前期的数据集构建工作

（1）将DAVIS、 BindingDB 和 BioSNAP数据集合并

（2）去除序列完全相同的蛋白质-小分子对

目前没有发现一篇论文将不同的药物-靶标相互作用（DTI）数据集直接合并，为什么？

合并数据集时，最需要注意两个问题：

（1）不同数据集中DTI的定义标准。

（2）不同数据集之间的冗余性。

1. 如何定义药物-靶标的相互作用

（1）计算药物-靶标的亲和力。

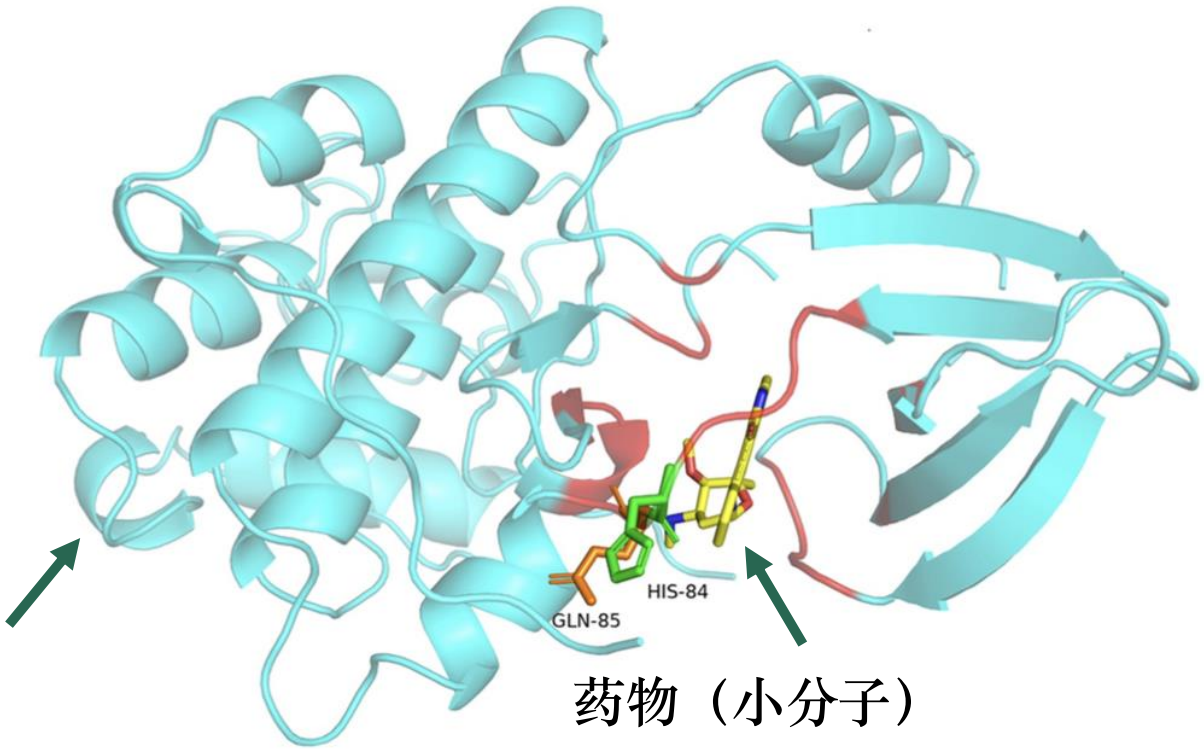
（2）给定阈值T：药物-靶标的亲和力高于阈值T，标记为相互作用（正样本）；

反之，标记为非相互作用（负样本）。

2. 如何定义药物-靶标的亲和力

（1）解离常数 (K_D)

（2）抑制常数 (K_i)



3. 解离常数 (K_D)

先看最经典的结合反应：

$$D + T \rightleftharpoons DT$$

- D：drug（配体 / 抑制剂）
- T：target（蛋白 / 激酶）
- DT：两者结合形成的复合物

解离常数 K_d 定义为：

$$K_d = \frac{[D]\,[T]}{[DT]}$$

在平衡时，

- $[D]$：游离药物浓度
- $[T]$：游离靶点浓度
- $[DT]$：结合在一起的复合物浓度

👉 直观理解：

> **Kd 越大，说明“解离”那边占优势（更爱分开）；**
> **Kd 越小，说明更容易“结合在一起”。**

常用的阈值：30 nM

Benchmarks Overview.

