Haotian Teng

E-mail: havens.teng@gmail.com haotiant@andrew.cmu.edu
Personal Website: https://haotianteng.github.io/

EDUCATION

Carnegie Mellon University, Pittsburgh, United States
Ph.D., Computational Biology, School of Computer Science 2019 - present

- Advisor: Ziv Bar-Joseph, Machine Learning Department, School of Computer Science, Carnegie Mellon University.

Featured courses: Probabilistic Graphical Models (10-708) A+, Deep Reinforcement Learning&Control (10-703), Convex Optimization (10-725) A+

University of Queensland, Brisbane, Queensland, Australia
M.S., Bioinformatics 2016 – 2018

- Advisor: Prof. Lachlan Coin, Institute of Molecular Bioscience, University of Queensland
- Advisor: Prof. Geoffrey Goodhill, Queensland Brain Institute, University of Queensland

Peking University, Beijing, China
B.S., Physics 2011 – 2015

PROFESSIONAL EXPERIENCE

Ph.D. Candidate, Carnegie Mellon University, Pittsburgh, PA 2019-09 to Present
Sarcasm detection on Tweets using RoBERTa language pre-trained model. 2022-09 to 2022-11

- Designed a model for automatic sarcasm detection, and achieved a relative 1.5% F1 score improvement comparing to the state-of-art model on the iSarcasm tweets dataset.
- Deployed the model and training dataset on Huggingface for one-off installation and inference.
- Designed a scale-adapter layer for fast fine-tuning pretrained RoBERTa language model.

Text mining and data annotation for named entity recognition 2022-07 to 2022-09

- Designed an automation pipeline to collect text from topic-specified online academic articles.
- Extracted and tokenized the text from PDF using existing python package and regular expression. Labeled the data using label-studio.
- Trained SciBERT model on our labeled dataset and verified it on the CoNLLpp dataset.

Target-based drug design using generative probabilistic diffusion model 2022-11 to present

- Generated a ligand-protein embedding by contrastive training on binding datasets.
- Designed a E(3)-equivalent graph neural network to generate drugable ligands based on protein targets on probabilistic diffusion model.

RNA methylation basecalling in Nanopore sequencing. 2020-10 to 2022-12

- Designed a novel hybrid non-homogeneous HMM and Convolutional Recurrent Neural Network to achieve accurate unsupervised signal segmentation by restricting the time-dependent transition matrix from neural network output, combining the interpretability of the HMM model and the prior knowledge learned by the NN model.
- Conducted data augmentation with score-dependent random walk graph sampling on a directed kmer graph constructed from the segmented signal.
- Developed a vector quantized variational autoencoder (VQ-VAE) based model to discover subtle signal differences that corresponding RNA post-transcript modification. A graphical model is used as encoder for model interpretability and a convolutional recurrent neural network (CRNN) is used as decoder for high classification accuracy.
- Achieved state-of-art accuracy in RNA methylation detection and is the first kind of model established methylation-aware basecalling.

Clustering spatial transcriptomics data at single cell level with hybrid NN and PGM. 2019-09 to 2021-02

- Conducted dimensional reduction using regularized denoising auto-encoder for gene expression profile.
- Developed a probabilistic Graphical generative model to cluster the cell from the spatial gene expression data (Program page: https://github.com/haotianteng/FICT).
- Produce a simulation pipeline for validating the spatial transcriptomics clustering tools and benchmark and visualization of the clustering result using a Jupyter notebook.
- Conduct differential gene expression analysis and GO term annotation.

Masters, University of Queensland, Brisbane, QLD, Australia 2016-02 to 2018-07
Using Deep Learning in Nanopore Basecalling 2017-02 to 2018-07

- Built a deep learning-based basecaller **Chiron** using Tensorflow, for Oxford Nanopore sequencer basecalling (Program page: https://github.com/haotianteng/Chiron)
- Developed a preprocessing tool **Nanopre** to identify the polyA region in the Nanopore RNA sequencing platform.
- Prepared training dataset of DNA and RNA Nanopore basecalling reads, using Nanoraw and Graphmap to label the data.
- Implemented a pipeline in Google Cloud and Google Compute Engine for end-to-end genome analysis.

The development of spontaneous neural activity in the zebrafish 2016-03 to 2017-02

- Built a pipeline for laboratory automation and data analysis in Zebrafish neuron experiment with Arduino, LabVIEW, and MATLAB.
- Constructed PHANTOM toolbox for projecting visual stimulation with conformal transformation, used for zebrafish tectum research. Program page in Github: https://github.com/haotianteng/PHANTOM-toolbox
- Developed algorithms for functional connectivity reconstruction using the regularization method under the scale-free assumption, correct the false positive correlation due to common ancestors, transition connection, and latent common input.

