maneuvers accurately. Lastly, the diameter of current robotic flexible instruments for sinus surgery typically ranges from 4 to 5 mm, which may pose challenges when operating in tight space and may limit their access to certain areas within the maxillary sinus. To overcome these challenges, we propose a multi-layer variable-curvature flexible section, allowing for improved dexterity, reachability, and enhanced stiffness for the HVHRI. The flexible section consists of three layers, including 1) a metal-woven mesh, 2) a variable-curvature flexible section, and 3) a spring backbone. ***First layer:*** The metal-woven mesh is woven with 24 braided steel cables of 0.1 mm diameter (see Fig. 4(b)). The mesh is tightly welded to the flexible section to enhance the torsional stiffness of the flexible section and ensure smooth insertion through the nasal cavity during the surgery. Based on the torsional stiffness test results, the metal-woven mesh decreased an average of 89.41% twisting angle of the original flexible section. ***Second layer:*** The flexible section is manufactured by laser-cutting with features of variable curvature, small diameter, and complex decoupling guide rings for driving cables. The variable curvatures are achieved by setting different joint limitations on the laser-cutting processed bendable portion. Note that the variable curvatures are optimized with the structure parameters optimization framework to improve the reachability and dexterity of the HVHRI in the maxillary sinus cavity. Based on the optimization results in Section IV-B, the flexible section has two segments with the same single joint length. The numbers of single joints in two segments are N_{F1}=5 and N_{F2}=7. And the joint limitations of the two segments are α_1=5° and α_2=30°. The flexible section is manufactured from a stainless-steel tube with a small outer diameter of 3.4 mm. And guide rings are manufactured by stamping on the laser-cutting processed tube. In addition, since the wristed forceps are attached at the distal end of the flexible section, the driving cables of the forceps are inevitably influenced by the bending motion of the flexible section. To avoid this coupling effect, two series of guide rings are positioned along the center axis of each side of the flexible section to pass through the two pairs of driving cables of the wristed forceps. This decoupling structure ensures that the wristed forceps are not significantly influenced by the bending motion of the flexible section, thus improving maneuverability during surgical

procedures. ***Third layer:*** The bending stiffness of the single flexible section is low, due to the small diameter of the flexible section. Therefore, a stiff spring (Φ 2 mm) is installed at the center of the flexible section, serving as a backbone to provide curvatures and enhance the bending stiffness for the flexible section. Besides, we use Nitinol wires as the driving cables of the flexible section to further improve its bending stiffness and make it easy to assemble.

3.2. Development of a compact actuation system (with weight < 800g) that can be easily handled by hand or integrated with robot arm.

In this project, we developed the actuation system of the HVHRI to be as compact and light as possible, as it is a handheld surgical device. To meet design criteria, as shown in Fig. 5(a), the BTM of the actuation system includes a pulley set decoupling mechanism that utilizes only four sets of capstans to control all 4-DOF motions of the hybrid-structure instrument of the HVHRI. This mechanism efficiently transfers the coupled translation motions of the six driving cables to the rotation motions of the four sets of capstans, which greatly improves the controllability of the HVHRI. Specifically, one set of capstans (C1) controls the bending of the flexible section, while the other three sets of capstans (C2, C3, and C4) control the pitch motion, the yaw motion of Jaw 1, and the yaw motion of Jaw 2 of the wristed forceps. With this pulley set decoupling mechanism, only four motors are used in the system (Fig. 5(b)). By contrast, without this pulley set decoupling mechanism, at least five motors would be required to control the wristed forceps and the flexible section, adding at least 20% unnecessary weight to the system. In addition, the quick-release mechanism on the BTM includes an adapter and two spring-return switches. During operation, the BTM is secured to the adapter by the switches. Pressing the switches allows easy detachment of the BTM from the actuation unit. The detached instrument unit only contains mechanical components, which can be sterilized individually through methods such as high temperature and liquid immersion.

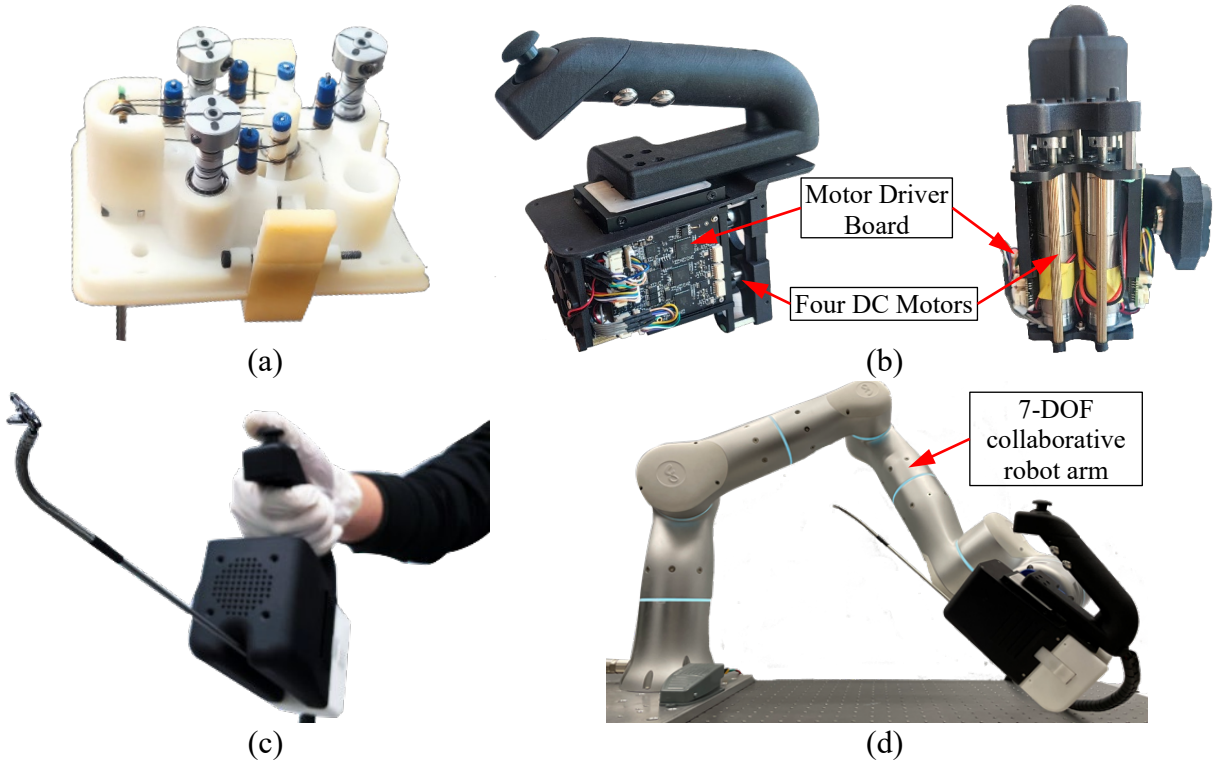


Fig. 5. (a) Prototype of the BTM. (b) Four-motor actuation unit and user interface of the HVHRI. (c) HVHRI works as a standalone handheld surgical instrument. (d) HVHRI works as a robot-arm-assisted surgical instrument.

The handheld user interface of the HVHRI is designed for easy to control during the MSS. As shown in Fig. 5(b), the interface includes a handle and a joystick with an integrated button for controlling the pitch and yaw motions of the wristed forceps, as well as the bending in the pitch direction of the flexible section. The integrated button serves as a switch to change the control mode. Two additional buttons are integrated into the interface to control the opening and closing of the forceps. In addition, a quick-release mechanism is designed between the

handle frame and the end-effector of the robot arm, enabling easy switching between robot-arm-assisted and standalone device as needed during the surgery (see Figs. 5(c)(d)).

3.3. Structure optimization for the 4-DOF HVHRI by maximizing the dexterity and reachability inside the maxillary sinus area.

We developed a structure optimization software for maximizing the dexterity and reachability of the HVHRI. The inputs of this software are position information sets of the environment cloud points ($\mathcal{P}_{envir}$), target space cloud points ($\mathcal{P}_{target}$), and waypoints ($\mathcal{P}_{way}$). In the development of the optimal flexible section curvature, we generated the anatomical structures of the nasal cavity and maxillary sinus cavity based on CT scan data obtained from a real phantom (Patient "Meyer", PHACON GmbH., Germany [15]) (see Figs. 6(a) and (b)). This process allows us to obtain the inputs including 1) a set of environment cloud points encompassing both the nasal cavity and the maxillary sinus cavity, 2) a set of target space cloud points representing the inner surface of the maxillary sinus cavity, and 3) a set of waypoints including the nasal entrance and the surgical incision opened between the nasal cavity and the maxillary sinus cavity.

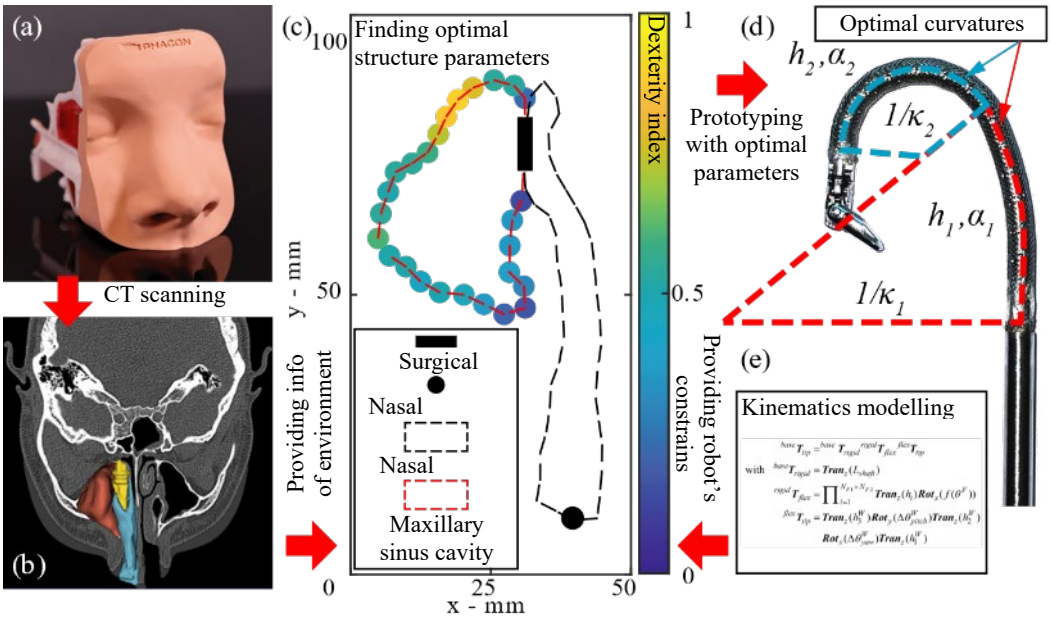


Fig. 6. Optimization procedure of the framework. (a) Sinus phantom constructed based on actual patient scans. (b) CT scan data based on the phantom. (c) Dexterity indices distribution based on the optimal flexible section curvatures. (d) Structure parameters with optimal flexible section curvatures. (e) Kinematics modeling of the flexible section serving as constrains.

As shown in Fig. 6(d), the output of this optimization framework is a parameter vector representing the optimal curvature of the flexible section ($\boldsymbol{s} = [N_{seg}, N_{F,1}, h_1, \alpha_1, \dots, N_{F,}, h_{N_{seg}}, \alpha_{N_{seg}}]^T$), which include the number of segments (N_{seg}), the number of joints in each segment ($N_{F,i},\ i = [1, N_{seg}]$), the joint length in each segment (h_i), and the joint limitation in each segment (α_i). Then, the joint in the i-th segment has a curvature of $\kappa_i \in [0, 2\tan(0.5\alpha_i)/h_i]$.

The optimization process of this framework is mainly maximizing the reachability and dexterity of the HVHRI (see Fig. 6(c)) under configuration limits and constrains of kinematics (see Fig. 6(e)). The reachability index (R) is defined as a ratio of the reachable surface area (A_{reach}) to the entire surface area (A_{all}) of the target space, which is $R = A_{reach}/A_{all}$. A commonly used rapidly exploring random tree algorithm is introduced to ensure the accurate deployment of the HVHRI without encountering collisions, thereby assessing its reachability effectively. The dexterity service spheres are used for every discrete point on the inner surface of the maxillary sinus cavity to show all orientations achieved by the tip of the HVHRI. The dexterity index ($D(\mathrm{P})$) of the point P is defined as a ratio of the area of the service region $\left(A^{\mathrm{P}}_{service}\right)$ to the total area of the void region without any obstacle $\left(A^{\mathrm{P}}_{void}\right)$ at the point P, which is $D(\mathrm{P}) = A^{\mathrm{P}}_{service}/A^{\mathrm{P}}_{void} \in (0, 1]$. Then, we get an objective

function by weighting and adding the reachability index and mean of the dexterity indices of the target space. Therefore, the flexible section curvature optimization framework is formulated as:

$$\max \quad \alpha_R R + \alpha_D \frac{1}{J}\sum_{j=1}^{J} D(\boldsymbol{p}_{target}^{j})$$

$$\text{s.t.} \begin{cases} \| \boldsymbol{p}_{target}^{j} - \boldsymbol{f}(\boldsymbol{T}_w, \boldsymbol{q}, \boldsymbol{s}) \| = 0, j \in [1, J] \\ \min(\| \boldsymbol{p}_{way}^{k} - \boldsymbol{g}(\boldsymbol{T}_w, \boldsymbol{q}, \boldsymbol{s}) \|) \le L_k, k \in [1, K] \\ q^m \in [q_{min}^m, q_{max}^m], m \in [1, M] \\ s^n \in [s_{min}^n, s_{max}^n], n \in [1, N] \\ inShape(alphaShape(\mathsf{P}_{envir}), \boldsymbol{g}(\boldsymbol{x}_0, \boldsymbol{q}, \boldsymbol{s})) = 1 \\ \boldsymbol{p}_w \in \mathsf{S}_{Limitation} \end{cases}$$

where α_R and α_D are two constants serving as weights of the reachability and dexterity indices, respectively. $\boldsymbol{p}_{target}^{j} \in \mathcal{P}_{target}$ is the position information of j-th point of the target space cloud points. $\boldsymbol{p}_{way}^{k} \in \mathcal{P}_{way}$ is the position information of k-th waypoint. $\boldsymbol{T}_w$ is the transformation matrix between the world coordinate system and the HVHRI's base coordinate system, which can be left multiplied by ${}^{base}\boldsymbol{T}_{tip}$, the transformation matrix between the Frame {base} to the Frame {tip} of the HVHRI in Fig. 6(e), to get the HVHRI's tip transformation matrix in the world coordinate system. $\boldsymbol{q} = [q^1, \ldots, q^M]^T$ is the kinematic configuration vector, which is $\boldsymbol{q} = \left[\theta^F, \theta_{pitch}^W, \theta_{yaw}^W\right]^T$ in the HVHRI. $\boldsymbol{f}(\boldsymbol{T}_w, \boldsymbol{q}, \boldsymbol{s})$ is the position vector of the HVHRI's tip point based on the flexible section curvatures and the kinematics in Fig. 6(e). $\boldsymbol{g}(\boldsymbol{T}_w, \boldsymbol{q}, \boldsymbol{s})$ is the position vector of HVHRI's body points based on the flexible section curvatures and the kinematics in Fig. 6(e). L_k is the distance tolerance between the k-th waypoint and the HVHRI's body. $alphaShape(\cdot)$ is a MATLAB function that constructs a bounding area or volume around a given set of 3D points. $inShape(\cdot)$ is a MATLAB function that determines whether a given point is inside an $alphaShape$. $\boldsymbol{p}_w$ is the position vector of the HVHRI's base in the world coordinate system constrained in space outside the human head $\mathsf{S}_{Limitation}$.

Based on the mechanical design of the HVHRI, we specified several constraints as:

$$\begin{cases} N_{seg} \le 2 \\ h_i \ge 4 \text{ mm}, \alpha_i \le 30^\circ, i = [1, N_{seg}] \\ L_1 = 5 \text{ mm}, L_2 = 5 \text{ mm} \\ \theta^F \in [-\sum_{i=1}^{N_{seg}} N_{Fi}\alpha_i, \sum_{i=1}^{N_{seg}} N_{Fi}\alpha_i] \\ \theta_{pitch}^W \in [-90^\circ, 90^\circ], \theta_{yaw}^W \in [-90^\circ, 90^\circ] \end{cases}$$

where h_i and α_i are constrained by the displacement of six guide rings in every single joint. L_1 is determined by the radius of the nasal entrance since the first waypoint locates at the center of the entrance. L_2 is determined by the radius of the surgical incision between the nasal cavity and the maxillary sinus cavity since the second waypoint locates at the center of the incision. Then, the entire framework was implemented in MATLAB, employing the Interior-Point algorithm to obtain a parameter vector representing the optimal flexible section curvatures for the HVHRI, which was $\boldsymbol{s}_{optimal} = [N_{seg} = 2, N_{F,1} = 5, h_1 = 4 \text{ mm}, \alpha_1 = 5°, N_{F,2} = 7, h_2 = 4 \text{ mm}, \alpha_2 = 30°]^T$ (see Fig. 6(d)). Utilizing these parameters, the dexterity index and mean dexterity indices of the proposed robot were calculated as 100% and 48%, respectively, based on sectional CT scan data of the maxillary sinus cavity (see Fig. 6(a)). For comparative purposes, we also computed the dexterity indices for a constant-curvature hybrid-structure forceps (CHF) with identical maximum bending angle (235°) which had a parameter vector, $\boldsymbol{s}_{CHF} = [N_{seg} = 1, N_{F,1} = 12, h_1 = 4 \text{ mm}, \alpha_1 = 19.58°]^T$. As shown in Fig. 7(b), the CHF exhibited a low reachability index of 35% and a low mean of the dexterity indices of 17%. In addition, the dexterity indices on a target point for both instruments are shown in Fig. 7. Notably, the HVHRI demonstrated a 122% higher dexterity index at the target point compared with the CHF.

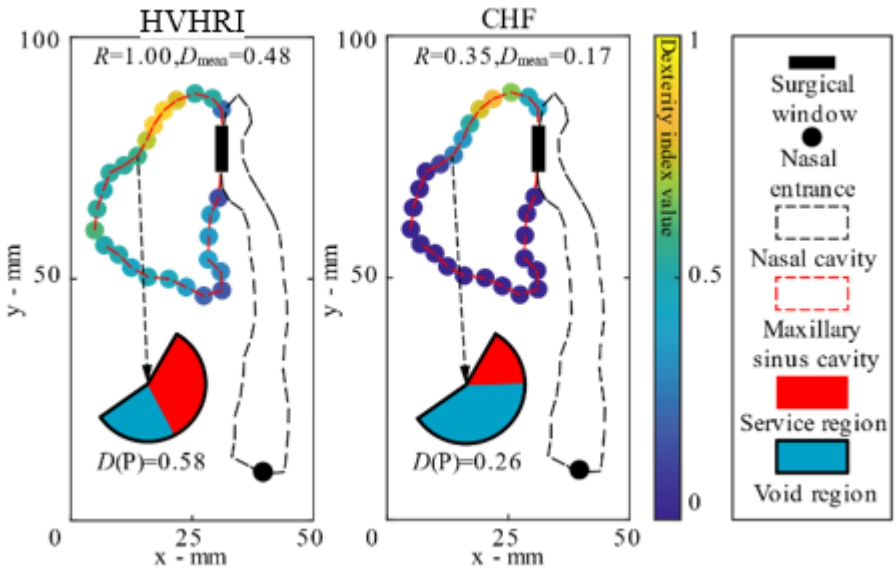


Fig. 7. Dexterity indices distributed in the maxillary sinus cavity of the HVHRI and the CHF

3.4. Experimental validation for the HVHRI.

3.4.1. TORSIONAL AND BENDING STIFFNESS TESTING

To assess the HVHRI's ability to maintain its shape during tissue manipulation, a bending stiffness test was conducted. As shown in Fig. 8(a), the actuation unit, along with the detachable instrument, was secured on a test bench. A 10 g weight hanger was suspended from the wristed forceps. The test involved gradually adding 20 g weights to the hanger until reaching a total weight of 70 g. The same stereo vision system used in the kinematics verification experiment was employed to measure the displacement of the base coordinate and the distal end of the wristed forceps. The test was carried out with the flexible section bent at 0°, 45°, and 90° to assess the bending stiffness under different configurations. The experimental results, as shown in Fig. 8(b), quantified the bending stiffness by representing the ratio of deflection to the total length under a 70 g (≈0.7 N) load. It was determined that the HVHRI exhibited a bending stiffness of 5.1%-11.1%. This finding indicates that the HVHRI fulfills the design requirements and is at least 197% stiffer than the instrument developed by Hong et al., which has a bending stiffness of 33% with a larger diameter of 4 mm.

