Kevin Kaichuang Yang

Senior Researcher, BioML
Microsoft Research New England

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

EDUCATION

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

INVITED TALKS

- Biomedical Informatics Seminar, Columbia University. October 2022.
- Bioinforange. October 2022.
- Merck. August 2022.
- Journal Club, OpenBioML. May 2022.
- Genomics Seminar, UC Riverside. May 2022.
- AI for Drug Discovery Seminar, University of Washington. May 2022.
- Bioinformatics Student Symposium, Boston University. May 2022.
- Molecular Maker Lab Institute Conference, UIUC. March 2022.
- AI4Science Seminar, Chalmers University. March 2022.
- Machine learning seminar, IBM. December 2021.
- Protein Engineering Congress Global. October 2021.
- GSK Data Forum. Feb 2021.
- Boğaziçi University Biotech Conference. Jan 2021.
- Ohio State University Society for Biological Engineering. Nov 2020.
- Janelia Research Center. May 2019.
- Gray-Hill Lecture, Occidental College, June 2018.

PUBLICATIONS

Peer-Reviewed Papers

[*co-first authors]

17. Exploring evolution-based &-free protein language models as protein function predictors. Mingyang Hu, Fajie Yuan, **Kevin K. Yang**, Fusong Ju, Jin Su, Hui Wang, Fei Yang, Qiuyang Ding. NeurIPs 2022, to appear.
16. Randomized gates eliminate bias in sort-seq assays. Brian L. Trippe, Buwei Huang, Erika A. DeBenedictis, Brian Coventry, Nicholas Bhattacharya, **Kevin K. Yang**, David Baker, Lorin Crawford. Protein Science, 2022.
15. 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
14. Machine learning modeling of family wide enzyme-substrate specificity screens. Samuel Goldman, Ria Das, **Kevin K Yang**, Connor W Coley. PLoS computational biology, 2022. 10.1371/journal.pcbi.1009853

13. A topological data analytic approach for discovering biophysical signatures in protein dynamics. Wai Shing Tang, Gabriel Monteiro da Silva, Henry Kirveslahti, Erin Skeens, Bibo Feng, Timothy Sudijono, **Kevin K. Yang**, Sayan Mukherjee, Brenda Rubenstein, Lorin Crawford. PLoS computational biology, 2022. 10.1371/journal.pcbi.1010045

12. Adaptive machine learning for protein engineering. Brian L. Hie and **Kevin K. Yang**. Current Opinion in Structural Biology, 2022. 10.1016/j.sbi.2021.11.002

11. 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

10. Protein sequence design with deep generative models. Zachary Wu, Kadina E. Johnston, Frances H. Arnold, and **Kevin K. Yang**. Current Opinion in Chemical Biology, 2021. 10.1016/j.cbpa.2021.04.004

9. Learned embeddings from deep learning to visualize and predict protein sets. Christian Dallago, Konstantin Schütze, Michael Heinzinger, Tobias Olenyi, Maria Littmann, Amy X. Lu, **Kevin K. Yang**, Seonwoo Min, Sungroh Yoon, James T. Morton, Burkhard Rost. Current Protocols, May 2021. 10.1002/cpz1.113

8. 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

7. 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.

6. 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.

5. Batched stochastic Bayesian optimization via combinatorial constraints design. **Yang KK**, Chen Y, Lee A, Yue Y. AIStats 2019. arxiv.

4. The Generation of Thermostable Fungal Laccase Chimeras by SCHEMA-RASPP Structure-Guided Recombination in Vivo. Mateljak I, Rice A, **Yang KK**, Tron T, Alcalde M. ACS Synthetic Biology, March 21, 2019. 10.1021/acssynbio.8b00509

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

2. 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.

1. Structure-Guided SCHEMA Recombination Generates Diverse Chimeric Channelrhodopsins. C. N. Bedbrook, A. J. Rice, **K. K. Yang**, X. Ding, S. Chen, E. M. LeProust, V. Gradinaru, F. H. Arnold. Proceedings of the National Academy of Sciences 114, E2624-E2633 (2017). 10.1073/pnas.170026911.

Preprints

4. 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

3. Deep self-supervised learning for biosynthetic gene cluster detection and product classification. C Rios-Martinez, N Bhattacharya, AP Amini, L Crawford, **KK Yang**.
2. Masked inverse folding with sequence transfer for protein representation learning. **Kevin K. Yang**, Niccolò Zanichelli, Hugh Yeh.
1. Convolutions are competitive with transformers for protein sequence pretraining. **Kevin K. Yang**, Alex X. Lu, Nicolo Fusi.

HONORS, AWARDS, AND FELLOWSHIPS

2017	Caltech Chemistry and Chemical Engineering Teaching Assistantship Award
2015	Caltech Biotechnology Leadership Program
2011	National Science Foundation Graduate Research Fellowship

TEACHING AND MENTORING

Teaching Assistantships

2018	Caltech ChE/BE 163 Introduction to biomolecular engineering
2016	Caltech ChE/BE 163 Introduction to biomolecular engineering
2015	Caltech ChE 101 Chemical reaction engineering

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 Biomedicines, 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)

- Used Gaussian process models (github.com/yangkky/gpmodel) to design channelrhodopsins with improved properties
- 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 on PyTorch to predict signal peptides from their corresponding mature 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 uses paralogs and model to predict variant outcomes