Chi Yen Tseng chiyen_tseng@lanl.gov

A multidisciplinary data scientist/applied statistician with expertise in experimental data, modeling, bioinformatics, chemistry, toxicology, and science communication.

SKILLS

Computer Skills: R, Linux, HPC, SQL, Python, Microsoft Excel, PowerPoint, and Word

Data Skills

Data analysis

- Bayesian statistics for data simulation, modeling, and uncertainty quantification.
- Adept in the use of R, RStudio, and HPC for data cleaning, transformation, integration, statistical analysis, and **data visualization**.
- Parallel computing on HPC for geonomics data processing.
- Github and Gitlab version control.
- Developing AI-ML, Retrieval-Augmented Generation (RAG) for large language model (LLM)s-based **data/ metadata management** strategy and graph database building (Python).
- Specialized in sparse, low sample size, high dimensional data analysis for differentiated features identification and data mining.
- Developing statistical methods, pipelines, tools, and workflow for analyzing transcriptome/ metabolome.
    - Machine learning and statistical learning between NGS data/ Mass Spectrometry data to predict environmental toxicants and classify exposure chemicals.
- Science communications, presenting data as **dashboards** such as R/Shiny and Microsoft office.
- Machine Learning (Caret, randomForest, Statistical learning, sparse methods for classification and regression, Decision trees, Boosting, SVM, Clustering) to determine biomarker.

Project Skills

- Identifying **mission critical** differentiated features associated with chemical warfare agents exposure.
- Implementing RAG-LLM for Mass spectrometry data and meta data management
- Developing **data analytics and visualization pipeline** for multi-omics mass spectrometry data integration, preprocessing, **QC**, Bayesian feature simulation (probabilistic statistics), and in the process of building dashboards to present performance metrics.

Communication skills

- Support Mass Spectrometry Center for Integrated Omics team (B-TEK) chemists and chemistry customers with mission critical data preprocessing, mining, differentiated feature identification, data visualization, functional explanation, and data management.
- Support wildlife biologist from Upper Midwest Environmental Sciences Center, U.S. Geological Survey for genomics and environmental data preprocessing, data mining, and modeling.
- Strong science communications skill with **non-data scientist**.

Project management

- For the current DR project, I served as the data team leader to develop project plans and scopes, manage data processing timeline and personnels (two post-master), and progress report to meet competing deadlines.
- As a pre-U.S. Geological Survey government contractor, I involved in multiple EPA funded project to assist project development, documenting reports, and to meet deadlines.
- Managing project budgets, purchasing, and safety compliance.
- Provide training and mentorship for post-masters and undergraduate student employees.

Qualifications

- **Working with numerical modeling and experimental data**
- **Familiar with parallel computing on HPC**
- **Machine Learnin and Artificial Intelligence**

- **Carry out independent and collaborative research**
- **Statistical Analysis**

EDUCATION

Ph.D. in Environmental Science, The Institute of Ecological, Earth, and Environmental Sciences (TIE3S)
Baylor University, Waco, TX GPA 3.87 May 2023

Master of Science, Toxicology, Environmental Health Science, University of Georgia, Athens, GA GPA 3.49 Dec 2013

Bachelor of Science, Agricultural Chemistry, Taiwan University, Taipei, Taiwan; GPA 3.20 May 2008

EXPERIENCE

Postdoc at B-TEK, Los Alamos National Laboratory **June 2023 – Present**
Montor: Blackwell, Brett Reginald, blackwell@lanl.gov

- Data Team lead (20230084DR InCEPTIon): Biomarker discovery from geonomics and mass Spectrometry-based multi-omics datasets measured in cell lines exposed to chemical warfare agents and their surrogates and related compounds
    - Developing pipelines and workflow for data preprocessing, evaluation, data visualization, and functional clustering (Bayesian hierarchical model) models to simulate longitudinal variation and generate posterior predictions.
    - Provide functional determination of top differentiated features.
    - Geonomics data processing on HPC.
    - Developing dashboards to present integrated longitudinal variation of multi-omics features.
- PI (20258379CT-IST): AI Agent for Streamlined Metadata Standardization for Enhanced Usability
    - ISTI rapid response funding to use Retrieval-Augmented Generation (RAG) for large language model (LLM) to semi-automatically identify key ontology entities from meta data and use LLM to build knowledge graph.
    - Enable FAIR principle of mass spectrometry data, i.e., making data Findable, Accessible, Interoperable, and Reusable.
    - Knowledge retrieval to improve metadata richness through combines vector and keyword indexes with graph retrieval for RAG applications.
    - Managing project budgets, purchasing, and safety compliance.
- Co-Investigator (20250339ER): PeptideRx: Unveiling Nature's Antimicrobial Arsenal
    - Determine small peptide sequence from algae media.
- Co-Investigator (ECR 20220584ECR): High-throughput Conotoxin, protein docking simulations and screening
    - Parallel protein docking simulation on GPU
- Manage data team scopes, processing timeline, and provide mentorship for post-masters and undergraduate student employees.
- Data analysis assistance
    - Identifying top differentiated metabolites and peptides from different beer samples (Beer NMSBA).
    - Chemistry division: Identifying top differentiated lipids from lipid (Nanodisc).

