CS286 AI for Science and Engineering

Course Projects

Fall semester, 2023

Project 1. Weakly Supervised Learning for Pattern Classification in Single Particle Imaging

Persons of Contact:
PI: Prof. Ping HUAI (怀平), email: huaiping@shanghaitech.edu.cn
TA: Dr. Xiaofeng ZHANG (张晓峰), email: zhangxf2@shanghaitech.edu.cn

Introduction:

X-ray Free Electron Laser (XFEL) so far is one of the most advanced X-ray sources, which makes it possible to achieve higher spatial and temporal resolution in structure determination than synchrotron radiation facility. Shanghai HIgh repetitioN rate XFEL and Extreme light facility (SHINE) currently is a new generation of XFEL that is now under construction by ShanghaiTech University in collaboration with Shanghai Advanced Research Institute and Shanghai Institute of Optics and Fine Mechanics.

With unprecedented brilliance, repetition rate and ultrashort pulse, SHINE will generate huge amount of data during experiments. The estimated maximum data rates may reach hundreds of gigabytes per second and the storage will reach tens petabytes per year. To deal with the grand challenge in XFEL data science, machine learning algorithms are supposed to play an important role in data analysis and data management. In recent years, machine learning has been demonstrated in European XFEL and LCLS in US as an important tool to deal with automatic optimization of experimental setup, anomaly detection, diagnostic prediction for some important parameters beyond direct measurement, online data analysis etc.

To achieve even higher resolution in structure determination in XFEL, millions of X-ray diffraction patterns need to be generated in experiments such as single particle imaging (SPI). Efficiently analysis of such huge amount of data is challenging and machine learning algorithms are supposed to be a promising tool. Among all the SPI patterns there are single hit, multiple hits, water diffraction, background and blank frames (due to the defect of detector), however only the single hit frames are useful in downstream analysis. So the selection of single hit frames from a bulk of diffraction patterns is a very important step and has a significant influence on the determination of eventual structure. Traditionally, the selection can only be done manually, which is not only slow but also requires advanced professional knowledge. Here we want to train deep learning models to solve this problem. Furthermore, it is very costly to annotate a large size of diffraction patterns, which not only need a long time but also require deep expertise, therefore we hope the models are weakly supervised, namely the size of labeled images used in the training should be as small as possible. Candidate methods include dimensionality reduction, feature engineering, pre-training with generative algorithms like VAE and GAN. Particularly, we encourage the pre-training of networks contain transformer architecture.

Goals:

In this project, students are expected to implement the whole data processing procedure as well as the pattern classification that covers data loading, data pre-processing, and model's developing etc. The models should be trained with less labeled data as far as possible. A numbered list is given as follows:

1. Data loading
2. Data pre-processing
3. Train models with enough labeled dataset
4. Train models with less labeled data as far as possible, but achieve comparable performance with the fully supervised models.

References:

1. Sun, Zhibin et al., Current Status of Single Particle Imaging with X-ray Lasers, *Appl. Sci.* **8** (2018), 132.
2. Haoyuan Li et al., Diffraction data from aerosolized Coliphage PR772 virus particles imaged with the Linac Coherent Light Source, *Sci. Data* **7** (2020), 404.
3. Yingchen Shi et al., Evaluation of the performance of classification algorithms for XFEL single-particle imaging data, IUCrJ (2019). **6**, 331-340.
4. Eduardo R. Cruz-Chú et al., Selecting XFEL single-particle snapshots by geometric machine learning, *Struct. Dyn*. **8**, 014701 (2021).
5. Cong Wang et al., SpeckleNN: a unified embedding for real-time speckle pattern classification in X-ray single-particle imaging with limited labeled examples, IUCrJ (2023). **10**.

Note: Ref. 1 gives a complete introduction to single particle imaging experiment, as well as the data analysis steps; Ref. 2 gives a detail description of the dataset used in this projection; Ref. 3~5 are related work with the dataset described in ref. 2.