The torsional stiffness of the HVHRI was tested across various bending configurations of the flexible section. The testing setup, as shown in Fig. 9(a), comprised a rotation stage and a torque sensor (Nano 17-E, ATI Industrial Automation, USA). Each test involved bending the flexible section to angles of 0°, 45°, and 90°, while the rotation stage was utilized to twist the HVHRI to specific angles ranging from 0° to 10°. The torque sensor was used to measure the twisting torques exerted on the robot. The results, as shown in Fig. 9(b), revealed that when a torque load of 12 N·mm was applied, the twisting angles of the HVHRI in the three bending configurations were 10°, 9°, and 8°, respectively. For easier comparison with other instruments, the torsional

stiffness is represented as a twisting angle per flexible section length under 6 N·mm load. Then, we can obtain that the torsional stiffness of the HVHRI is 0.1 deg/mm, which fulfills the design requirement and is 150% stiffer than the flexible tool developed by Zhang et al., which has a 0.25 deg/mm torsional stiffness.

3.4.2. KINEMATICS TESTING

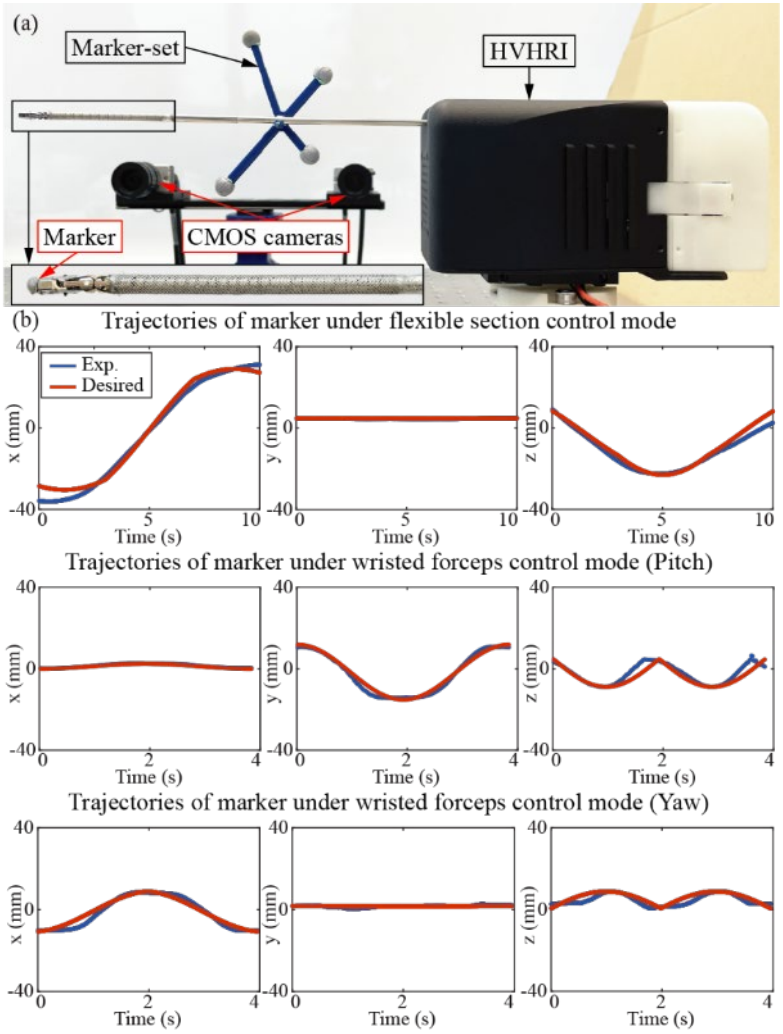


Fig. 10. Kinematics verification. (a) Experiment setup. (b) Measured and desired trajectories of the marker under two control modes.

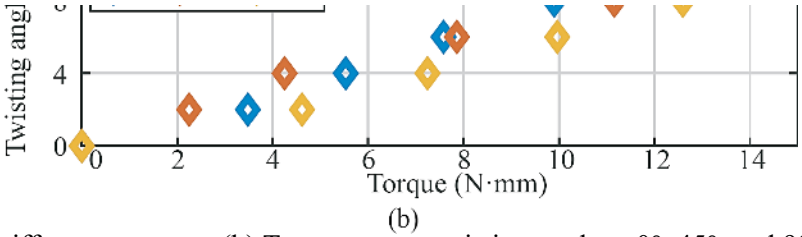


Fig. 9. (a) Torsional stiffness test setup. (b) Torque versus twisting angle at 0°, 45°, and 90° bending angles of the flexible section, which are marked in blue, red, and yellow, respectively.

In this experiment, we verified the kinematics of the HVHRI with the proposed two-mode control method. As shown in Fig. 10(a), the actuation unit with the detachable instrument of the HVHRI was fixed on the test bench. A stereo vision system including two CMOS cameras (DALSA Nano-m2420, Teledyne, Canada) (accuracy < 0.1 mm) was used to measure the 3D positions of various markers attached to the marker set, and the distal end of the wristed forceps. Four markers of the marker set formed a rigid body, which was fixed on the rigid shaft of the HVHRI to obtain the system's base coordinate.

As shown in Fig. 10(b), the position errors between the measured and desired trajectories of the marker were calculated for evaluating the kinematics model with the two-mode control method. Under the flexible section control mode, the 1-DOF flexible section was bent from -150° to 150°. The minimum error, the mean error, and the maximum error were 0.35 mm, 3.35 mm, and 7.26 mm, respectively. Then, under the wristed forceps control mode, the pitch and yaw joints of the wristed forceps were rotated from -90° to 90°, separately. The minimum

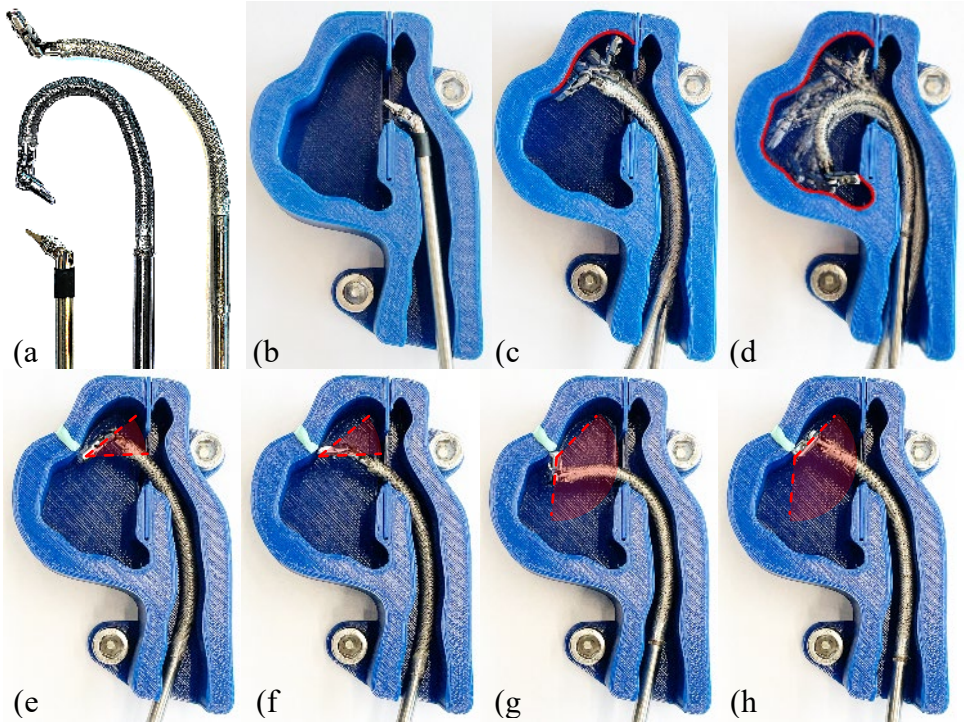


Fig. 11. (a) Prototypes of the WF (left), CHF (right), and HVHRI (middle). (b), (c), and (d) are the reachability test results of the WF, CHF, and HVHRI, respectively. CHF's reachable edge is marked yellow. HVHRI's reachable edge is marked red. (e) and (f) are the dexterity test results of the CHF, which has a yellow service region. (g) and (h) are the dexterity test results of the HVHRI, which has a red service region.

errors of the pitch and yaw motions were 0.25 mm and 0.11 mm. The mean errors of the pitch and yaw motions were 2.26 mm and 2.24 mm, respectively. And the maximum errors of the pitch and yaw motions were 7.07 mm and 5.34 mm, respectively.

3.4.3. REACHABILITY AND DEXTERITY COMPARISON

In this experiment, we compared the reachability and dexterity of the HVHRI with a pair of 3-DOF wristed forceps (WF) and a pair of 4-DOF CHF (see Fig. 11(a)) in a 3D-printed maxillary sinus and nasal cavity phantom, which was created based on a sectional CT scan data obtained in Section IV. Notably, all three instruments shared identical wristed forceps. Besides, the CHF featured a constant-curvature flexible section capable of achieving a maximum bending angle of 235°, which matched the capabilities of the HVHRI.

In the reachability test, an incision was made with a width of 10 mm between the nasal cavity and the maxillary sinus. All three instruments were operated to insert from the nasal entrance, pass through the incision, and try to touch the edge of the phantom's inner surface. In this 2D case, the reachability (R_{2D}) can be calculated as a ratio between the length of the reachable edge (L_{reach}) and the length of the entire edge (L_{all}), namely $R_{2D} = L_{reach}/L_{all}$. The test results are shown in Figs. 11(b)-(d). With a rigid shaft, the WF could not even reach the maxillary sinus, and therefore its reachability was 0. As the CHF was stuck in the nasal cavity to keep a constant curvature, its reachability was only 31%. The HVHRI had the largest reachability of 88%.

In the dexterity test, as the WF could not even reach the maxillary sinus cavity, only the CHF and the HVHRI were operated to insert from the nasal entrance, pass through the incision, and try to touch a point on the edge of the phantom's inner surface with different approaching angles. As shown in Figs. 11(e)-(f), the result showed that the CHF had a dexterity of 18% while the HVHRI had a dexterity of 69%.

3.4.4. USER STUDY FOR COMPARING THE WORKLOAD FOR DIFFERENT INTERFACES

As shown in Fig. 12, we designed three different handheld interfaces to conduct user study for comparing the workload. There were 7 male volunteers from the Multiscale Medical Robotics Centre, all right-handed, with normal or corrected to normal vision, and ages ranging from 24 to 31 years, participated the study. One of them had a medical background. The average collision number of style II (0.43 times) was fewer than the other two styles (style I: 1.71 times, style III: 1.57 times). The results of the NASA-TLX were shown in Fig. 13. The physical demand,
temporal demand, effort, and frustration of gripping style I were significantly lower than style II and III ($p < 0.05$). And the performance of style I was significantly better than style II and III.
The results indicated the gripping style II was the most suitable in these three styles.

3.4.5. FEASIBILITY TEST IN PHANTOM

To evaluate the feasibility of the HVHRI in the MSS, a realistic 3D maxillary sinus phantom (SN-as, PHACON GmbH., Germany) was employed to simulate polyps removal. As shown in Fig. 14 (c), two simulated polyps were placed in the anterior and posterior areas of the maxillary sinus cavity of the phantom. To visualize and guide the operation of the HVHRI inside the phantom, a hybrid-structure handheld robotic endoscope (HHRE) with robot-arm assistance was employed (see Fig. 14(a)). A surgical background researcher, who had taken a 15-minute training session on the operation of the HVHRI and the HHRE, performed the simulated polyps removal task. During the task, the researcher inserted the HVHRI and the HHRE together through the nasal cavity. Once the simulated polyps were shown in the endoscope's view, the position of the HHRE was fixed using the robot arm. Subsequently, the researcher focused on operating the HVHRI to grasp the simulated polyps. Ultimately, both the HVHRI and the HHRE were operated to extract the polyps from the phantom. The entire experiment was carried out three times. And it took an average of 130 s to locate, grasp, and retrieve the first simulated polyp and an average of 180 s for the second simulated polyp. As shown in Fig. 14(b), the results demonstrated that the HVHRI was effectively navigated through the 10 mm diameter surgical incision and successfully removed the two polyps from the maxillary sinus cavity.

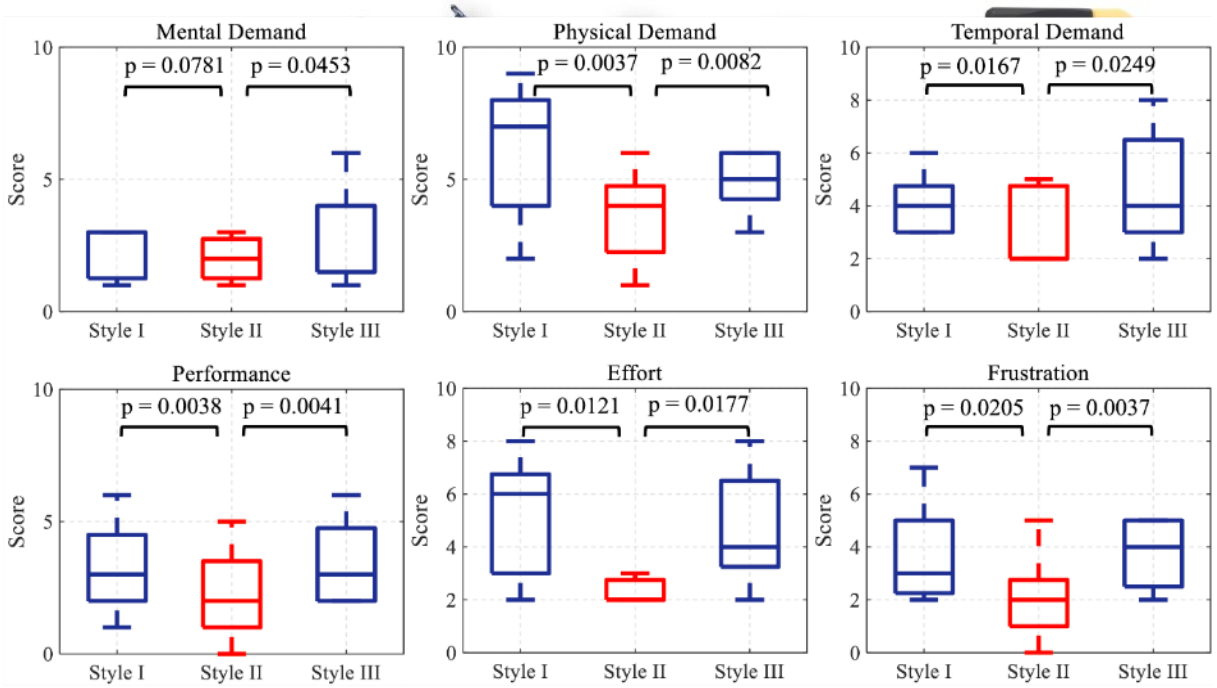


Fig. 13. The results of the NASA-TLX. The statistical analysis was conducted and the significant level was set at 0.05.

4. PUBLICATION AND AWARDS

J[1] Yi Yang, Puchen Zhu, Weibing Li, Richard M. Voyles, **Xin Ma*, "**A Fractional-Order Gradient Neural Solution to Time-Variant Quadratic Programming With Application to Robot Motion Planning", *IEEE Transactions on Industrial Electronics*, accepted, April, 2024.

J[2] Mei Liu, Kun Liu, Puchen Zhu, Guoqian Zhang, **Xin Ma***, and Mingsheng Shang*, "Data-Driven Remote Center of Cyclic Motion (RC2M) Control for Redundant Robots with Rod-Shaped End-Effector", *IEEE Transactions on Industrial Informatics*, accepted, February, 2024.

J[3] Xuchen Wang, **Xin Ma***, Puchen Zhu, Wee Shen Ng, Huayu Zhang, Xianfeng Xia, Russell H. Taylor, and Kwok Wai Samuel Au, "Design, Optimization, and Experimental Validation of a Handheld Nonconstant-Curvature Hybrid-Structure Robotic Instrument for Maxillary Sinus Surgery", *IEEE-ASME Transactions on Mechatronics,* accepted, 2024.

C[1] Yi Yang, Weibing Li*, Jianshu Zhou, Junda Huang, Jinfei Hu, Richard M. Voyles, **Xin Ma***, "PTC-FOZNN: A Strictly Predefined-Time Convergent Fractional-Order Recurrent Neural Network for Solving Time-Variant Quadratic Programming", *IEEE ICCA*, accepted, April, 2024.

C[2] Huayu Zhang, Jiajun An, Tianle Pan, Upinder Kaur, Zhijian Wang, Qiguang He*, **Xin Ma***, "Miniature Reconfigurable M odu lar Soft Robot s Using Liquid Crystal Elastomer Actuation", 6th International Conference on Reconfigurable Mechanisms and Robots, accepted, 2024.

C[3]Yi Yang, Puchen Zhu, Weibing Li, Richard M. Voyles, **Xin Ma***, "A Fractional-Order Recurrent Neural Network Model for Time-Variant Quadratic Programming in Robot Motion Planning", IEEE/ASME International Conference on Advanced Intelligent Mechatronics, accepted, 2024.

A[1] Internet + National Competition Gold Medal," 多層疊剛柔混合結構機械人系統", Instructor: **Xin MA** and Kwok Wai Samuel Au, December, 2023.

A[2] Third Prize, Professor Charles Kao Student Creativity Award, "應用於耳鼻喉手術中狹窄空間的掌上型混合結構的手術機器人系統", Instructor: **Xin MA** and Kwok Wai Samuel Au, July, 2023.

P[1] **Ma Xin**, Zhang Zihao, Wang Xuchen, Kwok Wai Samuel Au, "Cam-based Backend Transmission Mechanism for Driving Four-cable Wristed Instrument", 2023-8-7, US Patent Pending, US 63/531,314.

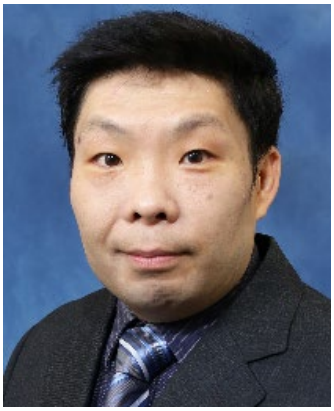
WORK IN PROGRESS

SMART BANDAGE WITH INTEGRATED ORGANIC ELECTRONIC SENSOR AND IONTRONIC DRUG DELIVERY PLATFORM FOR ADVANCED CHRONIC WOUND CARE

Principal Investigator: Professor MAK Wing Cheung
Department of Biomedical Engineering, CUHK

Research Team Members:
Junning QIAN, PostDoc[(1)]

[(1)] Department of Biomedical Engineering, CUHK



Reporting Period: 1 July 2023 – 31 May 2024

INNOVATION AND PRACTICAL SIGNIFICANCE:

"Theranostics" are emerging biomedical strategies that integrate therapeutic and diagnostic/biosensing components for the monitoring and management of chronic clinical conditions. From the innovative technology aspect, our project aims to develop a novel wearable theragnostic smart bandage platform with integrated sensors and drug delivery patches to monitor and actively react to modulate wound healing progress revolutionizing existing chronic wound addressing technologies with improved patient healthcare outcomes. From the material science engineering aspect, our project explores the unique properties of conductive polymers with mixed electronic and ionic conductivities as organic electronic platforms to perform both the electrochemical biosensing and ionic drug release feature, with additional advantages of being flexible, organic in nature, processable for large-scale fabrication and biocompatible, which has not been demonstrated in theragnostic smart bandage platform. The success of this project would open new prospects for the next generation of smart bandages with sensing and therapeutic features for the betterment of wound management for pressure ulcers (disabled person/neuromuscular diseases/stroke patients with long-term beds), diabetic foot ulcers (diabetes and elderly) ischemic ulcers (elderly) and venous ulcers (elderly). In fact, Asia is the hotspot for diabetes and facing huge challenge in ageing population. Therefore, we expect the new theranostics smart bandage developed in this project could offer new healthcare tools for the betterment of mankind in Hong Kong, Greater Bay Area, China and beyond. Moreover, as the skin is the largest organ in the body, the developed smart bandage platform may potentially as new wearable biomedical tools for monitoring and treatment of other skin diseases.