Low coverage benchmarks. We evaluate our framework on three broad-scale, low-coverage benchmark datasets. Two datasets, **DAVIS** (68) and **BindingDB** (69), consist of pairs of drugs and targets with experimentally determined dissociation constants (K_D). Following ref. 13, we treat pairs with $K_D < 30$ as positive DTIs, while larger K_D values are negative. The third dataset, ChG-Miner from **BIOSNAP** (70), consists of only positive DTIs. We create negative DTIs by randomly sampling an equal number of protein–drug pairs, making the assumption that a random pair is unlikely to be positively interacting. The DAVIS dataset represents a few-shot learning setting: It contains only 2,086 training interactions, compared to 12,668 for BindingDB and 19,238 for BIOSNAP. The rest of the data preparation follows (13). The datasets are split into 70% for training, 10% for validation, and the remaining 20% for testing. Training data are artificially subsampled to have an equal number of positive and negative interactions, while validation and test data are left at the ratio originally in the dataset.

Contrastive learning in protein language space predicts interactions between drugs and protein targets. PNAS, 2023.

3. 解离常数 (K_D)

3 为什么大家喜欢用 pKd？

因为 Kd 经常是这些值：

- 0.1 nM = 1×10^{-10} M
- 10 nM = 1×10^{-8} M
- 1 μM = 1×10^{-6} M

不好直接比较，所以经常用：

$$\mathrm{pKd} = -\log_{10}\left(\frac{\mathrm{Kd}}{10^{-9}}\right)$$

- 例如，**Kd = 30 nM**，则：

$$\mathrm{pKd} = -\log_{10}\left(\frac{30\times10^{-9}}{10^{-9}}\right) = -\log_{10}(30) \approx 7.52$$

pKd 越大，亲和力越强，数值直观又方便做回归，是 DAVIS 等数据集中常用的标签。

常用的阈值：7.5

While Pahikkala *et al.* (2014) used the K_d values of the Davis dataset directly as the binding affinity values, we used the values transformed into log space, pK_d, similar to He *et al.* (2017) as explained in Equation (1).

$$pK_d = -\log10\left(\frac{K_d}{1e9}\right) \quad (1)$$

Figure 1A (left panel) illustrates the distribution of the binding affinity values in pK_d form. The peak at pK_d value 5 (10 000 nM) constitutes more than half of the dataset (20 931 out of 30 056). These values correspond to the negative pairs that either have very

DeepDTA: deep drug–target binding affinity prediction. Bioinformatics, 2018.

4.抑制常数 (K_i)

1. Ki的定义

Ki 是 **抑制常数（Inhibition constant）**，用于描述抑制剂与酶或受体结合的亲和力。**Ki 越小**，表示抑制剂对靶标的亲和力越强。

2. Ki与Kd、Km的关系

Ki 和 Kd（解离常数）之间是密切相关的，尤其在 **竞争性抑制** 的情况下。**Ki** 是通过 **Kd** 和 **Km** 来计算的，具体计算公式如下：

$$K_i = \frac{K_d}{1 + \frac{[S]}{K_m}}$$

其中：

- **Ki**：抑制常数，表示抑制剂的亲和力。
- **Kd**：药物与靶点的解离常数，表示药物与靶点的结合亲和力。
- **[S]**：底物浓度。
- **Km**：米氏常数，表示酶与底物的亲和力。

(4) Binary labeling of DTIs: Bioactivity measurements (first K_d, then K_i, then IC50) were converted into binary interactions based on a threshold. When multiple bioactivity measurements have a difference of less than one log unit: if the average bioactivity value was less than 100 nM (10^{-7} M), the interaction was labeled as a positive DTI (binding). If the average bioactivity value was greater than 100 μM (10^{-4} M), the interaction was labeled as a negative DTI (nonbinding). If the average bioactivity value was in the intermediate range, i.e. between 100 nM

Drug–Target Interactions Prediction at Scale: The Komet Algorithm with the LCIdb Dataset. Journal of Chemical Information and Modeling, 2022.

5. DAVIS 数据集

靶标：442 激酶(kinases)，超过80%来源于人类。

药物：72激酶抑制剂(kinase inhibitors)；实验成功数量：68。

靶标-药物总对数：30056

靶标-药物亲和力： K_D 值

常用阈值： 30 nM

Davis, Mindy I., et al. "Comprehensive analysis of kinase inhibitor selectivity."

