Zhaoxiang (Simon) Cai

PhD (Cancer Data Science), MBusA (First Class Hons), BCompSc (First Class Hons)
Conjoint Associate Lecturer, University of Sydney
Senior Data Scientist (Cancer Data Science), Children's Medical Research Institute
+61 421 992 704 | scai@cmri.org.au
https://zhaoxiangsimoncai.github.io/SimonCaiCV.io/

Career Profile

I am a cancer data scientist specialising in artificial intelligence and machine learning for large-scale molecular data analysis. I hold a first-class honours degree in Computer Science from Monash University, where I graduated as the top student, and a PhD in Cancer Data Science from the University of Sydney. After gaining software engineering experience at Goldman Sachs, I transitioned to cancer research and now serve as a Conjoint Associate Lecturer at the University of Sydney and Senior Data Scientist at the Children's Medical Research Institute (CMRI).

My research focuses on developing and applying deep learning methods to cancer multi-omics data, with particular expertise in federated learning, generative models, and transformer architectures. As co-first author on the landmark Cancer Cell study mapping the proteomic landscape of 949 human cancer cell lines, I contributed to one of the field's most widely used resources (230+ citations). My subsequent first-author publications in Cancer Discovery and Nature Communications have introduced novel approaches in federated deep learning for privacy-preserving cancer subtyping, and generative AI for synthetic augmentation of multi-omic datasets. I maintain active international collaborations with the Wellcome Sanger Institute (UK) and the University of Lisbon (Portugal).

In 2025, I was awarded a Cancer Institute NSW Early Career Fellowship ($597,732 over three years) to develop proteomic foundation models integrated with federated learning for multi-hospital cancer research. My work to date has generated over 825 citations (h-index: 7) and I have contributed to more than $1.59 million in competitive research funding. I am driven by the goal of translating advanced computational methods into practical tools that improve cancer diagnosis and prognosis through precision medicine.

Citation Metrics

Citation metrics obtained from Google Scholar. Field-Weighted Citation Impact (FWCI) from Scopus/SciVal.

7	7	825	5 of 9
h-index	i10-index	Total Citations	First/Co-first Author

Qualifications

Doctor of Philosophy (Cancer Data Science) — *University of Sydney / Children's Medical Research Institute*
March 2020 – February 2023
Thesis: *Large-Scale and Pan-Cancer Proteogenomic Analyses with Machine Learning*
Sydney Cancer Partners PhD Scholarship recipient

Master of Business Analytics (First Class Honours) — *Melbourne Business School, University of Melbourne*
January 2018 – December 2018

KPMG-MBS Data Challenge: 1st Prize Winner (Natural Language Processing)
MBS Scholarship in Business Analytics; Co-President of Business Analytics Club

Bachelor of Computer Science (First Class Honours) — *Monash University*
January 2011 – July 2014
Dux of Bachelor of Computer Science (highest overall ranking in graduating cohort)
Bellamy Awards (top student in each academic year, 2011 and 2012)
International Merit Scholarship (2011–2013)

Professional Training

Higher Degree Research Student Supervisor Training Course, University of Sydney (2024)

Professional Appointments

Conjoint Associate Lecturer, Faculty of Medicine and Health, University of Sydney
June 2023 – present

Senior Data Scientist (Cancer Data Science), Children's Medical Research Institute, Westmead
February 2023 – present

PhD Candidate (Cancer Data Science), Children's Medical Research Institute / University of Sydney
March 2020 – February 2023

Data Scientist / Bioinformatician, Children's Medical Research Institute, Westmead
January 2019 – March 2020

Analyst Programmer, Goldman Sachs, Melbourne
November 2014 – January 2018
Electronic Trading (GSET) division. Delivered Shenzhen-Hong Kong Stock Connect project (Federation Award 2016). Established India GSET clearing clients business flow.

Grants and Fellowships

Fellowships

Cancer Institute NSW Early Career Fellowship (2026–2028; $597,732)
Role: Sole Chief Investigator
Improving Cancer Diagnosis and Prognosis via Generative Proteomic Foundation Model and Federated Learning
Awarded 2025. This competitive fellowship supports a three-year programme developing proteomic foundation models integrated with federated learning for multi-hospital cancer research collaboration without sharing sensitive patient data.

Research Grants

Medical Research Future Fund (MRFF) Grant (2024–2026; $993,500)
Role: Associate Investigator
Targeting the Dysregulated Epigenome to Enhance Immunotherapy Response
Multi-institutional collaboration between Hudson Institute of Medical Research, CMRI, and Children's Cancer Institute. Contribution: computational biology expertise in multi-omic analysis and artificial intelligence.