ABSTRACT

Chronic wound management is a major healthcare challenge that affects millions of patients (1.5-2.0% of the world population) suffering from slow healing processes, susceptibility to infection and risk of amputation, together with financial burdens to central healthcare system. Incidents are expected to rise with ageing population and increase in diabetic cases (i.e. diabetic foot ulcers). Emerging smart bandages with integrated sensors have demonstrated to support monitoring of wound healing progress with improved healthcare outcomes, yet there is limited development in integrating both sensor and on-demand drug delivery for wound modulation. Our project aims to develop a novel closed-loop smart bandage platform with integrated electronic sensor to track the wound conditions and iontronic on-demand drug delivery patch for modulation of infection incidents at early stage. We plan to develop wearable organic electronics with conducting polymers feature with unique mixed electronic-and-ionic conductivity properties for fast-response and reversible pH monitoring (indicator for infection), and iontronic-controlled release of antibiotics via doping/de-doping mechanism to modulate infections. The success of this project would open new prospects for next generation of smart bandages with sensing and therapeutic features for better wound management,

reducing the frequency of dressing replacement and visits to medical facilities with improved patient compliance.

1. OBJECTIVES AND SIGNIFICANCE

Objectives:
1. To develop a conducting polymer-based wearable platform for selective pH sensing and optimize the pH response with a wide pH window covering pH 4-10. (Completed on schedule)
2. To achieve high performance in fast pH response, reversible pH monitoring and long-term stability based on the smart bandage. (Completed on schedule)
3. To develop and optimize conducting polymer iontronic delivery patches and demonstrate the electrochemical triggered release. (On Track, Continue in 2nd year)
4. Integrate both pH monitoring and drug delivery (On Track, Continue in 2nd year)

Significance:
The outcome of the project will immediately impact the field of smart bandages with innovative sensing and therapeutic features for the advancement of chronic wound management which remains a global challenge affecting millions of patients. Moreover, chronic wounds are often associated with the ageing population and diabetes communities with Asia as the hotspot (including Hong Kong), and the developed smart bandages could potentially benefit and support the long-term healthcare system. Furthermore, as the skin is the largest body organ, the developed smart bandage platform may further innovate as new wearable biomedical tools for monitoring and treatment of other skin diseases.

For the first step, measuring the pH of wound exudate is the most important indicator for assessing bacterial infection and monitoring the wound healing progress. In the 1st phase of the project, we have successfully developed a PANI-based wearable pH sensor on a wearable and breathable fabric substrate tailored for the smart bandage application. The performance of the pH sensor delivered good sensitivity, wide pH sensing range (4- 10), good reversibility for pH cycling, fast response time (~1sec) and good stability over 12 hours of pH measurement. The details of the achievements are presented in Section 3.

2. RESEARCH METHODOLOGY

2.1 Materials and instruments

Screen-printed electrodes (SPEs, DropSens 110) were purchased directly from Metrohm. Polyimide film (PI Film: thickness of 75 μm) was purchased from ubisoft tape Co., Ltd. Polyimide fabric (PI Fabric: density of 60 g·cm-2, W-60) was purchased from Jiangsu xiannuo new materials technology Co., Ltd. Aniline and hydrochloric acid (HCl) were purchased from Macklin and Sigma-Aldrich, respectively. Solid-state silver/silver chloride (Ag/AgCl) was purchased from Dupont. All chemicals were of analytical grade and used without any further purification. Computer-controlled CO2 laser platform (HL 40-5g, 10.6 μm, Full Spectrum Laser LLC) was used to produce laser-induced graphene (LIG), which was controlled by the software Lightburn. Electrochemical workstation (PalmSens4, Europe) was used for electrochemical tests. The microstructure of electrodes was observed on a optical microscope (Axioscope, Zeiss, Germany). A commercial pH meter (FiveEasy Plus pH meter FP20-Std-Kit, Mettler Toledo, Switzerland) was used as a standard for pH measurement.

2.2 Fabrication of LIG:

LIG was prepared in different patterns using a commercial 10.6 μm CO2 laser cutter system (Gweike Cloud RF, China) on both PI Film and PI Fabric under ambient conditions. The resulting LIG on two substrates were noted as PI-Film-LIG and PI-Fabric-LIG, respectively. The laser parameters were optimized to be 8.5% of the full laser power (full power is 60 W), with a scan speed of 170 mm·s-1 and a defocus distance of 15 mm.

2.3 Electrodeposition of polyaniline (PANI) on LIG:

Firstly, a three-electrode system was used for PANI preparation. Pt wire and Ag/AgCl electrode were used as counter electrode and reference electrode, accordingly. SPEs, PI-Film-LIG, and PI-Fabric-LIG were used as

the working electrodes, respectively. The electrodeposition of PANI was carried out at room temperature. This three-electrode system was immersed in the 1 M HCl solution containing 0.1 M aniline. The electrochemical polymerization of aniline was then carried out using the cyclic voltammetry (CV) method in the potential range from -0.2 to 1 V vs. Ag/AgCl with a scanning rate of 50 mV·s-1. The number of CV cycles was five. After the electrochemical polymerization was completed, the prepared working electrodes were washed with deionized water and alcohol, and then dried under 60 °C for further use.

2.4 Fabrication and integration of a wearable pH sensor on bandage:
The two-electrode system was employed as the pH sensor based on the PI-Fabric-LIG. The working electrode was the as-prepared PANI decorated PI-Fabric-LIG, and the reference electrode was a solid-state Ag/AgCl electrode. This reference electrode was stencil printed with Ag/AgCl ink on the surface of PI-Fabric-LIG and dried under 60 °C for 1 hour. Then 5 µL of PVB polymer containing KCl fine powder (30% w/v) was dropped cast on top of the Ag/AgCl using a mask, followed by importing the reference electrode in 3 M KCl for 24 hours. After that, the surface of the reference electrode was coated with 5 µL of PVB polymer to further improve the stability of the fabricated pH sensor. Finally, the wearable pH sensor was assembled on the bandage by fixing the pH sensor layer between the lower dressing and the upper adhesive layers.
2.5 Electrochemical tests:
The response of the pH sensors was recorded and characterized by measuring the electromotive force (EMF) between the working electrode and the Ag/AgCl reference electrode in 0.1 M PBS with varied pH of 4-10.

3. RESULTS ACHIEVED SO FAR

We have successfully completed the content on building a conducting polymer-based wearable platform in objective 1. As shown in Figure 1, the conventional substrates such as screen-printed electrodes (SPE) are thick, rigid, non-porous and non-breathable, while the flexible PI films are flexible, however, the compacted PI film is a non-porous and non-breathable substrate which is not an ideal substrate material for bandage application. In contrast, we developed a PANI-based pH sensor on PI fabric that is flexible, thin, comfortable, porous and breathable suitable for wearable bandage devices. Additionally, the electrodeposition technique employed for synthesizing the conducting polymer-PANI pH sensing layer meticulously retains the porous morphology and integrity of the PI fabric substrates and their inherent architecture. As noted, the morphological characteristics of various substrates remain predominantly unaltered after the electrodeposition of PANI.

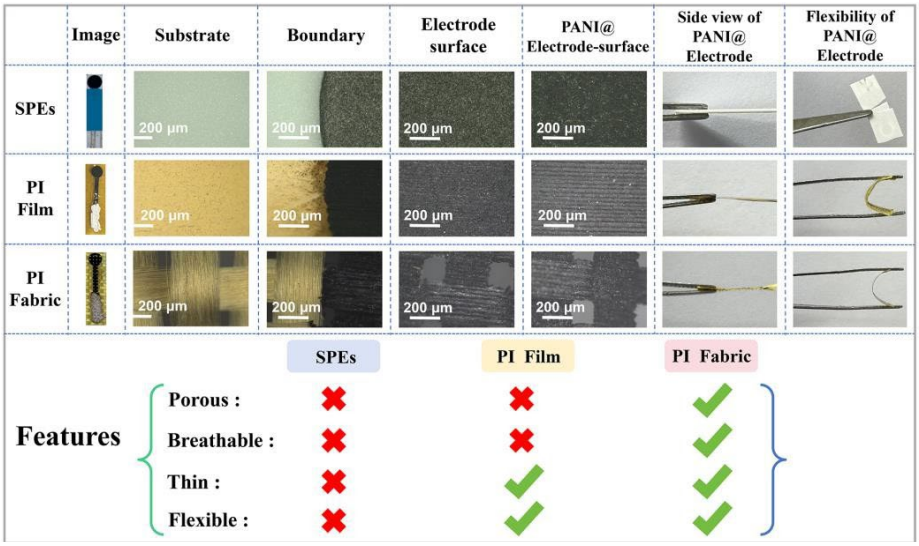


Figure 1. Microstructure, optical images and structural features of the SPEs, PI-film, PI-fabric based electrodes.

Next, we completed the optimization of the pH response with a wide pH window covering pH 4-10 in objective1.1.1. By analyzing the electrochemical characteristics and pH sensing performance of

different electrode platforms (Figure 2a-b), the platform of PI fabric decorated with PANI exhibits the best sensitivity of 70.592 mV/pH with a linear pH range covering pH 4-10 which is outperformed compared with other conventional platforms (Figure 2c-d).

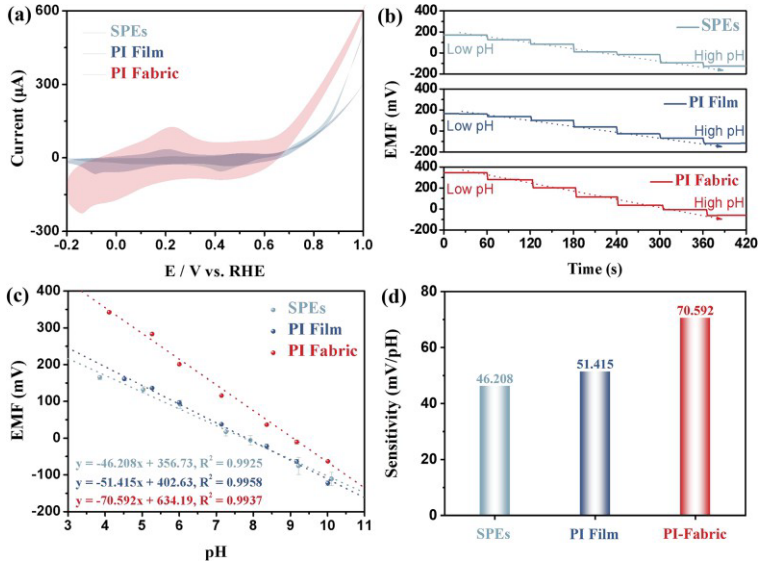


Figure 2. (a) CV curves of PANI decorated SPEs, PI Film and PI Fabric; (b) Potentiometric curves of EMF response stepwise increasing with pH 4-10; Corresponding calibration plots (c) and sensitivity values (d) of EMF versus pH.

We have achieved to develop a high-performance pH sensor of PI fabric with fast pH response, reversible pH monitoring and long-term stability for smart bandage in objective 1.1.2. The PANI-fabric pH sensor is porous, breathable, thin (0.06 mm), flexible and lightweight (6.38 mg·cm^{-2}) (Figure 3a-f). The integrated smart bandage comprises three components: dressing material, PANI-fabric pH sensor, and adhesive unit (Figure 3g), showcasing comfortable wearable features on the skin (Figure 3h). The PANI-fabric pH sensor showed a fast response to pH change within one second (Figure 3i). The pH sensitivity of the PANI-fabric pH sensor reaches 65.279 mV/pH with a linear pH range covering pH 4-10 (Figure 3i-j). The high sensitivity and wide pH sensing window of the PANI-fabric pH sensor surpasses most literature-reported pH sensors. [1-3] After completing the cycle of pH sensing from 4-10-4, it can retain 98.6% of its initial EMF value (Figure 3k). During a 12-hour stability test, the PANI-fabric pH sensor maintains the original signal of 95.1%, 93.7% and 96.9% compared with the starting points in pH values of 4, 7 and 10, accordingly (Figure 3l). These results demonstrated the excellent stability of the PANI-fabric pH sensor compared to most literature-reported pH sensors only showing a relatively short stability duration, typically within 1 hour. [4-6] Furthermore, after a pH value mutation from 10 to 4, the PANI-fabric pH sensor maintains 99.3% of the initial EMF value (Figure 3m) showing a high repeatability in pH sensing.

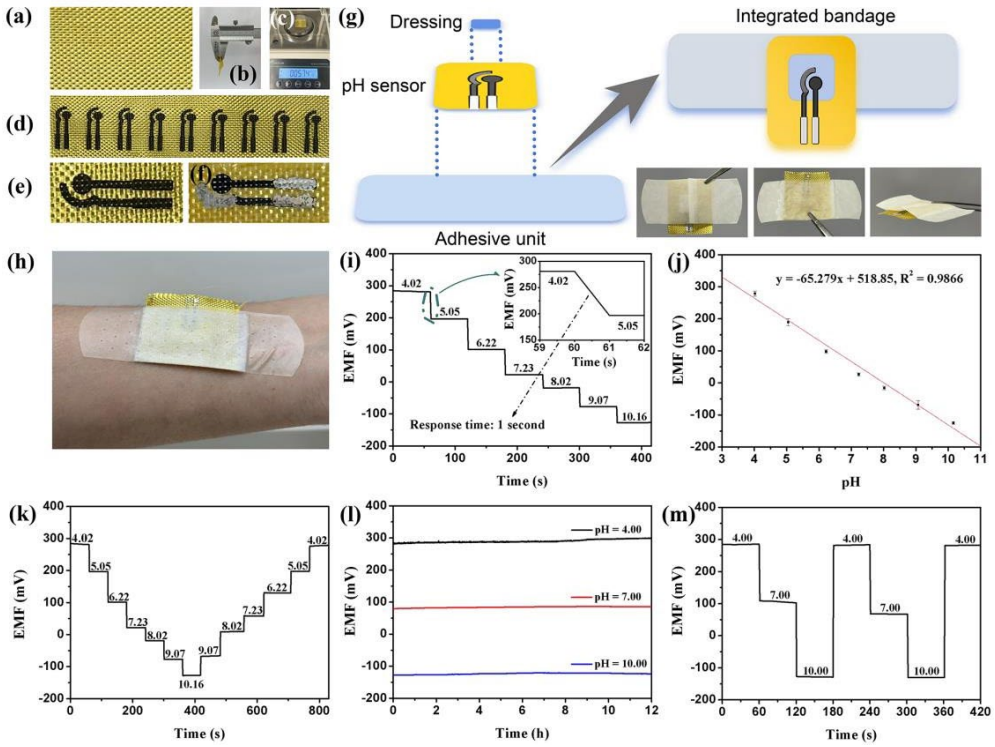


Figure 3. (a) Image, (b) thickness and (c) weight of PI Fabric; (d) Scalable synthesis of PI-Fabric-LIG; (e) enlarged area in Figure 3d; (f) Image of integration of PI-Fabric-LIG with solid-state Ag/AgCl, PANI and silver ink; (g) Schematic of craft for smart bandage and physical images from different perspectives; (h) wearability of smart bandage; (i) Potentiometric curve of EMF response stepwise increasing with pH 4.02-10.16, insert: enlarged area from pH values of 4.02-5.05; (j) Corresponding calibration plot of EMF versus pH; (k) Reversibility of smart bandage to stepwise increase pH value from 4 to 10 followed by stepwise decreasing from 10 to 4; (l) Long-term stability of smart bandage at pH values of 4, 7 and 10; (m) Repeatability of smart bandage at pH values of 4, 7 and 10.

Summary:

We have completed the development of the pH sensor on a breathable fabric substrate for smart bandage with good sensitivity, wide pH sensing range (4-10), good reversibility for pH cycling, fast response time (~1sec) and good stability over 12 hours pH measurement (completed on schedule). We will continue in the 2nd phase of the project to develop the conducting polymer iontronic delivery patches and finally integrate the smart bandage with the pH sensor and drug delivery patch (to be continued).

References

[1] X.Q. Cui, et. al, A wearable electrochemical sensor based on β-CD functionalized graphene for pH and potassium ion analysis in sweat, Talanta, 2022, 245, 123481.

[2] J.H.Yoon, et. al., Highly self-healable and flexible cable-type pH sensors for real-time monitoring of human fluids, Biosensors and Bioelectronics, 2020, 150, 111946.

[3] C.H. Zhu, et. al., A dual-functional polyaniline film-based flexible electrochemical sensor for the detection of pH and lactate in sweat of the human body, Talanta, 2022, 242, 123289.

[4] K. Chawang, et. al., Porous polypropylene membrane based pH sensing for skin monitoring," IEEE Access, 2022, 10, 111675-111687.

[5] K. Singh, et. al., Impact of yttrium concentration on structural characteristics and pH sensing properties of sol-gel derived Y2O3 based electrolyte-insulator-semiconductor sensor," Materials Science in Semiconductor Processing, 2020, 105, 104.

[6] L.R. Wang, et. al., Skin-like hydrogel-elastomer based electrochemical device for comfortable wearable biofluid monitoring,
Chemical Engineering Journal, 2023, 455, 140609.

COMPLETED

COUPLINGMOS2 FIELD-EFFECT BIOSENSORS WITH HYBRIDIZATION CHAIN REACTION SELF-ASSEMBLY AMPLIFICATION FOR HIGHLY SENSITIVE AND LABEL-FREE NUCLEIC ACID DETECTION

Principal Investigator: Professor Zhaoli GAO
Department of Biomedical Engineering, CUHK

Research Team Members:
Dong Wook JANG [(1)]
Jiawen YOU [(1)]

[(1)] Dept. of Biomedical Engineering, the Chinese University of Hong Kong

Project Start Date: 1st July 2022
Completion Date: 31st August 2024

INNOVATION AND PRACTICAL SIGNIFICANCE:

The objective of this project is to develop a highly sensitive and label-free platform for nucleic acid detection, based on MoS2 transistor arrays coupled with an HCR-based signal amplification scheme. The proposed project is novel and highly notable for multiple reasons: 1) The application of large-scale MoS2 transistor arrays in biosensing, which has not yet been reported, could result in a substantial enhancement in the limit of detection, compared to conventional 2D devices without a band gap, e.g., graphene-based devices; 2) Scalable fabrication of biosensor arrays based on MoS2 has yet to be achieved; and 3) Involving signal amplification in MoS2-based biosensors is a novel methodology for nucleic acid detection. Project success would bring accuracy to nucleic acid-based diagnosis, which is critically needed for the management of various diseases, such as early-stage cancers, and viral infectious diseases. We expect the all-electronic biosensors developed through this project would offer a diagnostic platform that would be widely applicable in the healthcare system of Hong Kong and that of the world over.

ABSTRACT

Nucleic acid testing plays a crucial role in clinical diagnoses. Existing approaches, such as polymerase chain reaction (PCR), are sensitive to inhibitors due to their enzyme application and fluorescent labeling. The past decade has witnessed the development of 2D bioelectronics for highly sensitive testing of biomolecules, such as nucleic acid associated with various diseases. 2D nanomaterials hold tremendous promise for low-level and label-free detection of nucleic acids, offering the prospects of simple and multiplexed testing, with high accuracy and specificity. However, the zero-bandgap nature of traditional 2D nanomaterials, such as graphene, and the binding-affinity-dependent limit of nucleic acid detection inhibit further advancement in the label-free detection of ultralow-level nucleic acid. The objective of our proposal is to develop a highly sensitive and label-free biosensing system for nucleic acid detection, through the combined use of gapped monolayer MoS_2 and hybridization chain reaction-based signal amplification scheme. The project success would open a new prospect for next generation ultrasensitive MoS_2-field-effect-transistor (MoS_2-FET) biosensors that could potentially exceed the detection limit of graphene-FET.

1. OBJECTIVES AND SIGNIFICANCE

1) To synthesize monolayer MoS_2 with controlled uniformity on a large scale, through chemical vapor deposition
2) To develop scalable fabrication of MoS_2-FET sensor arrays functionalized with hairpin probe DNA
3) To develop a target recycling and hybridization chain reaction (TRHCR) scheme that can work on MoS_2-FET devices, for sensitive detection of nucleic acid at aM levels
4) To test and optimize the sensing protocols with MoS_2-FET biosensors for maximum sensitivity, reproducibility, and reliability in physiological conditions

The proposed project is novel and highly notable for multiple reasons: 1) The application of large-scale MoS_2 transistor arrays in biosensing, which has not yet been reported, could result in a substantial enhancement in the limit of detection, compared to conventional 2D devices without a band gap, e.g., graphene-based devices; 2) Scalable fabrication of biosensor arrays based on MoS_2 has yet to be achieved; and 3) Involving signal amplification in MoS_2-based biosensors is a novel methodology for nucleic acid detection. Project success would bring accuracy to nucleic acid-based diagnosis, which is critically needed for the management of various diseases, such as early-stage cancers, and viral infectious diseases.

