Guojie Zhong

PhD candidate, Department of Systems Biology, Columbia University
1130 St. Nicholas Avenue, New York, NY, USA | Phone: (+01)7187305993 | Email: gz2294@cumc.columbia.edu
Website: https://zhongguojie1998.github.io/GuojieZhong/

RESEARCH INTERESTS

I create deep learning methods to understand the mechanisms of variants effects on protein functions utilizing state-of-the-art protein language models and geometry-aware transformers. I am particularly interested in its application to two fields: Genetics of missense variants in human diseases and directed-evolution based protein design problems.

PUBLICATIONS

- **Zhong, G.**, Choi, Y.A. & Shen, Y. VBASS enables integration of single cell gene expression data in Bayesian association analysis of rare variants. *Commun Biol* **6**, 774 (2023).
- Zhao, Y., **Zhong, G.**, Hagen, J., Pan, H., Chung, W.K., and Shen, Y. A probabilistic graphical model for estimating selection coefficient of nonsynonymous variants from human population sequence data. *medRxiv*, (2023).
- **Zhong, G.**, and Shen, Y. Statistical models of the genetic etiology of congenital heart disease. *Curr Opin Genet Dev* **76**, 101967 (2022).
- Edwards, N., **Zhong, G.**, Ahimaz, P., Kenny, A., Kingma, P., Wells, J., Shen, Y., Chung, W., and Zorn, A. Discovering the Developmental Basis of Trachea-Esophageal Birth Defects: Evidence for Endosome-opathies. *The FASEB Journal* **36** (2022).
- **Zhong, G.**, Ahimaz, P., Edwards, N.A., Hagen, J.J., Faure, C., Lu, Q., Kingma, P., Middlesworth, W., Khlevner, J., El Fiky, M., et al. Identification and validation of candidate risk genes in endocytic vesicular trafficking associated with esophageal atresia and tracheoesophageal fistulas. *HGG Adv* **3**, 100107 (2022).
- Wang, Y., Tiruthani, K., Li, S., Hu, M., **Zhong, G.**, Tang, Y., Roy, S., Zhang, L., Tan, J., Liao, C., et al. mRNA Delivery of a Bispecific Single-Domain Antibody to Polarize Tumor-Associated Macrophages and Synergize Immunotherapy against Liver Malignancies. *Adv Mater* **33**, e2007603 (2021).
- Ren, X.#, **Zhong, G.**#, Zhang, Q., Zhang, L., Sun, Y., and Zhang, Z. Reconstruction of cell spatial organization from single-cell RNA sequencing data based on ligand-receptor mediated self-assembly. *Cell Res* **30**, 763-778 (2020). (#, Contributed Equally)
- Zhang, Q., He, Y., Luo, N., Patel, S.J., Han, Y., Gao, R., Modak, M., Carotta, S., Haslinger, C., Kind, D., Peet G.W., **Zhong, G.**, Lu, S., Zhu, W., Mao, Y., Xiao, M., et al. Landscape and Dynamics of Single Immune Cells in Hepatocellular Carcinoma. *Cell* **179**, 829-845 e20 (2019).

CONFERENCES

- **Zhong, G.**, and Shen, Y. (2023). Predicting mode of action of missense variants by graph representation of protein structural context. *American Society of Human Genetics 2023 Annual Meeting, Washington, D.C.*
- **Zhong, G.**, and Shen, Y. (2022). Representation of missense variants for predicting modes of action. *Machine Learning in Structural Biology, Workshop at the 36th Conference on Neural Information Processing Systems.*
- **Zhong, G.**, Choi, Y.A., and Shen, Y. (2022). Integration of gene expression data in Bayesian association analysis of rare variants. *American Society of Human Genetics 2022 Annual Meeting,* Los Angelas, CA.
- Sewda, A., **Zhong, G.**, Chung, W. and Shen, Y. (2022). Enrichment patterns of damaging de novo variants in patients with isolated and complex congenital heart disease. *Bench to Bassinet (B2B; PCGC) Program Annual Meeting*, Arlington, VA.
- **Zhong, G.**, Wang, J., He, S. and Fu, X. (2021). Towards better understanding of developmental disorders from integration of spatial single-cell transcriptomics and epigenomics. *The 2021 ICML Workshop on Computational Biology.*

