ZIYANG "Claude" **HU**

zh4nh@virginia.edu | www.linkedin.com/in/ziyang-claude-hu | https://claudehu.github.io/
Charlottesville, VA – 22901, USA

SKILLS

Programming Languages: Python, R, C, Rust, Java, Bash/Shell, SQL
Machine Learning: PyTorch Lightning, Transformers, PyTorch, TensorFlow, scikit-learn
Natural Language Processing: SentenceTransformers, LangChain, FastEmbed, CoreNLP, Stanza
Data Analysis and Management: NumPy, Pandas, SQLite, Hugging Face Datasets, Qdrant
Visualization and Design: ggplot2, matplotlib, Inkscape

EDUCATION

DUKE UNIVERSITY, School of Medicine, Durham, NC
Master of Biostatistics, May 2023
Relevant coursework: Software Tools for Data Science, Statistical Programming for Big Data, Probabilistic Machine Learning

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

EXPERIENCE

UNIVERSITY OF VIRGINIA, Charlottesville, VA 2023-Present
Sheffield Lab, Department of Genome Sciences
Scientific Programmer
- Developing Python and Rust packages for genomic interval data analysis and machine learning applications.
- Investigating cross-modal text-file retrieval for ChIP-seq genomic interval output in bed narrowPeak format.
- Optimized genomic interval data storage and reduced database reindex 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.
- Participated in systematic review by extracting and summarizing key information from relevant scientific literature.
- Performed medical concept extraction and entity mapping from electronic health records with NLP tools.
- Retrieved and preprocessed raw data from the 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 using the annotation of Stanford CoreNLP for efficient relationship extraction.
- Preprocessed raw data for an interdisciplinary project to analyze statements from HIV patients.
- Utilized open-source Python libraries to perform sentiment analysis on interviews.

OTHER EXPERIENCE

Auto-Arrangement for Acoustic Guitar 2024-Present
Side project for interest with collaborator
Full-stack Developer
- Investigating genre-specific arrangement of songs for acoustic guitar by audio-text transcription/translation models.
- Extending the DadaGP Python package to tokenize guitar tab files with alternative tunings and process resulting tokens.
- Implementing BPE tokenizers and vocabularies to process the text-based output of guitar tabs generated by DadaGP.
- Curated a genre specific dataset of guitar tablature for future training and evaluation.

PUBLICATIONS

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.
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

PRESENTATIONS

Huang, H., Tong, J., **Hu, Z.**, Li, Y., Pencina, M., Chen, Y., & Hong, C. . "Enabling Inclusive Systematic Reviews: Incorporating Preprint Articles based on Semantic Learning and Large Language Model". Poster presentation at the ENAR Spring Meeting, Baltimore, MD, USA. March 10-13, 2024.
Xue, B., Khoroshevskyi, O., Stolarczyk, M., Mosquera, J. V., Campbell, D., **Hu, Z**., Tambe, S., LeRoy, N., Gharavi, E., Duzlevski, O., & Sheffield, N. C. . "BEDbase: A web application and API for genomic region sets" Poster presentation At the Biological Data Science Conference, Cold Spring Harbor, NY, USA, November 13-16, 2024.