Nature biotechnology 29.11 (2011): 1046-1051.

5. DAVIS 数据集

Table 1. Details of benchmark datasets.

Dataset	Drugs	Targets	Training	Validation	Test
Davis	68	442	24 044	3006	3006
PDBbind	6487	5266	7512		207
TDC-DG	140 469	476	182 905		48 992
BindingDB	7165	1254	12 657		13 272

- DTIAM: a unified framework for predicting drug-target interactions, binding affinities and drug mechanisms. Nature Communications, 2025.
- PMMR: generalizability of drug–target binding prediction by pre-trained multi-view molecular representations. Bioinformatics, 2025.

总样本：30056，训练集：24044（80%），

验证集：3006（10%），测试集：3066（10%）

Table 4. Full specification of benchmark datasets

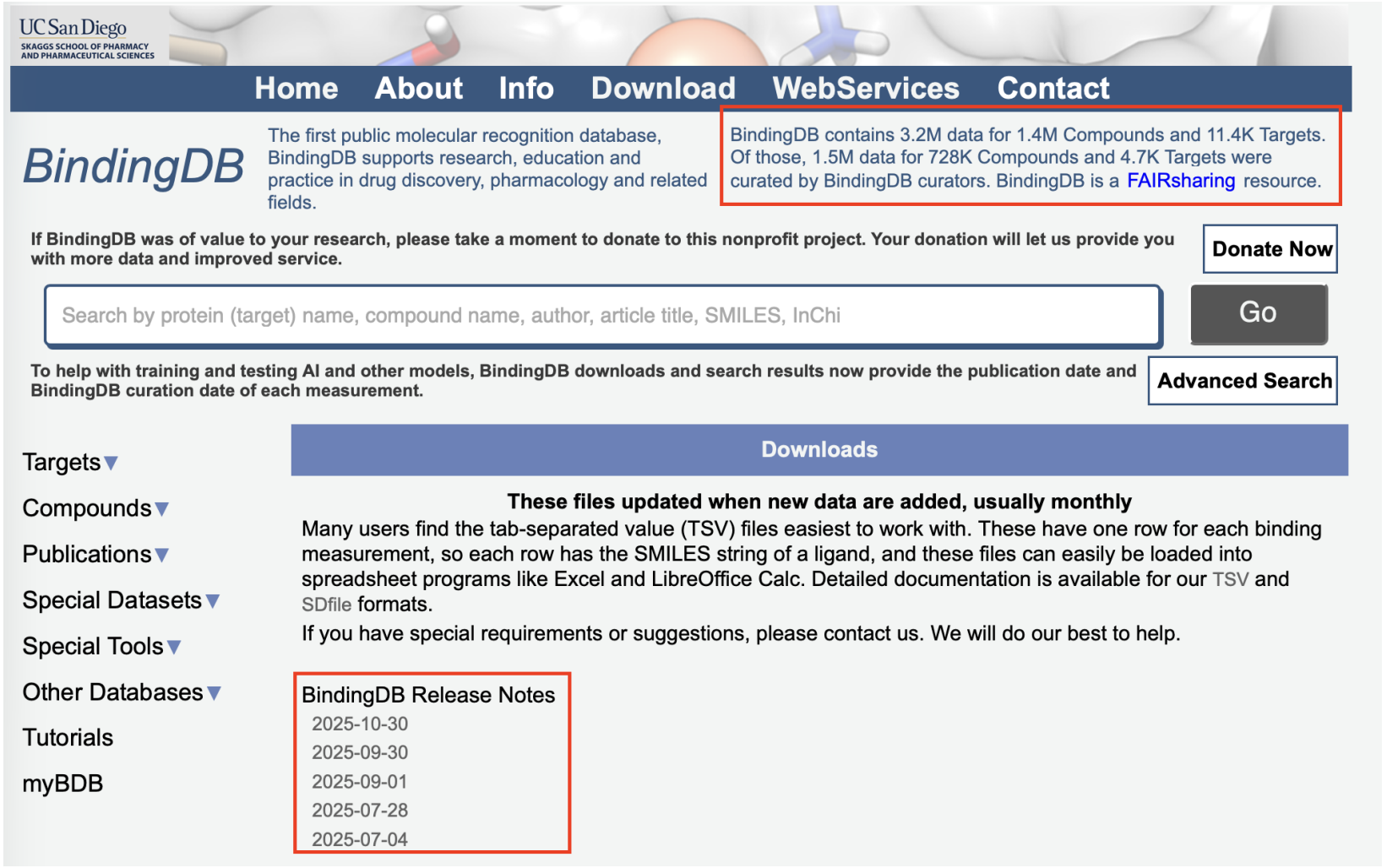
Dataset	Drugs	Targets	Median Coverage	# Training	# Validation	# Test
BIOSNAP	4,510	2,181	0.0023/0.0020	9,670/9,568	1,396/1,352	2,770/2,727
Unseen Drugs				9,535/9,616	1,383/1,353	2,918/2,675
Unseen Targets				9,876/9,499	1,382/1,386	2,578/2,762
BindingDB	7,165	1,254	0.0008/0.0010	6,334/6,334	927/5,717	1,905/11,384
DAVIS	68	379	0.3707/0.3676	1,043/1,043	160/2,846	303/5,708
TDC-DG	140,746	477	0.0021/0.0005	146,891	36,539	49,028
Phosphatase	165	218	1.0/1.0	5,054/27,286	—	370/3,260
Esterase	96	146	1.0/1.0	2,150/10,426	—	926/514
Glycosyltransferase	89	54	0.9259/0.9778	725/3,042	—	113/417
Halogenase	62	42	1.0/1.0	303/1,991	—	20/290
BKACE	17	161	1.0/1.0	255/2,193	—	19/270
DUD-E †				8,996/406,208	—	11,430/521,132
GPCR	99,671	5	18,563			
Kinase	315,399	26	15,409			
Protease	286,089	15	9,271			
Nuclear	151,133	11	16,257			

