Myungsun Kang

sunny.myungsun.kang@gmail.com
https://www.linkedin.com/in/myungsun-sunny-kang/

Computational scientist with expertise in developing large language models, geometric neural networks, molecular dynamics simulations, and quantum mechanics calculations for designing small and large molecule drugs in structure-based drug discovery

RECENT EXPERIENCE

2023 Jun- **Computational Chemist II**, Relay Therapeutics

- Contributed to identifying and confirming hits for two preclinical stage projects through combined use of MD simulations for understanding protein motions, large scale virtual screening, AI-based structure prediction, protein design and SAR analysis.
- Elucidated binding pockets by leveraging enhanced sampling methods for a target with no prior reported compound bound structure. The binding pocket later validated in in-house x-ray crystal structure
- Built de-novo protein and peptide design and optimization workflow and platform, routinely used for large tool molecules and protein targets
- Contributed for building new virtual screening pipeline encompassing large scale MD simulations including ligand stability simulation and enhanced sampling methods
- Built statistical analysis pipeline for NGS sequencing dataset obtained from phage display library

2021 Feb- **Principal Machine Learning Scientist,** Immuneering Corporation

- Benchmarked and deployed AI algorithms for learning meaningful representations of proteins and molecules to enhance hit-identification and lead optimization of in-house oncology drug programs
- Built a chemical property prediction model by training representation of chemical compound library and leveraging fine-tuning on property dataset
- Productionalized upgrades for the in-house hit-identification AI algorithm by overhauling the codebase with newer version of Tensorflow and migrating the pipeline to AWS ecosystem

2022 Jan- **Core contributor/Scientist, BigBio (BigScience, HuggingFace)**

- Contributed to the creation of BigBio - an open library of more than 120 biomedical dataloaders build using HuggingFace's datasets library for applications in NLP by designing harmonized dataset schemas by task type, writing and reviewing several data-loaders, which led to a publication in ACL
- Contributed to training and evaluating a large -scale Multi-task learning model on 106 different bioNLP tasks with fine-tuning (work accepted for NeuralIPS 2022)
- Contributed to release of BLOOM, the world's largest open multi-lingual language model

2020 Apr- Oct **Core contributor/Scientist, PathCheck Foundation (MIT)**

- Built 1D Convolutional neural networks that predict the distance between two devices from the time series records of Bluetooth and various sensors
- Took the third place in the competition held by NIST and the accompanying paper was accepted to a workshop in Neurips 2020

2019 Mar- **Senior data scientist**, Wayfair, Algorithms and Analytics

- Evaluated the efficacy of a new logistics strategy implemented for shipping by leveraging techniques in causal inference
- Upgraded a major machine learning pipeline that predicts Wayfair's profitability for   million orders per day by leveraging HIVE, PySpark and Airflow
- Built a machine learning models that predict a special category of Wayfair's product, which increased accuracy by 65 percent in comparison to the previous model

2018 Sep-Nov **Fellow**, Insight Data Science

- Built a web app that recommends anti-depressants pills tailored to patients' symptoms
- Extracted side effects from patient survey data using topic modeling, and infered their prevalence from the corpus of patient experience
- Engineered features from 17,000 subreddit comments collected through Pushshift Reddit API using NLP methods including TF-IDF, word2vec and sentiment analysis
- Built a recommendation system that is trained on words associated with positive experiences for a given drug. Employed Logistic regression and linear SVC for the classification
- Built an inter-active user interface with Flask, Bootstrap and AWS

2012 -2019 Feb **Graduate researcher**, Chemical Engineering, MIT

- Built modeling and analytical tools to predict immune response from HIV vaccine prototypes
- Employed non-linear regression to built a time-dependent deterministic nonlinear differential equation model in Matlab, which can predict serum Ab production upon vaccination
- Built a stochastic model of Ab response by implementing Tau-leap gillespie algorithm with partial deterministic approximation, which led to 50X speed enhancement with <1 percent accuracy tradeoff
- Simulated HIV vaccine prototypes and proposed prospective candidates, which is now being tested in non-human primates

2014 Feb-May **Consultant**, Cabot and SGCEnergia, David Koch School of Chemical Engineering Practice, MIT

- Completed two month-long projects at each of Cabot and SGCEnergia.
- For each project: prepared three formal talks, a proposal, and a final report that communicated progress to project managers.
- Improved the silica treatment process and the graphene manufacturing process (Cabot).
- Designed and improved "Fischer-Tropsch" product upgrading and reactor modeling process (SGCEnergia).