Project 2. Deep Learning for Host-Guest Recognition in Porous Molecular Materials

Persons of Contact:

PI: Prof. Shan Jiang (姜珊), email: jiangshan@shanghaitech.edu.cn
TA：Zidi Wang (王子頔), email: wangzd@shanghaitech.edu.cn

Introduction:

Supramolecular chemistry, often referred to as "chemistry beyond the molecule," is a scientific discipline that focuses on exploring molecular recognition and the formation of complex assemblies through noncovalent interactions[1]. Unlike traditional chemistry, which primarily deals with covalent bonds, supramolecular chemistry delves into the realm of non-covalent interactions between molecules, including hydrogen bonding, van der Waals forces, electrostatic attractions, and π-π interactions[2, 3]. A central theme in supramolecular chemistry is host-guest chemistry, where specific guest molecules selectively bind to host molecules through non-covalent interactions, resulting in unique functional properties. The ability of the host molecule's cavity size and its internal chemical environment to facilitate selective binding is crucial. The former ensures a geometric match, while the latter provides the driving force for binding. Within the realm of host-guest chemistry, porous molecular materials like organic cages[4], metal-organic cages[5], and macrocycles[6] have gained significant attention due to their distinctive structural characteristics and diverse applications in molecular sorting, biomolecular recognition, and catalysis. Consequently, they have become a focal point of research in supramolecular chemistry.

However, host-guest recognition is a complex area with various considerations, including the selection of precursor molecules, the design of host architectures, and the comprehension of host-guest interactions. These factors impose constraints on the design of molecular materials. Fortunately, advancements in artificial intelligence (AI) techniques have opened up opportunities to predict host-guest interactions, enabling the reverse design of host molecules. In the field of drug design, numerous powerful deep learning models have emerged, such as DeepDock[7], Uni-Mol[8], and DiffDock[9], capable of accurately predicting the binding poses of proteins and ligands and quantitatively assessing their binding affinities. Host-guest recognition shares similarities with protein-drug binding, as both are primarily governed by non-covalent interactions as the driving force for binding. Consequently, drug design-based deep learning models offer the potential for precise predictions in the field of host-guest recognition. This project aims to apply these deep learning models from drug design to the realm of host-guest recognition, enabling accurate predictions of binding poses and quantitative scoring of binding affinities.

Goals:

In this project, you are expected to implement the whole data processing that covers data loading, data pre-processing, model predictions, etc. The deep learning model should be trained or fine-tuned with less labeled data as far as possible. A numbered list is given as follows:

1. **Data Collection:**

- Collect Crystallographic Information Files (.cif) containing information about host-guest recognition from the Cambridge Crystallographic Data Centre (CCDC) or any other relevant data sources.
- Organize and store the collected data in a standard format for further processing, such as .mol or .pbd file.

2. **Data Pre-processing:**

- Clean and preprocess the collected data:
  - Removing duplicates and irrelevant entries.
  - Handling missing data or incomplete records.
  - Splitting the data into training, validation, and test sets.

3. **Aligning Input Data:**

- Prepare the input data for the deep learning models. In the context of host-guest recognition, this might involve:
  - Standardizing the molecular structures or representations to ensure consistency.
  - Generating molecular descriptors or features that represent the host and guest molecules.
  - Aligning host-guest pairs so that the input data is structured properly for training.

4. **Train the Deep Learning Model:**

- Choose an appropriate deep-learning architecture for host-guest recognition. This could be a pretrained model (such as Uni-Mol or DiffDock).
- Utilize techniques like fine-tuning with pre-trained models if available.
- Train the model using the preprocessed data, considering best practices such as:
  - Implementing data augmentation techniques if the dataset is small.
  - Monitor the training process and adjust hyperparameters as needed to achieve the desired performance.

5. **Evaluate and Fine-Tune:**

- Evaluate the trained model's performance with metrics.

6. **Testing and Predictions:**

- Assess the model's generalization by testing it on other datasets.
- Use the model to make predictions on new host-guest recognition scenarios, including predicting binding poses and binding affinities.
- Analyze the predictions and assess their accuracy and reliability.

References:

1. Concepts. In Supramol. Chem., 2009; pp 1-48.

2. Latimer, W. M.; Rodebush, W. H., Polarity and Ionization from the Standpoint of the Lewis Theory of Valence. In A Source Book in Chemistry, 1900-1950, Henry, M. L., Ed. Harvard University Press: Cambridge, MA and London, England, 1968; pp 110-112.

3. Zimmerman, S. C.; Wendland, M. S.; Rakow, N. A.; Zharov, I.; Suslick, K. S., Synthetic hosts by monomolecular imprinting inside dendrimers. Nature, 2002, 418 (6896), 399-403.

4. Tozawa, T.; Jones, J. T.; Swamy, S. I.; Jiang, S.; Adams, D. J.; Shakespeare, S.; Clowes, R.; Bradshaw, D.; Hasell, T.; Chong, S. Y.; Tang, C.; Thompson, S.; Parker, J.; Trewin, A.; Bacsa, J.; Slawin, A. M.; Steiner, A.; Cooper, A. I., Porous organic cages. Nat. Mater., 2009, 8 (12), 973-8.

