# Jianan Canal Li

GitHub | LinkedIn | canallee.github.io | jl3789 [at] cornell.edu

## RESEARCH INTERESTS

Deep Learning, Generative Models, Bayesian Learning, Synthetic Biology, Protein Engineering

## EDUCATIONS

**Cornell University** Aug 2019 - Dec 2022
Bachelor of Science in Computer Science and Bioengineering, GPA: 4.04/4.3

**University of Illinois Urbana-Champaign** Aug 2018 - May 2019
Bioengineering, GPA: 3.83/4.0

## PUBLICATIONS

(* denotes equal contribution.)

- Jianan Canal Li, Tianhao Yu, Yunan Luo, Huimin Zhao. **"PseudoMSA: Towards High-fitness Protein Variant Generation Guided by Protein Language Models."** In Preparation. [project]
- Tianhao Yu*, Haiyang Cui*, Jianan Canal Li, Yunan Luo, and Huimin Zhao. **"Enzyme Function Prediction using Contrastive Learning."** Under Review of *Science*. [paper] [poster] [code]
- Tao Yu*, Wentao Guo*, Jianan Canal Li*, Tiancheng Yuan*, Christopher De Sa. **"MCTensor: A High-Precision Deep Learning Library with Multi-Component Floating-Point."** In *ICML 2022: Workshop on Hardware Aware Efficient Training*. [paper] [poster] [code]
- Jianan Canal Li*, Yimeng Zeng*, Wentao Guo*. **"Cyclical Kernel Adaptive Metropolis."** Under Review of *5th Symposium on Advances in Approximate Bayesian Inference (2023)*. [paper] [code]

## RESEARCH EXPERIENCE

### University of Illinois Urbana-Champaign May 2022 - present

***Research Intern***

- **PseudoMSA: Towards High-fitness Protein Variant Generation Guided by Protein Language Models (advised by Prof. Huimin Zhao and Prof. Yunan Luo)**
  PseudoMSA is a novel zero-shot pipeline that generates high-order, high-fitness protein variants with only the wild-type protein sequence as input. To evaluate the fitness of generated sequences, we trained probabilistic surrogate models on large-scale deep mutational scanning datasets. PseudoMSA can generate high-order mutants with comparable or better fitness to experimentally curated mutants, drastically reducing the cost of manual screening.

- **Enzyme Function Prediction using Contrastive Learning (advised by Prof. Huimin Zhao)**
  Proposed CLEAN, a contrastive learning algorithm for enzyme function annotation that outperforms all existing tools. CLEAN fine-tunes embeddings from protein language models with contrastive losses. Refactored existing code base to achieve 10x training and 20x inferencing speedup. Designed a new contrastive loss. Designed two enzyme function calling algorithms: Max-separation and p-value, and proved the optimality of Max-Separation (see [notes]).

**Cornell University** May 2020 - present

***Research Assistant***

- **MCTensor: A High-Precision Deep Learning Library with Multi-Component Floating-Point (advised by Prof. Christopher De Sa)**

  We developed MCTensor, the first library based on PyTorch for efficient high-precision computations. We implemented operators from basic summation to optimizer completely. MCTensor models in 16-bit float can match or outperform the PyTorch model with float32 or float64 precision. We implemented Hyperbolic Embedding MCTensor models, including Half Space and Poincaré Ball model, with improved performance on reconstructing *WordNet* mammals hierarchy.

- **DeepSudoku: inferencing combinatorial pooling for whole-genome knockout collection through self-supervised learning (Advised by Prof. Buz Barstow)**

  We proposed DeepSudoku, a new inferencing algorithm for Knockout Sudoku, a method for curating knockout collections through combinatorial pooling. DeepSudoku utilizes unambiguous location mappings from Next-Gen Sequencing through self-supervision and infers on ambiguous mappings. Compared to the original algorithm, DeepSudoku gives higher-confidence predictions and recovered 10% more location mappings on the *Shewanella oneidensis* knockout collection. [ELI Report]

- **MC3: Massively Capable Markov Chain Monte Carlo (Advised by Prof. Chris De Sa)**

  We developed MC3, a benchmark package for scalable Markov Chain Monte Carlo. We implemented highly optimized Julia code for gradient-based MCMC and Metropolis-Hastings algorithms such as SGLD, HMC, and Tuna-MH. We designed experimental suites for large-scale radio velocity datasets and Bayesian matrix factorization for protein-protein interaction datasets.

## AWARDS

- 2021 Engineering Learning Initiative Research Award, Cornell University
- Dean's List, Cornell University
- Dean's List, University of Illinois Urbana-Champaign
- Silver Prize (Team Leader), 2017 Intl. Genetically Engineered Machine (iGEM) Competition [wiki]

## TEACHING EXPERIENCE

- **BEE 3600 Molecular and Cellular Bioengineering** Fall 2020 - present

  Course developer and teaching assistant, supervised by Prof. Buz Barstow.
  Help re-design the entire curriculum, design Python-based homework, and hold office hours.

- **CS 4670 Introduction to Computer Vision** Spring 2022

  Teaching assistant, supervised by Prof. Bharath Hariharan.

## SELECTED COURSES

### CS Courses

- CS 6787 Advanced Machine Learning Systems
- CS 4740 Natural Language Processing
- CS 4670 Introduction to Computer Vision

### Biology Courses

- BIOMG 3300 Principles of Biochemistry
- BEE 4570 Biorobotics
- BEE 3600 Molecular and Cellular Bioengineering