CS177 Bioinformatics: Software Development and Applications

Project Guidelines

School of Information Science and Technology, ShanghaiTech University

Spring, 2025

Lecturer: Associate Professor Jie Zheng (郑杰), Email: zhengjie@shanghaitech.edu.cn

Teaching Assistants:

- Mr. Mutian Hong (洪沐天), Email: hongmt2022@shanghaitech.edu.cn
- Miss Ting Lu (卢婷), Email: luting2024@shanghaitech.edu.cn

1. General information and rules

The projects are mainly about implementing, comparing and improving algorithms and deep learning models in bioinformatics. Through a hands-on project, students can test their understanding of computational problems and ideas in bioinformatics, and learn how to make innovations in AI and data science. Moreover, students will be trained to present their project ideas and results clearly through oral presentations and written reports, paving the way for becoming successful researchers or engineers in the future. Hereafter, "you" refers to every student who has been officially enrolled in the course.

Each project is **group-based** and each group comprises **2-3 students**, and you should not rely on anyone outside of your team for the project. The instructors (Dr. Jie Zheng, and his TAs, Ting Lu and Mutian Hong) are the persons who will make judgment and evaluation of your performance in the project.

2. Timeline

- **Projects start:** April 2, 2025
- **Mid-term personal interviews:** Week 11 (date TBD)
- **Oral representations:** Week 16 (date TBD)
- **Submission of final reports (with other project files):** Week 16 (date TBD)

3. Description of projects

There are 3 project options as described below. Each group is required to choose one and only one project. In principle, each project is to be chosen by at most 3 groups. *If more groups happen to choose the same project, adjustment would be made by interviews and discussion.* Once choosing a project, you should not switch to another one, unless absolutely necessary.

Project 1: Leveraging Multi-omics Data for Prediction of Context-specific Synthetic Lethality

Background

Synthetic lethality (SL) is a type of genetic interaction in which the simultaneous inactivation of two genes leads to cell death, while the inactivation of a single gene does not affect the cell viability, as illustrated in Figure 1. Therefore, SL can be leveraged to selectively kill cancer cells by targeting SL partners of cancer-specific genetic abnormalities while keeping normal cells alive. However, high-throughput wet-lab screenings of SLs are often costly and face various challenges (e.g., off-target effects). Therefore, deep learning (DL) methods for SL prediction have become popular in the past decade.

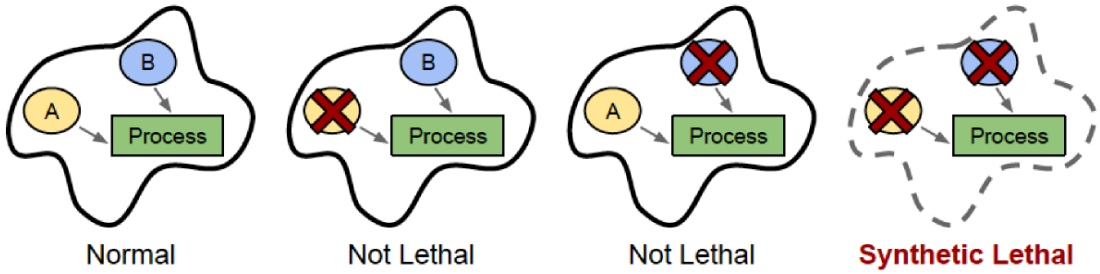

Figure 1. The concept of synthetic lethality.

Many machine learning approaches [1-8] have been proposed for predicting SLs. However, recent studies have shown that many SL relationships occur specifically in certain cell lines, i.e., two genes may have SL relationship in one cell line but not in other cell lines. In other words, SL gene pairs are significantly heterogeneous among cell lines. Therefore, the heterogeneity of SL pairs across cell lines necessitates models for more precise SLs prediction. Multi-omics data have been demonstrated to reflect the context-specific genetic background of cell lines or cancer types.

Goal: In this project, we aim to design a powerful DL model to learn from the multi-omics data

and thereby enhance the performance of predicting context-specific SLs.

Sub-goals and recommended steps:

1. Gather and integrate multi-omics data for capturing context-specific information. You are recommended (but not required) to select the multi-omics data from the following list:
   - Biomedical Knowledge Graphs (BKGs): PrimeKG, SynLethKG;
   - Gene expression data, copy number variation, and mutation: CCLE, TCGA;
   - Textual descriptions of genes or cell lines: NCBI gene card summary of human genes (GenePT provides the data and the embeddings here); cell line descriptions from Cellosaurus.

