Kevin Kaichuang Yang

614 260 9454 | yangkky@gmail.com | https://yangkky.github.io

EDUCATION

PhD in Chemical Engineering	California Institute of Technology	2018
B.S. in Chemical Engineering	Ohio State University	2011

RESEARCH EXPERIENCE

Senior Researcher, BioML

Microsoft Research New England, Cambridge, MA April 2020 – Present

- Develop methods for protein sequence pretraining and generation
- Develop benchmarks for protein function prediction
- Lead project teams that include interns, data scientists, and researchers

Machine Learning Scientist

Generate Biomedicine, Cambridge, MA January 2019 – April 2020

- Develop methods for quantifying the uncertainty of protein function predictions
- Develop methods for model-based optimization of protein function
- Build and train a language model for proteins that can be fine-tuned for specific prediction tasks.

Graduate Research Assistant, Professor Frances Arnold's Group

California Institute of Technology, Pasadena, CA August 2014 – December 2018

- Developed batched stochastic Bayesian optimization method for designing site-saturation libraries
- Developed interpretable and accurate Gaussian process kernels tailored to biological sequences
- Used Gaussian process models (github.com/yangkky/gpmodel) to design channelrhodopsins
- Designed embedded representations of protein sequences based on doc2vec to streamline machine-learning pipelines (github.com/fha_lab/embeddings_reproduction)
- Built neural machine translation models to predict signal peptides from protein sequences

Computational Intern

Ambry Genetics, Aliso Viejo, CA June 2017 – September 2017

- Developed and implemented neural network models in Keras and PyTorch to predict outcomes of genetic variation by transferring information across paralogous proteins
- Incorporated model into a pipeline that finds paralogs for variants of interest and then predicts outcomes

SELECTED PUBLICATIONS AND PRESENTATIONS

Invited Talks

- Biomedical Informatics Seminar, Columbia University. October 2022.
- Bioinformatics Student Symposium, Boston University. May 2022.
- Molecular Maker Lab Institute Conference, UIUC. March 2022.
- AI4Science Seminar, Chalmers University. March 2022.
- Protein Engineering Congress Global. October 2021.
- Janelia Research Center. May 2019.

Peer-Reviewed Papers

Evolutionary velocity with protein language models. Brian L. Hie, **Kevin K. Yang**, and Peter S. Kim. Cell Systems, 2022. 10.1016/j.cels.2022.01.003

FLIP: Benchmark tasks in fitness landscape inference for proteins. Christian Dallago, Jody Mou, Kadina E. Johnston, Bruce J. Wittmann, Nicholas Bhattacharya, Samuel Goldman, Ali Madani, **Kevin K. Yang**. NeurIPS 2021 Datasets and Benchmarks Track. 10.1101/2021.11.09.467890

Signal Peptides Generated by Attention-Based Neural Networks. Zachary Wu, **Kevin K. Yang**, Michael J. Liszka, Alycia Lee, Alina Batzilla, David Wernick, David P. Weiner, and Frances H. Arnold. ACS Synthetic Biology, 10 July 2020. 10.1021/acssynbio.0c00219

Machine learning-guided channelrhodopsin engineering enables minimally-invasive optogenetics. Bedbrook CN, **Yang KK**, Robinson JE, Gradinaru V, Arnold FH. Nature Methods, October 14, 2019. 10.1038/s41592-019-0583-8.

Machine-learning-guided directed evolution for protein engineering. **Yang KK**, Wu Z, Arnold FH. Nature Methods, July 15, 2019. 10.1038/s41592-019-0496-6.

Learned protein embeddings for machine learning. **Yang KK**, Wu Z, Bedbrook CN, Arnold FH. Bioinformatics. 23 March 2018. 10.1093/bioinformatics/bty178.

Machine learning to predict eukaryotic expression and plasma membrane localization of engineered integral membrane proteins. Bedbrook CN*, **Yang KK***, Rice AJ, Gradinaru V, Arnold FH. PLOS Computational Biology 13(10): e1005786 (2017). 10.1371/journal.pcbi.1005786.

Preprints

Protein structure generation via folding diffusion. Kevin E. Wu, **Kevin K. Yang**, Rianne van den Berg, James Y. Zou, Alex X. Lu, Ava P. Amini

Deep self-supervised learning for biosynthetic gene cluster detection and product classification. C Rios-Martinez, N Bhattacharya, AP Amini, L Crawford, **KK Yang**.

Masked inverse folding with sequence transfer for protein representation learning. **Kevin K. Yang**, Niccolò Zanichelli, Hugh Yeh.

Convolutions are competitive with transformers for protein sequence pretraining. **Kevin K. Yang**, Alex X. Lu, Nicolo Fusi.