2. RESEARCH METHODOLOGY

Task 1: Controlled large-scale growth of MoS_2

We intend to develop a feasible CVD method to grow uniform MoS_2 thin films with precise control over their layer number and crystal orientations. A Si/SiO_2 growth substrate will be coated with a layer of NaCl (1% aqueous solution), followed by spin coating of saturated ammonia heptamolybdate (AHM) solution as the Mo source and being heated to ~ 800 °C. A sulfur source will then be placed upstream of the furnace, and a flow of 500 sccm Ar will carry the sulfur vapor into the chamber for a 30-60 min growth period. After growth, the sample will be rapidly cooled to room temperature.

Our hypothesis is that CVD growth could lead to the controlled synthesis of a uniform MoS_2 monolayer across the entire material. The feasibility of our proposal was verified by trial growths of triangular MoS_2 crystals on SiO_2 substrates, and we plan to achieve more precise control, with further exploration of the governing mechanism of CVD conditions on MoS_2 growth.

Task 2: Scalable fabrication and functionalization of MoS_2 sensor arrays

We will carry out photolithography to fabricate the MoS_2 biosensor arrays in a scalable way. A sub-monolayer of gold will be deposited onto the MoS_2 by physical vapor evaporation to a nominal thickness of 2 Å. The MoS_2-Au heterostructure will be transferred onto a SiO_2 substrate with prefabricated Cr/Au electrodes. Fabrication of the MoS_2 transistor arrays will be implemented with the use of photolithography to pattern the transferred MoS_2, followed by plasma etching. The sensor array will be annealed at 225 °C to allow the formation of monodispersed AuNPs. An Al_2O_3 passivation layer will be deposited *via* atomic layer deposition to prevent leakage current from the source to drain. Finally, the arrays of MoS_2-AuNP sensors will be functionalized with thiolated hairpin probe DNA through Au−S bonds.

Task 3: MoS_2-FET performance testing and characterization

We will design, develop, and validate a TRHCR scheme, to quantify DNA using the MoS_2 sensing platform with accuracy and precision. We will verify the improved LOD and enhanced specificity of TRHCR with commercially available DNA (ThermoFisher Scientific), using four different concentrations (in the clinically relevant range of 10 fM, 1fM, 100 aM, and 10 aM), each in 1× SSC buffer solutions. We will also examine whether the biosensor responses reflect specific binding of the target DNA, by testing the MoS_2 biosensor against various negative controls, i.e., target DNA with a single-base mismatch at the 3′ or 5′ end.

3. RESULTS ACHIEVED

Result 1: Controlled large-scale growth of continuous monolayer MoS_2 thin film

We designed a custom atmospheric pressure chemical vapor deposition (APCVD) system to achieve precise control over the uniformity of MoS_2 thin films. As shown in **Fig. 1a**, a sapphire growth substrate coated with a metal halide promoter, specifically sodium chloride (NaCl), was positioned at the center of the CVD furnace. Molybdenum oxide (MoO_3) and elemental sulfur pellets served as growth precursors for MoS_2. These precursors were separately loaded onto a quartz boat and placed upstream of the gas flow, away from the central heating zone.

To prevent inhomogeneous MoS_2 growth, often caused by the reduction of the molybdenum source due to high concentrations of gaseous sulfur precursor [1-3], we isolated the source by inserting a small open-ended quartz tube inside the growth chamber's quartz tube. Once the precursors were positioned, the growth process commenced with an Ar gas purging step at room temperature, followed by ramping at 50°C per minute until reaching the growth temperature of 850°C. After a growth period of 10 minutes, the furnace was turned off to allow rapid cooling.

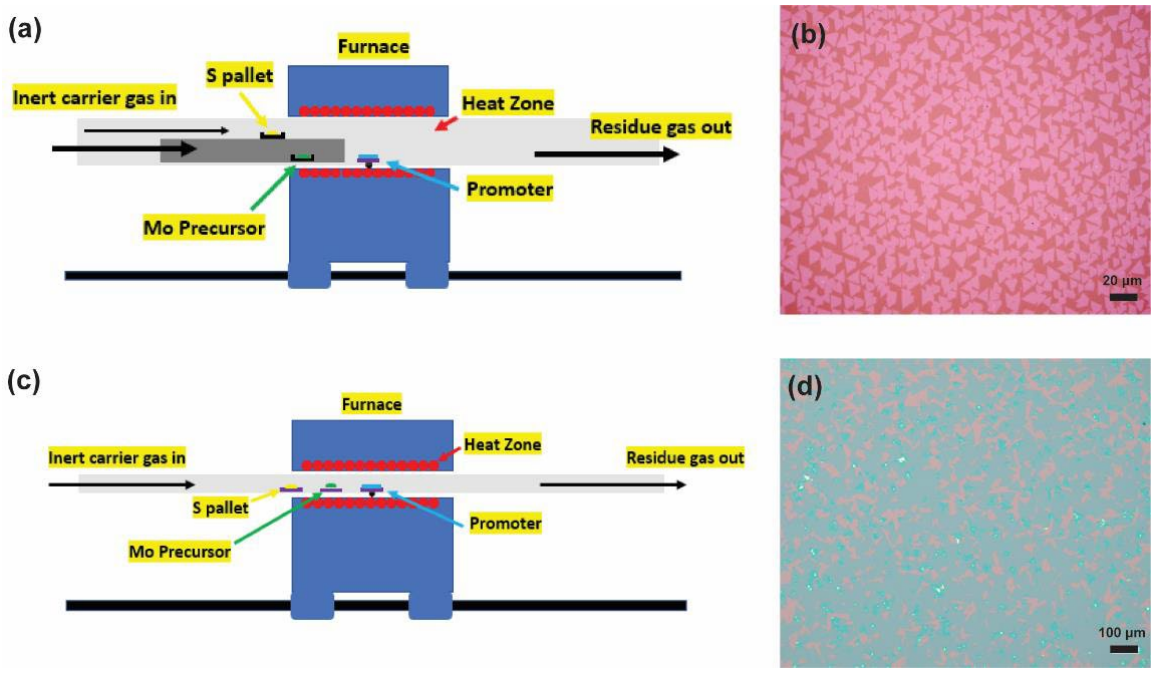


Figure 1. Illustration of the CVD setup for synthesizing continuous MoS_2, showing configurations (a) with source isolation, where the molybdenum oxide precursor is placed inside a small quartz tube, and (c) without source isolation. Growth results are shown in (b) with source isolation and (d) without.

Using the proposed CVD setup, we successfully synthesized a highly uniform and continuous monolayer MoS_2 thin film. **Fig. 2a** shows monolayer MoS_2 grown on a sapphire substrate after a 5-minute growth period. Extending the growth time to 8 minutes allowed triangular crystal domains to merge into a continuous thin film (**Fig. 2b**), with a highly uniform thickness observable under an optical microscope.

The film was further characterized by Raman spectroscopy, photoluminescence (PL) spectroscopy, selected area electron diffraction (SAED), and atomic force microscopy (AFM). Raman analysis, shown in **Fig. 2c**, revealed E_2g and A_1g peaks at 384 cm^{-1} and 403 cm^{-1}, respectively, with a peak difference of 19 cm^{-1}, confirming the characteristic monolayer MoS_2 peaks. The absence of 1T-phase peaks at J_1, J_2,

and J_3 acoustic modes, along with a high full-width at half maximum (FWHM) of the E_{2g} peak, indicated the high crystallinity of the semiconducting 2H-phase monolayer MoS_2 grown via CVD. Supporting these results, an intense PL peak at 661 nm (**Fig. 2d**), the trigonal prismatic lattice structure observed in aberration-corrected TEM (**Fig. 2e**), and an AFM height measurement of 0.73 nm (**Fig. 2f**) all confirmed the monolayer and 2H phase nature of the CVD-grown MoS_2 film.

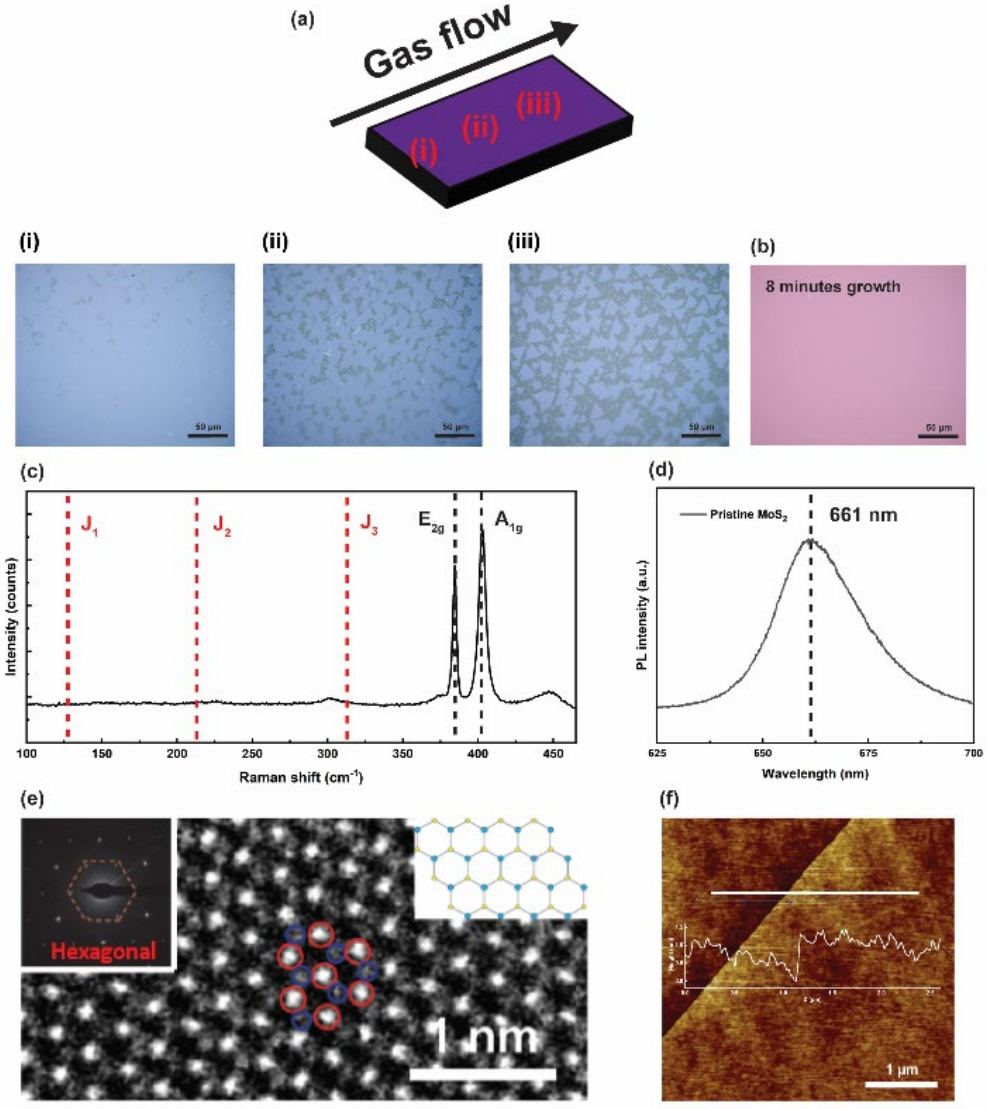


Figure 2. CVD growth and characterization of monolayer MoS_2. (a) Optical image of MoS_2 grown on a sapphire wafer at different positions after a 5-minute growth period. (b) Optical image of continuous monolayer MoS_2 formed by extending the growth time to 8 minutes. (c) Raman spectroscopy results showing characteristic E_{2g} and A_{1g} peaks confirming monolayer MoS_2. (d) PL spectroscopy further indicating monolayer quality. (e) High-resolution TEM image showing the trigonal prismatic structure of the 2H-phase MoS_2. (f) AFM image of the monolayer MoS_2 synthesized via the proposed CVD method.

Result 2: Scalable fabrication and functionalization of MoS_2-FET sensor array

We performed a standard photolithography process, as shown in **Fig. 3**, to fabricate the MoS_2-FET sensor array in a scalable manner. After defining the MoS_2 channel using O_2 plasma etching, a 3 Å layer of elemental gold was deposited onto the MoS_2 via e-beam evaporation under an ultra-low vacuum of 3 ×

10^{-6} Torr. This was followed by annealing in an argon-rich inert atmosphere at 250 °C to promote the formation of monodispersed gold nanoparticles (AuNPs). Finally, the annealed device was incubated with a 100 nM solution of thiolated hairpin probe A for 20 hours to immobilize the hairpin DNA probe onto the MoS_2 surface through Au-S bonds.

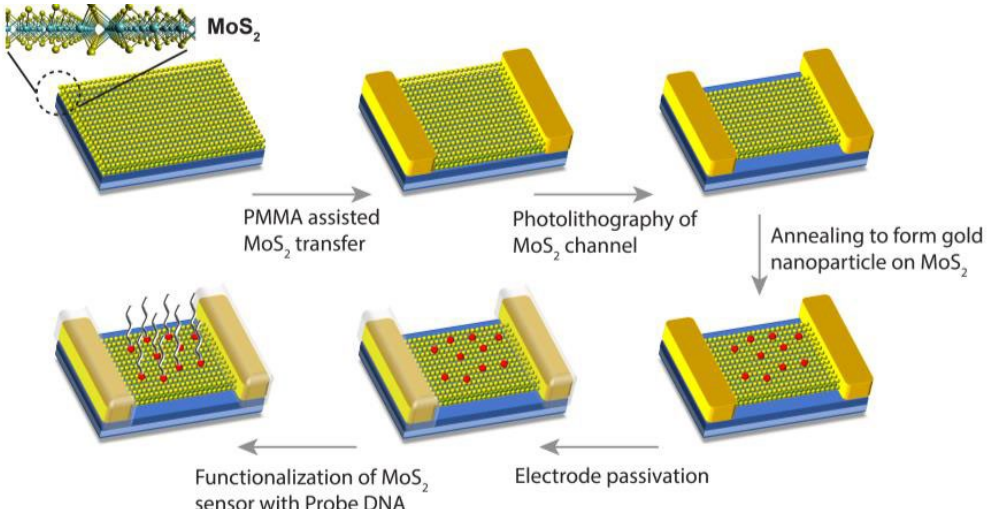


Figure 3. Schematic representation of the fabrication process for an AuNP-MoS_2 FET biosensor array immobilized with hairpin DNA probes.

The results are presented in **Fig. 4**, where we fabricated FET devices using CVD-grown MoS_2 and successfully functionalized them with monodispersed AuNPs. As shown in **Fig. 4a**, gold nanoparticles formed after depositing a 10 Å layer of Au followed by thermal annealing at various temperatures. Our findings indicate that lower annealing temperatures promote nanoparticle nucleation. Specifically, when comparing the size distributions of nanoparticles formed at annealing temperatures of 200 °C and 300 °C, we observed that annealing at 300 °C for 2 hours yielded the narrowest size distribution (**Fig. 4d**).

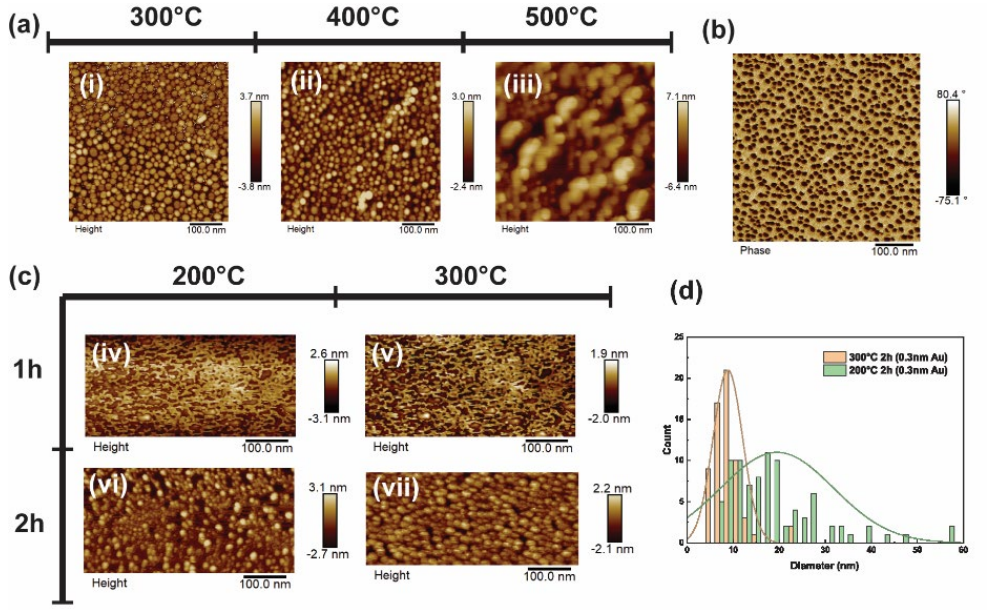


Figure 4. Schematic of AuNP-MoS_2 heterostructure FET devices fabricated via photolithography. (a) AFM images of the AuNP-MoS_2 composite with a 10 Å deposition thickness, following thermal annealing at (i) 300 °C, (ii) 400 °C, and (iii) 500 °C for 1 hour. (b) AFM phase image of the AuNP-MoS_2 composite. (c) AFM images of AuNPs on MoS_2 with a deposition thickness of 3 Å, followed

by thermal annealing at 200 °C and 300 °C for 1 hour and 2 hours. (d) Histogram showing particle size distribution at a 3 Å deposition thickness after 2 hours of annealing at 200 °C and 300 °C.

Result 3: MoS_2-FET performance testing and characterization

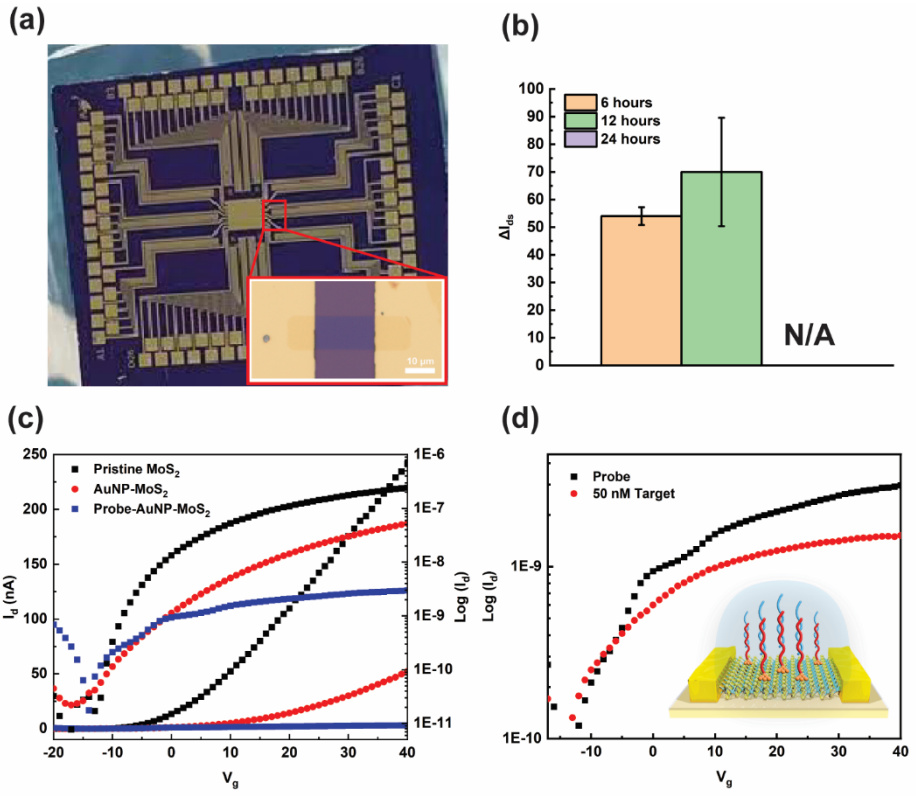


Figure 4. Transfer characteristics of the hairpin probe-functionalized MoS_2-FET sensor. (a) Optical image of the MoS_2-FET array fabricated via photolithography. (b) Measurements showing varying degrees of current degradation when the AuNP-functionalized MoS_2-FET array is incubated with a 100 nM solution of the thiolated hairpin probe for 6, 12, and 24 hours. (c) Transfer curves for the MoS_2-FET before and after functionalization with the hairpin DNA probe. (d) Response curve of the hairpin-functionalized MoS_2-FET sensor during detection of a 50 nM target DNA.