   Tip: Align gene identifiers across different datasets.
2. Select and implement the DL methods for learning the representations of the multi-omics data:
   - KG embeddings: ConvE, GNN-based encoders;
   - Gene/cell line descriptions: BioBERT, GenePT, or other biomedical LLMs.
   - Omics data (e.g., gene expression): Geneformer or scGPT.

   To assist you in constructing your model more effectively, we provide an example model framework for your reference (shown in Figure 2).

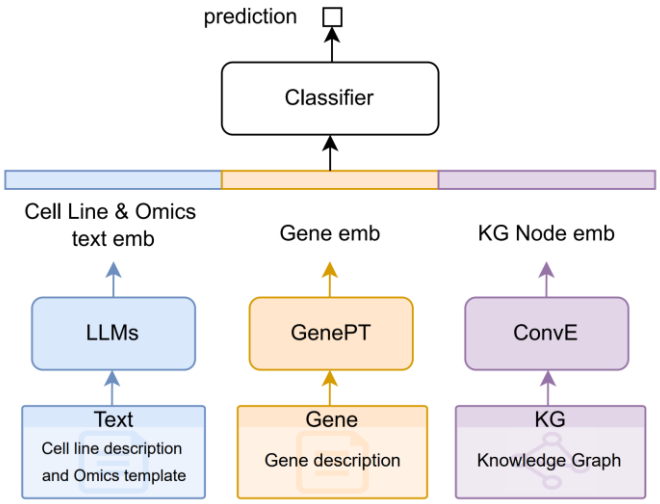

Figure 2. An example model framework.

3. Evaluate the performance of your proposed model and the baselines compared with it in the context-specific SL prediction scenarios. Note that we formulate the context-specific SL prediction as a binary classification task, i.e., 1 for positive SL pairs and 0 for non-SL pairs.

   In this project, you need to complete ***at least*** the first of the following two experiments:

   1) **Cell-line-specific SL prediction:** For each cell line, use 5-fold cross-validation (CV) to randomly split the gene pairs in this cell line. All the compared models are to be trained and

tested independently for the cell line.

2) **Cell-line-adapted SL prediction:** In this experiment, we evaluate the models in a cross-cell-line prediction scenario, for example, (IPC298, A375)→A549, i.e., the first two cell lines provide the training data, and the last cell line the test data.

Here, we provide the SL label data (SLKB):

https://epan.shanghaitech.edu.cn/l/mFFQJp (提取码：CS177)

The baselines to be compared (you can include more):

1) KG4SL [1]
2) SLMGAE [2]
3) MVGCN-iSL [4]

Besides, it is recommended to use the following classification metrics:

- Area under the receiver operating characteristic curve (AUC),
- Area under the precision-recall curve (AUCPR),
- F1 score,
- Balanced accuracy (BACC).

4. A discussion about the contributions of the multi-omics data combinations of the proposed methods is recommended. For example, which features contribute most to improving SL prediction and why?

Useful links:

SynLethDB 2.0: http://synlethdb.sist.shanghaitech.edu.cn/v2

PrimeKG: https://zitniklab.hms.harvard.edu/projects/PrimeKG/

ConvE: https://pykeen.readthedocs.io/en/stable/api/pykeen.models.ConvE.html

GenePT: https://github.com/yiqunchen/GenePT

SLKB: https://slkb.osubmi.org/

SL Benchmark:
https://github.com/JieZheng-ShanghaiTech/SL_benchmark?tab=readme-ov-file#benchmarking-of-machine-learning-methods-for-predicting-synthetic-lethality-interactions

References

[1] Wang S, Xu F, Li Y, et al. KG4SL: knowledge graph neural network for synthetic lethality prediction in human cancers. Bioinformatics, 2021, 37(Supplement_1): i418-i425.

[2] Hao Z, Wu D, Fang Y, et al. Prediction of synthetic lethal interactions in human cancers using multi-view graph auto-encoder. IEEE Journal of Biomedical and Health Informatics, 2021, 25(10): 4041-4051.

[3] Wang S, Feng Y, Liu X, et al. NSF4SL: negative-sample-free contrastive learning for ranking synthetic lethal partner genes in human cancers. Bioinformatics, 2022, 38(Supplement_2):

ii13-ii19.

