IDG-DREAM Drug-Kinase Binding Prediction Challenge ThinNguyen Writeup

Thin Nguyen
Applied Artificial Intelligence Institute, Deakin University, Australia
thin.nguyen@deakin.edu.au

In this challenge given a training drug-target affinity matrix, the task is to estimate the empty cells in the matrix. We evaluate several approaches on benchmark databases of similar problems where the ground-truth is available and suggest to use the best on the validation datasets, deep learning models, to tackle the challenge.

1 Introduction

Protein kinases are enzymes that catalyze phosphorylation reactions within the cells, thus regulating cell function. More than 500 kinases have been identified, representing approximately 2% of the human genome [6]. About 30% of human proteins may be modified by kinase activity [8], making kinases attractive targets for drug interventions. Measuring drug-kinase interactions through clinical trials is costly and time-consuming [3, 7]. Estimating the strength of the interactions for novel couples of drug-kinase based on the interactions already measured becomes an important alternative, where the challenge is a crowd effort.

Apparently the challenge could be considered as a collaborative filtering problem (CF). For example, in movie ratings as in the Nexflix competition[1], the rating for a couple of movie-user is learned, or collaboratively filtered, from the ratings by the movies/users similar to the given movie/user. The lesson from Nexflix competition is that if the number of training user-movie ratings is big enough, external information for users or movies does not make significantly contribution to the recommendation systems. However this is not always the case for drug-target binding prediction problem, where the affinity available is often sparse.

Another approach is kernel based, as in [2, 1]. In these work, kernels for drugs and targets are built from their molecular descriptors, input into a regularized least squares regression model (RLS) to predict the binding affinity.

For the challenge, the information of drugs, which are novel, is limited, making it difficult to compute biologically sensible kernels or similarity matrices among drugs, and hence the performance of CF or kernel-based methods could be compromised. It

[1] https://www.netflixprize.com/rules.html

is worse, or even inapplicable, when the drugs/proteins in the couples to be predicted the affinity are not in the training sets.

On the other hand, deep learning based modeling becomes a suitable approach when only 1D representation for drugs (SMILES) and proteins (sequences) is provided. Then the model learned from training data can predict the affinity for couples of drugs-targets in testing data with their 1D representation provided, regardless they are in the training data or not.

2 Methods

2.1 Collaborative filtering (CF)

For drug-target binding prediction problem, from the training affinity matrix we could build similarity matrices for both drugs and targets. Any distance measure can be used to calculate the similarity, e.g., cosine or correlation. With these similarity matrices at hand we can run drug-based or target-based filtering to estimate the binding power for unknown drug-target couples.

For example, for drug-based collaborative filtering, for an input couple (*d,t*), its affinity is estimated as the weighted sum of all other drug's affinity for target *t* where the weighting is the cosine similarity between *d* and other drugs. Similarly, for target-based collaborative filtering, for the same input couple (*d,t*), the affinity for the couple is estimated as the weighted sum of all other targets' affinity for drug *d* where the weighting is the cosine similarity between *t* and other targets.

$$\hat{a_{dt}} = \frac{\sum_{d'} sim(d,d')a_{d't}}{\sum_{d'} sim(d,d')}$$

$$\hat{a_{dt}} = \frac{\sum_{t'} sim(t,t')a_{dt'}}{\sum_{t'} sim(t,t')}$$

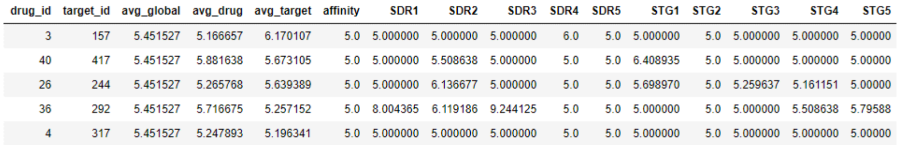
To void noise, instead of all, top K most similar drugs or targets can be used to estimate the affinity.

Another problem is that the scale of affinity differ among drugs and targets, i.e., some drugs have extreme low or high affinity for all targets. So a relative difference from the average can be used instead of the absolute affinity.

$$\hat{a_{dt}} = \bar{a_d} + \frac{\sum_{d'} sim(d,d')(a_{d't} - \bar{a_{d'}})}{\sum_{d'} sim(d,d')}$$

$$\hat{a_{dt}} = \bar{a_t} + \frac{\sum_{t'} sim(t,t')(a_{dt'} - \bar{a_{t'}})}{\sum_{t'} sim(t,t')}$$

where $\bar{a_d}$ and $\bar{a_t}$ is the average affinity for drug *d* and target *t*, respectively.