EDUCATION

2012 –2019 **Massachusetts Institute of Technology (MIT)**
Doctor of Philosophy candidate, Chemical Engineering, Institute for Medical Engineering & Science
(Minor: Statistics and Computer Science)
Master of Science in Chemical Engineering Practice, Chemical Engineering

2008–2012 **Korea Advanced Institute of Science and Technology (KAIST)**
Bachelor of Science, Chemical Engineering, Minor: Biology, *summa cum laude*

PUBLICATIONS

BigScience Workshop, "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model", arXiv (2022), **1403 citations**

Jason Fries, Leon Weber, Natasha Seelam, Gabriel Altay, Debajyoti Datta, Samuele Garda, **Sunny Kang** and other authors, "Bigbio: A framework for data-centric biomedical natural language processing", Advances in Neural Information Processing Systems (2022), **44 citation**

Jason Alan Fries, Natasha Seelam, Gabriel Altay, Leon Weber, **Myungsun Kang** and other authors "Dataset debt in biomedical language modeling, Challenges and Perspectives in Creating Large Language Models" ACL (2022), **8 citations**

Ramesh Raskar Sheshank Shankar, Rishank Kanaparti, Ayush Chopra, Rohan Sukumaran, Parth Patwa, **Myungsun Kang** and other authors "Proximity Sensing: Modeling and Understanding Noisy RSSI-BLE Signals and Other Mobile Sensor Data for Digital Contact Tracing", Machine Learning for Mobile Health workshop at NeurIPS (2020), **21 citations**

Rohan Sukumaran, Parth Patwa, TV Sethuraman, Sheshank Shankar, Rishank Kanaparti, Joseph Bae, Yash Mathur, Abhishek Singh, Ayush Chopra, **Myungsun Kang** and other authors "COVID-19 outbreak prediction and analysis using self reported symptoms", Journal of Behavioral Data Science (2021), **2 citations**

Hok Hei Tam*, Mariane B. Melo*, **Myungsun Kang*** and other authors, "Sustained antigen availability during Germinal Center initiation enhances antibody responses to vaccination", Proceedings of the National Academy of Sciences, 201606050 (*equal contributors) **401 citations**

Kang M, Eisen TJ, Eisen EA, Chakraborty AK, Eisen HN (2015), "Affinity inequality among serum antibodies that originate in lymphoid Germinal Centers", PLoS ONE 10 (10): e0139222. doi:10.1371/journal.pone.0139222 **17 citations**

Brett Hall Praveen Nair, Jason Funt, Sarah Kolitz, Jan de Jong, Peter King, Amy Yamamura, Mai Johnson, **Myungsun Kang** and other authors , "Humanized 3D tumor models that are mutually aligned with AACR GENIE patients predict IMM-1-104 activity in RAS-addicted tumors", American Association for Cancer Research (2023)

SKILLS

, Technical expertise	Discrete stochastic simulation, Machine learning, Causal inference, NLP, Deep Learning, Exploratory statistics, Inferential statistics, Bayesian statistics
Programming Languages	Python, PySpark, Pytorch, Tensorflow, Keras, SciPy, NumPy, Pandas, Seaborn, StochPy, C, Bash
Query Languages	SQL, HIVE
Applications	Git (source control), Docker, Airflow, BigQuery, GCP
Languages	English and Korean (fluent), Japanese (basic working proficiency)