TONG CHEN

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

EDUCATION

Fudan University, Shanghai Sept. 2021 - Present

Bachelor in Computer Science, *GPA:* 3.79/4.0 (Major: 3.93/4.0), *Rank:* 10/106

Selected Courses: Object-oriented Programming (A), Data Structure (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 Science, Large Language Models

PUBLICATION

- **moPPIt: De Novo Generation of Motif-Specific Binders with Protein Language Models**
  Tong Chen, Yinuo Zhang, Pranam Chatterjee. **(Under Review)**
- **Demonstration Distillation for Efficient In-Context Learning.**
  Tong Chen, Qirun Dai, Zhijie Deng, and Dequan Wang.

RESEARCH EXPERIENCES

Chatterjee Lab, Duke University June 2024 - Present

Research Intern, advised by Prof. Pranam Chatterjee

- **Mutant-specific Binder Design:**
  - Developed a graph-based model, muPPIt (**mu**tation **PPI t**argeting), to capture subtle differences between wild-type and mutant proteins, predicting the binding preference for a given binder using only protein sequences as inputs.
  - 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. Training muPPIt on this dataset resulted in zero test loss.
  - Adapted muPPIt to predict binding affinities for mutant-binder and wild-type-binder interactions, achieving a 62% accuracy when fine-tuned with pre-trained weights on the SKEMPI dataset.
  - Developed a GPT-architecture decoder to predict protein sequences from graph node representations.
  - Designed a mutant-specific binder design algorithm based on muPPIt and the decoder using a gradient update method.
  - Benchmarked this algorithm on designing binders to unseen mutant and wild-type protein sequences.
  - Evaluated the designed binders by predicting binder-target structures using AlphaFold3 and assessing binding energy using pyRosetta.
  - Benchmarking results showed that the designed binders had an average of 20 units lower free energy when binding to mutant proteins compared to wild-type.
- **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.
  - Benchmarked moPPIt on both structured proteins, with or without pre-existing binders, and intrinsically disordered proteins.
  - 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.

Shanghai AI Lab 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 using multiple sequence alignments (MSA) constructed from a diverse set of uricase sequences.
  - Leveraged the pre-trained MSA-VAE 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 experimentally validated COMPSS (Composite Metrics for Protein Sequence Selection) framework to filter these 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.
  - Conducted non-RuvC domain scaffolding with RFdiffusion to create candidate backbone structures.
  - Employed ProteinMPNN to generate novel sequences with optimized structures.
  - Utilized AlphaFold2 to predict the tertiary structures of the engineered sequences and TM-align to confirm the integrity of non-RuvC domains post-modification.
  - 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:**
  - Previous in-context learning demonstrations usually contain superfluous tokens, which result in high computation costs and hurt the robustness of LLM.
  - Proposed a training-free framework that leverages LLM-powered agents to iteratively refine the given demonstrations with meticulously designed prompts.
  - Resolved the computation overheads without compromising the performance and enhanced the generalization ability with streamlined demonstrations.
  - Experimented the framework on three public datasets and the results show that our method can achieve up to 4.3x compression ratio and 5% improvement in accuracy.
  - 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.

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