drug_id	target_id	avg_global	avg_drug	avg_target	affinity	SDR1	SDR2	SDR3	SDR4	SDR5	STG1	STG2	STG3	STG4	STG5
3	157	5.451527	5.166657	6.170107	5.0	5.000000	5.000000	5.000000	6.0	5.0	5.000000	5.0	5.000000	5.000000	5.00000
40	417	5.451527	5.881638	5.673105	5.0	5.000000	5.508638	5.000000	5.0	5.0	6.408935	5.0	5.000000	5.000000	5.00000
26	244	5.451527	5.265768	5.639389	5.0	5.000000	6.136677	5.000000	5.0	5.0	5.698970	5.0	5.259637	5.161151	5.00000
36	292	5.451527	5.716675	5.257152	5.0	8.004365	6.119186	9.244125	5.0	5.0	5.000000	5.0	5.000000	5.508638	5.79588
4	317	5.451527	5.247893	5.196341	5.0	5.000000	5.000000	5.000000	5.0	5.0	5.000000	5.0	5.000000	5.000000	5.00000

Figure 1: Represent a couple of drug-target through its neighbors in both drugs and targets.

Joint similarity collaborating filtering (joint-sim CF) While the above CF models suggest use either drugs or targets similarity to recommend affinity for a novel couple of drug-target, we suggest to use both for the task. In particular, given a couple of drug-target needs evaluating the affinity, we can represent it through the affinity scored by its *K* neighbors, in both drugs-based and targets-based. An example is shown in Figure 1, where *K*=5: the couple of (*drug_id*, *target_id*) is represented by *SDR1*,...,*SDR5* – the affinity with *target_id* scored by *K* drugs closest to *drug_id* and *STG1*,...,*STG5* – the affinity with *drug_id* scored by *K* targets closest to *target_id*.

2.2 Kernel based (KronRLS)

Given the problem is to predict the affinity for *n* drugs and *m* targets, there would be *n***m* combinations of them and the kernel would be in the size of $(n*m)^2$. To speed up model training, Cichonska et al. [2, 1] suggest to use KronRLS (Kronecker regularized least-squares). In KronRLS, a pairwise kernel *K* is computed as the Kronecker product of compound kernel of size *n***n* and protein kernel of size *m***m*.

2.3 Deep learning

2.3.1 Auto-encoder (DL_AE)

Auto-encoder based modeling is claimed to be state-of-the-art model for Netflix dataset[2]. In this approach, the training affinity matrix is assumed to be fully filled and is the input, and output, for an auto-encoder, as shown in Figure 2. Then the loss function is adjusted for ignoring not-in-the-train cells.

2.3.2 Embedded nodes in bipartite graph (DL_bipartite)

We consider the training affinity matrix as a bipartite graph and learn a continuous feature representation for drugs and targets by node2vec [5] for the graph, as illustrated in Figure 3a. Then the node vectors are concatenated in a neural network to predict the affinity for testing data, as shown in Figure 3b.

[2] https://paperswithcode.com/sota/collaborative-filtering-on-netflix

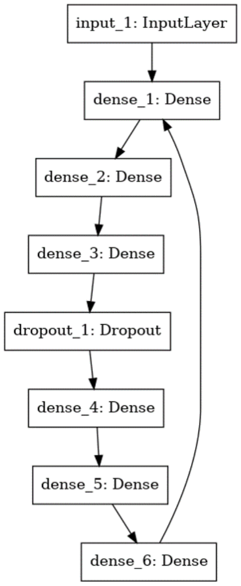


Figure 2: Auto-encoder of drug-target affinity matrix.

2.3.3 External information: SMILES and sequences (DL_1D)

No external information is used in the two deep learning models DL_AE and DL_bipartite. However, in the testing data provided by the organizer, for drugs, SMILES strings is given, and for targets, protein sequences can be retrieved with the UniProt_Id given [3]. These strings can be seen as 1D representation for drugs and proteins, input into a neural network to learn a model to predict the binding affinity for novel drug-kinase couples, as shown in Figure 4. In the figure, input_1 and input_2 are drugs and targets, respectively. As these are in 1D representation, layers of 1D convolutions and pooling are used to capture potential patterns in the inputs. They are then concatenated, sent through regularized layers of Dropout, and finally regressed with the training affinity.

