




Henry W. Leung Ph.D.

AI x Astrophysics Scientist

 [henrysky.github.io](https://github.com/henrysky) henryskyleung@gmail.com [henrysky](https://twitter.com/henrysky) [Henry Leung](https://www.linkedin.com/in/henry-leung) Bilingual in English & Chinese Python & C Canadian & Hong Konger Toronto, Canada

SUMMARY

Recent PhD graduate and former Data Science Institute doctoral fellow, applying **GenAI** methods to build **multi-modal foundation models for science**. I have 6+ years experience to adapt machine learning techniques to solve real world science problems with multi-terabytes datasets and 9+ year of Python programming and software development experience.

PROFESSIONAL EXPERIENCE

University of Toronto

Sept 2019 – Oct 2024

Graduate Researcher & Data Science Institute Doctoral Fellow

- Developed prototypes of multi-modal foundation models for astronomical data with **Transformers** architecture and **denoising diffusion** probabilistic models. Presented as spotlight talks and posters at **NeurIPS** and **ICML**.
- Built **self-supervised** Transformers models implemented in **PyTorch** trained on large **multi-terabytes datasets** with billions of stellar objects, leveraging tools such as **Docker** and **PostgreSQL** trained on national supercomputer equipped with Nvidia A100s. Created an online natural language interface using **LLMs** hosted on a home computer with **flask**.
- Built an encoder-decoder model implemented in **Tensorflow** for **unsupervised learning** to solve a cross-domain data rich but label poor problem, by reducing dimensionality of the data guided by physical knowledge and label regression on the latent space.
- Curated catalogues of millions of stellar parameters and associated uncertainties derived with data-driven models, with more than 10% improvement on stellar parameters accuracy and a few orders of magnitude faster to **low signal-to-noise data** compared to traditional physics-driven pipeline.
- Collaborated with the community to curate a 100 TB astronomical dataset for training large model in the future. Created and maintained **well-documented** and **thoroughly tested** open-source software mainly written in **Python**, **C** and **SQL**, contributing both to personal projects and to the wider scientific community.

EDUCATION

Ph.D., Astronomy & Astrophysics, University of Toronto

2020 – 2024

Dissertation: “Exploring the Milky Way with Deep Learning” with Prof. Jo Bovy

M.Sc., Astronomy & Astrophysics, University of Toronto

2019 – 2020

H.B.Sc., Physics & Astronomy, University of Toronto

2014 – 2019

PUBLICATION OVERVIEW

I am the first/second author on **9 refereed papers** that have **570+** citations. In total, I am an author on **16 refereed papers** that have **2730+** citations (h-index=11). My research has been presented at international conferences and workshops. Here are some of the highlights (first-author unless noted as part of a collaboration):

NeurIPS (2023, 2024)

- Talk on “Towards an Astronomical Foundation Model for Stars”
- Collaboration poster on “The Multimodal Universe: Enabling Large-Scale Machine Learning with 100TBs of Astronomical Scientific Data”

ICML (2024)




- Poster on “Estimating Probability Densities with Transformer and Denoising Diffusion”

Artificial Intelligence for Astronomy (2019)

- Talk on “Mapping the Milky Way Galaxy with Deep Learning”

SOFTWARE OVERVIEW

I am proficient in **Python** and **C** programming and familiar with tools around high-performance computing and **SQL** databases. I am learning **Rust** and **C++** by taking initiatives to implement wishlist features in other open-source projects. Most of my research are open-sourced including codes for publications are hosted on my [GitHub](https://github.com/henrysky). This includes a few software packages used by the community that are well tested using continuous integration with GitHub Actions and well documented with docstrings and user guides, for example:

- **astroNN**  - Deep Learning for Astronomers with **Keras**
- **Galaxy10**  - A CIFAR10-like galaxy image dataset for educational and research purposes
- **MyGaiaDB**  - A data management package to setup local serverless multi-terabytes astronomical databases using **SQLite** and run query locally with **Python**