Yan Chak (Richard) Li

✉ richardpokard@gmail.com | ⌂ huhrichard.github.io/ | github.com/huhrichard | linkedin.com/in/yan-chak-li-865b6b124/

Profile

Data-driven researcher and developer in machine learning for biomedical data, driving clinical diagnosis and biomedical discovery through innovative solutions. Expertise in designing and evaluating automated methods for complex data analysis, including mass spectrometry, medical imaging, protein function prediction and disease outcome prediction. Proficient in transforming research into actionable insights via user-friendly web applications and data visualization tools. Skilled in Python, R, SQL, Vue.js, and cloud platforms, with a passion for harnessing technical expertise to advance biomedical and machine learning research and applications.

Work Experience

Icahn School of Medicine at Mount Sinai

New York City, U.S.A.

Data Science Analyst I

Nov 2025 - Now

- Analyze clinical cohorts by machine learning and statistical techniques
- Build data visualization portal for clinical cohort on different omics data

Icahn School of Medicine at Mount Sinai

New York City, U.S.A.

Bioinformatician

Nov 2019 - Nov 2024

- Develop ensemble machine learning methods for multimodal biomedical data
- Analyze clinical cohorts by machine learning techniques
- Build data visualization portals for clinical cohorts with multi-omics data
- Teaching Assistant of 'Machine Learning for Biomedical Data Science' (Spring 2020 & 2021)

Education

The Hong Kong University of Science and Technology

Clear Water Bay, Hong Kong

M.Phil. in Bioengineering

Sep 2017 - Aug 2019

- Thesis: Deep Learning Enables Instance Edge Detection of Vertebral Bodies on X-ray Images
- Teaching Assistant of IELM/IEDA 2100E
- Courses: Computer Vision, Mathematical Foundations of Imaging, Topological and Geometric Data Reduction and Visualization etc.

The Hong Kong University of Science and Technology

Clear Water Bay, Hong Kong

B.Eng. in Computer Engineering

Sep 2013 - Aug 2017

- Undergradudate Research Project: Improving the Efficiency of Spectral Library Searching in Mass Spectrometric Data Analysis
- Courses: Introduction to Bioinformatics Algorithms, Medical Imaging, Heterogeneous Parallel Programming etc.

Projects

ProTrack: Melanoma PTRC - multi-omics data visualization across multiple cohorts

Ongoing

ProTrack is a series of web data visualization portals for different clinical studies. It supports advanced queries, visualization, and downloads of multi-omics data from comprehensive proteogenomic studies by the Clinical Proteomic Tumor Analysis Consortium (CPTAC).

SunBEAm-ABC web portal - multi-omics data visualization

Link: public version, beta version

The SunBEAm Analysis & Bioinformatics Center (SunBEAm-ABC) is assaying biosamples using omics and will apply integrative systems biology to identify novel determinants of food allergy and atopic dermatitis. The web portal aims to provide data visualization of multi-omics data via different types of plots, such as boxplot, barplot, heatmap, etc. On top of it, we also build network visualization, which allows users to upload & download network data to explore their interests.

KiNet - Kinase-Substrate Interaction Network Visualization

Link: Paper, KiNet website

The KiNet web portal aggregates and visualizes the network of interactions between protein-kinases and their substrates in the human genome. Each tab provides different ways to select proteins and display the known kinase-substrate interactions between them. We also provided detailed information on interactions by selecting the edges.

POND - Prediabetes/diabetes youth ONline Dashboard

Link on shinyapps.io / hpc.mssm.edu

POND is an interactive dashboard for exploring factors associated with prediabetes and diabetes mellitus (preDM/DM) among youth (aged 12-19 years) in the United States. Raw data were obtained from the National Health and Nutrition Examination Survey (NHANES) and processed into a multi-domain dataset that is the foundation of our study and this portal.

Ensemble Integration - multimodal machine learning

Link: Paper / eipy python package documentation

Ensemble Integration (EI, ensemble-integration/eipy as a python package) is a multimodal machine learning package for generating diverse ensembles of heterogeneous classifiers, as well as the accompanying metadata needed for ensemble learning approaches utilizing ensemble diversity for improved performance.

Data-driven ExposurE Profile (DEEP) - feature combination extraction from tree-based models

Link: Paper / github repository

DEEP uses the XGBoost algorithm to identify air toxic combinations associated with health outcomes. The combinations identified using XGBoost were then adjusted for potential confounders to identify early-life multi-air toxic combinations.