[4] Fan K, Tang S, Gökbağ B, et al. Multi-view graph convolutional network for cancer cell-specific synthetic lethality prediction. Frontiers in Genetics, 2023, 13: 1103092.

[5] Xing Y, Pu M, Tian K, et al. Cell context-specific synthetic lethality prediction and mechanism analysis. bioRxiv, 2023: 2023.09. 13.557545.

[6] Tepeli Y I, Seale C, Gonçalves J P. ELISL: early–late integrated synthetic lethality prediction in cancer[J]. Bioinformatics, 2024, 40(1): btad764.

[7] Zhao X, Liu H, Dai Q, et al. Multi-omics Sampling-based Graph Transformer for Synthetic Lethality Prediction. 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2023: 785-792.

[8] X. Liu, S. Tao and J. Zheng. Meta Learning for Low-Data Prediction of Cancer-Specific Synthetic Lethality as Drug Targets. 2023 IEEE International Conference on Knowledge Graph (ICKG), Shanghai, China, 2023, pp. 255-262, doi: 10.1109/ICKG59574.2023.00037.

Project 2: Assessing the Reliability of AlphaFold3 Predictions for Protein-Ligand Affinity Prediction via Sfcnn

Background

Protein-Ligand Affinity (PLA) prediction is a critical area in computational drug discovery, enabling efficient screening of drug candidates and understanding their interactions with target proteins. Traditional methods rely heavily on time-consuming and costly experimental approaches, and as such computational models have gained popularity with their ability to accelerate this process. In particular, deep learning models leveraging multi-modal features of drugs and proteins (e.g., sequences, structures, and physicochemical properties) have shown promise in PLA prediction tasks (Figure 3).

A key challenge in structure-based PLA methods, which rely on 3D structures of proteins and ligands, is the scarcity of experimentally resolved protein–ligand complexes. Recent advances in AI for protein structure prediction, particularly AlphaFold2 (AF2) [1] and AlphaFold3 (AF3) [2], enable accurate modeling of protein structures and biomolecular interactions, offering potential alternatives to experimentally determined structures. However, the reliability of AF3-derived models for PLA prediction remains to be validated more thoroughly.

In this project, you are required to reproduce an open-source PLA model using PyTorch [3], and verify the correctness of your implementation by evaluating its performance on the original dataset. Then, further test the model on a dataset constructed from AlphaFold3 predicted structures, to explore whether AF3-derived structures can reliably substitute for experimental structures in PLA prediction.

Project objective

The first goal is to reproduce the PLA prediction model named **Sfcnn** [4] on an open-source dataset **PDBbind** [5]. Sfcnn is a structure-based PLA prediction model based on 3D-CNN, while PDBbind collects PLA data from Protein Data Bank (PDB) [6] database together with their 3D structures.

The second goal is to construct a new test set by applying AF3 to the protein targets in the Sfcnn test set, and evaluate the performance of the replicated Sfcnn model on this AF3-derived test set.