Total competitive research funding contributed to: over $1.59 million

Awards and Honours

Sydney Cancer Partners PhD Scholarship (2022)
CMRI Peter Rowe PhD Scholarship (2020)
University of Sydney Innovation Challenge: 1st Prize (2019, 2020; $7,500 each)
3rd Prize, ODIR-2019 International Computer Vision Competition (approx. $20,000)
Dux of Bachelor of Computer Science, Monash University (2013)
International Merit Scholarship, Monash University (2011–2013)
Bellamy Award (Top Student), Monash University (2011, 2012)

Publications

Peer-Reviewed Journal Articles

Author name in **bold underline**. * denotes equal contribution. FWCI = Field-Weighted Citation Impact (Scopus/SciVal); FWCI > 1.0 indicates above world average.

1. **Cai, Z.**, Boys, E. L., Noor, Z., Aref, A. T., Xavier, D., Lucas, N., Williams, S. G., Koh, J. M., Poulos, R. C., Wu, Y., Dausmann, M., MacKenzie, K. L., Aguilar-Mahecha, A., Armengol, C., Barranco, M. M., Basik, M., Bowman, E. D., Clifton-Bligh, R., Connolly, E. A., ... Reddel, R. R. (2025). *Federated deep learning enables cancer subtyping by proteomics. Cancer Discovery*, OF1–OF16. **FWCI: 3.58.** [5 citations]

First author. IF 29.7. First application of federated deep learning to cancer proteomics, enabling model training across 7,500 cancer proteomes from multiple centres without sharing raw data.

2. Wu, Y., **Cai, Z.**, Cross, D., Noble, J. R., Prest, K., Littleboy, J., Cohen, S. B., Edlundh, B., Koh, J. M. S., Xu, R., Noor, Z., Bastami, M., Valentini, S., Richardson, L., Barthorpe, S., Arymanesh, N., Robinson, P. J., Hains, P. G., Garnett, M. J., ... MacKenzie, K. L. (2025). *Large-scale drug sensitivity, gene dependency, and proteogenomic analyses of telomere maintenance mechanisms in cancer cells. Nature Communications*, 16(1), 11337. **FWCI: N/A.** [1 citation]

Co-author. IF 14.7. Contributed computational analyses for proteogenomic characterisation of telomere maintenance mechanisms.

3. Baião, A. R., **Cai, Z.**, Poulos, R. C., Robinson, P. J., Reddel, R. R., Zhong, Q., ... & Gonçalves, E. (2025). *A technical review of multi-omics data integration methods: from classical statistical to deep generative approaches. Briefings in Bioinformatics*, 26(4), bbaf355. **FWCI: 6.98.** [77 citations]

Co-author. International collaboration with University of Lisbon.

4. Osteil, P., Withey, S., Santucci, N., Aryamanesh, N., Pang, I., Salehin, N., ... **Cai, Z.**, Wolvetang, E. & Tam, P. P. (2025). *MIXL1 activation in endoderm differentiation of human induced pluripotent stem cells. Stem Cell Reports*. **FWCI: 0.76.** [1 citation]

Co-author.

5. **Cai, Z.**, Apolinário, S., Baião, A. R., Pacini, C., Sousa, M. D., Vinga, S., ... & Gonçalves, E. (2024). *Synthetic augmentation of cancer cell line multi-omic datasets using unsupervised deep learning. Nature Communications*, 15(1), 1–12. **FWCI: 2.83.** [32 citations]

First author. IF 14.7. Introduces MOSA (Multi-Omics Synthetic Augmentation), a generative AI approach using multi-view variational autoencoders. International collaboration with University of Lisbon.

6. **Cai, Z.**, Poulos, R. C., Aref, A., Robinson, P. J., Reddel, R. R., & Zhong, Q. (2024). *DeePathNet: a transformer-based deep learning model integrating multi-omic data with cancer pathways. Cancer Research Communications*, 4(12), 3151–3164. **FWCI: 3.29.** [33 citations]

First author. Novel transformer-based architecture integrating multi-omic data with cancer pathway knowledge for improved interpretability and prediction accuracy.

7. Gonçalves, E.*, Poulos, R. C.*, **Cai, Z.***, ..., Robinson, P., Zhong, Q., Garnett, M., Reddel, R. R. (* equal contribution) (2022). *Pan-cancer proteomic map of 949 human cell lines. Cancer Cell*, 40(8), 835–849. **FWCI: 10.50.** [230 citations]

Co-first author. IF 50.3. The world's largest pan-cancer proteomic resource at time of publication. Collaboration with Wellcome Sanger Institute (Cambridge, UK). Featured on ABC Radio and institutional media.

8. **Cai, Z.**, Poulos, R. C., Liu, J., & Zhong, Q. (2022). *Machine learning for multi-omics data integration in cancer. iScience*, 103798. **FWCI: 5.73.** [288 citations]

First author. IF 5.8. Widely cited review for researchers entering the field of computational multi-omics integration.

9. Poulos, R. C., **Cai, Z.**, Robinson, P. J., Reddel, R. R., & Zhong, Q. (2022). *Opportunities for pharmacoproteomics in biomarker discovery. Proteomics*, 2200031. **FWCI: 2.39.** [27 citations]

Co-author. Invited Viewpoint article on pharmacoproteomics.

