Xin Cao

- (+86) 156 2390 6058 - https://xincao02.github.io/ - xincao@link.cuhk.edu.cn

EDUCATION

The Chinese University of Hong Kong, Shenzhen **Shenzhen, China**
Bachelor of Science in Data Science and Big Data Technology *Sep 2020 - Dec 2024*

Notable Courses
Advanced Machine Learning, Reinforcement Learning, Deep Learning and Applications, Speech and Natural Language Processing, Data Mining, Database Systems, Bioinformatics, Cloud Computing, Bayesian Statistics, Stochastic Simulation, Stochastic Processes, Optimization, Data Structures, Operations Management, C/C++

Honors
2024 Annual Outstanding Student Organization Leader, Harmonia College Student Representative

South Grove Elementary School **New York, United States**
Project Beyond Class (School Top 10% students) *Jan 2010 - Dec 2012*

ACADEMIC ACHIEVEMENTS

[1] **X. Cao**, J. Li, F. Meng, Y. Zou, Z. Wan, G. Cao (2024). Deep Learning-Based Prediction of Protein-Protein Interactions on Short Protein Datasets. *(Manuscript in preparation)*
[2] Y. Chen, **X. Cao**, Y. Pan, X. Chen, A. Gao, P. Chao, S. Huang, K. Xiao, M. Li, H. Huang (2024). Bayesian Approach Towards Transcriptional Factor Activity Prediction based on miRNA Data. *(Model in Development)*
[3] Z. Wan, A. Gao, P. Chao, **X. Cao**, J. Song, M. Ran (2024). Variation Method for Map Matching. *(Manuscript in preparation)*
[4] **X. Cao**, J. Li (2024). PASIR: A Proactive Social Interaction Robot Empowered by Multimodal Data Fusion and BCI-Driven Emotion Prediction Models. *International Conference on Social Robotics (ICSR)*. *(Oral presentation)*
[5] Reviewed 1 paper for SIGKDD Conference (2025) on behalf of official program committee member: Large Language Models for Next Group Point-of-Interest Recommendation.

RESEARCH EXPERIENCE

CUHKSZ - Warshell Institute of Computational Biology **Supervisor: Prof. Hsien-Da HUANG**
Transcription Factor Activity Prediction *Oct 2024 - Present*

- **Model Development:** Reviewed previous literature and generated ieas for model architecture based on own dataset. Chose and developed a Bayesian Model with Attention mechanism for more effective and detailed capture of miRNA and TFs. Trained and finetuned the model architecture based on own dataset.

Shenzhen University of Advanced Technology **Supervisor: Prof. Gang CAO**
Protein Peptide Interaction Deep Learning Prediction *June 2024 - Present*

- **Model Training and Optimization:** Led a project to train advanced deep learning models (GNNs, VQ-VAE, GRU-LSTM) by leveraging spatial and sequence information of proteins/peptides. Enhanced model performance through architecture modifications, milestone adjustments, and optimized dataset composition, achieving 99.94% accuracy and 98.85% F1-score, surpassing previous models.
- **Dataset Investigation and Data Analysis:** Conducted comparative experiments, revealing that pre-training on short protein datasets with fine-tuning on long+short datasets improved protein-peptide interaction predictions. Incorporated biological significance into statistical methods (frequency distribution, correlation, diversity analysis)
- **Data Preprocessing and Automation:** Responsible for automated data collection, deduplication, encoding, dimensionality reduction, cross-species dataset control, and multimodal data fusion for neural network inputs.

The Chinese University of Hong Kong, Shenzhen **Supervisor: Prof. Zhizheng WU**
Drumbeat Generation Based on Non-Percussive Music *April 2024 – June 2024*

- **Data Collection and Analysis:** Collected and analyzed model codes and datasets for audio enhancement and generation. Compared model performance for track generation, finding Latent Diffusion and Transformer models most effective. Trained and generated 90+ drumbeat segments for music tracks using Latent Diffusion.
- **Data Processing and Augmentation:** Utilized Demucs to seperate drum tracks from full music pieces. Adjusted dataset composition to balance musical styles, enhancing the diversity and quality of generated drumbeats.
- **Model Deployment:** Responsible for the local deployment and testing of the Demucs model (hybrid transformer) as part of audio editing research, adjusting dataflow I/O for current pipeline.

Southern University of Science and Technology **Supervisor: Prof. Feng ZHENG**
Multimodal Computer Vision Adversarial Robustness *May 2022 – Sep 2022*

- **Evaluation of System Robustness:** Investigated and organized metrics to evaluate and enhance the robustness of image-text retrieval systems. Contributed to data preprocessing, model debugging, and performance evaluation, emphasizing the effects of preprocessing on outcomes.
- **Research and Model Re-implementation:** Independently surveyed Optical Music Recognition and Music Generation. Refined recent codes, optimized data label for model accuracy, presented technical analyses & summaries.

