# ZUOLIN CHENG

• zuolin8@vt.edu • (+1)571-385-8606 • Cupertino, CA 95014
• https://zuolincheng.github.io/

## SUMMARY

Highly skilled Ph.D. in Bioinformatics with 8+ years of experience in designing/implementing multi-omics data analysis pipelines & developing novel machine-learning/statistical models. Proficient in multiple programming languages, machine-learning/data-analysis libraries, and bioinformatics tools. Proven track record of successful data analysis enabling significant biological discoveries in cross-functional collaborative research. Winner of the Best Annual Paper Award of Neuron journal.

## CORE COMPETENCIES

- **Expertise:** Bioinformatics, Machine Learning, Data Analysis, Statistical Modeling
- **Programming languages:** R, Python, MATLAB, C++, SQL
- **Pipeline design:** scRNA-seq, spatial transcriptomics, ATAC-seq, ChIP-seq, CUT&RUN, clinical data
- **Tools:** Git, PyTorch, Scikit-Learn, NumPy, Matplotlib, Scanpy, Seurat, MACS3, STAR, etc.

## WORK EXPERIENCE

**Virginia Tech** Cupertino, CA, USA
***Postdoctoral Research Associate*** Feb. 2023 – Present

Lead omics data analysis in collaborative biological research studies, effectively facilitating scientific discoveries. Develop state-of-the-art Next Generation Sequencing (NGS) data analysis models and methods, providing user friendly tools to benefit a larger audience. Mentor and train junior graduate students in research practices.

- ***Project 1: GO-BERT, Cell Type Annotation for scRNA-seq Data Using Large Language Model (LLM)***
  - *Data*: scRNA-seq data, Gene Ontology (GO); *Language*: Python, PyTorch
  - Innovatively introduced GO knowledge into neural network architecture during supervised fine tuning, significantly enhancing cell type annotation performance, increasing F1 score to 0.76 (previously 0.69)
  - **Reduced 31% parameters** in fine tuning without loss of performance compared to the state-of-the-art
  - Propose a **domain-insight-based pretraining strategy** for large-scale single-cell data, overcoming limitations of existing models (e.g. scBERT, scGPT, Geneformer) in capturing awareness of global context

- ***Project 2: CMC, A Machine Learning Model Unifying Various Omics Analysis Tasks***
  - *Data*: various types of omics/multi-omics data; *Language:* C++, R, MATLAB
  - Unified various omics analysis tasks (e.g. scRNA-seq data normalization, cancer gene detection, single-cell GO term activity inference, etc.) by pinpointing the shared key mathematical problem behind them
  - Developed **a novel model of joint inference of multiple hidden factors** by integrating sophisticated statistical/machine-learning techniques, enhancing confounding factor control in downstream analyses
  - Proposed a highly efficient customized optimizer leveraging model structure & code optimization (C++), achieving **scalability on large-scale real-world data** (reduced running time: days to minutes)
  - Developed user-friendly R packages on different OS (Windows, Linux, iOS), and shared them via GitHub

- ***Project 3: Therapy Data Analysis for Non-Small Cell Lung Cancer Patients***
  - *Data*: clinical data (medical record, RNA-seq, spatial transcriptomics), TCGA; *Language:* R, MATLAB
  - Design a pipeline of treatment effectiveness analysis for non-small cell lung cancer
  - Successfully identified two patient clusters that both respond to “chemotherapy + immunotherapy” but have distinct highly active GO terms, revealing difference in underlying response mechanisms

**Virginia Tech, Computational Bioinformatics and Imaging Lab** Arlington, VA, USA
***Graduate Research Assistant*** Aug. 2015 – Dec. 2022

Developed innovative AI/ML and statistical tools for omics data analysis, including TySim, scRNA-seq normalization method, and algorithms for driving transcription factor inference. Designed and implemented pipelines for analyzing data in cross-functional projects, collaborating with biologists and clinicians from leading institutions such as Stanford University, Cincinnati Children's Hospital Medical Center, and the Salk Institute.

- ***Project 1: scRNA-seq Data Analysis to Uncover Microglial Heterogeneity During Development***
  - *Data*: scRNA-seq data; *Language*: R
  - Designed and implemented scRNA-seq data analysis pipeline, including sequence alignment (STAR, etc.), scRNA-seq normalization, clustering (graph-based), data visualization (tSNE), DEG analysis (rank-sum test), pathway enrichment analysis (Fisher's exact test), pseudotime analysis (monocle3), etc.
  - **Customized** some modules in the pipeline, achieving best solutions to **project-specific requests**
  - Enabled biological discoveries, collaborating with Stanford University: **revealed a novel microglia subset**

- ***Project 2: Driving Transcription Factor Prediction & Glial Progenitor Development Study***
  - *Data*: scRNA-seq, ChIP-seq (Unibind DB), DNase-seq (CistromeDB DB), Motif (JASPAR DB); *Language*: R
  - **Discovered a long-ignored key factor** and integrated it into ML model, boosting the best record of disease driving TF detection based on large TF-gene binding database (**38.8% to 49.0%** on benchmark)
  - Identified and published a pivotal TF (Zfp36l1) in collaboration with Cincinnati Children's Hospital

- ***Project 3: Full-length scRNA-seq Data Normalization***
  - *Data*: scRNA-seq; *Language*: R, C, C++
  - Identified a key confounding factor (effective cDNA-length) ignored by previous analyses in scRNA-seq raw counts, by **combining empirical data mining** (high-throughput intermedia data of sequencing alignment) and **theoretical reasoning** (**deep understanding of sequencing technologies**)
  - Proposed a ML algorithm to learn and address the factor during scRNA-seq data normalization, enhancing the accuracy of downstream differential expressed gene (DEG) detection (AUC: 0.886 to 0.938)

## EDUCATION

**Virginia Tech** Arlington, VA, USA
PhD, Electrical Engineering Aug. 2015 – Dec. 2022
- ***Research areas****:* Bioinformatics; Machine Learning

**Peking University** Beijing, China
M.S., Electronic Science and Technology
- GPA: 3.87/4.0; Top Ranked: 1/30

## PUBLICATIONS

**TOTAL PUBLICATIONS: 13; TOTAL CITATIONS: 1400 +**

- Li Q*, **Cheng Z* (co-first author)**, Zhou L, … & Tony Wyss-Coray, Ben A Barres. Developmental heterogeneity of microglia and brain myeloid cells revealed by deep single-cell RNA sequencing. ***Neuron***. 2019; 101(2)
- Barclay K, Nora A, **Cheng Z**, Kim M, Zhou L, Yang J, Rustenhoven J, Mazzitelli J, et al. An inducible genetic tool for tracking and manipulating specific microglial states in development and disease. ***Immunity***. 2024.
- **Cheng Z**, Wei S, Wang Y, Wang Y, Lu R, Wang Y, Yu G. An Efficient and Principled Model to Jointly Learn the Agnostic and Multifactorial Effect in Large-Scale Biological Data. *bioRxiv* (2024): 2024-04.
- **Cheng Z,** Wei S, Yu G. A Single-Cell-Resolution Quantitative Metric of Similarity to a Target Cell Type for scRNA-seq Data. *In 2022 IEEE International Conference on Bioinformatics and Biomedicine (**BIBM**)* 2022.
- **Cheng Z**, Cui X, Cui X, Lee CL. Self-heating burn-in pattern generation based on the genetic algorithm incorporated with a BACK-like procedure. ***IET Computers & Digital Tech****.* 2015; 9(6):300-10.

## AWARDS

- **Best papers** 2018-2019 published in **Neuron**