

# Jianan Canal Li

 [GitHub](#) |  [LinkedIn](#) |  [canallee.github.io](https://canallee.github.io) |  [jl3789@cornell.edu](mailto:jl3789@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 AND MANUSCRIPTS

---

(\* denotes equal contribution.)

- [Jianan Canal Li](#). “**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.**” In review. [\[poster\]](#) [\[code\]](#)
- 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.**” Preprint. [\[paper\]](#) [\[code\]](#)

## RESEARCH EXPERIENCE

---

**Carl R. Woese Institute for Genomic Biology, UIUC**

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 [\[notes\]](#)).

**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 pretrained on unambiguous location mappings from NGS and infers on ambiguous mappings through transfer learning. DeepSudoku recovers ~10% more ambiguous mappings while predicting ~50% fewer false positives. [\[ELI Report\]](#)

- **MC3: A Scalable Software Toolkit for Massively Capable MCMC in the Biological Sciences (advised by Prof. Christopher 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.

**WET-LAB EXPERIENCE**

---

- **CRISPR-AID system and Chemoenzymatic catalysis** Fall 2018 - Fall 2019  
Advised by Prof. Huimin Zhao. Helped to mutate *Pichia pastoris* for CRISPR-AID system. Purified photo-active proteins for Chemoenzymatic tandem catalysis.
- **Site-directed Mitochondrial RNA editing with ADAR** Fall 2019 - Spring 2020  
Advised by Prof. Zhenglong Gu. Constructed plasmids to express ADAR (Adenosine deaminases acting on RNA) proteins with various mitochondrial targeting sequence.

**AWARDS**

---

- 2021 Engineering Learning Initiative Research Award, Cornell University
- Tau Beta Pi Engineering Honor Society
- 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 AND COMMUNITY SERVICE**

---

- **CS 4670 Introduction to Computer Vision** Spring 2022  
Teaching assistant, supervised by Prof. Bharath Hariharan.
- **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 homework, and hold office hours.
- **Volunteer college application tutor** Fall 2019 - Fall 2021  
Volunteer to advise college applicants with special circumstances.