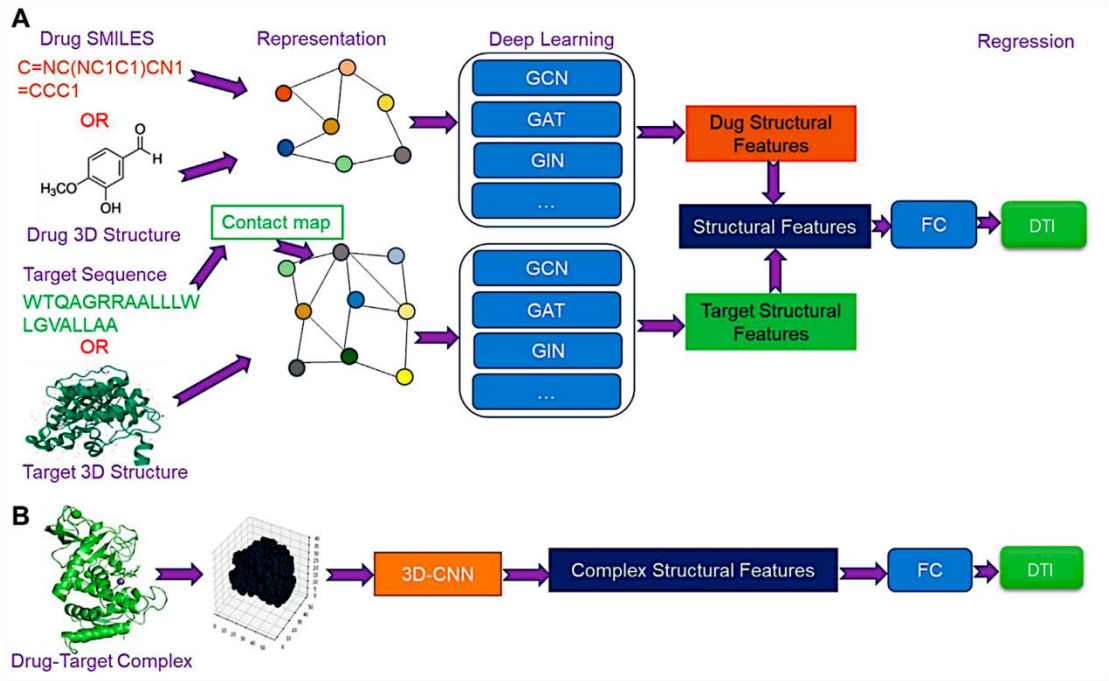

Figure 3. The overall architectures of structure-based deep learning methods. (A) The extraction of structural features from molecular graphs of drugs and targets; (B) the extraction of structural features of drug-target complexes from their 3D structures [7].

Recommended steps:

1. Data Collection and Preprocessing:
   - Use the same training and test sets as Sfcnn by collecting data from the PDBbind v2019 refined set for training and the core set (CASF-2016 [8]) for testing. These datasets are publicly available at (http://www.pdbbind.org.cn/).
   - Run the local AlphaFold3 or AlphaFold3 web server to predict protein-ligand complex in CASF-2016 dataset (https://github.com/google-deepmind/alphafold3).
2. Reproduce Sfcnn (https://github.com/bioinfocqupt/Sfcnn):
   - Reproduce the model of Sfcnn with PyTorch.
   - Reproduce the result of Sfcnn on the CASF-2016 dataset.
3. AlphaFold3 Performance Test:
   - Run Sfcnn on the predictions of AlphaFold3 and analyze its performance in comparison with the original CASF-2016 dataset.

References

[1] Jumper et al. Highly accurate protein structure prediction with AlphaFold. Nature, 2021, 596(7873): 583-589.

[2] Abramson et al. Addendum: Accurate structure prediction of biomolecular interactions with

AlphaFold 3. Nature, 2024: 1-1.

[3] Ansel J, Yang E, He H, et al. Pytorch 2: Faster machine learning through dynamic python bytecode transformation and graph compilation. Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2. 2024: 929-947.

[4] Wang et al. Sfcnn: a novel scoring function based on 3D convolutional neural network for accurate and stable protein–ligand affinity prediction. BMC Bioinformatics, 2022, 23(1): 222.

[5] Wang et al. The PDBbind database: methodologies and updates. Journal of Medicinal Chemistry, 2005, 48(12): 4111-4119.

[6] Berman et al. Announcing the worldwide protein data bank. Nature Structural & Molecular Biology, 2003, 10(12): 980-980.

[7] Zeng, X. et al. A comprehensive review of the recent advances on predicting drug-target affinity based on deep learning. Frontiers in Pharmacology, 2024, 15: 1375522.

[8] Su et al. Comparative assessment of scoring functions: the CASF-2016 update. Journal of Chemical Information and Modeling, 2018, 59(2): 895-913.

Project 3: Drug Synergy Prediction through Reproduction and Improvement of SynergyX

Background

Drug synergy refers to pharmacological interactions where the combined therapeutic effect of two drugs exceeds the sum of their individual effects. Combinations demonstrating such synergistic effects hold significant clinical promise, in contrast to those exhibiting antagonistic or additive effects, where combined efficacy is less than or equal to individual effects (Figure 4) [1-2]. Identifying synergistic drug pairs enables optimized combination therapies that enhance treatment efficacy while reducing toxicity, dosage requirements, and drug resistance -- a pivotal strategy in modern targeted cancer therapies [3-5].

