Francesca-Zhoufan Li

AI for Science & Engineering

@ fzl@caltech.edu
📞 +1 510-708-9551
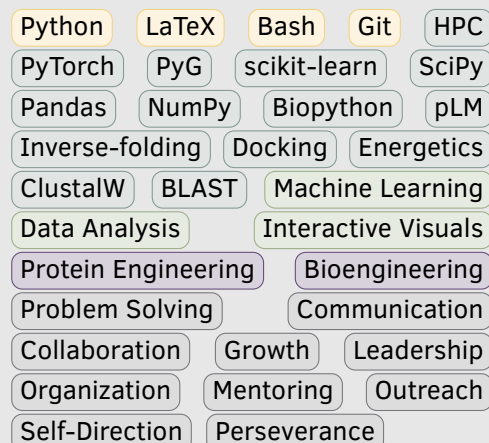in francescazhoufanli
github francescazfl
🌐 www.francescazfl.com
🏳 US Citizen, Shanghainese American

Goal

Advance and implement impactful, accessible ML tools with a focus on protein engineering, leveraging a computational skill set grounded in experimental experience, along with an interdisciplinary, adaptable, and collaborative mindset

Skills & Tools

Python LaTeX Bash Git HPC
PyTorch PyG scikit-learn SciPy
Pandas NumPy Biopython pLM
Inverse-folding Docking Energetics
ClustalW BLAST Machine Learning
Data Analysis Interactive Visuals
Protein Engineering Bioengineering
Problem Solving Communication
Collaboration Growth Leadership
Organization Mentoring Outreach
Self-Direction Perseverance

Selected Talks & Posters

- ICML, 2024
- ML Protein Engineering Seminar, 2024
- SynBioBeta, 2023
- Caltech Bioscience Futures Day, 2023
- Seagate-Minnesota AI/ML Virtual Distinguished Speaker Series, 2023
- Google Research invited talk, 2022

Other Experiences

- Co-organized the GEM Bio workshop at ICLR 2024 with over 108 paper submissions and 300 attendees
- Code with Young Legends: Led an introductory coding workshop
- i-STEM: Mentored under-invested high school students on research projects
- Bioengineering Honor Society: Mentored high school bioengineering research competitions, with one team winning 2nd place
- Biotech Connection Los Angeles: To grow the local biotech landscape
- Biology Scholars Program: To challenge who can do STEM

Education

09/20-Present · **Ph.D. in Bioengineering, GPA: 4.0** · California Institute of Technology
- NSF Graduate Research Fellowships Program • Amazon AI4Science Fellowship • Biotechnology Leadership Training Program

08/15-05/19 · **B.S. in Bioengineering, GPA: 3.96** · University of California, Berkeley
B.S. in Chemical Biology
- Highest Honors • Jack & Birthe Kirsch Prize • Tau Beta Pi Scholarship
- John Gorton Davis Scholarship • T. Dale Stewart Scholarship
- Genentech Outstanding Student Award Runner-Up

Industry & Academic Experience

06/22-09/22 · **BioML Research Intern** · Microsoft Research
- Performed a systematic analysis of protein language model transfer learning via 370 experiments across downstream tasks, architectures, model sizes, model depths, and pretraining time (ICML 2024)

01/21-Present · **Machine Learning for Proteins** · Arnold Lab & Yue Group, Caltech
- Developing zero-shot predictors for non-native enzyme activities
- Systematically analyzed multiple ML-assisted directed evolution strategies, including active learning and focused training using six distinct zero-shot predictors, across 16 protein fitness landscapes
- Contributed to the development of interactive visualization web app for rapid sequence-function data generation and analysis
- Provided structural and ligand docking insights to collaborators
- Led multiple grant writing efforts and cross-group collaborations
- Mentored 2 master's students and 1 undergrad on their theses

09/20-01/21 · **Extremophile Genetic Component Discovery** · Murray Lab, Caltech
- Constructed an RNA-seq analysis pipeline in R to discover novel genetic circuit components in non-canonical cell-free extracts
- Delivered results to academic and government organizations

06/20-08/20 · **RNA-Seq Sample Preparation Pipeline Optimization** · Zymergen
- Developed a Python package to design DNA oligos for RNaseH-based ribosomal RNA depletion for 8 strains in 7 programs
- Wrote R scripts to quality control and preprocess RNA extraction data from industry-standard electrophoresis instruments
- Delivered talks, collaborated within and outside the company

06/19-05/20 · **Bioinformatics Tool Development** · Koide Lab, NYU Langone Health
- Developed MATLAB software for SARS-CoV-2 mutation analysis from GISAID database, covering 25k global sequence entries
- Wrote user-friendly scripts to identify monobody and antibody complementary-determined region mutations for protein engineering
- Automated chromatogram visualization with customizable features

05/18-07/18 · **Cell-Free Platform Streamline** · Tierra Biosciences, QB3 Program
- Optimized non-standard protein production in cell-free expression systems using Design Of Experiments methodology

01/16-05/19 · **Independent Bioengineering Researcher** · Dueber Lab, UC Berkeley
- Automated time-course betaxanthin production analysis in MATLAB
- Engineered yeast to increase benzylisoquinoline alkaloids yield
- Improved beta-glucosidase stability & activity in a basic solution for indigo bio-production in *E. coli* via error-prone PCR libraries
- Honor thesis: A "Microbial Factory" Toolkit: Yeast Spheroplast Transformation Method Development for CRISPR-Cas9 Multiplexing

Featured Publications

2024 — **Li F-Z**, *et al*. Evaluation of Machine Learning-Assisted Directed Evolution Across Diverse Combinatorial Landscapes. *bioRxiv*.

2024 — **Li F-Z**, *et al*. Feature Reuse and Scaling: Understanding Transfer Learning with Protein Language Models. *PMLR*. **235**, 27351-27375.

2024 — Yang J, **Li F-Z**, & Arnold FH. Opportunities and Challenges for Machine Learning-Assisted Enzyme Engineering. *ACS Cent. Sci.* **10**, 226–241.

2024 — Long Y, Mora A, Gürsoy E, Johnston KE, **Li F-Z**, & Arnold FH. LevSeq: Rapid Generation of Sequence-Function Data for Directed Evolution and Machine Learning. *bioRxiv*.

2023 — Yang J, Ducharme J, Johnston KE, **Li F-Z**, *et al*. DeCOIL: Optimization of Degenerate Codon Libraries for Machine Learning-Assisted Protein Engineering. *ACS Synth. Biol.* **12**, 2444-2454.