We successfully functionalized the AuNP-MoS_2-FET with hairpin DNA probes, validating the FET biosensor's performance by incubation with a 50 nM target DNA solution. Transport measurements on the CVD-grown MoS_2-based device demonstrated a satisfactory on-off ratio of 10^5 and a field-effect mobility of 2.27 $cm^2/V \cdot s$ under a 0.1 V source-drain bias. Following AuNP functionalization, the device's on-state current decreased by an order of magnitude, accompanied by a p-doping shift to the right in the transfer curve, consistent with previous reports [4]. To immobilize hairpin probes on the MoS_2-FET array, the AuNP-functionalized device was incubated with 100 nM of thiolated probe DNA for 12 hours, resulting in further p-doping in the transfer curve and a decrease in on-state current due to charge scattering. The as-fabricated FET biosensor was then tested with a 50 nM target DNA solution by incubating for 30 minutes. As expected, the results confirmed successful target capture, as evidenced by additional p-doping and current degradation after the target incubation period.

4. PUBLICATION AND AWARDS

J[1] T. Huang et al., "Rapid miRNA detection enhanced by exponential hybridization chain reaction in graphene field-effect transistors," ***Biosensors and Bioelectronics***, vol. 266, p. 116695, 2024.

J[2] H. Sun et al., "Minimizing Contact Resistance and Flicker Noise in Micro Graphene Hall Sensors Using Persistent Carbene Modified Gold Electrodes," ***ACS Applied Materials & Interfaces***, vol. 16, no. 24, pp. 31473-31479, 2024.

J[3] J. Li et al., "Emergent Moiré fringes in direct-grown quasicrystal," ***arXiv***, arXiv:2406.07068, 2024.

J[4] T. Kang et al., "Epitaxial Growth of Two-Dimensional MoO_2–$MoSe_2$ Metal–Semiconductor Heterostructures for Schottky Diodes," ***Nano Letters***, vol. 24, no. 27, pp. 8369-8377, 2024.

J[5] J. You *et al.*, "Epitaxial Growth of 1D Te/2D $MoSe_2$ Mixed-Dimensional Heterostructures for High-Efficient Self-Powered Photodetector," ***Advanced Functional Materials***, vol. 34, no. 10, p. 2311134, 2023.

5. REFERENCES

[1] T. Li *et al.*, "Epitaxial growth of wafer-scale molybdenum disulfide semiconductor single crystals on sapphire," *Nature Nanotechnology,* vol. 16, no. 11, pp. 1201-1207, 2021/11/01 2021. [Online]. Available: https://doi.org/10.1038/s41565-021-00963-8.

[2] J.-H. Fu *et al.*, "Oriented lateral growth of two-dimensional materials on c-plane sapphire," *Nature Nanotechnology,* vol. 18, no. 11, pp. 1289-1294, 2023/11/01 2023, doi: 10.1038/s41565-023-01445-9.

[3] H. Yu *et al.*, "Wafer-Scale Growth and Transfer of Highly-Oriented Monolayer MoS2 Continuous Films," *ACS Nano,* vol. 11, no. 12, pp. 12001-12007, 2017/12/26 2017. [Online]. Available: https://doi.org/10.1021/acsnano.7b03819.

[4] J. Liu *et al.*, "Ultrasensitive Monolayer MoS2 Field-Effect Transistor Based DNA Sensors for Screening of Down Syndrome," *Nano Letters,* vol. 19, no. 3, pp. 1437-1444, 2019/03/13 2019. [Online]. Available: https://doi.org/10.1021/acs.nanolett.8b03818.

Multimedia Technologies & AI Track

Research Reports (2024-2025) In Multimedia Technologies and AI

Newly Funded Projects (2025-2027)	* iCoT: Inference-Time Chain-of-Thought Reasoning of Large Language Models for Scientific Discovery: Framework and Algorithms * Development of Implicit-Based Neuron-Driven Computer-Aided Manufacturing (CAM) Kernel for Multi-Axis Smart Manufacturing
Continuing Projects (2024-2026)	* Building Personalized Multi-modal Large Language Models
(2023-2025)	* Robustifying Decentralized Training of Deep Learning Models with Fully Stochastic Optimization Algorithms * Intelligent Analysis of User Psychological Traits via Online Multi-Modal Social Fingerprint

Completed Project (2022-2024)	* Intelligent Mobile Robot-Augmented Self-Supervised Wi-Fi Sensing

(Funded Year)

ICOT: INFERENCE-TIME CHAIN-OF-THOUGHT REASONING OF LA RGE LANGUAGE MODELS FOR SCIENTIFIC DISCOVERY: FRAMEWORK AND ALGORITHMS

Principal Investigator: Professor Songtao LU
Department of Computer Science and Engineering
CUHK

Project Start Date: 1 July 2025

ABSTRACT

This research focuses on developing a novel inference-time Chain-of-Thought (iCoT) alignment framework to enhance the reasoning capabilities of large language models (LLMs) without requiring modifications to their pretrained parameters. Traditional alignment methods, such as reinforcement learning with human feedback and fine-tuning, are computationally expensive and require extensive labeled data. In contrast, iCoT leverages an inference-time iterative optimization approach, successive policy iteration, to refine policy decisions dynamically. By integrating multilevel optimization techniques, this research aims to enable LLMs to generalize across diverse downstream reasoning tasks, even in the absence of explicit step-wise rewards.

The proposed methods introduce theoretical advancements in policy optimization, offering a provably efficient alternative to conventional fine-tuning, with improved sample efficiency and computational scalability. Practical applications span scientific discovery, theorem proving, and complex problem-solving in specialized domains. The research will result in a software package that allows seamless adaptation of LLMs to domain-specific tasks through inference-time reinforcement learning, fostering real-world applicability. This work bridges machine learning theory and practical deployment, making LLM adaptation more efficient, interpretable, and accessible for a wide range of industries.

INNOVATION AND PRACTICAL SIGNIFICANCE:

To include a paragraph to highlight specifically the innovation and practical significance of your work. Both VC and the donor would like to see more research endeavors be directed to innovation and technology transfer for the betterment of mankind.
This research presents a paradigm shift in adapting LLMs for reasoning tasks by introducing inference-time successive policy iteration, which optimizes decision-making dynamically without retraining. Unlike conventional fine-tuning methods, iCoT enables real-time refinement of LLM outputs, making the adaptation process more efficient, cost-effective, and scalable.

The practical significance of this work lies in its broad applicability and transferability. The developed software package can integrate LLMs into scientific research, healthcare, finance, and industrial automation, accelerating innovation across disciplines. Moreover, this approach allows alignment with privacy-sensitive or domain-specific data without requiring extensive retraining, making it highly relevant for commercial and enterprise AI solutions. By eliminating computational bottlenecks and enhancing LLM adaptability, this research sets the foundation for next-generation AI deployment strategies that prioritize efficiency, generalization, and real-world usability.

PROJECT OBJECTIVES:

The proposed research aims to enhance the reasoning capabilities of pretrained large language models (LLMs) for downstream tasks without requiring model parameter modifications. The key objectives include:
1. Parameter-Free Adaptation. Develop an inference-time adaptation framework that does not require modifying pretrained model parameters.
2. Versatility Across Reasoning Tasks. Enable LLMs to handle diverse downstream reasoning tasks, even in the absence of explicit intermediate process rewards.
3. Theoretical Guarantees. Provide formal justification that the proposed inference-time approach outperforms fine-tuning in terms of sample efficiency and computational complexity.

Expected Outcomes and Long-Term Impact:
• Practical Deployment. Development of a software package enabling seamless adaptation of LLMs to domain-specific applications via inference-time reinforcement learning.
• Theoretical Insights. Establishing iCoT's theoretical foundation, demonstrating its ability to refine reasoning quality under limited training data constraints.
• Advancing Post-Training Techniques. Contributing novel post-training methodologies applicable across real-world applications, unlocking more efficient, scalable, and transferable LLM alignment strategies.
• Expanding AI Reasoning Capabilities. Showcasing the potential of inference-time reasoning algorithms to generate high-quality solutions without requiring extensive data coverage in the pretrained base model.
This research will bridge theoretical advancements with practical applications, making post-training adaptation more efficient, interpretable, and broadly applicable across industries.

DEVELOPMENT OF IMPLICIT-BASED NEURON-DRIVEN COMPUTER-AIDED MANUFACTURING (CAM) KERNEL FOR MULTI-AXIS SMART MANUFACTURING

Principal Investigator: Professor Guoxin FANG
Department of Guoxin Fang, Engineering
CUHK

Project Start Date: 1 July 2025

ABSTRACT

Computer-aided manufacturing (CAM) is the core computational engine bridging digital innovation and physical production, making it a fundamental computational kernel to support smart manufacturing. Despite advancements in robot-assisted additive manufacturing (AM) hardware, CAM development has lagged in supporting further enhancement systems due to challenges in three-dimensional decomposition for spatial toolpath generation. Existing mesh-based solutions are computationally inefficient and cannot leverage GPU parallelism, resulting in high error rates and low scalability for large-scale systems. This research introduces a neuron-based optimization pipeline as the kernel of the CAM system. It leverages implicit geometry representation with signed distance fields to enable support-free toolpath generation and differentiable collision detection. Unlike traditional mesh-based approaches, our method supports GPU acceleration through specially designed algorithms, reducing computation time from several hours to seconds. The CAM kernel is sufficient to be integrates into the design-manufacturing co-optimization pipeline, optimizing the mechanical performance of fabricated components. The project advances robot-assisted AM practices by enabling rapid manufacturing of large-scale, complex structure components while reducing material waste to support sustainability goals. The success of this research will drive future advancements in the academic fields of computational design and digital manufacturing. Meanwhile, we plan to commercialize the CAM kernel as a software package to strengthen Hong Kong's leadership in smart manufacturing.

INNOVATION AND PRACTICAL SIGNIFICANCE:

Smart manufacturing driven by AI and robotics relies on advanced CAM systems to translate digital designs into physical production. However, existing mesh-based CAM solutions create computational bottlenecks, limiting the fabrication of complex, high-performance components. This research introduces a next-generation CAM kernel that leverages implicit geometry representation and a neuron-based optimization pipeline to overcome these limitations. By enabling support-free toolpath generation, real-time collision checking, and motion planning for robotic systems, this breakthrough will make it possible to manufacture previously unattainable structures—such as lightweight, high-performance models optimized through topology optimization. These advancements will significantly expand the capabilities of robot-assisted additive manufacturing (AM) in aerospace, automotive, and construction industries. Beyond manufacturability, this project directly improves the sustainability and production efficiency in current smart manufacturing practice. Additive manufacturing inherently reduces material waste compared to traditional subtractive methods, and this research further enhances sustainability by optimizing material distribution and minimizing unnecessary support structures with the help of robot motion. The proposed CAM system will also drastically reduce manufacturing lead time through a GPU-accelerated computational pipeline, cutting toolpath generation and robotic motion planning from hours to seconds—enabling faster production cycles and greater scalability for large-scale applications. By driving innovation in computational manufacturing, this research will establish a

high-efficiency, sustainable production framework, reinforcing Hong Kong's leadership in next-generation smart manufacturing.

PROJECT OBJECTIVES:

Key issue addressed: Traditional CAM systems rely heavily on heuristic algorithms and explicit geometric representations (e.g., voxels, meshes, point clouds), which are often tailored to specific geometries or manufacturing scenarios. They lack the flexibility to handle complex and diverse designs in modern robot-assisted smart manufacturing systems, which benefit from enhanced multi-axis motion capabilities. Furthermore, these CAM systems are poorly suited for integration into advanced design-manufacturing co-optimization pipelines due to their incompatibility with shape representations, slow computational performance, and limited scalability for large-scale problems. As a result, they lead to suboptimal mechanical performance and excessive material waste. ***There is an urgent need to develop an advanced AI-driven CAM kernel*** that enables high-level automation of manufacturing practices and fully leverages multi-axis motion capabilities to support the manufacturing of high-performance components by computational innovation.

Project objectives and outcome: The outcome of this project is to develop an general AI-driven CAM kernel utilizing implicit geometry representation and neuron-based optimization pipeline to enhance precision, scalability, and computational efficiency in toolpath generation and motion planning for robot-assisted manufacturing. The key project objectives include:

• Fabrication-Aware SDF Training and Neuron-Based Field Optimization for Toolpath Generation: Develop an implicit-based solid and guidance field learning framework for curved working surface slicing and toolpath generation, unifying diverse geometric inputs into a single differentiable representation.
• Design-Fabrication Co-Optimization: Create a stress-aware infill pattern optimization pipeline that dynamically generates spatially adaptive truss structures. This will enhance mechanical strength while enabling support-free printing for components that require high-performance and lightweight structures.
• Real-Time Collision Detection and Motion Optimization for Robot System: To ensure efficient and adaptive robot-assisted 3D printing, we will implement GPU-accelerated collision checking and neuron-driven motion planning, taking advantage of implicit representation.
• Experimental Validation for Industrial Applications: Conduct physical fabrication experiments to evaluate performance across diverse cases, demonstrating real-world applicability and commercial potential.

In the long term, the development of this CAM kernel will establish a new computational foundation for smart manufacturing, providing an end-to-end solution with minimal manual input and seamless automation of complex fabrication workflows. T***he CAM kernel will be commercialized as a software package*** and integrated with state-of-the-art multi-axis manufacturing hardware through industry collaborations. Fostering AI-driven design-fabrication optimization will transform production processes across industries. For example, it will accelerate the development of lightweight, high-performance aerospace components to reduce overall energy consumption. It can also support the fabrication of patient-specific implants in healthcare for better treatment and energy-efficient structures in construction to support sustainable urbanization. Ultimately, this technology will shape a more sustainable and equitable future, reinforcing global leadership in smart manufacturing to address humankind's evolving needs.

WORK IN PROGRESS

BUILDING PERSONALIZED MULTI-MODAL LARGE LANGUAGE MODELS

Principal Investigator: Professor YUE Xiangyu
Department of Information Engineering
CUHK

Research Team Members: Wei-Hong LI, Postdoctoral Fellow [(1)]

[(1)] Dept. of Information Engineering, CUHK

Reporting Period: 1 July 2024 – 30 April 2025

INNOVATION AND PRACTICAL SIGNIFICANCE:

The proposed research aims to enhance Large Language Models (LLMs) for personalized applications across various domains. Currently, LLMs e.g. ChatGPT have limitations in understanding multiple data modalities, hindering their effectiveness in real-world scenarios where information exists in diverse formats such as images, text, audio, and more. This project seeks to address this challenge by developing a Multi-modal LLM capable of processing and comprehending different data types. By designing a unified data encoder and implementing a mixture of experts approach, the scalability and adaptability of the model will be significantly improved. Furthermore, the project focuses on enhancing the context-awareness and relevance of responses generated by Multi-modal LLMs through cross-modal retrieval-based approaches and cross-modal contrastive learning methods. Additionally, a domain-adaptive learning approach will be developed to ensure the model's robustness to data distribution shifts. The innovations proposed in this research will enable LLMs to understand and process more data modalities (text, image, video, audio, infrared, MRI, IMU, etc.), and generate personalized responses tailored to individuals. This will have significant practical implications, with applications spanning various sectors including healthcare, education, finance, and beyond.

ABSTRACT

Large Language Models (LLMs), *e.g.* ChatGPT, have become increasingly popular and play a crucial role in Artificial Intelligence. These models excel at understanding and acting upon instructions conveyed in natural language. However, current LLMs can only understand a very limited number of data types (image, text), and cannot provide personalized responses tailored to individual users (*e.g.* all users use the same ChatGPT model). To address this, our proposal focuses on developing personalized LLMs in three key areas. Firstly, we aim to create a Multi-modal LLM capable of understanding various data types such as images, text, audio, video, etc. This involves designing a system that can adapt and process multiple data types efficiently. Secondly, we want to improve how these models respond to user queries by incorporating personal knowledge and context into their answers. This will involve retrieval-augmented generation and aligning features across different types of data. Finally, we aim to ensure that these models perform consistently across different domains by developing a method to adapt to changes in data distribution. Overall, our goal is to make LLMs understand a wide range of data modalities and generate personalized responses, ultimately enhancing their utility in various real-world applications.

1. OBJECTIVES AND SIGNIFICANCE

The proposed research aims to enhance Large Language Models (LLMs) for personalized applications across various domains. Currently, LLMs such as ChatGPT have limitations in understanding multiple data modalities, hindering their effectiveness in real-world scenarios where information exists in diverse formats such as images, text, audio, and more. This project seeks to address this challenge by developing a Multi-modal LLM capable of processing and comprehending different data types. By designing a unified data encoder and implementing a mixture of experts approach, the scalability and adaptability of the model will be significantly improved. Furthermore, the project focuses on enhancing the context-awareness and relevance of responses generated by Multi-modal LLMs through cross-modal retrieval-based approaches and cross-modal contrastive learning methods. Additionally, a domain-adaptive learning approach will be developed to ensure the model's robustness to data distribution shifts. The innovations proposed in this research will enable LLMs to understand and process more data modalities (text, image, video, audio, infrared, MRI, IMU, *etc.*), and generate personalized responses tailored to individuals. This will have significant practical implications, with applications spanning various sectors including healthcare, education, finance, and beyond.

Project Objectives:

1. **Develop a Personalized Multi-modal LLM**: This model will cater to diverse data modalities and user contexts, setting a foundation for AI technologies that are finely tuned to individual preferences and needs, ultimately transforming user interactions across sectors such as education, healthcare, and public services.
2. **Create a General-purpose Multi-modal LLM**: By supporting a large array of data modalities with computational efficiency, this model will expand the application of AI in creating personalized services and products, significantly impacting technology adoption and the user experience in both personal and professional settings.
3. **Incorporate User-Specific Knowledge Efficiently**: With this objective, the project aims to make AI interactions more relevant and insightful, enriching user experience and fostering advancements in AI applications that are sensitive to the nuances of individual user histories and preferences.
4. **Make the Multi-modal LLM Domain-Adaptive**: By considering individual data distribution shifts, the project will result in AI models that are robust to changes in user environments and preferences, ensuring long-term reliability and effectiveness of AI systems in dynamic real-world scenarios.

2. RESEARCH METHODOLOGY

This project investigates personalized multi-modal instruction-following models, valuable both theoretically and practically. We introduce an efficient Multi-modal LLM framework, supporting more data modalities than current models. Additionally, we present a retrieval-based method for fast personal knowledge integration and a domain-adaptive approach for effective data adaptation.

(i) **Unified Multi-modal large language models framework**

Current MLLMs employ independent feature encoders and projection modules for each data modality, resulting in significant model size and computational complexity. Our proposed MLLM framework introduces a unified encoder and mixture of experts to address these challenges efficiently. Given a sample of N modalities, we employ a lightweight tokenizer to convert it into token sequences. Utilizing a pretrained transformer model as a universal encoder, we encode token embeddings from various modalities. Unlike previous methods with modality-specific projections, we introduce a unified projection module with a mixture of projection experts and a light modality router for token projection. The concatenated modality tokens and text prompt serve as input for multimodal-to-text generation in the LLM.

Multi-modal alignment. We propose to progressively ground all modalities into LLMs. We first train the tokenizer and projection module on image-text data to align the image with text considering image is the most common modality. We then divide other modality data into multiple training stages according to their data

scale.

Multi-modal instruction tuning. The model after multi-modal alignment still lacks the ability to follow human instructions and generate detailed responses. We propose to tune the model on both unimodal and multi-modal instruction data with parameter-efficient methods, such as Bias Tuning and LoRA, keeping the encoder frozen.

(ii) **Incorporating personal knowledge into MLLMs**

Current LLMs typically struggle to incorporate personal knowledge into their responses. To deliver more relevant and context-aware responses tailored to individual users, we propose two methods that incorporate personal knowledge into MLLMs effectively: Retrieval-Augmented Generation with Cross-modal Contrastive Alignment and Parameter-Efficient Finetuning.

Cross-modal contrastive alignment. To explicitly align different modalities for further improving information retrieval, for any paired input (x_i, x_j, x_T), *e.g.*, image, audio and video subtitle, we generatively map the input signal to text: $x_i/x_j \rightarrow x_T$, and contrastively align x_i, x_j and x_T into a unified space using the InfoNCE loss.

