TONG CHEN

(+86) 133-2182-7829 ◇ tchen21@m.fudan.edu.cn ◇ Website

EDUCATION

Fudan University, Shanghai Sept. 2021 - Present

Bachelor in Computer Science, *GPA:* 3.74/4.0 (Major: 3.93/4.0), *Rank:* 10/106

Selected Courses: Object-oriented Programming (A), Set and Graph Theory (A), Computer Organization and Architecture (A), Database System (A), Algebraic Structure and Mathematical Logic (A), Pattern Recognition and Machine Learning (A), Computer Vision (A), Artificial Intelligence (A), Parallel Distributed Computing (A)

RESEARCH INTERESTS

AI for Biology, Large Language Models

PUBLICATION

- **PTM-Mamba: A PTM-Aware Protein Language Model with Bidirectional Gated Mamba Blocks**
  Zhangzhi Peng, Chentong Wang, **Tong Chen**, Sophia Vincoff, Pranam Chatterjee. **(Under Review)**
- **moPPIt: De Novo Generation of Motif-Specific Binders with Protein Language Models**
  Tong Chen, Yinuo Zhang, Pranam Chatterjee. **(Accepted by NeurIPS 2024 workshop and MoML)**
- **Synergizing sequence and structure representations to predict protein variants**
  Tong Chen, Pranam Chatterjee. **(Accepted for publication in Cell Research, Research Highlights)**
- **Demonstration Distillation for Efficient In-Context Learning.**
  Tong Chen, Qirun Dai, Zhijie Deng, and Dequan Wang. **(Submitted to ICLR 2023)**

RESEARCH EXPERIENCES

Chatterjee Lab, Duke University June 2024 - Present

Research Intern, advised by Prof. Pranam Chatterjee

- **Mutant-specific Binder Design: (Ongoing)**
  - Developed an ESM-2-based model to capture subtle differences between wild-type and mutant proteins, and to predict the distance between mutant-binder and wild-type-binder joint embeddings.
  - Constructed a novel, large-scale dataset derived from PPIRef, systematically mutating binding site residues on target sequences into the least likely alternatives using BLOSUM62 matrices, ensuring that mutant proteins do not bind to the binders.
  - Employed ESM-Fold for structure prediction and Prodigy for binding affinity calculation.
  - Trained this model on this custom dataset, achieving 61% test accuracy.
  - Developing an MDLM model for binder generation.
- **PTM-Mamba Benchmark:**
  - Conducted an evaluation of PTM-Mamba embeddings for predicting post-translational modification (PTM) effects, comparing their performance to other embedding methods.
  - Developed a neural network that is compatible with all embedding types for a fair comparison.
  - Demonstrated that PTM-Mamba consistently outperformed other embeddings across most metrics, highlighting its superior ability to capture PTM-specific insights.
- **Motif-specific Binder Design:**
  - Developed BindEvaluator, a model for accurately predicting binding sites in both protein-protein and peptide-protein interactions using only sequence information.
  - Pre-trained BindEvaluator on the extensive PPIRef dataset to establish foundational knowledge of protein-protein binding sites, achieving a test AUC score of 0.94.
  - Fine-tuned BindEvaluator on a custom peptide-protein binding dataset, combining PepNN and BioLip2 datasets, resulting in an improved test AUC score of 0.97 for peptide-protein binding site prediction.
  - Designed the moPPIt (**mo**tif-specific **PPI** **t**argeting) algorithm by integrating BindEvaluator with a genetic algorithm, enabling the design of motif-specific binders tailored to any target protein sequence.

- Assessed the binding affinity of moPPIt-designed binders using AlphaFold2-Multimer and validated their binding specificity with PeptiDerive.
- Benchmarking results validated moPPIt's capability to design motif-specific binders for both structured and disordered proteins with high specificity.
- This work resulted in a first-author paper.

Qing Yuan Research Institute Nov. 2022 - June 2024
Research Intern, advised by Prof. Dequan Wang and Prof. Zhijie Deng

- **Homo-UOX Variants Design:**
  - Human uricase (UOX) is non-functional because the gene responsible for uricase production is a pseudogene.
  - Trained an MSA-VAE model to generate uricase variants based on the Macaca mulatta uricase sequence, selecting those with over 97% sequence similarity to the human uricase.
  - Computed ESM-1v and ProteinMPNN scores for all selected sequences and applied the COMPSS (Composite Metrics for Protein Sequence Selection) framework to identify functional variants.
  - The resulting uricase variants are expected to exhibit high enzymatic activity in vivo.
- **Customized CRISPR-Cas9 Base Editor Design:**
  - Streamlined spCas9 for base editing applications, aiming to retain full functionality outside the RuvC domain while minimizing the enzyme length for enhanced delivery efficiency.
  - Employed RFdiffusion to scaffold non-RuvC domain and ProteinMPNN to optimize sequences.
  - Utilized AlphaFold2 and TM-align to confirm the sequences' structural integrity of non-RuvC domains.
  - Developed an iterative model for sequence minimization, selectively mutating or excising amino acids at RuvC locations to shorten the sequence without altering the non-RuvC domain structure.
  - Successfully reduced the amino acid count by 132, 10% of the original length, maintaining the structural and functional fidelity of the non-RuvC domains.
- **Activity Prediction Model:**
  - Developed a model using RoseTTAFold and GearNet to predict the activitiy of any given TadA protein sequence with high accuracy.
- **Efficient In-Context Learning:**
  - In-context learning (ICL) demonstrations often contain redundant tokens, leading to unnecessary costs.
  - Proposed a training-free framework that leverages LLM-powered agents to iteratively refine the given demonstrations with meticulously designed prompts.
  - Experimented the framework on three public datasets and the results show that our method can achieve up to 4.3x demonstration compression ratio and 5% improvement in ICL performance.
  - The work resulted in a first-authored paper.
- **Dataset Distillation:**
  - Adapted a novel generative distillation model for medical image classification, achieving performance comparable to existing SOTA methods.
  - Distilled a dataset with over 400000 medical images into 20 synthetic images, enabling a classification model trained on this distilled dataset match the accuracy of one trained on the original dataset.

AWARDS

- Honorable Mention Prize, Interdisciplinary Contest In Modeling 2023
- Second Prize, Academic Excellence Scholarship, Fudan University 2022 - 2023

SKILLS

Programming Languages: C/C++, Python, Java, R, Linux, Git, LaTeX, SQL, HTML, JavaScript, TypeScript
Tools: Pytorch, PyMOL, pyRosetta, PeptiDerive, Pytorch Lightning, Weight&Bias, Docker, Django, Vue
English Test: TOEFL IBT 108