U.S. Geological Survey Contractor **Jan 2021 – Sep 2022**
Supervisor: Natalie Karouna-Renier, nkarouna@usgs.gov

- Determine if land-use conditions and environmental contaminant mixtures were correlated with biological effects, especially in immune system, in tree swallow nestlings.

- RNA-Seq data collection and preparation from tree swallow nestling spleen, data cleaning, and integration with contaminant profiles (LC-MS/MS) with Linux and HPC (e.g., Trimmomatic, STAR, and FeatureCounts), R (dplyr, tidyr, trimmomatic) to establish the correlation between the variation in contaminant profiles and gene expression.
- Project timeline management, communication, and coordination with multiple U.S. Geological Survey entities.

Graduate Research Assistant **2014 - 2023**
Baylor University - Department of Environmental Science, The Institute of Ecological, Earth, and Environmental Sciences (TIE3S)
Supervisor: Cole W. Matson, cole_matson@baylor.edu

- **Nanotoxicity** − Assess the killifish embryo acute toxicity of silver nanoparticles with varying surface modifications in early killifish embryos.
- **Mixture toxicity** − Assess the interactions between contaminant profiles, transcriptome, and metabolome from tree swallows in the Great Lakes contaminated and clean sites with U.S. Geological Survey, Upper Midwest Environmental Sciences Center, Eastern Ecological Science Center, and Great Lake Restoration Initiative.
    - Data cleaning, integration, and reformatting for tissue chemistry, transcriptome, and metabolome data.
    - *De novo* assembly tree swallow genome and annotation using 10 X Genomics linked reads technology with Supernova and Maker2 in Linux.
    - Exploring the Connections between Land-Use, Contamination Profiles, and Omics Signals on a Regional Scale: Maumee River, OH (Trimmomatic, STAR, and FeatureCounts in Linux; GLMM, Vegan, MixOmics, and EdgeR in R).
    - Contaminant Mixtures and Gene Expression in Tree Swallows of the Great Lakes: Machine learning and predictive modeling to support management decisions (glmnet and caret, R) for initial assessment and for assessing Remedy effectiveness.
    - Comparison of Predictive Models for Monitoring Contamination in Tree Swallow Nestlings Using Non-targeted Global Gene Expression, Targeted Gene Panels, and Bioindicators in the Great Lakes Region.

University of Georgia – Environmental Health Science **2011- 2013**
Supervisor: Marsha Black (retired)

- Assess the fathead minnow embryotoxicity of silver nanoparticles and determine the LC50 level and metabolomic responses with US Environmental Protection Agency, National Exposure Research Laboratory, Athens, GA.

Publications, Thesis, and Presentations

- **Tseng, C. Y.**, Salguero, J. A., Breidenbach, J. D., Rivera, E. S., Harvey, T., Solomon, E., Sanders, C. K., ... & Glaros, T. G. (2025) Evaluation of normalization strategies for mass spectrometry-based multi-omics datasets. *Metabolomics*. Submitted.
- Breidenbach, J. D., Rivera, E. S., Harvey, T., Mikolitis, A. S., **Tseng, C. Y.**, Sanders, C. K., ... & Glaros, T. G. (2024). The Addition of Transcriptomics to the Bead-Enabled Accelerated Monophasic Multi-Omics Method: A Step toward Universal Sample Preparation. *Analytical Chemistry*, *96*(46), 18343-18348.