EDUCATION

Department of Systems Biology, **Columbia University**, New York, NY, USA. *2019.08 – present*
PhD in Systems Biology, Cellular, Molecular and Biomedical Sciences program
Thesis Advisor: Yufeng Shen

Department of Integrated Biology, **University of California, Berkeley**, CA, USA. *2017.08 – 2017.12*
Exchange Student in Berkeley Bioscience Study Abroad Program (BBSA)

Yuanpei College, **Peking University**, Beijing, China. *2015.09 – 2019.07*
Bachelor of Science in Integrated Science Program
Thesis Advisor: Zemin Zhang
Thesis: Reconstruction of cell spatial organization from single-cell RNA sequencing data based on ligand-receptor mediated self-assembly

AWARDS

Dean's fellowship, Graduate School of Arts and Science, Columbia University 2019

2014-2015 academic year Outstanding Freshman Scholarship, Peking University 2015
The 28th National Olympiad in Chemistry in Provinces, 1st prize, Zhejiang Province, China 2014

RESEARCH EXPERIENCES

PhD Student, Department of Systems Biology, Columbia University
Advisor: Dr. Yufeng Shen

PreMode: Predict mode of action of missense variants by graph representation of protein sequence and structural context. *2022.02-present*

- Accurate prediction of the functional impact of missense variants is one of the bottlenecks in discovering genetic causes of diseases and implementing genomic medicine. Current methods focused on generating an binary prediction score to distinguish pathogenic and benign variants while overlooking the fact that variant effect should be multi-dimensional. Pathogenic missense variants in the same gene may act through different modes of action, such as loss of folding stability, binding affinity or enzymatic activity at molecular level, or gain or loss of functions (G/LoF) at genetics level.
- Defined several benchmarking mode-of-action prediction tasks at both molecular and genetics levels with data from clinical and experimental measurements of variant effects.
- Developed PreMode, a foundation method to learn universal representation of sequence variation from protein context utilizing pre-trained protein language models, protein structures and multiple sequence alignments.
- Demonstrated the utility of PreMode in predicting a range of benchmarking tasks of several protein and protein families through transfer learning, where it achieved the state-of-the-art performances.

Integration of gene expression data in Bayesian association analysis of rare variants. *2019.08-2022.02*

- The statistical power to identify risk genes by rare *de novo* variants is generally low due to rarity of genotype data. Previous studies have shown that disease risk genes usually have high expression in relevant cell types, although for many diseases the identity of these cell types are largely unknown. Recent efforts in single cell atlas in human and model organisms produced large amount of gene expression data.
- Developed VBASS, that integrate expression data to improve power of rare variants association analysis. With two versions that optimized for bulk RNA-seq and single-cell transcriptomics data respectively, VBASS models the association of disease risk as a function of expression profiles of relevant tissue or cell types in Bayesian frameworks. VBASS uses both analytical likelihood function and neural network approximations in joint probability calculation, and it learns the importance of cell types jointly from expression and genetics data. On simulated data, both methods show proper error rate control and better power than extTADA, the state-of-the-art Bayesian method. We applied the methods to published datasets and identified more candidate risk genes than extTADA with supports from literature or data from independent cohorts.

Undergraduate Researcher, Biomedical Pioneering Innovation Center (BIOPIC), Peking University.
Advisor: Dr. Zemin Zhang, Dr. Xianwen Ren

3D single cell interaction network reconstruction based on ligand-receptor mediated self-assembly. *2018.03-2019.07*

- Developed and utilized a new algorithm, CSOmap, which can give an inference of cellular spatial organization as well as cellular interaction. Different from mapping-based algorithms which require references generated from imaging technologies such as *in situ* sequencing, this algorithm only takes single cell transcriptome data as input, and uses an unsupervised machine learning model to predict ligand-receptor based cell spatial organization.
- Applied this new algorithm to five published cancer datasets, including liver carcinoma, lung carcinoma, carcinoma of colon and rectum, melanoma, head and neck cancer. CSOmap can successfully recapitulate spatial characteristics of tumor microenvironment for multiple cancers, prioritize molecular determinants of cellular interactions, and generate biological insights consistent with literature via *in silico* interference.

SKILLS

- Solid background in applications of Machine Learning, Deep Learning, Geometric Deep Learning and Statistical Learning to biological questions.
- Solid programming experience in Python (PyTorch), R, Matlab.
- Solid experiences single cell functional genomics analysis and whole genome sequencing (WGS) analysis.