2024 Bremen Big Data Challenge (BBDC)

Supervisor: **Prof. Haizhou LI**

- **Data Visualization and Model Development:** Conducted visual analysis & data augmentation using DS methods (polynomial regression, correlation analysis, weighted dimensionality reduction) on heart rate, ECG, Ibi, and PPG data. Trained 2D-CNN and KNN models for emotion/action prediction, securing a global top 10 ranking.
- **Team Leadership and Collaboration:** Led team role definition and resource allocation, boosting efficiency and completing the project within one month, outperforming competing teams in the same track.

Machine Learning Course Projects

March 2023 - Present

- **Advanced Machine Learning:** Co-developed models to distinguish real human faces from GAN-generated using DenseNet121, ResNetV2, and VGG19. Implemented and compared models with/without data augmentation (RGB jitter, Gaussian noise). Integrated models into ensemble classifier, improving accuracy from 80% to 99.1%.
- **Deep Learning Applications:** Built and fine-tuned object detection models (Fast R-CNN, EfficientDet) using mmcv and mmdet toolkits for VOC2012 object recognition competition. Trained on dataset subset and reported results.
- **Speech and Natural Language Processing:** Processed audio with varying timbres, extracted Mel spectrograms & pitch, applied resampling & dynamic time warping. Used K-means and t-SNE for visualization, achieving clear classification of audio based on timbre.

INVENTION PATENTS

- TelCentSpec: Chromosome-Specific Telomere and Centromere Repeats Detection and Specificity Filtering System *(Under Review)*
- Audio-Visual System for Sleep Promotion Based on 40Hz Gamma Frequency Stimulation. *(Under Review)*

INTERNSHIP EXPERIENCES

Kunyu Biotechnology Co., Ltd.

Wuhan, China

Algorithm Engineer

July 2021 – Sep 2021

- **Gene Alignment Program Development:** Developed gene alignment program using Python and tools (BLAST & Bowtie2) to identify high-repetition DNA fragments in telomere and centromere regions of target chromosone. Performed second round of alignment in non-target chromosomes to eliminate non-unique sequences, adjusted thresholds for both rounds of alignment. Quantitatively analyzed the impact of threshold changes on final results, optimized them to identify effective chromosome-specific DNA sequences. Secured invention patent *TelCentSpec*.
- **Program Optimization:** Thoroughly assessed & reformed algorithm logic and comparatively experimented different toolkits, reducing execution time from 48h to under 10h, memory usage by threefold.
- **Data Processing and Report Writing:** Developed Linux scripts to process and analyze output data, integrated the results with biological information to select detectable DNA sequences without fold spatial structures. Compiled final sequences into result file, wrote a user-friendly manual for the program.

AWARDS

- **Finalist (4th Place)**, ICSR 2024 Robot Design Competition - Sep 2024
- **Global Top10 Ranking**, 2024 Bremen Big Data Challenge - Mar 2024
- **Outstanding Leader (10,000 RMB)**, 2023 Chinese University of Hong Kong, Shenzhen Harmonia Annual Scholarship - Sep 2024
- **Champion**, 2022 Shenzhen Annual Soccer League (Team Core Player) - Dec 2022
- **Silver**, 7th Aunual Sports Festival of Chinese University of Hong Kong, Shenzhen (Team Captain) - June 2024

LEADERSHIP & EXTRACURRICULAR ACTIVITIES

- **Captain of Soccer Team:** Led training & matches, improving team annual rank from 7th to 2nd. Collaborated with other university teams administrators for event organization and managed team budget.
- **Annual University Music Festival:** Secured sponsorships with business professionals, coordinated venue arrangements in plazas, organized & participated performances with hundreds of live audiences on & off campus.
- **President of Campus Broadcast Station:** Led station reform: introduced new collaboration-competition system, added new departments, secured additional equipment & permissions via negotiations with officials. Established & taught audio editing for special pre-produced weekly program. Number of applicants tripled the following year.

SKILLS

Technical: Python, Linux, PyTorch, Tensorflow, SQL, Spark, RegEx, LaTeX, C/C++, Java, Stable Diffusion, Ollama
English: Five years of school in America (Native Speaker); TOEFL 100 (Listening: 29)
Music: Logic Pro, Reaper; Band vocal & guitar; University vocal contest 1st place (Duet); Piano ABRSM Grade 5
Soccer: Captain of college team; Core player of university team; Champion of Shenzhen University soccer League