SHANE K. CHU

555 Melville Ave, Apt 3N, St. Louis, MO 63130

+1 917-640-7301 ⋄ skchu@wustl.edu ⋄ kchu25.github.io

EDUCATION

Ph.D., Computer Science   Aug 2018 - Dec 2023 (Expected)
Washington University in St. Louis (WUSTL)
Advisor: Gary D. Stormo
Thesis: Interpretable Deep Learning via Sparse Representations for Protein-DNA Interactions.

Bachelor of Arts in Mathematics   Aug 2014 - May 2018
University of Kansas

WORK EXPERIENCE

Predoctoral Researcher (Machine Learning)   Jan 2019 - Dec 2023 (Expected)
WUSTL
- Designed and implemented interpretable deep learning models for biological sequence analysis.
  - Quantified a new framework that significantly accelerates training (approximately 5-10 minutes for tens of thousands of 100-200bp DNA sequences) and offers full interpretability on model parameters.
  - Uncovered new protein-DNA binding patterns in well-known databases such as JASPAR, FactorBook, and ReMap that traditional methods could not identify.
- Quantified a sequence-to-expression mapping framework through the utilization of sparse representations and Bayesian nonparametrics.

Machine Learning Research Intern   May 2023 - Aug 2023
Regeneron Pharmaceuticals
- Designed and implemented a Bayesian machine learning method for synthetic promoter DNA in gene therapy. This GPU-accelerated method efficiently generates approximately 1,000 DNA promoter candidates in five hours.

Graduate Research Assistant (Bioinformatics)   Jan 2019 - Present
WUSTL
- Implemented learning-based motif discovery method **MOTIFs.jl**, approximation algorithm **MotifPvalue.jl**, and random walk on de-Bruijn graph **SeqShuffle.jl** for bioinformatics applications.

Research Assistant (Machine Learning)   Nov 2017 - May 2018
Center for Remote Sensing of Ice Sheets (CReSIS)
- Conducted hyperparameter optimization for large-scale ice bottom detection models.
- Automated model selection through integrated parameter tuning subroutines.

Research Intern (Software Development)   May 2017 - Aug 2017
ZOLOZ
- Constructed parallel data processing pipelines for data collections.
- Integrated ZOLOZ's software development kit into web and mobile technologies.

TEACHING EXPERIENCE

Assistant instructor   Jan 2021 - Jun 2022
WUSTL
- Created lecture materials, led interactive recitations, and offered dedicated office hours for courses on mathematical optimization and Bayesian machine learning.

SELECTED PUBLICATIONS

Published:

- Chu, S. K., & Stormo, G. D. (2023). Finding motifs using DNA images derived from sparse representations. Oxford Bioinformatics, 39(6), btad378.
- Paden, J. D., Berger, V., Al-Ibadi, M., Chu, S. K., Xu, M., Crandall, D., & Fox, G. (2018, December). Subglacial bed topography using machine learning and geostatistical analysis applied to 2D and 3D radar sounding. In AGU Fall Meeting Abstracts (Vol. 2018, pp. C43A-07).
- Berger, V., Xu, M., Chu, S. K., Crandall, D., Paden, J., & Fox, G. C. (2018, July). Automated tracking of 2D and 3D ice radar imagery using VITERBI and TRW-S. In IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium (pp. 4162-4165). IEEE.

Preprints:

- Chu, S. K., & Stormo, G. D. (2022). Deep unfolded convolutional dictionary learning for motif discovery. bioRxiv, 2022-11.

In preparation:

- Chu, S. K., Stormo, G. D., "Interpretable sequence-to-expression map via sparse representations"

CONFERENCE PRESENTATIONS

- Poster, Machine Learning in Computational and Systems Biology, International Society for Computational Biology (ISMB)   July 2022

SOFTWARE ENGINEERING SKILLS

- Machine learning frameworks: PyTorch, Jax, Flux.jl
- Programming languages: Python, Julia, C/C++, Matlab, R, JavaScript, LaTeX.
- Data preprocessing: Pandas, DataFrames.jl
- Mathematical Tools: FFTW
- High-performance/parallel computing: CUDA.jl, Slurm
- Data visualizations: Matplotlib, Seaborn, Makie.jl, Luxor.jl, D3.js, HTML
- Pipeline Development: Snakemake

OPEN SOURCE CONTRIBUTIONS

Github: https://github.com/kchu25
Packages:

- **MOTIFs.jl** A motif discovery method that is capable of identifying long, gapped, and cooperative binding patterns, in addition to finding the primary binding sites.
- **SeqShuffle.jl** Shuffle strings such that it preserves the k-mer frequency in each string.
- **MotifPvalue.jl** An approximation algorithm that estimates the score thresholds of position weight matrices.
- **SimDNA.jl** Create DNA strings that simulate motifs with variable spacing, multimeric bindings, and multiple binding domains.

KEY COURSES

- Washington University in St. Louis: Large-scale optimization for data science, machine learning, Bayesian methods for machine learning, Bayesian inference, advanced algorithms, stochastic processes.
- University of Kansas: Operating systems, parallel computing, numerical methods, mathematical analysis.

LANGUAGES

English (Fluent) and Mandarin (Native).