Joseph Bloom

jbloomaus@gmail.com | https://www.linkedin.com/in/joseph-bloom1

CAREER PROFILE

An independent researcher working in mechanistic interpretability, funded by LTFF, Manifund and LightSpeed Grants. Currently using Sparse Autoencoders to study neural networks in Neel Nanda's stream of the ML Alignment and Theory Scholars (MATS). I've been a maintainer of the TransformerLens package for Mechanistic Interpretability (700+ stars on github). I've developed a library for training GridWorld Decision Transformers (50+ stars on github). Before working in technical AI safety, I earned a double degree in computational biology and statistics and stochastic processes at Melbourne University. I worked at a SaaS startup for 2 years as a Data Scientist.

EXPERIENCE

Alignment Research

Remote *Jan 2023 – Present*

I've been an independent research for approximately a year. I spent most of this time working on mechanistic interpretability of gridworld models and recently pivoted to study Sparse AutoEncoders. I also maintained transformerLens until November 2023. I also TA'd at ARENA 2.0 during this time.

Independent Researcher

- Published 3 blog posts on Decision Transformer Interpretability.
- Published 1 blog post on Sparse Autoencoders.
- While maintaining TransformerLens, I supported the release of multi-GPU support, import of numerous new models and architectures (eg: Llama and Bert) and various minor features/bug fixes.
- Started the Open Source Mechanistic Interpretability Slack (>40 participants) and have begun maintaining key open-source infrastructure in the field (TransformerLens Package)

Alignment Research Engineering Accelerator (ARENA) - Participant

London, UK *Oct 2022 – Jan 2023*

I am attending a 9-week course on LLM Alignment and related topics such as interpretability, modelling objectives, scaling laws, RL, alignment and adversarial training.

- Implemented and fine tuned models in PyTorch, such as a fluent "Shakespeare" model.
- Completed a capstone on Decision Transformer Interpretability.

FTX Future Fund Re-grant Recipient

Remote *Jul 2022 – Dec 2022*

I received six months' worth of funding that enabled me to upskill in AI safety, deep learning and explore possible roles in biosecurity and AI alignment.

Mass Dynamics - Data Scientist

Melbourne, Australia *Jul 2020 – Jul 2022*

MassDynamics is a Proteomics Software-as-a-Service (SaaS) solution for academic researchers and industry.

- Wrote numerous python/R packages utilising machine learning, statistics and visualisation tools to automate customer analysis of complex datasets and provide publishable and actionable insights.
- Planned, communicated and coordinated whole-of-company efforts to deliver novel services on tight timelines.
- Engineered and benchmarked a novel approach to the protein inference problem (see protein inference package)
- Benchmarked, documented and published the MassDynamics 1.0 platform (Journal of Proteome Research)

Tools Used: Python, R, AWS, Azure, CI/CD, Docker

Buckle Protein Engineering Lab - Research Assistant

Melbourne, Australia *June 2017 – July 2019*

Operating within the Monash Biodiscovery Institute, the Buckle Lab combines physics, high-performance computing and protein expertise to do research that underpins the development of novel drugs to fight disease.

- Designed novel protein (protease) using Hidden-Markov models, expressed and characterised in the wet lab.
- Perform unsupervised learning of molecular dynamics simulations (protein conformational structures).

EDUCATION

THE UNIVERSITY OF MELBOURNE - MELBOURNE BUSINESS SCHOOL *Melbourne, Australia*
Master of Business Analytics (differed) *Jan 2020 – August 2020*

- Official Commendation in Module 1 for "asking good questions."
- Grade average: 81%
- Differed with two modules remaining to take up a position at Mass Dynamics

THE UNIVERSITY OF MELBOURNE – FACULTY OF SCIENCE *Melbourne, Australia*
Bachelor of Science – Major in Computational Biology *Jan 2016 – Dec 2019*

- Breadth Sequence/Minor in Economics
- Founder, President Melbourne University Biological Society
- President, Coach, Melbourne University Quidditch Club
- Grade average: 85%

THE UNIVERSITY OF MELBOURNE – FACULTY OF SCIENCE *Melbourne, Australia*
Diploma in Mathematical Sciences – Statistics and Stochastic Processes *Jan 2017 – Dec 2019*

- Grade average: 72%
- Member of Mathematics Society/Physics Society

Publications

- [Blog] **Bloom J**. 2024. Open Source Sparse Autoencoders for all Residual Stream Layers of GPT2 Small. Available at: https://www.alignmentforum.org/posts/f9EgfLSurAiqRJySD/open-source-sparse-autoencoders-for-all-residual-stream.
- [Blog] Watkins M, **Bloom J**. Linear encoding of character-level information in GPT-J token embeddings. LessWrong. 2023 Nov 10. Available at: https://www.lesswrong.com/posts/GyaDCzsyQgc48j8t3/linear-encoding-of-character-level-information-in-gpt-j.
- Nanda N, Bloom J. TransformerLens. 2022. Available at: https://github.com/neelnanda-io/TransformerLens.
- **Bloom J,** Triantafyllidis A, Quaglieri A, Burton Ngov P, Infusini G, Webb A. Mass Dynamics 1.0: A Streamlined, Web-Based Environment for Analyzing, Sharing, and Integrating Label-Free Data. J Proteome Res. 2021 Nov 5;20(11):5180-5188. doi: 10.1021/acs.jproteome.1c00683. Epub 2021 Oct 14. PMID: 34647461.
- Szeto C*, **Bloom JI***, Sloane H, Lobos CA, Fodor J, Jayasinghe D, Chatzileontiadou DSM, Grant EJ, Buckle AM, Gras S. Impact of HLA-DR Antigen Binding Cleft Rigidity on T Cell Recognition. International Journal of Molecular Sciences. 2020; 21(19):7081. https://doi.org/10.3390/ijms21197081
- 4 Medium articles published on Medium by Towards Data Science and the Computational Biology Magazine, including a front-page feature, editors' picks on TDS adding to more than 5k views and 2k reads.

Courses

- Deep Learning Specialization (17 weeks), DeepLearning.AI, Coursera (March 2021)
- Network Analysis in System Biology, Icahn School of Medicine at Mount Sinai, Coursera (April 2021)
- Information Visualization: Programming with D3.js, New Your University, Coursera (May, 2021)
- Graph Analytics for Big Data, University of California San Diego, Coursera (May 2021)
- Colours for Data Science A- Z: Data Visualization Color Theory, Kiril Eremenko, Udemy (June 2019)

ADDITIONAL

- **Visa Status: Australian Citizen, Right to work in UK for 2 years from July 2023**
- Languages: English
- Behavioural Skills: Strong communication skills (Writing, Presentations), Critical reasoning, Collaboration
- Technical Skills: Python, R, Bash, Statistics, Machine Learning, DeepLearning, CI/CD, AWS, Azure
- References:
    - Neel Nanda (Google Deepmind - Head of Mechanistic Interpetability): neelnanda27@gmail.com.
    - Aaron Triantafyllidis (Mass Dynamics - Technical): aaron@massdynamics.com
    - Pr. Ashley Buckle (Monash University): ashley@ptngconsulting.com + 61 430 913 031