1.DataSet 2

The second data set was introduced in 2012 by Yamanishi et al[1]. This dataset includes three txt files of drug-target interactions, drug features, and target features, and target features, and all data in this data set are numeric.

The second data set is described in two parts; In the first part, information about drug-target interactions and in the second part, information about the drug and target used and how they are represented as feature vectors are described.

1-1. Drug-target interaction data

Drug-target interaction file as a two-dimensional matrix $Y \in \mathbb{R}^{n\times m}$ with n drug rows and m target columns. That is, $Y_{ij} = 1$ if drug d_i and target $t_j$interact and $Y_{ij} = 0$ otherwise. In other words, the one indicates the presence of interaction and the zero indicates the lack of interaction between the drug d_i which is in the matrix row and the target t_j which is located in the matrix column. Also, the features of drugs are in the form of $D \in \mathbb{R}^{n\times f}$ matrix and the features of targets are in the form of $T \in \mathbb{R}^{m\times g}$ matrix, where f and g are the number of features of drugs and targets. The interaction data were obtained from the DrugBank [2, 3]. In total, there are 4809 drug-target interactions between 1862 drugs and their 1554 protein interaction partners.

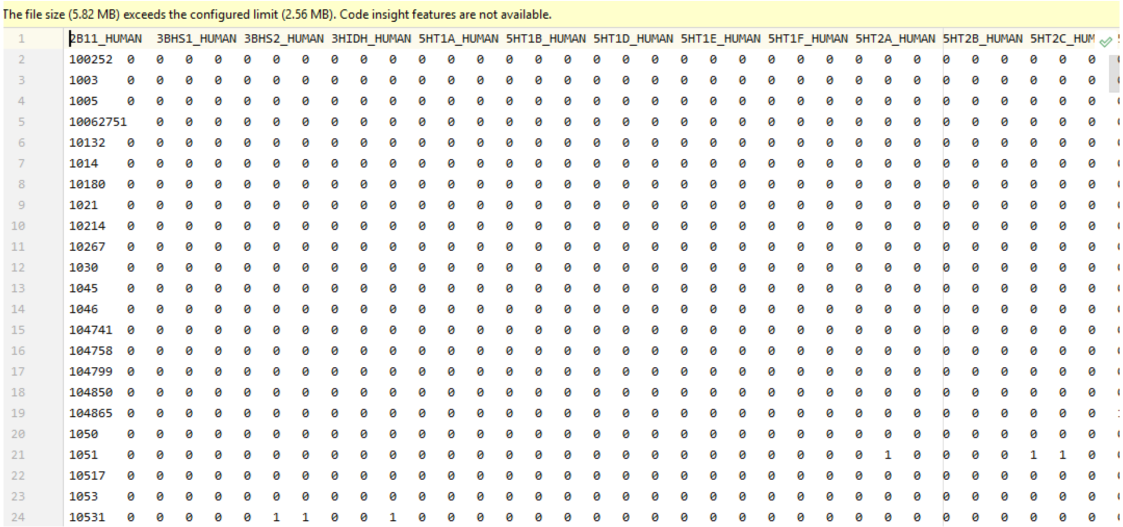
1-2. Drugs and targets information

Chemical structures of drugs were encoded by a chemical fingerprint corresponding to 881 chemical substructures defined in the PubChem database. Chemically identical drugs with the same structures (duplicates) are removed, so structures of all drugs in the above interaction data are unique. Each drug was represented by 881-dimensional binary vector whose elements encode for the presence or absence of each PubChem substructure by 1 or 0, respectively. Genomic information about target proteins was obtained from the UniProt database, and associated protein domains were obtained from the PFAM database [4]. Target proteins in our dataset were associated with 876 PFAM domains. Each target protein was represented by 876-dimensional binary vector whose elements encode for the presence or absence of each of the retained PFAM domain by 1 or 0, respectively

The file of drug features is in the form of $D \in \mathbb{R}^{n\times f}$ matrix and the target features are in the form of $T \in \mathbb{R}^{m\times g}$ matrix, where f and g are the number of drug features and targets features. Formation of drug-target vectors for use in machine learning model as $[d_1, d_2, \dots, d_p, t_1, t_2, \dots, t_q, I]$, which label I is a binary number and indicates the presence or absence of interaction Is between the drug and the target in the vector. Label I is identified by the drug-target interaction matrix. Details of this dataset are given in Table 1. Figures 1 and 2 show a view of the drug-target interaction file and the drug features file, respectively.

