# **ZIYANG** “Claude” **HU**

Charlottesville, VA, 22901 | zh4nh@virginia.edu | https://claudehu.github.io/

## PROFILE

Data science professional with background in machine learning, natural language processing, and bioinformatics. Bilingual communicator passionate about exploring and solving real-world data challenges.

## SKILLS

| | |
|---|---|
| **Programming:** | Python, R, C, Java, Rust, shell scripting |
| **ML libraries:** | Huggingface Transformers, PyTorch, TensorFlow, scikit-learn |
| **NLP libraries:** | SentenceTransformers, Langchain, FastEmbed |
| **Data analysis:** | NumPy, Pandas, SQLite, ggplot2, matplotlib, Huggingface Datasets |
| **Development:** | Docker, Slurm |

## EDUCATION

**DUKE UNIVERSITY**, School of Medicine, Durham, NC
***Master of Biostatistics,*** May 2023
Relevant coursework: Software Tools for Data Science, Bayesian & Modern Statistics, Probabilistic Machine Learning

**EMORY UNIVERSITY**, College of Arts and Sciences, Atlanta, GA
***Bachelor of Science*****, Double Major: Computer Science, Neuroscience and Behavioral Biology**, May 2021
Relevant coursework: Analysis of Algorithms, Machine Learning, Numerical Analysis, Big/Small Data and Visualization

## EXPERIENCE

**University of Virginia**, Charlottesville, VA 2023-now
Sheffield Lab, Department of Genome Sciences
***Scientific Programmer***

- Contributed to the development of a Python package for machine learning models of genomic interval data.
- Extended & tested a Rust crate for genomic interval data analysis and processing.
- Optimized genomic interval data storage pipeline, reduced runtime by 40%.
- Built a file retrieval pipeline combining semantic search with genomic interval representation learning.
- Curated ad-hoc datasets with biomedical ontologies.
- Fine-tuned open-source language models for information retrieval.

**DUKE UNIVERSITY**, Durham, NC 2022-2023
Department of Biostatistics & Bioinformatics
***Graduate Research Assistant***

- Developed Python/R functions to automate preprocessing of unstructured clinical data.
- Performed medical concept extraction and entity mapping from electronic health records with NLP tools.
- Retrieved and preprocessed raw data from Duke Clinical Research Data Mart (CRDM).
- Trained predictive models and analyzed model fairness across demographic groups.

**EMORY UNIVERSITY**, Atlanta, GA 2020-2021
Department of Sociology
***Research Assistant***

- Contributed to the development of a computational social science data analysis toolkit.
- Designed a pipeline that applied the annotation of Stanford CoreNLP to concise relationship extraction.
- Preprocessed raw data for an interdisciplinary project that analyzed statements from HIV patients.
- Utilized open-source Python libraries to perform sentiment analysis on interviews.

## PUBLICATIONS & PRESENTATIONS

Franzosi, R., Dong, W., **Hu, Z**., Dai, W., Cha, M., Piloto, R., & Wang, G. (2024). “Automatic information extraction of the narrative elements who, what, when, and where” [Manuscript submitted for publication]. Social Science Computer Review.

Xue, B., Khoroshevskyi, O., Stolarczyk, M., Mosquera, J. V., Campbell, D., **Hu, Z**., Tambe, S., LeRoy, N., Gharavi, E., Duzlevski, O., & Sheffield, N. C. (2024, November). “BEDbase: A web application and API for genomic region sets” [Poster presentation]. Biological Data Science Conference, Cold Spring Harbor, NY, USA.

Yang, R., Tong, J., Wang, H., Huang, H., **Hu, Z.**, Li, P., Liu, N., Lindsell, C. J., Pencina, M. J., Chen, Y., & Hong, C. (2025). “Enabling inclusive systematic reviews: Incorporating preprint articles with large language model-driven evaluations” [Manuscript submitted for publication]. NEJM AI. https://doi.org/10.48550/arXiv.2503.1385