3 Model validation

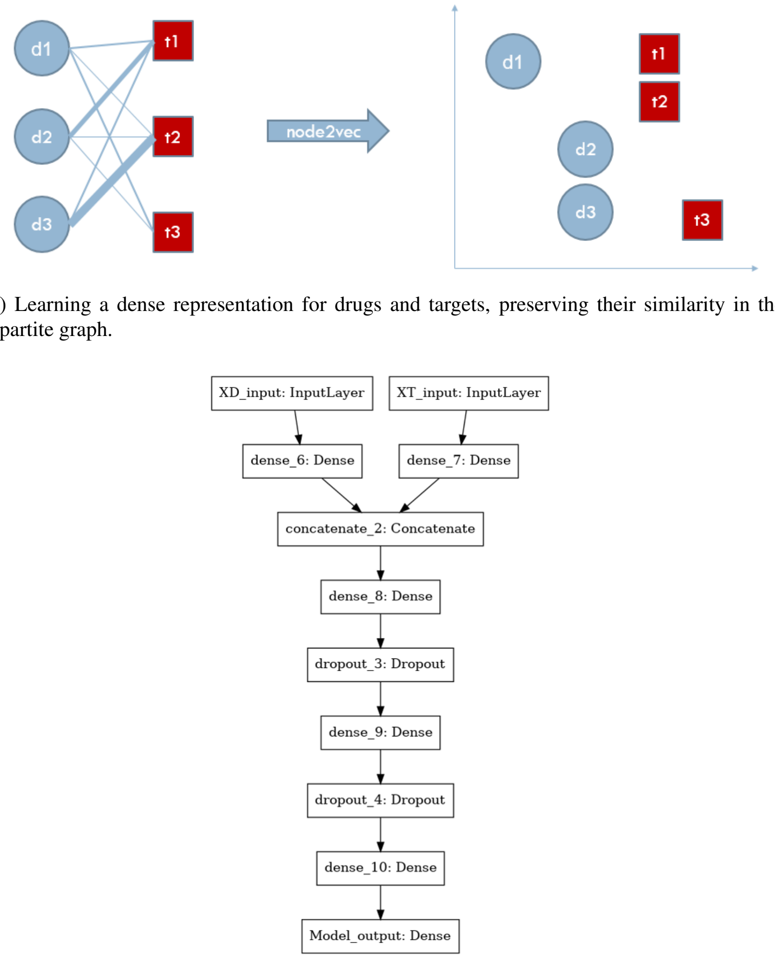
To seek a good model for the challenge we experimented the candidate models above with benchmark datasets of similar problems.

3.1 Datasets

Two datasets were used to evaluate the models:

- **Davis** dataset: binding affinities observed for all pairs of 72 drugs and 442 targets, measured by Kd value (kinase dissociation constant) [4].

[3] https://www.uniprot.org/uniprot



(a) Learning a dense representation for drugs and targets, preserving their similarity in the bipartite graph.

(b) Drug and target vectors as input for building a model to predict the affinity.

Figure 3: Predicting binding affinity through the bipartite graph.

Model	rmse	Spearman corr
CF	1.28	0.46
DL_AE	0.86	0.27
joint-sim CF	0.69	0.57
KronRLS	0.58	**0.69**
DL_bipartite	0.56	0.65
DL_1D	**0.51**	0.68

(a) For Davis dataset. Best result is in bold.

Model (Kiba)	rmse	Spearman corr
CF	4.29	0.21
DL_AE	2	0.31
KronRLS	0.6	0.74
joint-sim CF	0.55	0.76
DL_bipartie	0.53	0.78
DL_1D	**0.43**	**0.85**

(b) For Kiba dataset. Best result is in bold.

Table 1: Prediction performance.

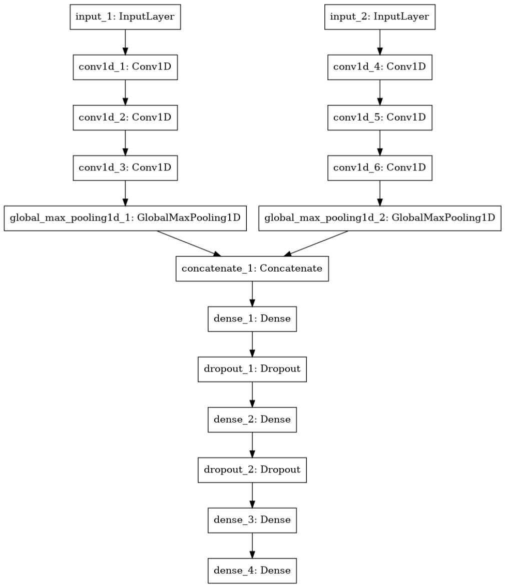


Figure 4: Predicting the affinity using external information.

- **Kiba** dataset: binding affinities for 2,116 drugs and 229 targets [9].

80% of data instances were used for training and 20% were for testing the models. Same data splitting was used for learning all the models.