- **Tseng, C. Y.**, Custer, C. M., Custer, T. W., Dummer, P. M., Karouna-Renier, N., & Matson, C. W. (2025). Integrated analysis and modelling of environmental mixtures and transcriptomic responses in tree swallow (*Tachycineta bicolor*) nestlings in the Great Lakes. Under review

- Rivera, E. S., LeBrun, E. S., Breidenbach, J. D., Solomon, E., Sanders, C. K., Harvey, T., **Tseng, C. Y.**, ... & Glaros, T. G. (2024). Feature-agnostic metabolomics for determining effective subcytotoxic doses of common pesticides in human cells. Toxicological Sciences, 202(1), 85-95.
- Breidenbach, J. D., Rivera, E. S., Harvey, T., Mikolitis, A. S., **Tseng, C. Y.**, Sanders, C. K., ... & Glaros, T. G. (2024). The Addition of Transcriptomics to the Bead-Enabled Accelerated Monophasic Multi-Omics Method: A Step toward Universal Sample Preparation. *Analytical Chemistry*, *96*(46), 18343-18348.
- Custer, C. M., Custer, T. W., Dummer, P. M., Schultz, S., Karouna-Renier, N., **Tseng, C. Y**., & Matson, C. W. (2024). Exposure to and biomarker responses from legacy and emerging contaminants along three drainages in the Milwaukee Estuary, Wisconsin, USA. Environmental Toxicology and Chemistry, 43(4), 856-877.
- **Tseng, C. Y.**, Custer, C. M., Custer, T. W., Dummer, P. M., Karouna-Renier, N., & Matson, C. W. (2023). Multi-omics responses in tree swallow (*Tachycineta bicolor*) nestlings from the Maumee Area of Concern, Maumee River, Ohio. *Science of The Total Environment*, *856*, 159130.
- Custer, C. M., Custer, T. W., Dummer, P. M., Schultz, S., **Tseng, C. Y.**, Karouna-Renier, N., & Matson, C. W. (2020). Legacy and contaminants of emerging concern in tree swallows along an agricultural to industrial gradient: Maumee River, Ohio. *Environmental Toxicology and Chemistry*, *39*(10), 1936-1952.
- **Tseng, C. Y.**, (2022) PRJNA835816: Tree swallows (*Tachycineta bicolor*) genome sequence and assembly, BioProject, NCBI
- **Tseng, C.-Y.** (2013). Effects of silver nanoparticles and hydroxylated fullerenes on early life stage of the fathead minnow (*pimephales promelas*): metabolomic approach. University of Georgia: Thesis, Fall 2013.
- Chen, P. J., Su, C. H., **Tseng, C. Y.**, Tan, S. W., & Cheng, C. H. (2011). Toxicity assessments of nanoscale zerovalent iron and its oxidation products in medaka (*Oryzias latipes*) fish. Marine pollution bulletin, 63(5-12), 339-346.

Platform & Poster Presentations

- **Tseng, C. Y.,** Rivera, E. S., Tipton, J. R., Glaros, T. G. Multi-omics Data Preprocessing and Functional Clustering. 2025 Conference on Data Analysis (CoDA).
- **Tseng, C. Y.**, Salguero, J. A., Breidenbach, J. D., Rivera, E. S., Harvey, T., Solomon, E., Sanders, C. K., ... & Glaros, T. G. The impact of omics normalization strategies affects the analysis of biological datasets. 2024 ASMS Conference on Mass Spectrometry and Allied Topics.
- **Tseng, C. Y.**, Custer, C. M., Custer, T. W., Dummer, P. M., Karouna-Renier, N., & Matson, C. W. Assessment of the transcriptome in tree swallow (*Tachycineta bicolor*) nestlings from the Great Lakes Maumee River Area of Concern. 2019 SETAC North America 40th Annual Meeting
- **Tseng, C. Y.**, Custer, C. M., Custer, T. W., Dummer, P. M., Karouna-Renier, N., & Matson, C. W. Assessment of the transcriptome in tree swallow (*Tachycineta bicolor*) nestlings from Great Lakes Areas of Concern. 2018 SETAC North America 39th Annual Meeting
- **Tseng, C. Y.**, Custer, C. M., Custer, T. W., Dummer, P. M., Karouna-Renier, N., & Matson, C. W. Assessment of gene expression and EROD activity in tree swallows (*Tachycineta bicolor*) nestlings collected from Great Lakes areas of concern. 2016 SETAC North America 37th Annual Meeting

Grants, Teaching and Outreach activities

- 20258379CT-IST: AI Agent for Streamlined Metadata Standardization for Enhanced Usability (2025) - $60K
- Glasscock Endowed Fund for Excellence in Environmental Sciences (2015) - $5,000
- Teach Intro-Environmental Analysis for one semester (2014).
- Serve as a judge for Central Texas Science and Engineering Fair (2017)
- Serve as a session teacher for Girl Scout STEMfest (2015)