5. Zhang, D.; Ronson, T. K.; Zou, Y.-Q.; Nitschke, J. R., Metal–organic cages for molecular separations. Nat. Rev. Chem., 2021, 5 (3), 168-182.

6. Ji, X.; Ahmed, M.; Long, L.; Khashab, N. M.; Huang, F.; Sessler, J. L., Adhesive supramolecular polymeric materials constructed from macrocycle-based host–guest interactions. Chem. Soc. Rev., 2019, 48 (10), 2682-2697.

7. Méndez-Lucio, O.; Ahmad, M.; del Rio-Chanona, E. A.; Wegner, J. K., A geometric deep learning approach to predict binding conformations of bioactive molecules. Nat. Mach. Intell., 2021, 3 (12), 1033-1039.

8. Zhou, G.; Gao, Z.; Ding, Q.; Zheng, H.; Xu, H.; Wei, Z.; Zhang, L.; Ke, G. Uni-Mol: A Universal 3D Molecular Representation Learning Framework. The Eleventh International Conference on Learning Representations. Kigali, Rwanda, May 1-5, 2023

9. Corso, G.; Stärk, H.; Jing, B.; Barzilay, R.; Jaakkola, T. DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking. 2022, arXiv:2210.01776

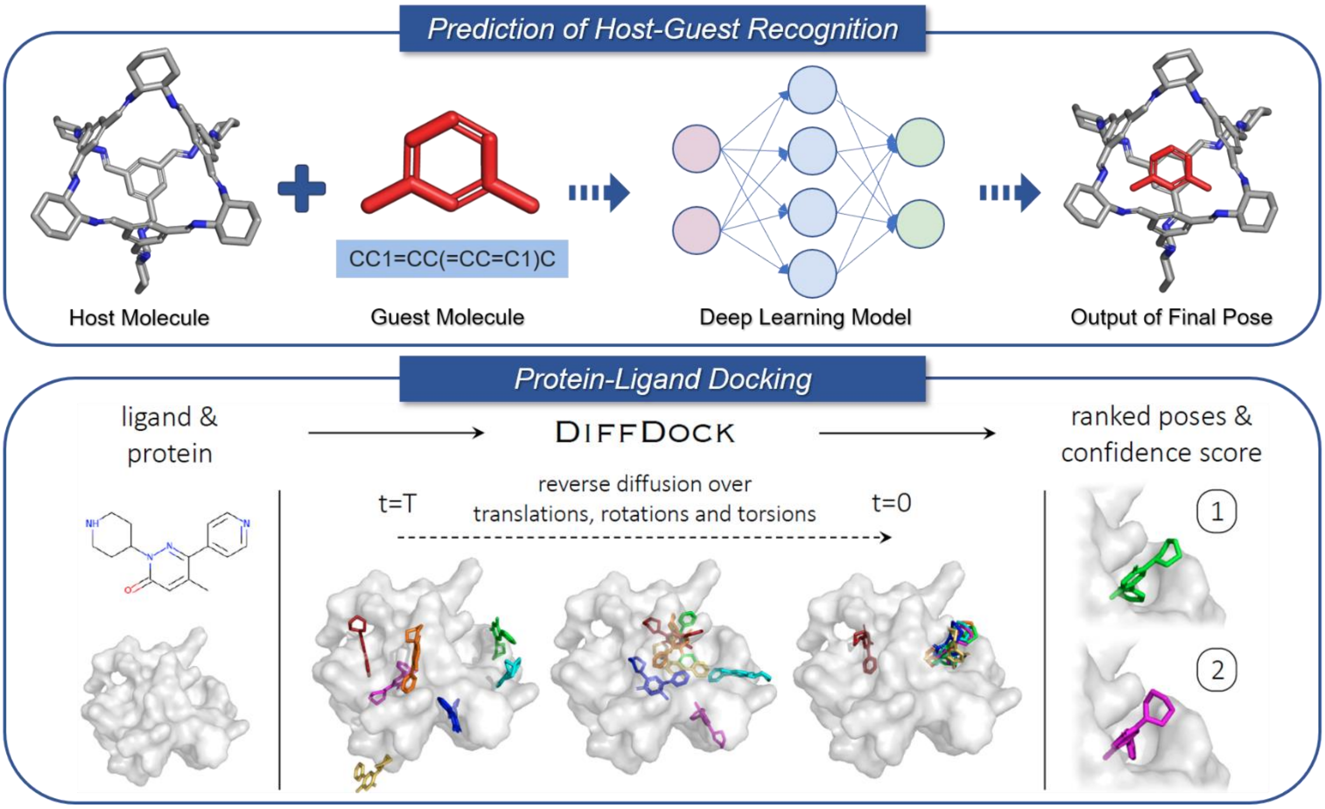

Figure 1. Workflow for Prediction of Host-Guest Recognition and Protein-Ligand Docking