Identifying clinical features of COVID-19 mortality

Link: Paper / github repository

We developed a machine learning model to predict COVID-19 mortality using clinical data from a large cohort of patients treated at Mount Sinai Health System. The model trained on data from 3,841 patients, achieved high accuracy (AUC=0.91) in predicting mortality when tested on retrospective and prospective datasets. The model relies on just three clinical features: patient age, minimum oxygen saturation during their medical encounter, and type of patient encounter (inpatient vs outpatient/telehealth).

Publications

A web portal for exploring kinase-substrate interactions *npj Systems Biol. and App.*

Sekar JAP, **Li YC**, Schlessinger A, Pandey G *2024*

Link: KiNet - web portal, paper, github repository

A comprehensive exploration of the druggable conformational space of protein kinases using AI-predicted structures *PLoS Comput. Biol.*

Herrington NB, **Li YC**, Stein D, Pandey G, Schlessinger A *2024*

Link: paper

A comprehensive youth diabetes epidemiological dataset and web portal: Resource Development and Case Studies *JMIR Public Health Surveill*

McDonough C, **Li YC**, Vangeepuram N., Liu B., Pandey G. *2024*

Link: POND - web portal, paper

Multi-omic integration reveals alterations in nasal mucosal biology that mediate air pollutant effects on allergic rhinitis *Allergy*

Irizar H, Chun Y, Hsu HHL, **Li YC**, Zhang L, Arditi Z, Grishina G, Grishin A, Vicencio A, Pandey G, Bunyavanich S *2024*

Link: paper

Machine learning-driven identification of air toxic combinations associated with asthma symptoms among elementary school children in Spokane, Washington, USA *Science of The Total Environment*

Amiri S, **Li YC**, Buchwald D, Pandey G *2024*

Link: paper

eipy: An Open-Source Python Package for Multi-modal Data Integration using Heterogeneous Ensembles *arXiv*

Bennett JJR, **Li YC**, Pandey G *2024*

Link: eipy package, preprint

Developing better digital health measures of Parkinson's disease using free living data and a crowdsourced data analysis challenge *PLOS Digital Health*

Sieberts SK, Borzymowski H, Guan Y, Huang Y, Matzner A, Page A, … , **Li YC**, … , Stanescu A, … , Pandey G, Shawen N, Synder P, Omberg L *2023*

Link: paper

Integrating multimodal data through interpretable heterogeneous ensembles *Bioinformatics Advances*

Li YC, Wang L, Law JN, Murali TM, Pandey G *2022*

link: Paper, github repository

Machine learning-driven identification of early-life air toxic combinations associated with childhood asthma outcomes *Journal of Clinical Investigation*

Li YC, Hsu HL, Chun Y, Chiu PH, Arditi Z, Claudio L, Pandey G, Bunyavanich S *2021*

Link: paper, github repository

Clinical features of COVID-19 mortality: development and validation of a clinical prediction model *Lancet Digital Health*
Yadaw AS, **Li YC**, Bose S, Iyengar R, Bunyavanich S, Pandey G *2020*
Link: paper, github repository

Conference Presentations

Integrating multimodal data through interpretable heterogeneous ensembles *Madison, Wisconsin, U.S.A.*
Li YC, Wang L, Law J, Murali TM, Pandey G *Jul 2022*
Oral and poster present at The 30th Conference on Intelligent Systems for Molecular Biology (ISMB)

Automatic Instance-edge Detection Network (AID-Net) - Vertebral Edge Detection by Deep Learning *Coimbra, Portugal*
Li RYC, Chin NJW, Wang Y, So RHY *May 2019*
Oral present at European Society for Clinical Investigation Congress (ESCI Congress) 2019

Fast Similarity Measure of SWATH-MS by Cosine Similarity of Random Pairs (CS-RP) *Biopolis, Singapore*
Li YC, Wu L, Lam H *Dec 2017*
Oral present at Asia Oceania Mass Spectrometry Conference (AOMSC) 2017

Skills

Data science	Python: Pandas, NumPy, Scikit-learn, PyTorch, BeautifulSoup, OpenCV, Keras, joblib, graphviz, statsmodels; R: statistical analyses, ggplot2, plotly, vistnetwork; Tableau, SQL
Web development	R Shiny, Vue.js, Flask, Django, D3.js, Plotly.js, Firebase, Google Analytics, JavaScript, HTML
Other computing skills	High-performance computing, AWS, Google Cloud, Oracle Cloud, Linux, LaTeX, Git, CUDA C, C, C++, Java

Awards

2017	**Young Scientist Travel Award**, Asia Oceania Mass Spectrometry Conference 2017	*Singapore*
2013	**Dean of Engineering Scholarship**, HKUST	*Hong Kong*