3.2 Results

The result for all models mentioned above on the two datasets is presented in Table 1. **DL_1D** is best in two measures for Kiba dataset. It is also best in RMSE and second best in Spearman correlation for Davis dataset.

4 Model and data for the challenge

As **DL_1D** performs really well in the two benchmark datasets, we choose it as the model for the challenge.

Model is now trained on Drug Target Commons (DTC) data [10]. Only those tuples with the standard_type of ["KD","Kd","KD’", "PKD"] is selected. For those with ["KD","Kd","KD’"], the standard value is converted to pKd unit. 55,816 compound-protein pairs were included in this training data for 13,651 distinct compounds and 1,489 distinct proteins.

Model is validated on all Davis data [4]. There are 30,056 compound-protein pairs in this validation data.

The model gaining smallest RMSE for validation data is then used to predict the affinity for testing data.

5 Running model

- On GPUs:

```
docker login docker.synapse.org docker run --runtime=nvidia -it -v ${PWD}/io_gpu:/output docker.synapse.org/syn18518883/my-model:gpu
```

- On CPUs:

```
docker login docker.synapse.org docker run -it -v ${PWD}/io_cpu:/output docker.synapse.org/syn18518883/my-model:cpu
```

- Notes:

For running on GPUs, 'nvidia-docker' should be installed[4].

For our computers, training on GPUs is about 4.5 times faster than on CPUs, 4,003 seconds versus 18,078 seconds.

References

[1] Anna Cichonska, Tapio Pahikkala, Sandor Szedmak, Heli Julkunen, Antti Airola, Markus Heinonen, Tero Aittokallio, and Juho Rousu. Learning with multiple pairwise kernels for drug bioactivity prediction. *Bioinformatics*, 34(13):i509–i518, 2018.

[2] Anna Cichonska, Balaguru Ravikumar, Elina Parri, Sanna Timonen, Tapio Pahikkala, Antti Airola, Krister Wennerberg, Juho Rousu, and Tero Aittokallio. Computational-experimental approach to drug-target interaction mapping: A case study on kinase inhibitors. *PLoS Computational Biology*, 13(8):e1005678, 2017.

[3] Philip Cohen. Protein kinases–the major drug targets of the twenty-first century? *Nature Reviews Drug Discovery*, 1(4):309, 2002.

[4] Mindy I Davis, Jeremy P Hunt, Sanna Herrgard, Pietro Ciceri, Lisa M Wodicka, Gabriel Pallares, Michael Hocker, Daniel K Treiber, and Patrick P Zarrinkar. Comprehensive analysis of kinase inhibitor selectivity. *Nature Biotechnology*, 29(11):1046, 2011.

[5] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 855–864. ACM, 2016.

[6] Gerard Manning, David B Whyte, Ricardo Martinez, Tony Hunter, and Sucha Sudarsanam. The protein kinase complement of the human genome. *Science*, 298(5600):1912–1934, 2002.

[7] Martin EM Noble, Jane A Endicott, and Louise N Johnson. Protein kinase inhibitors: Insights into drug design from structure. *Science*, 303(5665):1800–1805, 2004.

[4] https://github.com/NVIDIA/nvidia-docker

[8] Shawn J Stachel, John M Sanders, Darrell A Henze, Mike T Rudd, Hua-Poo Su, Yiwei Li, Kausik K Nanda, Melissa S Egbertson, Peter J Manley, Kristen LG Jones, et al. Maximizing diversity from a kinase screen: Identification of novel and selective pan-Trk inhibitors for chronic pain. *Journal of Medicinal Chemistry*, 57(13):5800–5816, 2014.

[9] Jing Tang, Agnieszka Szwajda, Sushil Shakyawar, Tao Xu, Petteri Hintsanen, Krister Wennerberg, and Tero Aittokallio. Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. *Journal of Chemical Information and Modeling*, 54(3):735–743, 2014.

[10] Jing Tang, Zia ur Rehman Tanoli, Balaguru Ravikumar, Zaid Alam, Anni Rebane, Markus Vähä-Koskela, Gopal Peddinti, Arjan J. van Adrichem, Janica Wakkinen, Alok Jaiswal, Ella Karjalainen, Prson Gautam, Liye He, Elina Parri, Suleiman Khan, Abhishekh Gupta, Mehreen Ali, Laxman Yetukuri, Anna-Lena Gustavsson, Brinton Seashore-Ludlow, Anne Hersey, Andrew R. Leach, John P. Overington, Gretchen Repasky, Krister Wennerberg, and Tero Aittokallio. Drug target commons: A community effort to build a consensus knowledge base for drug-target interactions. *Cell Chemical Biology*, 25(2):224 – 229.e2, 2018.