Figure for protein-ligand docking is adapted from *https://github.com/gcorso/DiffDock*

Project 3. Machine Learning Interatomic Potentials

Persons of Contact:
PI: Professor Peijun Hu (胡培君), email: hupj@shanghaitech.edu.cn
TA: Dr. Xie Wenbo (谢闻博), email: xiewb1@shanghaitech.edu.cn

Background

In the realm of surface reaction studies, quantum chemistry calculations, notably density functional theory (DFT), have revolutionized our approach since the 1990s. Serving as a cornerstone in catalysis research, DFT can give molecular-level insights that are often hard to obtain by experimental means. However, the scalability of DFT calculations has reached a plateau, as the simulations of larger reaction systems (exceeding 300 atoms) or extend our time scale to the nanosecond range, the computational demands escalate exponentially.

Recently, machine learning interatomic potential (MLIP) has emerged as a promising way to break the limitation. Utilizing the *ab intio* calculations results, MLIP effectively 'learning' the dynamics of atomic interactions. By applying latest ML models such as CNN, GNN, and transformers, these models can simulate atomic forces and energy changes as atoms move. This way avoids the expensive *ab intio* calculation while giving acceptable results. Very recently, Transformer has emerged as a powerful ML approach for a wide range of applications, such as image recognition, natural language processing and computer vision. In this project, we are going to explore the possibility of improving current MLIP models, tailing them for catalytic reactions and thereby increase the overall accuracy of the model predictions.

Goal:

In this project, students are expected to implement the Transformer or similar neural networks for machine learning interatomic potentials. The training data from DFT calculations will be given (https://github.com/HuGroup-shanghaiTech/CS286). Specific objectives are given as follows:

1. Learning how the atomic data is translated for ML model training.
2. Understanding mainstream MLIP models.
3. Evaluating the Transformer performance and tring to improve the current MLIP models.

MLIP methods:

DeepMD: https://docs.deepmodeling.com/projects/deepmd/en/master/index.html
Nequip: https://github.com/mir-group/nequip
Equiformer: https://github.com/atomicarchitects/equiformer

References

[1] Kang, P.-L.; Shang, C.; Liu, Z.-P. Large-Scale Atomic Simulation via Machine Learning Potentials Constructed by Global Potential Energy Surface Exploration. Acc. Chem. Res. 2020, 53 (10), 2119–2129.

[2] Batzner, S.; Musaelian, A.; Sun, L.; Geiger, M.; Mailoa, J. P.; Kornbluth, M.; Molinari, N.; Smidt, T. E.; Kozinsky, B. E(3)-Equivariant Graph Neural Networks for Data-Efficient and Accurate Interatomic Potentials. Nat Commun 2022, 13 (1), 2453.

[3] Zhang, Y.; Xia, J.; Jiang, B. Physically Motivated Recursively Embedded Atom Neural Networks: Incorporating Local Completeness and Nonlocality. Phys. Rev. Lett. 2021, 127 (15), 156002.

Project 4. Visual Analytics of Deep Learning Model Interpretability in Synthetic Lethality Prediction

Persons of Contact:
PI: Quan Li (李权), email: liquan@shanghaitech.edu.cn
Co-PI: Jie Zheng (郑杰), email: zhengjie@shanghaitech.edu.cn
TA: Haoran Jiang (姜浩然), email: jianghr@shanghaitech.edu.cn

Background

Synthetic lethality (SL) is a type of genetic interaction between a pair of genes, where the defects of both genes will significantly impair cell viability, but the defect of a single gene will not affect cell fitness, as shown in Figure 1. It is a promising strategy to target a non-essential gene in an SL interaction where the other gene is a cancer-specific defective gene, which would selectively kill the cancer cells without harming normal cells. However, high-throughput wet-lab screenings of SLs are often costly and face various challenges (e.g., high cost, off-target effects, and inconsistency across platforms or cell lines). Therefore, computational methods for SL prediction have been a practical and valuable complement to wet-lab techniques for SL identification.