- Contrastive learning in protein language space predicts interactions between drugs and protein targets. PNAS, 2023.
- Molecular Interaction Transformer for drug-target interaction prediction. Bioinformatics, 2021.

总样本：11103，正样本：1506（阈值30nM）

训练集：2086，验证集：3006（10%），测试集：6011（20%）

6. BindingDB数据集 (http://www.bindingdb.org)



Liu, Tiqing, et al. "BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities." Nucleic acids research 35.suppl_1 (2007): D198-D201.

6. BindingDB数据集 (http://www.bindingdb.org)

4.1 Dataset

BindingDB [Gilson *et al.*, 2016] is a public, web-accessible database of measured binding affinities, focusing chiefly on the interactions of small molecules (drugs/drug candidates) and proteins (targets/target candidates). We took a snapshot of BindingDB that contains 1.3 million data records, each of which contains information such as the identifiers of involved entities, the observed experiment results, etc. By the following criteria we construct a binary classification dataset[2] with 39,747 positive examples and 31,218 negative examples.

1. Record has chemical identifier (PubChem CID), and the small molecule has chemical structure represented by SMILES[3].
2. Record has protein identifier (Uniprot ID), and the protein has both sequence representation and Gene Ontology annotations [Ashburner *et al.*, 2000].
3. Record has IC50 value, a primary measure of binding effectiveness.
4. The chemical molecule weight is less than 1,000Da, due to our focus on small molecule drugs.
5. By following the activity threshold discussion in [Wang *et al.*, 2016], record is positive if its IC50 is less than 100nm, negative if IC50 greater than 10,000nm.

- Gao, Kyle Yingkai, et al. "Interpretable drug target prediction using deep neural representation." IJCAI. Vol. 2018. 2018.
- Wu, Yifan, et al. "BridgeDPI: a novel graph neural network for predicting drug–protein interactions." Bioinformatics 38.9 (2022): 2571-2578.

Table 1. Dataset statistics

Dataset	# Drugs	# Proteins	# Pos Interactions	# Neg Interactions
BIOSNAP	4510	2181	9619/1374/2748	9619/1374/2748
DAVIS	68	379	1043/160/303	1043/2846/5708
BindingDB	10 665	1413	6334/927/1905	6334/5717/11 384

Note: For the number of interactions columns, we include training/validation/testing interactions statistics in onefold of data.

