

## HIGHLIGHTS

- Bioinformatics, Machine Learning, Data Analysis, Statistical Modeling
- 8+ years' experience in multi omics data analysis & novel machine-learning/statistical model development

## TECHNICAL SKILLS

- **Programming languages:** R, Python, MATLAB, C++, SQL
- **Pipeline design:** scRNA-seq, spatial omics data, ATAC-seq, ChIP-seq, CUT&RUN, clinical data
- **Tools** (in-depth understanding & hand-on experience): Seurat, MACS3, STAR, etc.

## EDUCATION

|   |  |                              |
|---|--|------------------------------|
| <b>PhD, Electrical Engineering</b>                        | <b>Virginia Tech (Arlington, VA, USA)</b>    | <b>Aug. 2015 – Dec. 2022</b> |
| • <b>Research areas:</b> Bioinformatics; Machine Learning |  |                              |
| <b>M.S., Electronic Science and Technology</b>            | <b>Peking University (Beijing, China)</b>    | <b>Sep. 2012 – Jun. 2015</b> |
| • GPA: 3.87/4.0; Rank: 1/30                               |  |                              |
| <b>B.E., Electronic Science and Technology</b>            | <b>Shenzhen University (Shenzhen, China)</b> | <b>Aug. 2008 – Jul. 2012</b> |
| • GPA: 3.83/4.0; Rank: 1/117                              |  |                              |

## WORK EXPERIENCE

|  |                            |                            |
|--|----------------------------|----------------------------|
| <b>Research Associate</b>  | <b>CBIL, Virginia Tech</b> | <b>Feb. 2023 – Present</b> |
| • Managed the design and implementation of data analysis for various biological research projects      |                            |                            |
| • Developed state-of-the-art omics data analysis models and tools to facilitate biological discoveries |                            |                            |

## SELECTED PROJECTS

|  |  |
|--|--|
| <b>GO-BERT: Cell Type Annotation for scRNA-seq Data Using Large Language Model (LLM)</b>   |  |
| (Language: Python, Pytorch) Dec. 2023 – Present  |  |
| • <b>Data:</b> scRNA-seq data, Gene Ontology (GO)  |  |
| • Significantly enhanced the performance of cell type annotation ( <b>f1 score: 0.69 -&gt; 0.76</b> ) by <b>innovatively introducing GO knowledge</b> into the model architecture during supervised fine tuning            |  |
| • <b>Reduced 31% parameters</b> in fine tuning without loss of performance compared to state-of-the-art (scBERT)   |  |
| • Proposed a <b>domain-insight-based pretraining strategy</b> for large-scale single-cell data, <b>overcoming limitations of existing models</b> (e.g. scBERT, scGPT, Geneformer) in capturing awareness of global context |  |

### CMC: A Machine Learning Model Unifying Various Omics Analysis Tasks

(Language: C++, R, Matlab) Dec. 2020 – Dec. 2022

- **Data:** various types of omics/multi-omics data
- Pinpointed a shared key mathematical problem behind challenges of **various omics analysis tasks** (e.g. scRNA-seq data normalization, cancer gene detection, single-cell GO term activity inference, etc.): inferring the effect of multiple hidden factors in large-scale multi-omics data.
- Developed a **novel model + algorithms**, much enhancing **confounding factor control** in downstream analyses
- Integrated sophisticated **statistical/machine-learning techniques**: principle of maximum entropy, Lagrange multipliers, maximum likelihood, Newton's method, EM, saddle point approximation, etc.
- Proposed a highly efficient customized optimizer leveraging model structure & code optimization (C++), achieving **scalability on large-scale real-world data** (reduced running time: days -> minutes)
- Developed user-friendly R **packages** on **different OS** (Windows, Linux, iOS), and shared them via **GitHub**  
<https://github.com/yu-lab-vt/CMC>

## ***Heterogeneity of Microglia and Brain Myeloid Cells During Development***

(Collaborator: Stanford University)

(Language: R) Apr. 2017 – Jan. 2019

- Data: scRNA-seq
- Designed and implemented **scRNA-seq data analysis pipeline**, including sequence alignment (STAR, etc.), scRNA-seq normalization, clustering (graph-based), visualization (tSNE), DEG analysis (rank-sum test), pathway enrichment analysis (Fisher's exact test), pseudotime analysis (monocle3), etc.
- **Customized** some modules in the pipeline, accommodating to **project-specific requests**
- **Enabled biology discoveries**: characterized microglial heterogeneity; revealed a distinct microglia subset

## ***Driving Transcription Factor (TF) Prediction & Its Application to Glial Progenitor Development Study***

(Collaborator: Cincinnati Children's Hospital Medical Center)

(Language: R) May 2016 – Apr. 2019

- Data: scRNA-seq, ChIP-seq (Unibind database(DB)), DNase-seq (CistromeDB DB), Motif (JASPAR DB)
- Significantly boosted the best record of disease driving TF detection based on large TF-gene binding database (**38.8% -> 49.0%** on benchmark) by discovering a **long-ignored key factor** & integrating it into ML model
- Identified a pivotal TF (Zfp361l1) in a **collaborative interdisciplinary study**, published as biology discovery

## ***Therapy Data Analysis for Non-Small Cell Lung Cancer Patients***

(Language: R, MATLAB) Feb. 2024 – Present

- Data: clinical data (medical records, RNA-seq, spatial transcriptomics)
- Designed 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, suggesting different underlying response mechanisms

## ***Missing Value Imputation by Joint Inference of Global-Local Interaction Network***

(Language: MATLAB) Apr. 2018 – Feb. 2019

- Data: proteomics data, metabolomics data
- Proposed a novel model and algorithm of missing value imputation for proteomics/metabolomics data, utilizing **original insights in protein/metabolite network**: a compound of global and local network structures
- Achieved significantly lower imputation errors than 8 state-of-the-art methods on real-world data, with **at least 16.8% lower NMSE** in various metabolite datasets, regardless of global/local component ratio

## **SELECTED PUBLICATIONS (TOTAL PUBLICATIONS: 13; TOTAL CITATIONS: 1256)**

( <https://scholar.google.com/citations?user=6pHncW4AAAAJ&hl=en&oi=ao> )

- 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)
- Weng Q, Wang J, Wang J, He D, **Cheng Z**, Zhang F, et al. Single-cell transcriptomics uncovers glial progenitor diversity and cell fate determinants during development and gliomagenesis. **Cell stem cell**. 2019; 24(5).
- 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**. 2024-05.
- **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 Dec 6.
- **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 Nov;9(6):300-10.

## **SELECTED AWARDS**

- **Best papers** 2018 – 2019 published in **Neuron** (1 of the 11)