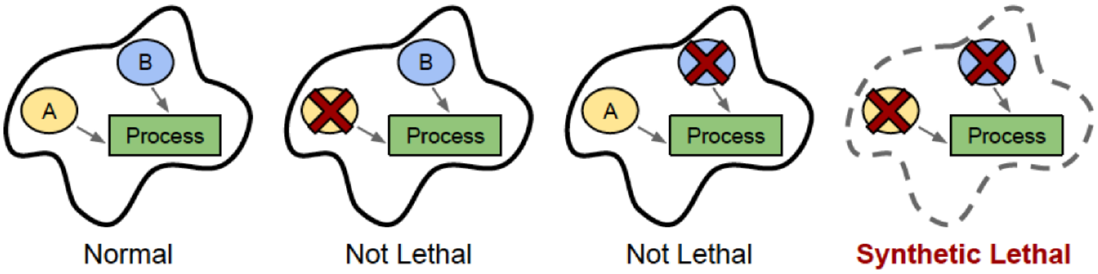

Figure 1. The concept of synthetic lethality.

Recently, deep learning methods for SL prediction have drawn some attention. For example, the problem of predicting SLs can be formulated as a link prediction task on a graph of interacting genes (i.e., the SL graph). Hence, graph-based methods, e.g., matrix factorization methods ([1], [2]) and graph neural network methods ([3], [4]), have thus been proposed for SL prediction.

While these models have achieved certain success in prediction accuracy, they still lack sufficient interpretability [8] to assist domain experts in further screening more experimentally valuable gene pairs or optimizing prediction models with their domain knowledges. Therefore, there is an urgent need for visual analytics techniques [7] to enhance the interpretability of neural network models used in synthetic lethality prediction.

Goal: In this project, we aim to develop a visual analytics system of deep learning model interpretability in synthetic lethality prediction.

Sub-goals and recommended steps:

1. Select and augment datasets. PI Jie Zheng's group have developed the world first

comprehensive database for SL, named SynLethDB [5]. The data in SynLethDB have been used to train and validate quite a few machine learning methods. Compared to the SynLethDB, SynLethDB 2.0 [6] now includes scoring for new CRISPR-based large-scale screening techniques, and in the second-step integration process, where custom weight parameters can be applied.

2. Prepare and integrate the implementations of the GNN model [4] for SL prediction.
3. Gather and implement suitable visual analysis system with other possible techniques (LLM, etc.) to improve the interpretability of GNN models for SL prediction.
4. An evaluation or discussion about the usability and utility of the proposed visual system or methods is recommended.

Useful links:
SynLethDB 2.0: http://synlethdb.sist.shanghaitech.edu.cn/v2
KG4SL: https://github.com/JieZheng-ShanghaiTech/KG4SL
Trustworthy Machine Learning: Trustworthy Machine Learning | Scalable Trustworthy AI

References

[1] Yong Liu, Min Wu, Chenghao Liu, Xiaoli Li, and Jie Zheng. "SL2MF: Predicting Synthetic Lethality in Human Cancers via Logistic Matrix Factorization." *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (2019).

[2] Jiang Huang, Min Wu, Fan Lu, Le Ou-Yang, and Zexuan Zhu. "Predicting synthetic lethal interactions in human cancers using graph regularized self-representative matrix factorization." *BMC Bioinformatics* 20, no. 19 (2019): 1-8.

[3] Ruichu Cai, Xuexin Chen, Yuan Fang, Min Wu, and Yuexing Hao. "Dual-Dropout Graph Convolutional Network for Predicting Synthetic Lethality in Human Cancers." *Bioinformatics* (2020).

[4] Shike Wang, Fan Xu, Yunyang Li, Jie Wang, Ke Zhang, Yong Liu, Min Wu, Jie Zheng. KG4SL: knowledge graph neural network for synthetic lethality prediction in human cancers[J]. Bioinformatics, 2021, 37(Supplement_1): i418-i425.

[5] Jing Guo, Hui Liu, and Jie Zheng. "SynLethDB: synthetic lethality database toward discovery of selective and sensitive anticancer drug targets." Nucleic Acids Research 44, no. D1 (2016): D1011-D1017.

[6] Jie Wang, Min Wu, Xuhui Huang, Li Wang, Sophia Zhang, Hui Liu, Jie Zheng. SynLethDB 2.0: a web-based knowledge graph database on synthetic lethality for novel anticancer drug discovery[J]. Database, 2022, 2022: baac030.