Positive: KD<30 nM

- Huang, Kexin, et al. "MolTrans: molecular interaction transformer for drug–target interaction prediction." Bioinformatics 37.6 (2021): 830-836.
- Singh, Rohit, et al. "Contrastive learning in protein language space predicts interactions between drugs and protein targets." Proceedings of the National Academy of Sciences120.24 (2023): e2220778120.
- Ouyangke, et al. "Improving generalizability of drug–target binding prediction by pre-trained multi-view molecular representations." Bioinformatics 41.1 (2025): btaf002.

A. Experimental setup

1) Dataset: we construct a low-bias version of binary BindingDB [4, 5] dataset in this experiment. Following the IC_{50} threshold used by Gao et al. [3], we consider a drug-target pair to be positive if its IC_{50} is less than 100 nm, and negative if its IC_{50} is greater than 10,000 nm, which is a 100-fold difference.

2) Bias-reducing preprocessing: Due to the drug-wise pair imbalance, 91% of drugs only have one type of pairs (positive or negative) in the binary BindingDB dataset. This implies that we can train a model to make right classification without considering protein information for DTI pairs associated with the 91% of drugs. High classification accuracy does not indicate successful learning of correct DTI patterns.

Therefore, we further process the data by removing all DTI pairs of drugs containing only one pair type. This gives us a low-bias dataset with 29,674 positive samples and 32,752 negative samples. Figure 2a shows the drug probability distribution in terms of log ratios of positive to negative samples in the dataset, which is calculated as:

- Bai, Peizhen, et al. "Hierarchical clustering split for low-bias evaluation of drug-target interaction prediction." 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2021.
- Luo, Zhengchao, et al. "Accurate and transferable drug–target interaction prediction with DrugLAMP." Bioinformatics 40.12 (2024): btae693.

7. BioSNAP数据集 (http://snap.stanford.edu/biodata)

By *Jure Leskovec* STANFORD UNIVERSITY

Drug-target interaction network

Dataset information

This is a drug-target interaction network that contains information on which genes (i.e., proteins encoded by genes) are targeted by drugs that are on the U.S. market. Drug targets are molecules that play a critical role in the transport, delivery or activation of the drug. Drug target information is widely used to facilitate computational drug target discovery, drug design, drug docking or screening, drug metabolism prediction, drug interaction prediction, and general pharmaceutical research.

- SNAP for C++
- SNAP for Python
- SNAP Datasets
- BIOSNAP Datasets
- What's new
- People
- Papers
- Projects
- Citing SNAP
- Links
- About
- Contact us

Open positions

Open research positions in **SNAP** group are available here.

Dataset statistics	
Nodes	3932
Drug nodes	284
Gene nodes	3648
Edges	18690
Nodes in largest SCC	3891
Fraction of nodes in largest SCC	1.000000
Edges in largest SCC	18660
Fraction of edges in largest SCC	0.998395
Diameter (longest shortest path)	9
90-percentile effective diameter	3.987420

This dataset considers only interactions between small chemicals (i.e., drugs) and target proteins that had been experimentally verified by biological experiments or formal pharmacological studies.

7. BioSNAP数据集 (http://snap.stanford.edu/biodata)

3.1 Experimental setup

Dataset. We use the MINER DTI dataset from BIOSNAP collection (Zitnik *et al.*, 2018a) as our main dataset of experiments. It consists of 4510 drug nodes and 2181 protein targets, and 13 741 DTI pairs from DrugBank (Wishart *et al.*, 2008). BIOSNAP dataset only contains positive DTI pairs. For negative pairs, we sample from the unseen pairs, following common practice (Zhang and Chen, 2018; Zitnik *et al.*, 2018b). We obtain a balanced dataset with equal positive and negative samples. In addition to BIOSNAP, we also include two benchmark datasets in the main predictive performance comparison experiment. DAVIS consists of wet lab assay K_d values among 68 drugs and 379 proteins (Davis *et al.*, 2011) and BindingDB consists of K_d values among 10 665 drugs and 1413 proteins (Liu *et al.*, 2007). DTI pairs that have K_d values <30 units are

Table 1. Dataset statistics

Dataset	# Drugs	# Proteins	# Pos Interactions	# Neg Interactions
BIOSNAP	4510	2181	9619/1374/2748	9619/1374/2748
DAVIS	68	379	1043/160/303	1043/2846/5708
BindingDB	10 665	1413	6334/927/1905	6334/5717/11 384

Positives: 13741, Negatives: 13741

- Molecular Interaction Transformer for drug-target interaction prediction. Bioinformatics, 2021.
- Contrastive learning in protein language space predicts interactions between drugs and protein targets. PNAS, 2023.
- Yu, Qinze, et al. "GS-DTI: a graph-structure-aware framework leveraging large language models for drug–target interaction prediction." Bioinformatics 41.8 (2025): btaf445.