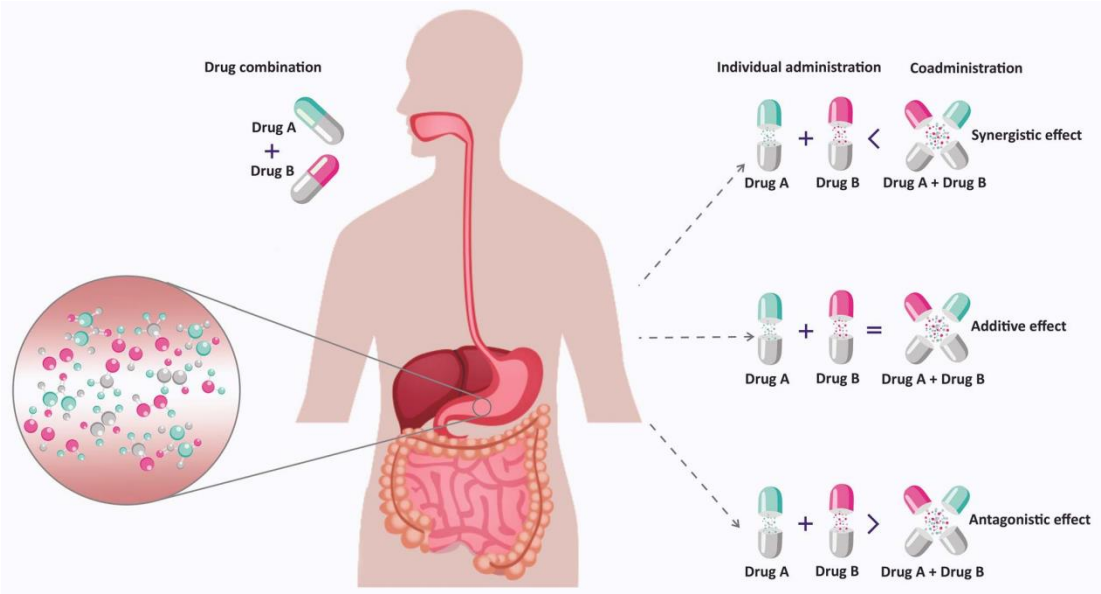

Figure 4. Conceptual diagram of drug synergy [1].

However, the combinatorial explosion of potential drug pairs renders exhaustive wet-lab experimentation prohibitively costly and time-consuming [6]. Computational prediction of drug synergy has thus emerged as a critical research frontier. Recent advances employ deep learning techniques with innovations in drug encoding, cell line characterization, information fusion architectures, and interpretability analysis [7]. Among these, SynergyX [8] represents a state-of-the-art deep learning framework notable for its simple and effective drug representation and elegantly designed information fusion mechanism, achieving superior performance. Nevertheless, persistent limitations exist in current methodologies including SynergyX: sub-optimal cell line characterization pipelines, inconsistent performance across experimental settings, and computationally intensive training processes requiring acceleration.

Objective

In this project, we aim to build a new drug synergy prediction method based on the SynergyX framework and the data it uses from the perspective of improvement or innovation, toward achieving performance improvement or solving pain points (such as long training time) for existing methods.

Sub-goals and recommended steps

1. Selection of Drug Synergy Labeled Data

Recommendation: Use SynergyX's preprocessed dataset derived from DrugComb [9], which has been rigorously filtered and cleaned. Given its substantial size, we recommend initial model tuning with a subset of the data followed by full-dataset benchmarking against baseline models.

Additional public datasets:

Table 1. A summary of drug synergy combination datasets [7].

Dataset	# of drugs	# of cell lines	# of drug pairs	# of combination
DrugComb	8397	2040	74421	751498
O'Neil	38	39	583	22737
DrugCombDB	2887	124	448555	6055926
ALMANAC	104	60	5232	304549
SYNERGxDB	1977	151	22507	536596

Note: While you may select any suitable dataset beyond the recommended list, please ensure to do the following when using raw datasets:

- Explicitly define the synergy score metric (e.g., Loewe, Bliss, or ZIP score).
- Perform necessary data cleaning and standardization.

2. Drug Encoding and Cell Line Representation

There are several methods you should use to encode drug and cell line information, and we present some suggestions for your reference (see Figure 5).

Drug encoding strategies:

- 1D: SMILES sequence encoding using transformer-based models.
- 2D: Molecular graph via Graph Neural Networks (GNNs) or molecular fingerprint.
- 3D: 3D structural encoding with frameworks like GeminiMol [10].