[7] Zhang, Q. and Zhu, S.-C., “Visual Interpretability for Deep Learning: a Survey”, arXiv e-prints, 2018. doi:10.48550/arXiv.1802.00614.

[8] P. E. Pope, S. Kolouri, M. Rostami, C. E. Martin and H. Hoffmann, "Explainability Methods for Graph Convolutional Neural Networks," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 10764-10773, doi: 10.1109/CVPR.2019.01103.

Project 5. Build a deep learning model from single cell RNA-seq data on cell life-span

Persons of Contact:
Dr. Lichun Jiang (蒋立春), email: jianglch@shanghaitech.edu.cn
Dr. Wei Wang (王玮), email: wangwei@shanghaitech.edu.cn

Introduction:

Multiple cellular organisms are composed of various of different types of cells. Each type of cell, for example, neurons and muscle cells can live for decades while the lifespan of neutrophils is only a few days. We know little about what makes this difference. Understanding the lifespan of cells may help us to better understand disease and aging. The scientific community has already accumulated many scRNA-seq datasets, and we know roughly the turn over or lifespan of major cell types. There have been some previous studies using RNA-seq data to study the characteristics of cells with different life span. We propose to apply scRNA-seq data to decipher cell life span using deep learning or other machine learning techniques.

Goals:

To explore whether we can build a deep learning model based on scRNA-seq that models cell lifespan. We can then apply this model to new datasets to predict lifespan of new type of cells

Tasks:

1. Data downloading, cleaning and organizing.
2. Data pre-processing with Geneformer or other large language model with scRNA-seq data such as scBert, scGPT etc.
3. Build and train supervised learning model on datasets with labeled data from human or mouse studies.
4. Apply model to datasets of human cell atlas datasets such as human lung cell atlas and pick out interesting cells predicted to have very short or long life-span.

Besides gene expression matrix, other variables can be considered for genes and cells:

- Length of most ubiquitously expressed isoforms of genes
- dN/dS rate of genes between mammalians
- Noise/entropy of gene expression within similar cells
- Average size of cells

Limitation of this project:

We don't know each cell's exact life-span in the study. We only know on average life span of various cell types. Life-span of cell is only meaningful for cells that no longer divide so we may need to distinguish those cells from cells capable of dividing at first place.

References：

1. Bulteau, R., Francesconi, M. Real age prediction from the transcriptome with RAPToR. Nat Methods 19, 969–975 (2022).
2. Gillooly JF, Hayward A, Hou C, Burleigh JG. Explaining differences in the lifespan and replicative capacity of cells: a general model and comparative analysis of vertebrates. Proc Biol Sci. 2012 Oct 7;279(1744):3976-80.
3. Rubin H. 1997. Cell aging *in vivo* and *in vitro*. Mech. Ageing Dev. 98, 1–35 10.1016/S0047-6374(97)00067-5
4. Ming Yang, Benjamin R. Harrison, Daniel E.L. Promislow. Cell age drives asynchronous transcriptome aging. bioRxiv 2023.05.31.543091
5. M.Elise Bullock, Thea Hogan, Sinead Morris, Maria Nowicka, Christiaan van Dorp, Andrew J. Yates, Benedict Seddon. Cell age, not chronological age, governs the dynamics and longevity of circulating CD4+ memory T cells. bioRxiv 2023.10.16.562650
6. Sender R, Milo R. 2021. The distribution of cellular turnover in the human body. *Nat Med* 27:644, pp. 45–48.
7. Zhang MJ, Pisco AO, Darmanis S, Zou J. Mouse aging cell atlas analysis reveals global and cell type-specific aging signatures. Elife. 2021 Apr 13;10:e62293. doi: 10.7554/eLife.62293.
8. Hatton IA, Galbraith ED, Merleau NSC, Miettinen TP, Smith BM, Shander JA. The human cell count and size distribution. Proc Natl Acad Sci U S A. 2023 Sep 26;120(39):e2303077120.
9. Han, X., Zhou, Z., Fei, L. et al. Construction of a human cell landscape at single-cell level. Nature 581, 303–309 (2020).
10. Sikkema, L., Ramírez-Suástegui, C., Strobl, D.C. et al. An integrated cell atlas of the lung in health and disease. Nat Med 29, 1563–1577 (2023).
11. Elmentaite, R., Kumasaka, N., Roberts, K. et al. Cells of the human intestinal tract mapped across space and time. Nature 597, 250–255 (2021).
12. Theodoris, C.V., Xiao, L., Chopra, A. et al. Transfer learning enables predictions in network biology. Nature 618, 616–624 (2023).
13. Yang, F., Wang, W., Wang, F. et al. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. Nat Mach Intell 4, 852–866 (2022).
14. Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Bo Wang. scGPT: Towards Building a Foundation Model for Single-Cell Multi-omics Using Generative AI. bioRxiv 2023.04.30.538439.
15. Spalding K. L., Bhardwaj R. D., Buchholz B. A., Druid H., Frisén J. 2005. Retrospective birth dating of cells in humans. Cell 122, 133–143 10.1016/j.cell.2005.04.028
16. Seim, I., Ma, S. & Gladyshev, V. Gene expression signatures of human cell and tissue longevity. npj Aging Mech Dis 2, 16014 (2016). https://doi.org/10.1038/npjamd.2016.14
17. Hatton IA, Galbraith ED, Merleau NSC, Miettinen TP, Smith BM, Shander JA. The human cell count and size distribution. Proc Natl Acad Sci U S A. 2023 Sep 26;120(39):e2303077120.