Retrieval-augmented generation. User devices usually have too limited computation resources to update the model locally and we cannot send user data to the server due to privacy considerations. With all modalities aligned into a unified space, information retrieval from the user knowledge database (Retrieval-augmented generation) becomes more accessible and can be used to generate a more accurate and personalized response.
Parameter-efficient finetuning. RAG is a tuning-free approach, but it might not precisely align with the input distribution, leading to subpar performance. To address this, we propose enhancing the model weights using parameter-efficient fine-tuning (PEF) methods such as LoRA [45]. These algorithms demand minimal computation, enabling local device tuning without necessitating extensive training data, making it feasible to personalize a MLLM with limited user data.

(iii) **Mitigating individual data distribution shift**

Moreover, the model must address data distribution shifts to avoid performance degradation. Given the MLLM's long-term assistance for individuals and the continuous nature of domain shifts, we introduce a test-time adaptation method involving marginal generalization and prototypical self-supervised learning. We consider two domains: a) a source domain consisting of public data that the initial MLLM model is trained on, and b) a target domain referring to personal data whose distribution differs across individuals and might change dynamically. We propose to align features of source and target domains.

Open-set prototype learning. Most previous work focuses on close-set settings, where the number of classes is known. Conversely, the general-purpose MLLM targets an open-set setting, where the number of categories is undefined. The data also tend to have a hierarchical semantic structure, *e.g.* “bird” is a broader category compared “pigeon”. We first perform prototype learning on the source domain by considering the hierarchical data semantic structure via k-means, constructing prototypes and classifier for cross-domain feature alignment.

Marginal generalization. Unlike previous test-time adaptation methods, we propose the marginal generalization. We initialize the target domain’s encoder with the source encoder’s weights and train it over the target domain with the proposed Marginal Adaptation Loss that aligns the encoder outputs of both source and target domains up to a predefined distance threshold.

Cross-domain prototypical self-supervised learning. To enhance the alignment of source and target features more effectively, we propose a prototypical self-supervised learning approach. We promote the target features with similar semantics to be more closely aggregated, aiming to learn a more discriminative target encoder.

3. RESULTS ACHIEVED SO FAR

During the reporting period from July 1, 2024, to April 30, 2025, significant progress has been made towards the project objectives of building personalized multi-modal large language models, laying a strong foundation for potential commercialization and technology transfer. The research team's work, detailed in the publications listed in Section 4, directly contributes to developing innovative technologies with broad applicability.

Specifically, the advancements in fundamental multi-modal learning and generative AI techniques are crucial for creating highly versatile AI systems. Research on 3D-aware image compositing with language instructions [3] and tuning-free multi-prompt longer video generation [5] demonstrate progress in enabling AI to understand and generate complex multimedia content, which is directly applicable to creative industries, entertainment, and virtual reality applications. Foundational work in areas like training matting models without alpha labels [8] and unified spatio-temporal learning [7] further enhances the models' ability to process and interpret diverse visual and temporal data, increasing their potential for use in areas like video surveillance, content editing software, and scientific analysis.

Furthermore, the project has made strides in addressing the critical need for personalization and adaptability in LLMs. The development of Retrieval-Augmented Personalization for Multimodal Large Language Models [1] (Figure 1) represents a key achievement in enabling models to effectively incorporate user-specific knowledge, paving the way for highly customized and relevant AI interactions in applications such as personalized tutors, healthcare assistants, and financial advisors. Progress in spatial understanding, including online vectorized HD map construction [2] and dynamic contextual human motion generation [6], contributes to the models' ability to operate effectively in diverse physical and virtual environments, essential for robotics, augmented reality, and simulation technologies. Investigations into room layout reconstruction from sparse views [9] also contribute to spatial understanding crucial for multi-modal personalized applications. Additionally, underlying research improving diffusion models [4] enhances the quality and efficiency of generative capabilities, crucial for producing high-fidelity personalized content. This comprehensive progress directly contributes to unlocking significant commercial opportunities in personalized AI applications, data integration solutions, and advanced content creation tools.

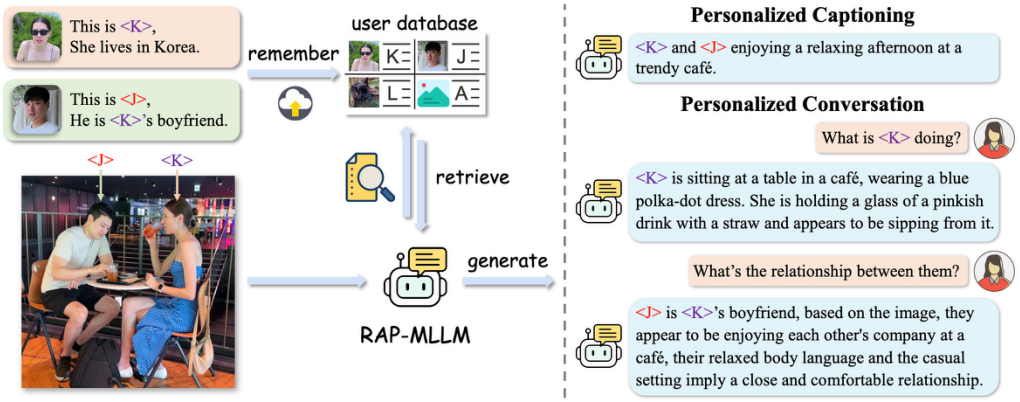


Figure 1. RAP: Retrieval-Augmented Personalization for Multimodal Large Language Models.

These achieved results represent tangible progress in building the core technologies required for personalized multi-modal LLMs. The potential for technology transfer and commercialization remains a central focus, with the current progress directly contributing to unlocking opportunities in areas such as personalized virtual assistants, data integration platforms, privacy-preserving applications, and customized content generation, driving innovation across industries. The project is actively refining these technological foundations and exploring engagement with stakeholders to facilitate future transfer and adoption.

4. PUBLICATION AND AWARDS

C[1] H. Hao, J. Han, C. Li, Y.-F. Li, and **X. Yue**, "RAP: Retrieval-Augmented Personalization for Multimodal Large Language Models," IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2025.

C[2] Z. Zhang, Y. Zhang, X. Ding, F. Jin, and **X. Yue**, "Online vectorized HD map construction using geometry," European Conference on Computer Vision (ECCV), Springer Nature, Switzerland, pp. 73-90, 2024.

C[3] L. Li, K. Gong, W. Li, X. Dai, T. Chen, X. Yuan, and **X. Yue**, "BIFRÖST: 3D-Aware Image compositing with Language Instructions," Annual Conference on Neural Information Processing Systems (NeurIPS), pp. 129480-129506, Canada, 2024.

C[4] L. Zhuo, R. Du, H. Xiao, Y. Li, D. Liu, Z. Ma, X. Luo, Z. Wang, K. Zhang, L. Zhao, S. Liu, **X. Yue**, W. Ouyang, Y. Qiao, H. Li, and P. Gao, "Lumina-Next : Making Lumina-T2X Stronger and Faster with Next-DiT," Annual Conference on Neural Information Processing Systems (NeurIPS), pp. 131278-131315, 2024.

C[5] M. Cai, X. Cun, X. Li, W. Liu, Z. Zhang, Y. Zhang, Y. Shan, and **X. Yue**, "DiTCtrl: Exploring Attention Control in Multi-Modal Diffusion Transformer for Tuning-Free Multi-Prompt Longer Video Generation," International Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, USA, 2025.

C[6] P. Cong, Z. Wang, Y. Ma, and **X. Yue**, "SemGeoMo: Dynamic Contextual Human Motion Generation with Semantic and Geometric Guidance," International Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, USA, 2025.

C[7] C. Tang, X. Ma, E. Su, X. Song, X. Liu, W.-H. Li, L. Bai, W. Ouyang, and **X. Yue**, "UniSTD: Towards Unified Spatio-Temporal Learning across Diverse Disciplines," International Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, USA, 2025.

C[8] W. Liu, Z. Ye, H. Lu, Z. Cao, and **X. Yue**, "Training Matting Models without Alpha Labels," Conference on Artificial Intelligence, AAAI, pp. 5604-5612, USA, 2025.

C[9] Y. Huang, X. Dai, J. Wang, Y. Yuan, and **X. Yue**, "Unposed Sparse Views Room Layout Reconstruction in the Age of Pretrain Model", International Conference on Learning Representations (ICLR), 2025.

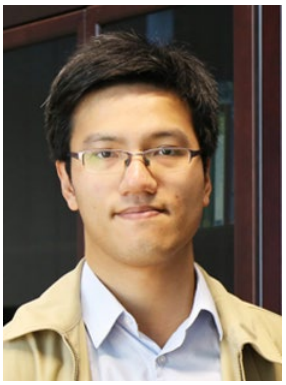
WORK IN PROGRESS

ROBUSTIFYING DECENTRALIZED TRAINING OF DEEP LEARNING MODELS WITH FULLY STOCHASTIC OPTIMIZATION ALGORITHMS

Principal Investigator: Professor Hoi-To Wai
Department of Systems Engineering and Engineering Management, CUHK

Research Team Members:
Chung Yiu Yau, PhD student [(1)],
Haoming Liu, Research Associate [(1)]

[(1)] Dept. of Systems Engineering and Engineering Management, CUHK



Reporting Period: 1 July 2023 – 31 May 2024

INNOVATION AND PRACTICAL SIGNIFICANCE:

This project entails major innovations on algorithm design principles in decentralized DNN training. This is both practical and timely with the recent advocation of federated learning by Google and other leading tech companies. The proposed innovative design tackles a challenging problem of adapting decentralized optimization algorithms to practical network environments via a general and flexible design principle. We anticipate that the results can greatly benefit downstream machine learning (ML) applications on emerging scenarios such as IoT devices, distributed computing, etc. Together with the strong technical component on theoretical analysis of the algorithms, which will be a significant contribution on its own right, the project is anticipated to help advance the state-of-the-art in ML.

ABSTRACT

The popularization of Internet-of-Things (IoT) networked devices has rendered our world more connected than ever. Simultaneously, it raised the challenge of how to leverage the collective intelligences of networked devices in a decentralized manner, e.g., can smart phones cooperate to classify images with their private dataset? can a fleet of autonomous cars coordinate among themselves to reach a certain destination? These challenges have prompted the development of decentralized optimization algorithms which combine training (stochastic gradient) with peer-to-peer communication. Meanwhile, the widespread use of deep neural networks (DNNs) has deployed large models with impressive performance. Training such large models in a distributive manner could waste a lot of bandwidth since these algorithms require devices to communicate the entire model frequently.

This project proposes a fully stochastic optimization framework that adapts to completely random communication graph and compression. Leveraging on the PI's recent works, this project aims to (A) robustify the decentralized training process of DNNs against links failure and/or long latency, (B) advance theories of decentralized stochastic optimization, especially for algorithm designs with compression, and (C) conduct numerical experiments on realistic time varying network to provide empirical evidence that supports the algorithmic/theoretical findings.

1. OBJECTIVES AND SIGNIFICANCE

To robustify decentralized training algorithms of deep neural networks (DNNs) against dynamical network environment, we address a key open issue of existing algorithms in training DNNs as network overheads slow down the training process. We develop a stochastic primal dual framework for decentralized optimization that adapt to dynamical network environments.

To advance theories of stochastic decentralized optimization. While partial results have been reported, a *fully stochastic* algorithm as Objective 1 that simultaneously adapt to compression and random graphs with link failures has not been studied. Our framework also offers new mechanisms for analyzing similar stochastic algorithms.

To deploy and test the proposed framework on realistic network environment to support the theories in Objective 2. Training of DNNs such as ResNet-50, VGGNet-16 will be simulated on a testbed accounting for network latency, etc.

2. RESEARCH METHODOLOGY

This project aims at improving the robustness of decentralized training of DNN via advances in **algorithm design**, **theories**, and **applications**. We inquire the following research questions:

> *Can we suggest a general design principle for **robustified** stochastic decentralized optimization with compression on dynamical graphs? What are the general convergence guarantees? How do the algorithms work in real network environments?*

To formulate the decentralized DNN training problem, we treat each device as a node and consider n nodes by $\mathcal{V} = \{1, ..., n\}$, linked through a connected graph denoted by $\mathcal{G} = (\mathcal{V}, \mathcal{E})$:

$$\min_{\boldsymbol{\theta}_i \in \mathbb{R}^d, i=1,...,n} \; F(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_n) := \tfrac{1}{n} \textstyle\sum_{i=1}^n f_i(\boldsymbol{\theta}_i) \quad \text{s.t.} \quad \boldsymbol{\theta}_i = \boldsymbol{\theta}_j, \; \forall \, (i,j) \in \mathcal{E}, \tag{1}$$

where $\boldsymbol{\theta}_i$ is the DNN's weights, $f_i(\boldsymbol{\theta}_i)$ is a local function such as the cross-entropy loss for the local data, both held by node i. Note that $f_i(\boldsymbol{\theta}_i)$ is non-convex in general and is difficult to compute as it encapsulates a lot of training samples.

To tackle (1), notice that the main challenge is the consensus constraint, which can be represented as a linear equality: set $\boldsymbol{\theta} = (\boldsymbol{\theta}_1; \cdots; \boldsymbol{\theta}_n) \in \mathbb{R}^{nd}$ and $\boldsymbol{B} \in \mathbb{R}^{|\mathcal{E}| \times n}$ be an incidence matrix whose rows correspond to the edges of $\mathcal{G}$. The requirement $\boldsymbol{\theta}_i = \boldsymbol{\theta}_j, \; \forall \, i, j \in \mathcal{E}$ is equivalent to $(\boldsymbol{B} \otimes \boldsymbol{I}_d)\boldsymbol{\theta} = \mathbf{0}$. The **key idea** is that $\boldsymbol{B} \otimes \boldsymbol{I}_d$ can be replaced by its randomization $\boldsymbol{B}(\xi) \otimes \boldsymbol{I}(\xi)$:

$$(\boldsymbol{B} \otimes \boldsymbol{I}_d)\, \boldsymbol{\theta} = \mathbf{0} \iff \mathbb{E}[\; \underbrace{\boldsymbol{B}(\xi)}_{\text{random graph}} \otimes \underbrace{\boldsymbol{I}(\xi)}_{\text{random compression}} \;]\, \boldsymbol{\theta} = \mathbf{0}. \tag{2}$$

The linear constraint can thus be interpreted as a *stochastic linear equality* encompassing random graph and compression. With $\boldsymbol{B}(\zeta)$ taken as a random graph, taking $\boldsymbol{I}(\zeta) = \mathrm{Diag}(\boldsymbol{i}(\zeta))$ as $\boldsymbol{i}(\zeta) \sim \mathrm{Ber}(p)^d$ models the case of random sparsification that communicates $pd \ll d$ numbers.

Leveraging on the stochastic equality (2), we consider the stochastic augmented Lagrangian function for (1): with $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1^\top, ..., \boldsymbol{\theta}_n^\top)^\top \in \mathbb{R}^{nd}$

$$\mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{\lambda}) := \mathbb{E}_\xi \left[\frac{1}{n} \sum_{i=1}^n f_i(\boldsymbol{\theta}_i; \xi) + \eta \langle \boldsymbol{\lambda}, \boldsymbol{B}(\xi) \otimes \boldsymbol{I}(\xi) \rangle + \frac{\gamma}{2} \| \boldsymbol{B}(\xi) \otimes \boldsymbol{I}(\xi)\, \boldsymbol{\Theta} \|^2 \right]. \tag{3}$$

Using (3), the *fully stochastic proximal primal-dual algorithm* (FSPPD) algorithm can be built as a primal-dual stochastic gradient method: let $\widetilde{\mathbf{L}}(\xi^{t+1}) = \boldsymbol{B}(\xi^{t+1})^\top \boldsymbol{B} \otimes \boldsymbol{I}(\xi^{t+1})$, for any $t \geq 0$,

$$\begin{cases} \boldsymbol{\Theta}^{t+1} = \boldsymbol{\Theta}^t - \gamma \widetilde{\mathbf{L}}(\xi^{t+1}) \boldsymbol{\Theta}^t - \eta \widehat{\boldsymbol{\lambda}}^t - \alpha \nabla \mathbf{f}(\boldsymbol{\Theta}^t; \xi^{t+1}), \\ \widehat{\boldsymbol{\lambda}}^{t+1} = \widehat{\boldsymbol{\lambda}}^t + \beta \widetilde{\mathbf{L}}(\xi^{t+1}) \boldsymbol{\Theta}^t, \end{cases} \tag{4}$$

where $\alpha, \beta > 0$ are additional step size parameters. Notice that (4) is adapted to random graph with coordinate-wise sparsification and is a special case of the *stochastic forward-backward* algorithm in [1]. Meanwhile, in the special case of (4) with static graph, the former can be used to recover classical algorithms such as ProxPDA [2], DIGing [3]. We anticipate FSPPD to achieve fast convergence as these existing algorithms.

To achieve our objectives and long term goals, the project will be divided into the following work packages to be achieved in order:

1. **Design principle for FSPPD**: In addition to studying the random sparisfication idea as in (2) which constitute a "linear compression" method, we will also study "nonlinear compression" method such as incorporating randomized quantization in the FSPPD framework.
2. **Convergence properties of FSPPD**: We will study the expected convergence of FSPPD using both empirical and theoretical analysis. Our main analysis technique is to investigate the algorithm using a Lyapunov function approach.

3. **Training large DNNs with FSPPD**: The FSPPD algorithm will also be tested on realistic communication network environment.

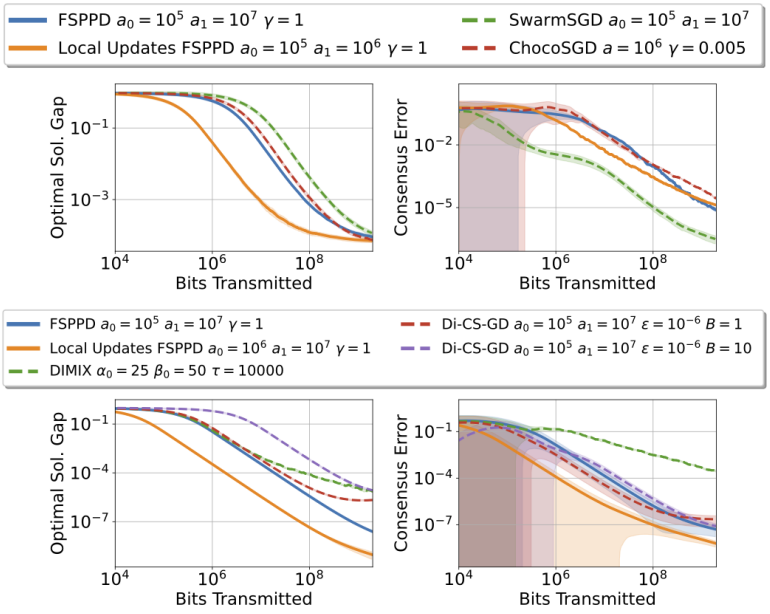
References

[1] P. Bianchi, W. Hachem, and A.Salim. A fully stochastic primal-dual algorithm. Optimization Letters, 15(2):701–710, 2021.
[2] D. Hajinezhad and M. Hong. Perturbed proximal primal–dual algorithm for nonconvex nonsmooth optimization. Mathematical Programming, 176(1-2):207–245, 2019.
[3] A. Nedic, A. Olshevsky, and W. Shi. Achieving geometric convergence for distributed optimization over time-varying graphs. SIAM Journal on Optimization, 2017.

3. RESULTS ACHIEVED SO FAR

The project team is composed of 1 PhD student (Mr. Yau Chung Yiu) and 1 Research Associate (Dr. Liu Haoming). Furthermore, an incoming PhD student (Mr. Rongxin Du from Zhejiang University) will join the team in summer 2024. Over the past year, we have made progresses in algorithm development and theoretical analysis. We list the achievements of this project so far:

1. The paper [C1] has been accepted and presented at the IEEE CDC 2023, Singapore in December 2023. In the paper, we proposed the FSPPD framework using the stochastic linear equality formulation and provided preliminary convergence analysis for the case of strongly convex objective function. Here are numerical experiments presented:

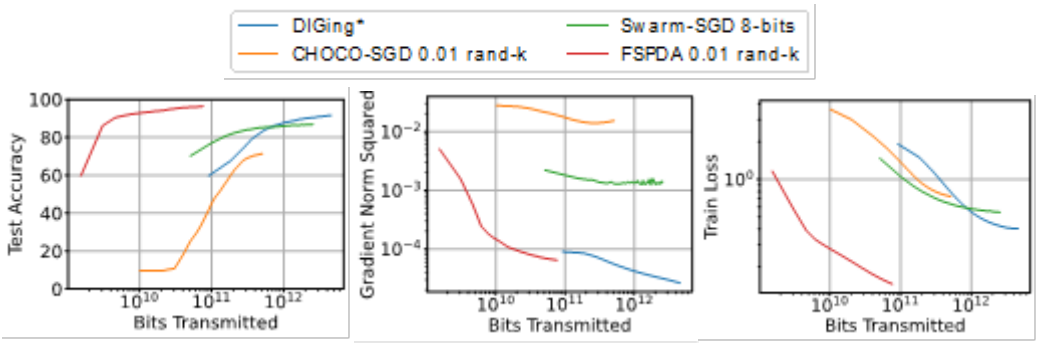


We have tested the algorithms on a simple linear regression problem and showed that it achieves a better communication efficiency compared to SOTA algorithms.