We only use proteins in our analysis that have at least one link in STRING or one association in PhenomeNET, and drugs with at least one side effect. Therefore, the intersection between these resources yields 1428 drugs and 7368 human proteins with 32 212 interactions for STITCH, 1837 interactions between 680 drugs and 1458 proteins for Yamanishi, and 6498 links between 949 drugs and 2221 proteins for BioSnap dataset. We provide links to and methods for obtaining and processing the necessary data in our Github repository.

Positives: 6498, Negatives: 6498

- Hinnerichs, Tilman, and Robert Hoehndorf. "DTI-Voodoo: machine learning over interaction networks and ontology-based background knowledge predicts drug–target interactions." Bioinformatics 37.24 (2021): 4835-4843.

8. 数据集之间的冗余性

蛋白质A-小分子B， 蛋白质C-小分子B

若A和C的同源性很高（>90%），是否属于冗余数据?

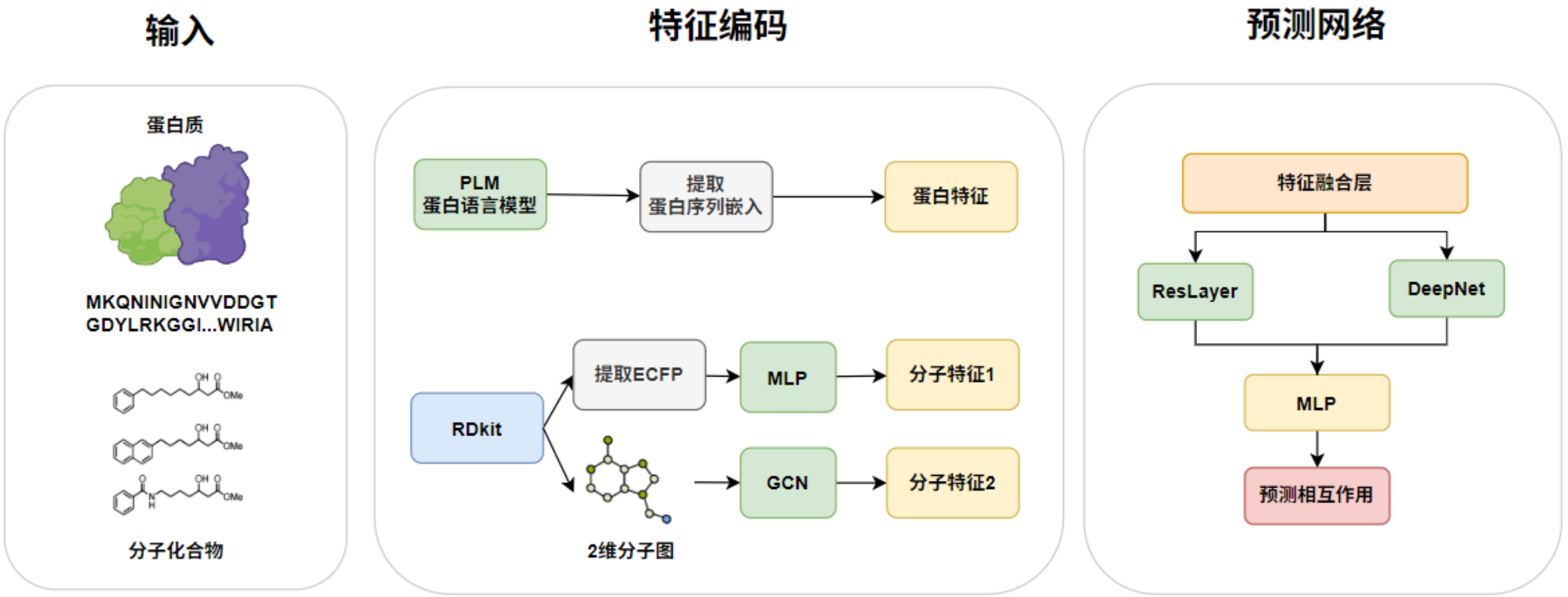
综上，不建议合并数据集!