Project 6. Subtomogram averaging particle picking – Repeated 3D particles detection from reconstructed 3D tomogram

Persons of Contact:
PI: Quan Wang (王权), email: wangq@shanghaitech.edu.cn
TA: An Mu (慕安), email: mua@ephysics.cn

Introduction：

In recent years, electron cryo-microscopy (cryo-EM) has allowed the 3D imaging of an increasing number of biological macromolecules at resolutions sufficient for de novo atomic modelling. Among these high-resolution structures most were resolved by single-particle analysis.

However, single-particle analysis is limited to investigating isolated purified protein complexes. To examine these complexes in their biological context, cryo-electron tomography (cryo-ET) has been used instead. In cryo-ET, the sample is tilted multiple times during image acquisition, yielding tilt series of images which can be aligned and reconstructed to acquire 3D tomogram. If repeated features in those tomograms were extracted, aligned and averaged, then higher-resolution reconstructions of these (particle-like) features can be acquired, which is referred to as subtomogram averaging.

Unlike single-particle analysis, many image processing tools used for subtomogram averaging still require experienced operators for manual inspection and tuning many parameters. One of the challenges is estimating the position of each repeated particle in 3D tomograms. Because the sample cannot be rotated 180 degrees in the microscope during data acquisition, the tomogram contains empty regions in Fourier space (so-called missing wedge). This artifact may affect the accuracy of template-matching particle picking which is relative robust in single-particle analysis. Besides, complex biological context and relative thicker sample often make the contrast of particles worse than those in single-particle micrographs.

Goals:

Perform subtomogram averaging particle picking in 3D tomograms from one of the following public datasets (https://www.ebi.ac.uk/empiar/) to locate ribosome particles utilizing deep learning.

(1) Sub-tomogram averaging in RELION (https://www.ebi.ac.uk/empiar/EMPIAR-10045/)
(2) VPP subtomogram averaging (https://www.ebi.ac.uk/empiar/EMPIAR-10064/)
(3) Tilt series of native M. pneumoniae cells treated with chloramphenicol (https://www.ebi.ac.uk/empiar/EMPIAR-10499/)

References：

[1] Himes B A, Zhang P. emClarity: software for high-resolution cryo-electron tomography and subtomogram averaging[J]. Nature methods, 2018, 15(11): 955-961.

[2] Chen M, Bell J M, Shi X, et al. A complete data processing workflow for cryo-ET and subtomogram averaging[J]. Nature methods, 2019, 16(11): 1161-1168.

Project 7. De Novo Design of Binding Protein

Persons of Contact:
PI: Jie Zheng (郑杰), email: zhengjie@shanghaitech.edu.cn
TA: Tao Zhang (张涛), email: zhangtao4@shanghaitech.edu.cn
Minzhang Li (李敏章), email: limzh2022@shanghaitech.edu.cn

Background

There are 20^{200} possible amino-acid sequences for a 200-residue protein, of which the natural evolutionary process has only sampled an infinitesimal subset. Exploring the 'dark region' within the protein sequence space appears to be both significant and challenging, with de novo protein design emerging as a prominent approach.

De novo protein design, in short, is an inverse problem of protein folding. It used to be a hand-drafted task which requires users to specify their requirements and design the desired protein structures and sequences from scratch. It was not easy to speed up this procedure until deep-learning techniques emerged and have been consistently evolving. Along with the rapid advancement of AI generative model, diffusion-based techniques are applied to sampling structures of proteins.