Internship, Center for Brain Science, Harvard University, Boston, MA
Feedback in AIY neurons in Thermotaxis behavior of C.elegans 2015-07 to 2015-12

- Studied thermotaxis in C.elegans with tracking microscopy and fluorescent marked neurons.
- Conducted experiment using a spinning disk confocal microscope and the afterward data acquisition & processing with the combination of ImageJ (Miji) and Matlab
- Proved the derivation dependence between AFD neuron and temperature, designed and conducted the experiment to measure the parameters of the AFD-temperature relationship with temperature signal input under different shapes.

Internship, Center for Bioinformatics, Peking University, Beijing 2011-09 to 2015-06
Locomotion and PH sensoring mechanism in C.elegans & fast reaction tracking System development 2012-07 to 2015-06

- Marked GCaMP6 into the C.elegans ASH, AWC and ASE neurons to testify and determine the neuron responsible for PH sensoring.
- Developed a neuro-muscle model of C.elegans motor system and proved the theoretical prediction of gait adaptation in C.elegans.
- Recorded and analyzed long-term locomotion parameters of the C.elegans by using a tracking and photographing system.
- Developed a visualization tool with openGL to describe and simplify the neuron network in C.elegans, and enabled the tool to search the whole neural pathway through any two given neurons.
- Built a tracking system as one of the contributors, which could achieve high-precision (accuracy below 1 micron) tracking and photographing and simultaneous data collection & processing
- Modified and developed a "snake" model-based algorithm for robust and precise C.elegans center line extraction.

WORKING EXPERIENCE

Algorithm Engineer Winter Intern 2019-01 to 2019-02
Alibaba, Hangzhou, China

- Intelligent cache prediction using deep learning models based on user's biometric information.

Bioinformatics Engineer
Novogene Europe, Beijing, China 2018-09 to 2019-01

- Optimized the human resequencing and laboratory automation pipeline.
- Designed and developed the long-read sequencing platform.

Senior Research Technician
Institute for Molecular Bioscience, University of Queensland, Australia 2017-06 to 2018-07

- Worked on Oxford Nanopore Technologies Long-read Nanopore direct RNA sequencing data processing, improved the sequencing accuracy and efficiency, improved the succeeded sequencing reads ratio by 15X compared to the original pipeline for long poly-A tail reads.

Intern 2014-07 to 2014-10
Biodynamic Optical Imaging Center, PKU, Beijing, China

- Micro-fluid chip preparation and fabrication.
- Developed a Computational Fluid Dynamics (CFD) module for the microfluid chips fluid field calculation in Fluent, which could draw the flow field from the CAD design sketch.

PUBLICATIONS

- **Teng, H.**, Yuan, Y. and Bar-Joseph, Z., 2021. Clustering Spatial Transcriptomics Data. *Bioinformatics.*
- Pitt, M. E., Nguyen, S. H., Duarte, T. P., **Teng, H.**, Blaskovich, M. A., Cooper, M. A., & Coin, L. J. (2020). Evaluating the genome and resistome of extensively drug-resistant Klebsiella pneumoniae using native DNA and RNA Nanopore sequencing. *GigaScience, 9(2), giaa002.*

- **Teng, H.**, Cao, M. D., Hall, M. B., Duarte, T., Wang, S., & Coin, L. J. (2018). Chiron: translating nanopore raw signal directly into nucleotide sequence using deep learning. *GigaScience, 7(5), giy037.*
- Avitan, L., Pujic, Z., Mölter, J., Van De Poll, M., Sun, B., **Teng, H.**, Amor, R., Scott, E.K. and Goodhill, G.J., 2017. Spontaneous activity in the zebrafish tectum reorganizes over development and is influenced by visual experience. *Current Biology, 27(16), pp.2407-2419.*
- **Teng, H.** "A neuron-muscle circuit model of C.elegans's locomotion." *Bachelor of Science Thesis: Peking University, 2015*

HONORS AND AWARDS

- The 1st Prize at 27th Chinese Physics Olympiad, Zhejiang Province (rank 1/1232 in theory part) 2011
- The Silver Medal at 27th Chinese Physics Olympiad, Finals 2011
- The 1st Prize at 29th Parts of the National College Students Physics Competition 2012

SKILLS

- Programming: Python, C, C++, Matlab, R, Linux, LaTeX,
- Packages&Platforms: Tensorflow, MXNet, Caffe, CUDA, cuDNN, OpenGL, BWA, SAMtools, Velvet, DIAMOND, BLAST+, Minimap2, H5py, Psychtoolbox, LabVIEW, Arduino.
- Software: PyMOL, Fluent(ANSYS), Origin, AutoCAD, Primer Premier, DNA Man, Microsoft Office,
- Wet-lab experiment skill: Molecular cloning, Microinjection
- Language: Chinese(Mother Language), English(Fluent), Spanish(basic), German (Pizza-orderable)
  TOEFL: Cumulative 103 (R 29, L 29, S 23, W 22); GRE: V 150, Q 169, AW 3.0
- Proficient in Piano playing, learned since 6 years old. Skillful in saxophone.