Cell line characterization:

- Protein-protein interaction (PPI) network embeddings.
- Multi-omics data integration.

Exploration encouraged: We encourage you to find or create new characterization methods to

improve the overall performance of prediction.

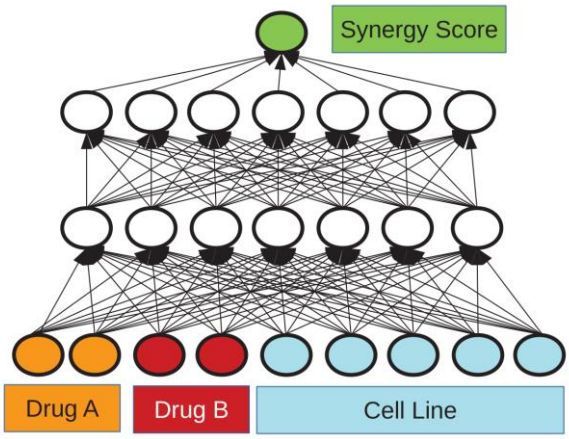

Figure 5. The basic framework of drug synergy prediction [11]

3. Local Deployment/Reproduction of SynergyX

SynergyX serves as a SOTA baseline due to its robust and stable performance. You are required to successfully deploy SynergyX's GitHub project (https://github.com/GSanShui/SynergyX) and reproduce its framework to achieve performance comparable to the original publication.

Key considerations:

- Resolve potential code inconsistencies or undocumented dependencies during deployment.
- Validate model stability and reproducibility using benchmark datasets.

Final Project Requirements

The implemented framework must fulfill:

- Comprehensive chemical structure compatibility: Encode all drugs with standard chemical notations.
- Dual-task adaptability: Support both regression (continuous synergy prediction) and classification (synergistic/non-synergistic binary prediction) tasks. Employ task-specific evaluation metrics (e.g., MSE for regression; AUROC for classification).
- Performance superiority: Exceed most baseline models (including SynergyX itself if your work is focused on improvement).

References

[1] Torkamannia A, Omidi Y, Ferdousi R. A review of machine learning approaches for drug synergy prediction in cancer. Briefings in Bioinformatics, 2022, 23(3): bbac075.

[2] Chou T C. Theoretical basis, experimental design, and computerized simulation of synergism and antagonism in drug combination studies. Pharmacological Reviews, 2006, 58(3): 621-681.

[3] Gottesman M M, Lavi O, Hall M D, et al. Toward a better understanding of the complexity of cancer drug resistance. Annual Review of Pharmacology and Toxicology, 2016, 56(1): 85-102.
[4] Zheng W, Sun W, Simeonov A. Drug repurposing screens and synergistic drug-combinations for infectious diseases. British Journal of Pharmacology, 2018, 175(2): 181-191.
[5] Goldoni M, Johansson C. A mathematical approach to study combined effects of toxicants in vitro: evaluation of the Bliss independence criterion and the Loewe additivity model. Toxicology in Vitro, 2007, 21(5): 759-769.
[6] Bansal M, Yang J, Karan C, et al. A community computational challenge to predict the activity of pairs of compounds. Nature Biotechnology, 2014, 32(12): 1213-1222.
[7] Wang Y, Wang J, Liu Y. Deep learning for predicting synergistic drug combinations: State-of-the-arts and future directions. Clinical and Translational Discovery, 2024, 4(3): e317.
[8] Guo Y, Hu H, Chen W, et al. SynergyX: a multi-modality mutual attention network for interpretable drug synergy prediction. Briefings in Bioinformatics, 2024, 25(2): bbae015.
[9] Zagidullin B, Aldahdooh J, Zheng S, et al. DrugComb: an integrative cancer drug combination data portal. Nucleic Acids Research, 2019, 47(W1): W43-W51.
[10] Wang L, Wang S, Yang H, et al. Conformational Space Profiling Enhances Generic Molecular Representation for AI-Powered Ligand-Based Drug Discovery. Advanced Science, 2024, 11(40): 2403998.
[11] Preuer K, Lewis R P I, Hochreiter S, et al. DeepSynergy: predicting anti-cancer drug synergy with Deep Learning. Bioinformatics, 2018, 34(9): 1538-1546.

END OF CS177 PROJECT GUIDELINES (SPRING, 2025)