9. 后续的工作

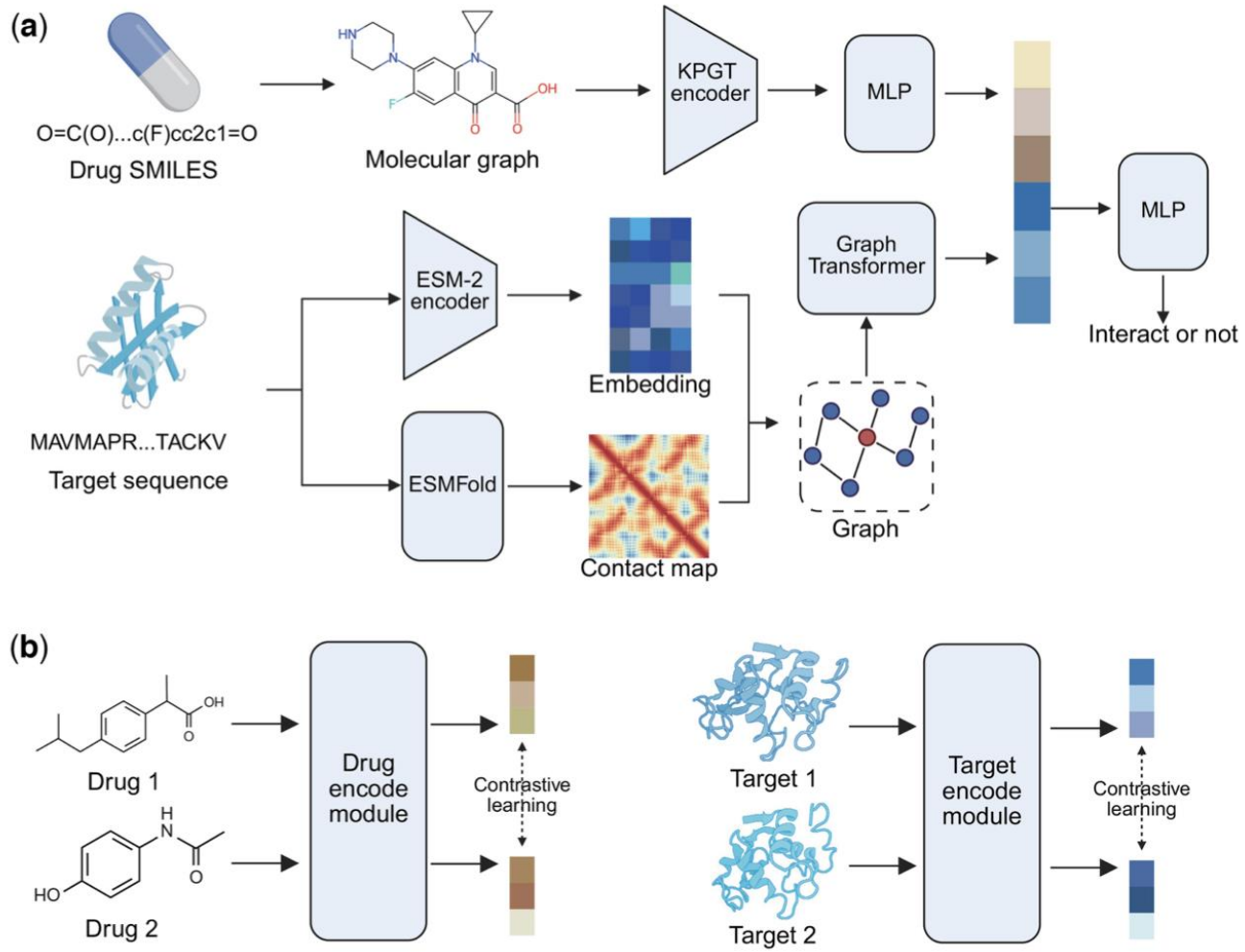
1. 单独在DAVIS、 BindingDB 和 BioSNAP数据集上做benchmark。