2. We have completed the convergence analysis of FSPPD in the **non-convex optimization** setting. Through a unique stochastic Lyapunov function design, we obtained the following result: under suitable and mild assumptions, it holds

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\bar{\boldsymbol{\theta}}^t)\right\|^2 = \mathcal{O}\left(\frac{L\sum_{i=1}^{n}\sigma_i^2}{\sqrt{nT}}\right)$$

where the omitted constant and step size criterion depends on the stochasticity of the random graph and sparsification properties. Currently the analysis works only for uniform sparsification and edge selection. We have also extended the numerical experiments for DNN training and achieved the following results:



As a preliminary example, the above experiments are performed on training a 2-layer feedforward neural network in classifying the handwritten digits in MNIST. Observe that the proposed approach outperforms several SOTA algorithms in terms of the communication complexity.

Observe that the proposed FSPPD algorithm has significantly outperformed the competing algorithms in terms of communication efficiency. It has thus partially achieved the planned **Objectives 1, 2, and 3**.

3. We are preparing a journal paper, to be submitted to the IEEE Transactions on Automatic Control, during the summer of 2024. In addition, our results will be presented at an invited special session at the Asilomar Conference, Asilomar, CA in October, 2024.

4. The above result has been further extended to incorporate nonlinear compression schemes (e.g., quantization), thus achieving the second component in **Objective 1.** The idea of the extension is to develop the algorithm through a majorization-minimization approach and integrating with nonlinear gossiping algorithm.

5. Currently, we are working on the convergence analysis of the extended algorithms and conducting numerical experiments, thus achieving the second component in **Objectives 2 & 3**. We plan to complete and submit a journal paper by the end of October in 2024.

4. PUBLICATION AND AWARDS

[C1] C.-Y. Yau, H.-T. Wai, “Fully Stochastic Distributed Convex Optimization on Time-Varying Graph with Compression”, in Proc. of IEEE CDC, 2023.

WORK IN PROGRESS

INTELLIGENT ANALYSIS OF USER PSYCHOLOGICAL TRAITS VIA ONLINE MULTI-MODAL SOCIAL FINGERPRINTS

Principal Investigator: Professor CHENG Hong
Department of Systems Engineering and Engineering Management, CUHK

Research Team Members:
SUN Xiangguo, PhD student [(1)],
FENG Zijin, PhD student [(1)]

[(1)] Dept. of Systems Engineering and Engineering Management, CUHK

Reporting Period: 1 July 2023 – 31 May 2024

INNOVATION AND PRACTICAL SIGNIFICANCE:

Innovation
Psychological analysis in online social networks is very challenging but definitely worthy of our investigation. Although traditional psychological studies provided many insightful observations to the inner personality of people, they heavily rely on users' offline surveys or questionnaires and the analyzed results may vary a lot by different experts. It can hardly meet the current demand because nowadays people transfer more of their activities online and they need faster, more accurate, and more timely estimation of their real-time mental status to help them for better decision. As an alternative, online social networks record users' real-time data with more personalized behavior traces and support highly intelligent psychological analysis in a self-service manner. However, online social networks are far more complicated than traditional questionnaires because they usually have no clear indicator of one's personality traits. How to uncover the underlying psychological motivations from various behavior data, and how to deal with the consistency and conflict of different types of data during the psychological study have become two intractable challenges. To this end, we propose to learn the private features of each kind of data, and the shared features across different data modalities. In this way, we can seamlessly integrate the fact that different people usually have different sensibilities to different expression forms, and further improve the reliability of our psychological analysis with the help of multi-modal fusion. To further study the impact of social environments, we introduce the hypergraph learning (a hypergraph allows one edge to connect multiple nodes, which is perfect to model various social environments) to study the interaction between user's inner traits and their online environments.

Significance
Although there are some existing studies trying to detect one's psychological status from textual or visual data, most of them ignore more profound influence among different data modalities, and users'

various social environments. By contrast, this project considers how to deal with the consistency and conflict of multiple data modalities (including but not limited to text, images, network, etc.), and further study the social impact from the hypergraph level. These technical innovations will guarantee our project more feasible to the real-world applications and obtain more insightful findings from user's psychological analysis. The outcome will benefit the whole society from different perspectives as illustrated in the following. For example, the developed system can be easily integrated in online interviewing so that the interviewer and interviewee can both recognize their matching degree between specific jobs and user's inner traits. The online studying system can analyze students' behaviors to customize more personalized teaching plans for different students. Universities can leverage this system to care about their students' psychological states and provide timely

counselling. Online social platforms can also integrate this system to find psychological abnormality to build a healthier cyberspace.

ABSTRACT

Caused by the long-term impact of COIVD-19, there are an increasing number of people suffering from serious mental diseases and students discouraged by low-effective remote study. Thus an important issue is how to improve people's mental healthiness in the post-pandemic era. Under this background, effective psychological analysis would be very crucial and beneficial as it can identify risks of mental diseases, help people understand themselves better, and improve their performance by more personalized suggestions. Unfortunately, traditional psychotherapy services are usually too specialized, expensive and with limited availability, which is far from flexible and efficient. There is a pressing need to develop more intelligent technologies which can be easily applied to various applications. By capitalizing on our previous research on online social data mining and interdisciplinary study, this project will develop an AI-driven system to analyze users' online social behavior for their psychological traits and prompt suggestions if they present the high risk of mental diseases. The outcome of this project can support large-scale deployment and usage for online interviewing, online education, remote mental intervention, and so on, and can be easily embedded in many online platforms.

1. OBJECTIVES AND SIGNIFICANCE

Objectives:

1. Learning the underlying psychological patterns from user's online behaviors.
2. Handling the consistency and conflicts of different data modalities to improve the reliability of psychological analysis.
3. Studying the social interaction influence with user's psychological traits.
4. Developing friendly systems/plug-ins for psychological self-evaluation and more applications.

Significance:

This project studies how to deal with the consistency and conflict of multiple data modalities (including but not limited to text, images, network, etc.), and further study the social impact from the hypergraph level. These technical innovations will make our project more feasible to the real-world applications and obtain more insightful findings from user's psychological analysis. The outcome will benefit the whole society from different perspectives. Online social platforms can integrate this system to find psychological abnormality to build a healthier cyberspace.

2. RESEARCH METHODOLOGY

2.1. Learning Psychological Patterns from Different Types of Records

To learn the underlying psychological patterns, we use deep learning-based models to obtain text, voice, and visual representations from the networks. We then design a prompting component for each channel to reconcile these AI models' prior knowledge and the specific psychological analysis task. In this way, the learning process can be conducted more efficiently with less training burden. This is particularly helpful considering that psychological signals are sparsely labeled in most online social networks.

2.2. Improving the Psychological Analysis by Multi-modal Fusion

We enhance psychological analysis reliability by merging different modal data, focusing on consistency, and addressing conflicts between modalities. We identify shared features across data by finding the largest overlapping subspace in their latent spaces and learning private features by identifying informative components, allowing for an integrated multi-modal data approach.

2.3. Modeling Social Impact and Online Environment for User Psychological Traits

We use hypergraphs to model social interactions and environments online, which can connect multiple nodes with one edge, mirroring real-world social groups. This approach enables us to study psychological transitions

and evolving environments, crucial for the early detection of depressive disorders and for aiding users in adjusting their self-perception for better decision-making in real-life social settings.

2.4. An Intelligent System for Self-service Psychological Analysis

We develop a self-service psychological analysis system, which can function independently or as a plugin for services like online education or psychotherapy. This system analyzes users' public information, generating psychological reports. Users can interact with the system to receive personalized advice, promoting immediate and improved psychological well-being.

3. RESULTS ACHIEVED SO FAR

3.1. Hypergraph Clustering

We designed a novel random hypergraph model, HEM, which simplifies the existing random hypergraph model while preserving the essential features of real hypergraphs. The simplification paves the way to an efficient modularity computation for quality clustering. On top of it, we proposed a new hypergraph clustering algorithm PIC. Our method substantially outperformed existing methods in terms of clustering quality, and achieved up to five orders of magnitude faster clustering time. It provides a useful tool for analyzing complex social structures for the analysis of social influence in personality analysis. This work has been published in Proceedings of the ACM on Management of Data 2023 (J[1]).

3.2. Graph Neural Network Prompting

We designed novel multi-task prompting techniques for graph neural networks to bridge the gap between general graph knowledge and specific application needs. By integrating NLP-inspired prompting mechanisms, we can enhance the adaptability of graph models to various tasks without the need for extensive re-training. Specifically, we first unify the format of graph prompts and language prompts with the prompt token, token structure, and inserting pattern. Then, to further narrow the gap between various graph tasks and state-of-the-art pre-training strategies, we study the task space of various graph applications and reformulate downstream problems to the graph-level task. Afterward, we introduce meta-learning to efficiently learn a better initialization for the multi-task prompt of graphs so that our prompting framework can be more reliable and general for different tasks. This approach is instrumental in setting new standards for graph analysis, particularly in how models are pre-trained and fine-tuned for diverse applications. The prompting techniques can be used to abstract important psychological knowledge from complex social relations in our project. This work has been published at KDD2023 Conference (C[1]) and received the Best Paper Award (Research Track).

3.3. Adversarial Reinforcement Learning

Our pioneering work in adversarial reinforcement learning for scoring systems introduced robust methodologies for evaluating and enhancing the scoring mechanisms. We proposed a “counter-empirical attacking” mechanism that can generate “attacking” behavior traces and try to break the empirical rules of the scoring system. Then an adversarial “enhancer” is applied to evaluate the scoring system and find the improvement strategy. By training the adversarial learning problem, a proper scoring function can be learned to be robust to the attacking activity traces that are trying to violate the empirical criteria. The “counter-empirical attacking” mechanism allows for the dynamic testing and improvement of scoring systems, making them more resilient against potential manipulations and better suited for real-world applications. Our approach has been proved to be effective across different platforms such as the financial and resource management systems, demonstrating its versatility and impact. Based on this technique, we can score the mental status for an online user and give a quantitative analysis. This work has been accepted by IEEE Transactions on Knowledge and Data Engineering (TKDE) (J[2]) and will appear soon.

3.4 Survey on Using LLMs in Graph Learning Tasks

Recently, Large Language Models (LLMs) have been leveraged in graph learning tasks to surpass traditional Graph Neural Networks (GNNs) based methods and yield state-of-the-art performance. In this survey, we presented a comprehensive review and analysis of existing methods that integrate LLMs with graphs. First of all, we proposed a new taxonomy, which organizes existing methods into three categories based on the role

(i.e., enhancer, predictor, and alignment component) played by LLMs in graph-related tasks. Then we systematically surveyed the representative methods along the three categories of the taxonomy. Finally, we discussed the remaining limitations of existing studies and highlight promising avenues for future research. This work has been accepted by IJCAI 2024 Conference (C[2]).

4. PUBLICATIONS AND AWARDS

4.1. Publications

J[1] Zijin Feng, Miao Qiao, Hong Cheng. "Modularity-based Hypergraph Clustering: Random Hypergraph Model, Hyperedge-cluster Relation, and Computation," Proceedings of the ACM on Management of Data, 1(3), 1-25, 2023.

C[1] Xiangguo Sun, Hong Cheng, Jia Li, Bo Liu, Jihong Guan. "All in One: Multi-task Prompting for Graph Neural Networks," KDD, ACM, Long Beach, pp. 2120-2131, 2023.

J[2]. Xiangguo Sun, Hong Cheng, Hang Dong, Bo Qiao, Si Qin, Qingwei Lin. "Counter-Empirical Attacking based on Adversarial Reinforcement Learning for Time-Relevant Scoring System," Accepted by IEEE Transactions on Knowledge and Data Engineering, 2024.

C[2] Yuhan Li, Zhixun Li, Peisong Wang, Jia Li, Xiangguo Sun, Hong Cheng, Jeffrey Xu Yu. "A Survey of Graph Meets Large Language Model: Progress and Future Directions," IJCAI, Jeju, 2024.

4.2. Awards

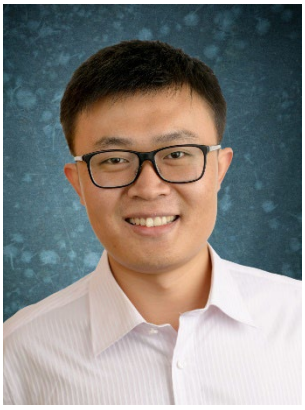
C[1] received the **Best Paper Award (Research Track)** at KDD 2023 Conference. This is the first time in Hong Kong and Mainland China to receive this award (**News Release**).

COMPLETED

INTELLIGENT MOBILE ROBOT-AUGMENTED SELF-SUPERVISED WI-FI SENSING

Principal Investigator: Professor He CHEN
Department of Information Engineering, CUHK

Co-Investigator (if any):
Dr. WU Chenshu, Assistant Professor [(1)]



Research Team Members:
- Dr. ZHANG Rui, Postdoctoral Research Fellow (1)
- Ms. WANG Qijia, Research Assistant (1)
- Ms. HUANG Bingyuan, Research Assistant (1)
- Ms. HE, Jing PhD student (1)
- Ms. KONG Ruiqi, PhD student (1)
- Mr. XU Ke, PhD student (1)
- Mr. WANG Shengqian, PhD student (1)

[(1)] Dept. of information Engineering, CUHK
[(2)] Dept. of Computer Science, HKU

Project Start Date: 1st July 2022
Completion Date: 30th June 2024

INNOVATION AND PRACTICAL SIGNIFICANCE:

Intelligent robots are entering our daily lives from laboratories and playing increasing roles in housework, home security monitoring, and indoor entertainment. These robots are typically equipped with Wi-Fi modules for connecting them to the Internet. Meanwhile, Wi-Fi sensing technologies are becoming more mature with the fast developments in the past ten years. The proposed marriage between intelligent mobile robots and Wi-Fi sensing has great potential to accelerate the broad applications of both technologies in homes for boosting people's safety and security. On the one hand, Wi-Fi sensing-augmented mobile robots could become more favorable thanks to the added value brought by the existing Wi-Fi module. On the other hand, the mobile capability of intelligent robots serves as an effective means to address Wi-Fi sensing's hurdle of complicated environment-dependent algorithm optimization for different homes, essential to make Wi-Fi sensing ubiquitous. Nevertheless, such a marriage is not technically straightforward. Instead, the marriage presents new technical challenges and calls for the design and validation of new signal processing and sensing algorithms.

To realize such a promising marriage between intelligent mobile robots and Wi-Fi sensing, we will develop a clean slate mobile robot-based self-supervised Wi-Fi sensing framework in this project. The proposed framework will leverage intelligent robot's onboard sensors and mobile capability to double-check the events detected by Wi- Fi sensing algorithms. The proposed algorithms can perform continuous and automated adaptation in a self- supervised manner towards higher and higher detection accuracy in homes with different environments. Furthermore, we will tackle a brand new and challenging problem unique to mobile robot-based Wi-Fi sensing– –sensing while moving. Specifically, the robot will be enabled to remove the impact of its self-movement to Wi- Fi signals before detecting other targeted movements.

Our preliminary research shows that the off-the-shelf Wi-Fi devices can detect indoor motions, and the optimization of Wi-Fi sensing algorithms for different environments can be self-supervised by fusing visual data. The proposed research, if successfully executed, will push the practicality of Wi-Fi sensing technologies to new heights, and accelerate the wide adaptation of intelligent mobile robots in homes, thereby boosting the safety and security of people whole live alone, particularly older adults during the pandemic.

ABSTRACT

This project aims to develop and validate a clean slate self-supervised Wi-Fi sensing framework through building prototypes. For the first time, our framework proposes to employ the increasingly popular in-home intelligent mobile robots for addressing a pain point of state-of-the-art Wi-Fi sensing technology– Sensing algorithms need to be manually calibrated for different environments (e.g., homes) to guarantee the detection accuracy. Wi-Fi sensing technologies, which leverages advanced machine learning (ML) algorithms to analyze and interpret Wi- Fi signals for the realization of human activity recognition (e.g., falls of older adults), health monitoring, and object detection, have advanced substantially in the past ten years and have been commercialized recently. However, today's Wi-Fi sensing systems, due to their environment-dependent nature, need to be carefully tuned for different homes to ensure their detection accuracy. Such tuning can be complicated and require frequent human interventions, hindering the broad penetration of Wi-Fi sensing. To circumvent the pain point, this project will devise new algorithms to leverage the onboard smart sensors (e.g., camera) and mobile capabilities of in-home intelligent robots to realize self-supervised optimization of Wi-Fi sensing algorithms with almost no human intervention. For example, a robot, which detects a human presence (i.e., a potential intrusion) in an empty home from its Wi-Fi sensing module, will move to the suspicious area to confirm it through other onboard sensors before raising the alarm. This project extends a seed project funded by the Worldwide Universities Network.

1. OBJECTIVES AND SIGNIFICANCE

Intelligent robots are entering our daily lives from laboratories and playing increasing roles in housework, home security monitoring, and indoor entertainment. These robots are typically equipped with Wi-Fi modules for connecting them to the Internet. Meanwhile, Wi-Fi sensing technologies are becoming more mature with the fast developments in the past ten years. The proposed marriage between intelligent mobile robots and Wi-Fi sensing has great potential to accelerate the broad applications of both technologies in homes for boosting people's safety and security. On the one hand, Wi-Fi sensing-augmented mobile robots could become more favorable thanks to the added value brought by the existing Wi-Fi module. On the other hand, the mobile capability of intelligent robots serves as an effective means to address Wi-Fi sensing's hurdle of complicated environment-dependent algorithm optimization for different homes, essential to make Wi-Fi sensing ubiquitous. Nevertheless, such a marriage is not technically straightforward. Instead, the marriage presents new technical challenges and calls for the design and validation of new signal processing and sensing algorithms.

To realize such a promising marriage between intelligent mobile robots and Wi-Fi sensing, we will design and prototype an intelligent mobile robot-based self-supervised Wi-Fi sensing system. More specifically, the project has the following three interconnected objectives:

1. To develop intelligent mobile robot-based self-supervised Wi-Fi sensing algorithms, which enable the robot to detect and localize indoor motions via analyzing Wi-Fi signals, confirm detected anomalies through other onboard sensors like cameras, and automatically calibrate the underlying Wi-Fi sensing model for different environments in a self-supervised manner.
2. To devise more robust Wi-Fi sensing algorithms that can detect human motions even when the robot is moving.
3. To design and prototype a self-supervised Wi-Fi sensing system based on the TurtleBot2 platform available in PI's lab to validate and demonstrate the effectiveness and robustness of the proposed algorithms.

This project will demonstrate a novel union between intelligent in-home mobile robots and Wi-Fi sensing. To the best of our knowledge, the proposed union will be first-of-its-kind and thus has great potential to open a new research line in the field. Furthermore, we believe this is a "win-win" union that will accelerate the widespread penetration of both technologies into every household, as Wi-Fi networks have achieved so far. The proposed research, if successfully executed, will push the practicality of Wi-Fi sensing technologies to new heights, and accelerate the wide adaptation of intelligent mobile robots in homes, thereby boosting the safety and security of people whole live alone, particularly older adults during the pandemic.