Teaching and Education Contributions

Higher Degree Research Supervision

Ms Fatemeh Mehdikhani (PhD candidate)
Commenced: July 2025 | Role: Co-supervisor
Faculty of Medicine and Health, University of Sydney
Non-Invasive Diagnosis and Classification of Nevi and Early-Stage Melanoma
Provides guidance on computational methodology, machine learning model development, and data analysis approaches.

Thesis Examination

Ms Rita Brito Gama (Master of Philosophy)
October 2025 | Role: External thesis examiner
University of Lisbon, Portugal
GAIN-DANN: A Domain-Adversarial Generative Model for Missing Data Imputation in Proteomics
Responsibilities included critical evaluation of the written thesis and participation in the oral defence. Ms Brito Gama successfully defended her thesis and was awarded her degree.

Mentorship of Researchers

Dr Emma Boys – Medical Oncologist and PhD candidate, CMRI. Guidance on machine learning methods for cancer classification research (Cancers of Unknown Primary). Co-authored publication in Cancer Discovery (2025).

Dr Liz Connolly – Medical Oncologist and PhD candidate, CMRI. Technical guidance on bioinformatic analysis, differential expression analysis, and survival analysis methods for prostate cancer research.

Dr Di Xiao – Junior researcher, Cancer Data Science team, CMRI. Support for integration into the team, guidance on computational platforms and analytical tools.

Formal Teaching Experience

Casual Tutor, Algorithms and Data Structures, Monash University (2013–2014)
Delivered tutorials in computational subjects supporting student learning in algorithm design and implementation.

Staff Supervision (Forthcoming)

Cancer Institute NSW Early Career Fellowship (2026–2028) includes funding for a junior postdoctoral position. Expected to serve as primary supervisor, providing direction on computational approaches to cancer proteomics and generative AI methods.

Conference Presentations and Invited Talks

Oral Presentations

Advancing Multi-Omics into the Clinic Symposium (2024)
70th ASMS Conference on Mass Spectrometry and Allied Topics (2022)
KCA Precision Medicine for Childhood Cancer Symposium (2021, **invited**)
Combined ABACBS and Phylomania Hybrid Conference (2021)

Poster Presentations

ABACBS Annual Conference (2023)
International Conference on Intelligent Systems for Molecular Biology (ISMB) (2023)
Human Proteome Organisation (HUPO) World Congress (2022)
ABACBS Annual Conference (2022)

Institutional Seminars

CMRI All Staff Seminar (2023): Presented research on computational approaches to cancer multi-omics data analysis to the broader institute community.

Media Engagement

ABC Radio interview on the Cancer Cell (2022) pan-cancer proteomic map publication, communicating the significance of the research to a public audience.

Service and Leadership

Internal Service

Leadership of Cancer Data Science Journal Club (2023–present)
Managed and hosted this monthly forum, bringing together 10–15 researchers from across the ProCan research initiative (Cancer Data Science, Software Engineering, and Oncology teams). Facilitates cross-team discussion, knowledge exchange, and critical evaluation of recent publications in cancer data science and machine learning.

Participation in CMRI Research Seminars and Scientific Meetings (2023–present)
Active engagement with presentations across diverse research areas, contributing to discussions that strengthen the collaborative research environment at CMRI.

Peer Review

Reviewed approximately five manuscripts per year since 2023, totalling over ten reviews. Journals include: *Nature Communications, Genome Biology, Briefings in Bioinformatics*, among others.

Grant Assessment

Innovation and Technology Commission (Hong Kong) (2025): Invited external expert reviewer for competitive national funding scheme.

Dutch Research Council (NWO, Netherlands) (2025): Invited external expert reviewer for competitive national funding scheme.

International and National Collaborations

Wellcome Sanger Institute (Cambridge, UK): Ongoing collaboration in cancer proteomics and drug screening. Resulted in the landmark Cancer Cell publication (2022) and the Nature Communications telomere maintenance study (2025).

University of Lisbon (Portugal): Ongoing collaboration in multi-omics data integration and generative AI methods. Produced the Nature Communications MOSA publication (2024) and the Briefings in Bioinformatics technical review (2025).

Sydney Cancer Partners / Multi-institutional collaborations: Contributing computational biology expertise to the MRFF-funded project *Targeting the Dysregulated Epigenome to Enhance Immunotherapy Response* (Hudson Institute of Medical Research, CMRI, and Children's Cancer Institute).

Key Technical Skills

Machine Learning and AI: Expertise in deep neural networks, transformer architectures, variational autoencoders, generative models, federated learning, and multi-view integration methods.

Bioinformatics and Multi-Omics Data Integration: Proficient in end-to-end proteomic data analysis (data QC, peptide-to-protein rollup, pre-processing, differential expression, pathway analysis, survival analysis), whole exome/genome sequencing (germline/somatic mutations, copy number variations, structural variants), and single-cell RNA-seq analysis.

Programming and Software Engineering: Python, R, PyTorch, SQL, C++, Linux, Perl, and bioinformatics software suites.

Research Communication: Scientific writing for high-impact journals, conference presentations, peer review, and public engagement.

Referees

Available upon request.