2. 升级ESM2的模型版本。

3. 用ESM2预测接触图，引入图神经网络（GCN和GAT）。

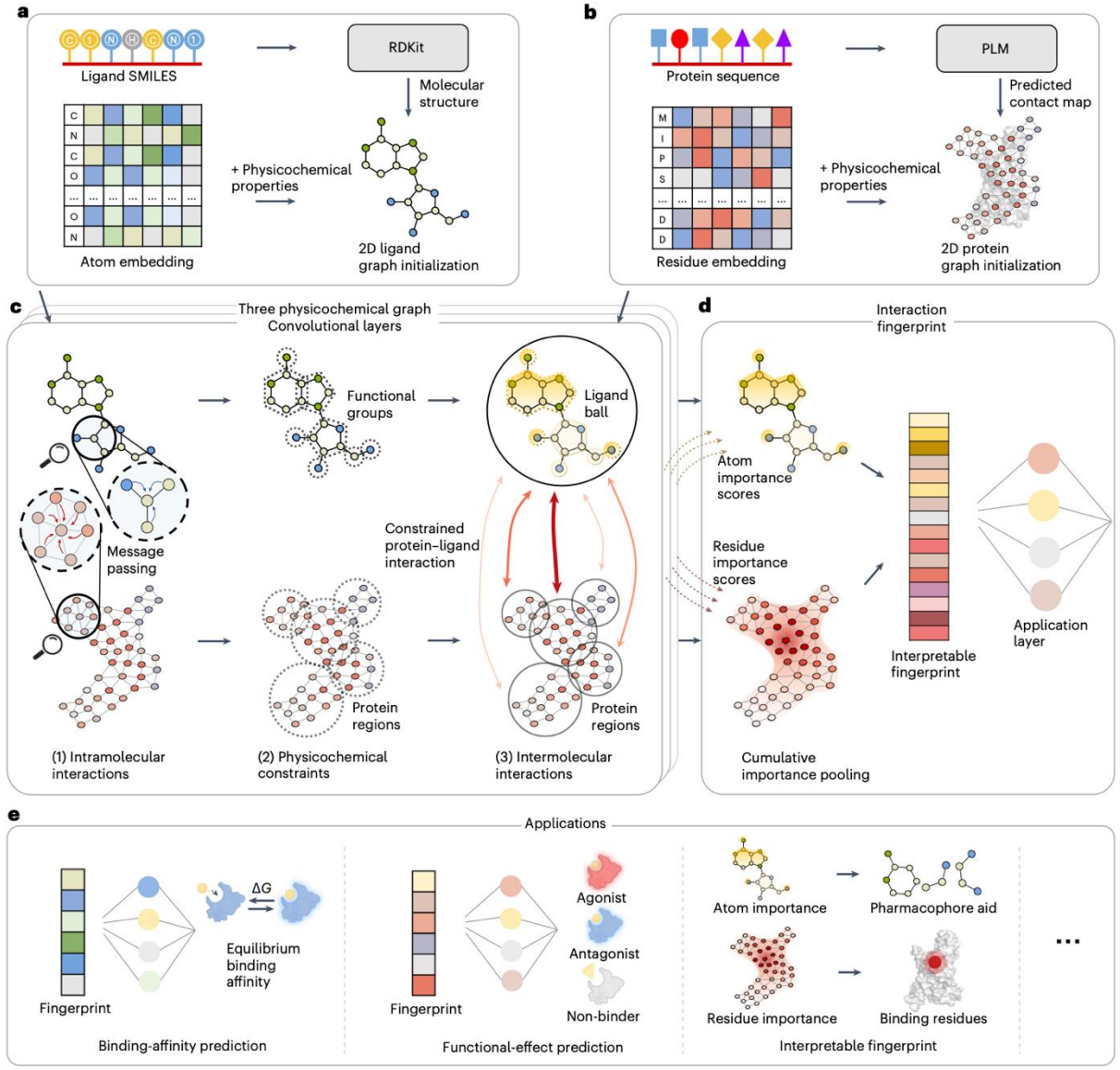


9. 后续的工作



Yu, Qinze, et al. "GS-DTI: a graph-structure-aware framework leveraging large language models for drug–target interaction prediction." Bioinformatics 41.8 (2025): btaf445.

9. 后续的工作



Koh, Huan Yee, et al. "Physicochemical graph neural network for learning protein–ligand interaction fingerprints from sequence data." Nature Machine Intelligence 6.6 (2024): 673-687.