2. RESEARCH METHODOLOGY

To circumvent the identified problems in the state-of-the-art Wi-Fi sensing systems, we will develop an interdisciplinary approach to automate the anomaly confirmation process in this project. Specifically, increasingly popular intelligent in-home robots will be employed to execute anomaly confirmation through leveraging their mobility and onboard sensors such as cameras. Our approach will enable a self-supervised calibration of wireless sensing algorithms for different environments with almost no human intervention. We will adopt the following research plan and methodology to accomplish the project if funded:

To conduct research for delivering Objective 1: The first objective is to develop intelligent mobile robot-based self-supervised Wi-Fi sensing algorithms, which facilitate the following functionalities of the robot: (1) detect and localize indoor motions, (2) confirm detected anomalies with onboard sensors, and (3) automatically calibrate the Wi-Fi sensing model for different environments in a self-supervised manner. We propose a two-step signal processing scheme to be computationally efficient for realizing the first functionality. In the first step, we estimate the target speed by evaluating the autocorrelation function of the channel state information (CSI) extracted from the Wi-Fi module. We then combine the speed values extracted from each subcarrier and receiving antenna with the maximum ratio combining (MRC) strategy. After that, we will apply a recurrent neural network model (RNN) with long short-term memory (LSTM) blocks to exploit the temporal relationship of target speed and thus detect the suspicious anomalies. Once an anomaly is detected, in the second step, we will execute the multiple signal classification (MUSIC) algorithm to estimate the angle-of-arrival (AoA) of incoming signals incurred by the suspicious motion. We will integrate open-source visual processing algorithms for realizing the second functionality in our framework for executing anomaly confirmations. When it comes to the third functionality, after each anomaly confirmation, the corresponding temporal speed sequence and anomaly confirmation result will self-supervise the training of RNN. Built on the initial training of the RNN that can be done based on standard templates generated by band-relaxed segmental local normalized dynamic time warping (SLN-DTW) and DTW barycenter averaging (DBA) algorithms, the proposed anomaly confirmation result can automatically calibrate the RNN- based Wi-Fi sensing model efficiently.

To conduct research for delivering Objective 2: Our framework raises a brand-new technical problem for Wi-Fi sensing—sensing while moving. This distinguishes our work from previous designs of Wi-Fi sensing systems, where the Wi-Fi signal transceivers are considered to stay in a fixed location. Any motion of the transceiver will make the information extraction much more challenging. This is because Wi-Fi sensing relies on extracting information from the signals reflected by the target (e.g., human), and reflections are weak. Nevertheless, in our project, the in-home robot can be a platform for multiple tasks, thus cannot be static for most time. Furthermore, when a suspicious motion is detected, and the robot is on the way to confirm it, the sensing functionality should be maintained in case another abnormal event occurs. We thus need to develop new sensing algorithms that enable the continuous monitoring capability for the intelligent robot. To that end, inspired by the successive interference cancellation (SIC) technique, which is used in wireless communications and allows a receiver to decode two or more data packets that arrived simultaneously, we will develop a SIC-alike Wi-Fi sensing algorithm that allows an intelligent robot to detect a potential human motion masked by its self-movement. More specifically, by leveraging the onboard sensor like IMU, the robot will be able to estimate its current velocity, carefully characterize the effects of its self-movement to the received Wi-Fi signals, and then remove the effects before detecting other targeted movements. We will design and compare both model-based and model-free approaches to characterize the impacts of self-movement.

To conduct research for delivering Objective 3: Built on the mobile robotic platform established in the PI's lab, the project team will design and implement a mobile robot-based self-supervised Wi-Fi sensing system prototype to evaluate and showcase the effectiveness and robustness of the proposed algorithms in Objectives 1-2 in real-world environments. To that end, the team will substantially modify the Linux

kernel of the mini-PC that serves as the “brain” of the mobile robot, so that it will output the CSI of each received Wi-Fi packet to the application layer. After that, we will implement the proposed algorithms on the application layer and build a tunnel between our application with the Robot Operating System (ROS) so that the outputs of our algorithms can be used to control the robots' movements.

3. RESULTS ACHIEVED

3.1 Research Results

At the time of writing this report, we have made significant progress in detecting human motion on mobile robots using Wi-Fi signals. The system we built allows a moving robot to detect the presence of an individual behind a corner and estimate their movement directions, effectively preventing potential collisions. This effort represents the first passive Wi-Fi detection system specifically designed for mobile robots, which does not require direct line-of-sight observation of the target. This technology can be applied to various applications, such as fall detection. By identifying anomalies beyond its field of view, the robot can respond to a diverse range of human actions in complex indoor environments and proactively approach suspicious areas when further investigation is necessary. A photo of the developed experimental platform is shown at the bottom of previous page.

The mobile robot detects a person behind the corner by estimating AoA of received Wi-Fi signals using the MUSIC algorithm. The presence of a person introduces a new value to the AoA estimation result compared to a scenario without a person. However, the AoA estimation results are influenced by both human and robot movements, making it challenging to determine the human's movement direction. The challenge of predicting potential collisions as the robot moves towards the corner lies in isolating human motion data from the received Wi-Fi signals. We employ principal component analysis (PCA) on the CSI extracted from the Wi-Fi module to identify the most stable amplitude path, designating it as the reference path consisting only of the robot's motion. Drawing inspiration from the SIC technique, we develop a robot movement removal algorithm that filters the robot's motion from the entire CSI dataset, resulting in CSI data containing solely human motion information. We then use the mean of the correlation coefficients between all adjacent subcarriers in the CSI dataset to revel the person movement directions. Specifically, the correlation coefficients reflect the fluctuation of the path through which signals propagate from the transmitter to the receiver. It increases as the propagation length shortens and decrease as the path length increases.1 We conducted experiments to compare the correlation coefficients of raw CSI data and the data with robot self-motion removal in three scenarios: no person behind the corner, a person moving towards the corner, and a person moving away from the corner. The mobile robot moves towards the corner from another direction in all three scenarios and thus has no line-of-sight view of the person. Our results revealed that human movements are obscured by the robot's motion in the raw data, causing the correlation coefficient to increase in all three scenarios. After implementing our proposed algorithm, the correlation coefficient remains constant when no person is present, increases as the person approaches the corner, and decreases when the person moves away, confirming the robot's passive sensing capabilities while in motion. We have written a paper based on this research work and submitted it to IEEE INFOCOM 2025 for possible publication in July 2024.

In addition to passive Wi-Fi sensing on mobile robots, we are tackling the project objectives from three other angles. Firstly, we have developed a new framework to estimate multipath parameters, facilitating more refined Wi-Fi sensing applications. We observed that due to the limited bandwidth of practical wireless systems, a single multipath component may appear as a discrete pulse comprising multiple taps in the digital delay domain. This phenomenon, known as channel leakage, complicates multipath delay estimation and has been largely overlooked in existing Wi-Fi sensing research. To address this issue, we exploit the limited number of paths in physical environments and frame the estimation problem as a sparse recovery challenge. We propose using a sparse Bayesian learning (SBL) method to estimate the sparse vector, effectively determining both the number of physical paths and their associated delay parameters. Our simulation results indicate that our algorithm accurately identifies path numbers and achieves superior precision in path delay estimation and channel reconstruction compared to two benchmarking schemes. The line of work has led to two conference papers [2], [4] and one journal paper [3]. Built on top of these research results, we further develop a radio-

based passive target tracking algorithm using multipath measurements, including the angle of arrival and relative distance. We focus on a scenario in which a mobile receiver (e.g., a mobile robot) continuously receives radio signals of opportunity from a transmitter located at an unknown position. The receiver utilizes multipath measurements extracted from the received signal to jointly localize the transmitter and the scatterers over time, with scatterers comprising a moving target and stationary objects that can reflect signals within the environment. We develop a comprehensive probabilistic model for the target tracking problem, incorporating the localization of the transmitter and scatterers, the identification of false alarms and missed detections in the measurements, and the association between scatterers and measurements. We employ a belief propagation approach to infer the posterior distributions of the positions of the scatterers and the transmitter. Additionally, we introduce a particle implementation for the belief propagation method. Simulation results demonstrate that our proposed algorithm outperforms existing benchmark methods in terms of target tracking accuracy. We have drafted a journal paper [5] and will further polish it before submitting it for possible publication.

Secondly, we explored the technical feasibility of recognizing robot identification through their transmitted WiFi signals, similar to how humans are identified through their voices. Along this line, we introduced CSI-RFF, a new framework that uses micro-signals present in Channel State Information (CSI) to facilitate radio-frequency fingerprinting of commodity off-the-shelf (COTS) WiFi devices for open-set authentication. These micro-signals, termed 'micro-CSI,' primarily arise from imperfections in the RF circuitry and are detectable across WiFi 4/5/6 network interface cards (NICs). The challenge in leveraging micro-CSI for authentication stems from its entanglement with distortions introduced by wireless channels, referred to as true CSI. This complex interplay makes separating these components non-trivial. To address this, we have developed a signal space-based extraction technique targeted for Line-of-Sight (LoS) scenarios, which effectively isolates the channel distortions from micro-CSI. Throughout an extensive data collection period spanning over a year, we observed that micro-CSI exhibits unique, device-specific characteristics that remain consistent over time, making it a robust candidate for device fingerprinting. Our experimental findings reveal that the micro-CSI-based authentication algorithm achieves an impressive average attack detection rate of close to 99% with a false alarm rate of 0% among 19 NICs, under both static and mobile conditions, using just 20 CSI measurements per fingerprint. We also explored how we can extend the CSI-RFF framework to non-line-of-sight (NLoS) conditions using deep learning techniques. This line of research has led to two conference papers [6], [8] and one journal paper [7].
Thirdly, besides WiFi sensing for robots, our team has also been investigating new visual sensing techniques for cooperative mobile robots, which loosely connects to the anomaly confirmation part of this project. In this line of research, we propose a new optical ISAC (OISAC) scheme for cooperative mobile robots by integrating camera sensing and screen-camera communication (SCC). As a case study, we consider the leader- follower formation control problem, an essential part of cooperative mobile robotics. The proposed OISAC scheme enables the follower robot to simultaneously acquire the information shared by the leader and sense the relative pose to the leader using only RGB images captured by its onboard camera. We design and conduct real-world experiments involving uniform and non-uniform motions to evaluate the proposed system and demonstrate the advantages of applying OISAC over a benchmark approach that uses extended Kalman filtering (EKF) to estimate the leader's states. Our results show that the proposed OISAC-augmented leader- follower formation system achieves better performance in terms of accuracy, stability, and robustness. This line of research has led to one conference paper [9].

3.2 Patent application and/or product commercialization plan

The results obtained during this project have enabled the PI to apply for an RGC grant (funded) and an ITF seed grant (invited for interview but not funded). Additionally, they helped the PI secure a research donation from Huawei, which will support further research.

The PI presented the core concept and demonstrated the experimental results to Computime Group Limited, a Hong Kong Stock Exchange-listed smart home/building company. They appreciated our technology, which integrates in-home intelligent robots and Wi-Fi sensing, and showed strong interest. The PI is working to secure their sponsorship for larger grant applications, such as the ITF platform project. This support would allow us to file patents and advance the prototype towards commercialization.

4. PUBLICATION AND AWARDS

During this project, we produced the following publications. PDF copies have been submitted with this report.

[1] J. He, R. Zhang, and H. Chen, “WiseBot: WiFi-based Passive Human Proximity Sensing at Corners for Mobile Robots,” submitted to IEEE INFOCOM 2025, July 2024.

[2] K. Xu, H. Chen and C. Wu, "Pulse Shape-Aided Multipath Delay Estimation for Fine-Grained WiFi Sensing," 2023 IEEE 24th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), Shanghai, China, 2023, pp. 181-185.

[3] K. Xu, R. Zhang and H. Chen, "Pulse Shape-Aided Multipath Parameter Estimation for Fine-Grained WiFi Sensing," in IEEE Transactions on Communications, vol. 72, no. 10, pp. 6116-6130, Oct. 2024.

[4] H. Hu, R. Kong, K. Xu and H. H. Chen, "Deep Learning-Based Pulse-Shaping Filter Estimation for Fine-Grained WiFi Sensing," ICC 2024 - IEEE International Conference on Communications, Denver, CO, USA, 2024, pp. 4421-4426.

[5] K. Xu, R. Zhang and H. Chen, “Radio-Based Passive Target Tracking by a Mobile Receiver with Unknown Transmitter Position," under preparation for possible journal submission, Oct. 2024.

[6] R. Kong and H. Chen, "Physical-Layer Authentication of Commodity Wi-Fi Devices via Micro-Signals on CSI Curves," 2023 IEEE 24th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), Shanghai, China, 2023, pp. 486-490.

[7] R. Kong and H. Chen, "CSI-RFF: Leveraging Micro-Signals on CSI for RF Fingerprinting of Commodity WiFi," in IEEE Transactions on Information Forensics and Security, vol. 19, pp. 5301-5315, 2024.

[8] R. Kong and H. Chen, "Towards Channel-Resilient CSI-Based RF Fingerprinting using Deep Learning," IEEE INFOCOM 2024 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), Vancouver, BC, Canada, 2024, pp. 1-6.

[9] S. Wang and H. Chen, "Optical Integrated Sensing and Communication for Cooperative Mobile Robotics: Design and Experiments," 2023 21st International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt), Singapore, Singapore, 2023, pp. 207-214.

Commercialization Endeavors

1. Segmented High-Entropy Thermoelectric Materials for Geothermal Heat Harvesting
by **Professor Ady SUWARDI** *(page 22)*

Professor Ady SUWARDI is planning to develop an optimized processing conditions to achieve a high average zT of >1.5 over a temperature range of 50 – 200 °C (Figure 1c-d). This phase of the project will build upon the previous phase by directly taking the optimized small batch melted sample and further process it by zone-melting.

The successful completion of the project will pave the way for the production of high performance segmented high-entropy materials that can be obtained with minor adjustments to commercial manufacturing processes. As a proof of concept, we also aim to optimize the assembly of the materials to make a unicouple TE device prototype with a power conversion efficiency of >10%. This will involve tuning the unicouple geometry and electrode materials to maximize heat transfer and minimize contact resistance.

2. Manufacturing Metasurface based All-optical Cnn for Real-time and Power-efficient Machine Vision
by **Professor Chaoran HUANG** *(page 38)*

Prof Chaoran HUANG's research work has applied for US patent with a Serial No. 63/643,973. The team has developed a metasurface-based optical neural network (meta-ONN) that rivals deep digital networks across multiple machine vision tasks using the same physical chip, with only a lightweight digital backend trained per task. It achieves performance comparable to large-scale models such as ResNet-50, Vision Transformer (ViT), and the Segment Anything Model (SAM), while drastically reducing computational load and energy consumption. In a clinically significant application, the system was deployed for breast cancer metastasis detection in gigapixel whole-slide images, achieving high accuracy (AUC up to 97.0%) and extreme acceleration-processing an entire slide in just 1.02 seconds compared to SAM's 1.48 hours- enabling scalable high-throughput diagnostic support.

3. Cost-efficient Highly Potent Antimicrobial Peptide Discovery for Livestock Farming Antibiotic Alternatives with Protein Language Model-powered Ai Methods
by **Professor Yu LI** *(page 48)*

Prof Yu LI has applied for a PCT patent (Publication No. WO 2024/169915 A1) entitled "Machine Learning Pipeline for Discovering Novel Antimicrobial Peptides," covering both the AMP discovery pipeline and the novel AMPs identified. The research involved constructing a mammalian digestive tract microbial genome database (MDTMGD) and retraining the ESM-2 protein language model to develop HMD-AMP—a hierarchical multi-label deep forest framework that achieves state-of-the-art performance in predicting evolutionarily remote AMPs and their antimicrobial spectra. From over 20 million candidate sequences, the team prioritized porcine and gut microbiota-derived AMPs by integrating metaproteomic validation, leading to the selection of 62 high-scoring peptides for synthesis. Among these, 52 showed strong antibacterial activity

against major porcine pathogens, with 8 exhibiting broad-spectrum efficacy comparable to conventional antibiotics and minimal cytotoxicity. These results have been submitted for peer review in Nature Biomedical Engineering.

4. Development of Mitochondria-Targeting, Single-Atom Nanozyme for Accelerated Bone Regeneration
by **Professor Zhong LI, Alan** *(page 53)*

Professor Zhong Li is going to advance the commercialization of a novel bone regeneration technology that employs a multifunctional nanozyme to simultaneously scavenge mitochondrial ROS and restore OXPHOS, effectively addressing oxidative stress and metabolic dysregulation in bone defects. Its precise mitochondria-targeting capability enables localized repair without systemic toxicity, offering advantages over conventional single-mechanism approaches. The technology also holds promise for treating other bone conditions like osteoporosis. Utilizing a cost-effective and scalable synthesis process, the team is preparing a patent application and seeking grants to support large-animal studies for further validation, paving the way for clinical translation and technology transfer.

5. Development of Implicit-based Neuron-Driven Computer-aided Manufacturing (Cam) Kernel for Multi-axis Smart Manufacturing
by **Professor Guoxin FANG** *(page 88)*

Professor Guoxin FANG is planning to commercialize a CAM kernel that establishes a new computational foundation for smart manufacturing. This end-to-end solution enables seamless automation of complex fabrication workflows with minimal manual intervention. The kernel will be launched as a standalone software package and integrated with advanced multi-axis manufacturing equipment through industry partnerships. By fostering AI-driven design-fabrication optimization, it aims to transform sectors such as aerospace (e.g., lightweight energy-saving components), healthcare (e.g., patient-specific implants), and construction (e.g., energy-efficient structures). Ultimately, this technology strives to build a more sustainable and equitable future while strengthening global leadership in smart manufacturing.

6. Intelligent Mobile Robot-Augmented Self-supervised Wi-Fi Sensing
by **Professor He CHEN** *(page 104)*

Professor He CHEN has successfully secured an RGC grant and was invited to interview for an ITF seed grant based on the project outcomes. He also obtained a research donation from Huawei to further advance the research. Professor CHEN presented the core technology- integrating in-home intelligent robots with Wi-Fi sensing- to Computime Group Limited, a listed smart home company, which expressed strong interest and potential sponsorship support for larger grants such as the ITF platform project. These partnerships will facilitate patent filings and prototype development toward commercialization.

Shun Hing Institute of Advanced Engineering

Distinguished Lecture Series 2025

Challenges and Progress in Automatic Speech-to-Speech Translation: Bridging the Gap to Real-Time Interpretation

by

Professor Satoshi Nakamura

The Chinese University of Hong Kong, Shenzhen

Date: 31 March 2025 (Monday)
Time: 10:30 a.m. – 12:00 noon
Venue: Room 222, 2/F, Ho Sin Hang Engineering Building (SHB), CUHK

Abstract

Automatic speech-to-speech translation has long been a dream technology for humanity. Through numerous breakthroughs from years of research, we have now reached the stage where this service can be used on smartphones. However, there are still many challenges remaining before we can achieve translation quality comparable to that of a trained simultaneous interpreter. Key issues include how to translate across languages with different word orders, such as between English and Japanese, without waiting for the end of a sentence or utterance; how to balance latency and content fidelity in translation; and how to extract the speaker's intent from their intonation.
Having conducted research on speech translation over an extended period, I would like to reflect on some of the past progress and introduce ongoing research addressing these challenges.

Biography of the Speaker

Dr. Satoshi Nakamura is Presidential Chair Professor at School of Data Science, The Chinese University of Hong Kong, Shenzhen. He is a Professor Emeritus at the Nara Institute of Science and Technology (NAIST) and an Honorary Professor at the Karlsruhe Institute of Technology, Germany. He is an IEEE Fellow, ISCA Fellow, Information Processing Society of Japan Fellow, and Advanced Telecommunications Research Institute International (ATR) Fellow. He got his Ph.D. by dissertation from Kyoto University in 1992. He was Department Head, Director of Spoken Language Communication Research Laboratories, and Vice President of ATR from 2000 till 2008. From 2009 to 2010, he served as the Director General of Keihanna Research Laboratories and the Executive Director of Knowledge Creating Communication Research Center, National Institute of Information and Communications Technology. He moved to the Nara Institute of Science and Technology as a full professor in 2011. He established the Data Science Center at NAIST and served as a director from 2017 to 2021. He became a professor at the Chinese University of Hong Kong Shenzhen in 2024. His research interests include modeling and systems of spoken language processing, speech processing, spoken language translation, spoken dialog systems, natural language processing, and data science. He is one of the world leaders in speech-to-speech translation research. He has been serving for various speech-to-speech translation research projects, including C-Star, A-Star, and International Workshop on Spoken Language Translation IWSLT. He is chairperson of the International Speech Communication Association Special Interest Group: Spoken Language Translation. He also contributed to standardizing the network-based speech translation at the International Telecommunication Union. He received the Antonio Zampolli Prize in 2012.

* * * * * * * * * **ALL ARE WELCOME** * * * * * * * * *

Shun Hing Institute of Advanced Engineering (SHIAE)
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong

Office : Room 702, William M.W. Mong Engineering Building
Tel : (852) 3943 4351
Fax : (852) 3943 4354
Email : applications@shiae.cuhk.edu.hk
Website : http://www.shiae.cuhk.edu.hk