In this project, you will get a first flavor of AI for de novo protein design from creating basic mini-protein binders.

Please note that the main model used here, named RFdiffusion [1], is computationally intensive. Students working on this project will additionally need high-performance computing (HPC) resource.

Project objectives:

- **Writing parts.**

  Read the selected papers in the references, and incorporate answers to the questions below in your report. However, your report should be in the form of a conference or journal paper, i.e. containing more than just answers to these questions.

  Q1: Briefly explain the distinct roles of 'protein backbone generation' and 'protein sequence optimization' in the process of protein design. Additionally, what are the advantages and disadvantages of treating these two aspects separately? ([1, 2, 4])

  Q2: Parameters of *de novo* protein backbone generation lie in the **SE(3) space**, while image parameters in diffusion process lie in the **pixel space (Euclidean space)**. What theoretical concerns does this difference raise when adapting diffusion techniques to protein backbone generation?

  Q3: The radius of gyration (Rg for short) is a metric that measures the folding degree of protein backbone. It describes important biophysical properties of generated backbones. Give its formula a short, concise description. What can we expect (or infer) from a large/small Rg score, in terms of protein conformations and stability, respectively? What about the Rg scores of protein backbones with many loop regions? What about 3-helix bundles? ([5])

- **Coding parts.**

  After reading the reference papers, you would already have a rough idea of the four steps of binding protein de novo design. The process of designing a protein that binds to a specific receptor involves a series of complex and exquisite steps. First, the **active site of the receptor**, where the protein will bind, must be selected and determined based on biophysical properties. This is a crucial part of the process. Secondly, the protein's **backbone structure** is designed, including determining its overall structure and folding pattern to ensure compatibility and stability with the target active site. Thirdly, the amino acid **sequence of the protein** is optimized, adjusting the sequence for optimal interaction and functionality. The final step involves using **computational tools to evaluate** the designed protein. This process combines advanced software tools, a deep understanding of protein structures and functions, and iterative testing and refinement. The goal is to create a mini-protein binding to the specified receptor, with solid *in silico* validation.

  Designing a protein, especially one that binds to a specific receptor, is a complex and fascinating process. Luckily, you do not have to implement all steps from scratch, as we will offer you an HPC account with images (see [3] for information about *docker* and *image*), containing a basic toolset and a scaffold protein library. Note that our aim is to enhance your comprehension of the interrelation between protein sequences, structures, and functions in the process of designing binding proteins, by using the tools. However, we encourage you to explore more deeply into the ideas behind the tools.

  A recommended list of steps (please feel free to try your own ideas to make innovation):

  (a) Preprocess and analyze a specified receptor protein. (Tip: you may find ‘biopython’ module useful for preprocessing and atomic analysis of the given receptor.)

  (b) Select active sites of the target proteins. (Tip: you may find hydrophobicity significant. [6])

  (c) Generate backbones through RFdiffusion. (Tip: you may find fold-conditioning strategy useful. Also, use the scaffold library that we offer wisely.)

  (d) Implement Rg code based on your own understanding and formula. Test it on results of (c).

  (e) Optimize protein sequences.

  (f) *In silico* evaluation. It is up to you to run (or implement) appropriate metric on your results. (Tip: [7])

You should submit a .zip/.tar file containing your report and results. More instructions will be given near the date for final presentations.

References

[1] Watson, J.L., Juergens, D., Bennett, N.R. et al. De novo design of protein structure and function with RFdiffusion. Nature 620, 1089–1100 (2023). https://doi.org/10.1038/s41586-023-06415-8

[2] Jason Yim, Brian L. Trippe, Valentin De Bortoli, Emile Mathieu, Arnaud Doucet, Regina Barzilay, & Tommi Jaakkola. SE(3) diffusion model with application to protein backbone generation. Proceedings of the 40th International Conference on Machine Learning (ICML), July 2023.

[3] https://yeasy.gitbook.io/docker_practice/basic_concept/image

[4] Huang, PS., Boyken, S. & Baker, D. The coming of age of de novo protein design. Nature 537, 320–327 (2016). https://doi.org/10.1038/nature19946

[5] https://en.wikipedia.org/wiki/Radius_of_gyration

[6] Cao, L., Coventry, B., Goreshnik, I. et al. Design of protein-binding proteins from the target structure alone. Nature 605, 551–560 (2022). https://doi.org/10.1038/s41586-022-04654-9

[7] https://github.com/nrbennet/dl_binder_design

END OF CS286 PROJECT LIST