Table 1. Statistics of second dataset

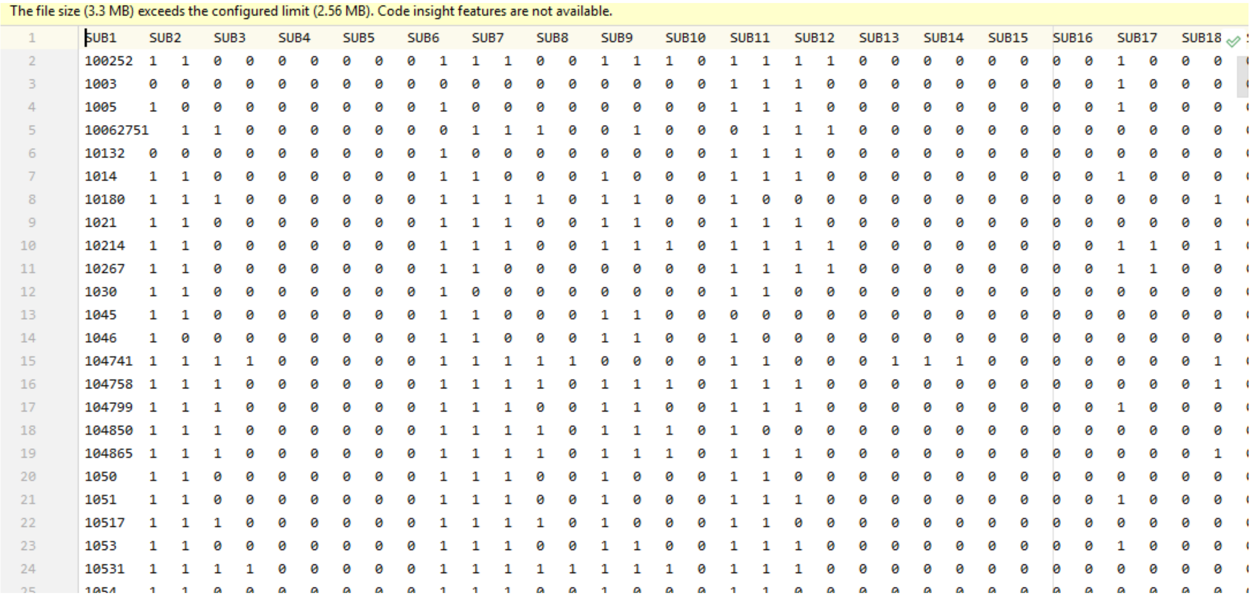
Drugs	Targets	Interactions
1862	1554	4809



The file size (5.82 MB) exceeds the configured limit (2.56 MB). Code insight features are not available.

```
2B11_HUMAN  3BHS1_HUMAN 3BHS2_HUMAN 3HIDH_HUMAN 5HT1A_HUMAN 5HT1B_HUMAN 5HT1D_HUMAN 5HT1E_HUMAN 5HT1F_HUMAN 5HT2A_HUMAN 5HT2B_HUMAN 5HT2C_HUM
100252  0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1003    0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1005    0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
10062751  0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
10132   0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1014    0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
10180   0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1021    0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
10214   0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
10267   0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1030    0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1045    0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1046    0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
104741  0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
104758  0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
104799  0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
104850  0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
104865  0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1050    0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1051    0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 1 1 0
10517   0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1053    0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
10531   0 0 0 0 0 1 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

Figure 1. view of the drug-target interaction file



The file size (3.3 MB) exceeds the configured limit (2.56 MB). Code insight features are not available.

```
SUB1    SUB2    SUB3    SUB4    SUB5    SUB6    SUB7    SUB8    SUB9    SUB10   SUB11   SUB12   SUB13   SUB14   SUB15   SUB16   SUB17   SUB18
100252  1   1   0   0   0   0   0   0   0   1   1   1   0   0   1   1   1   0   1   1   1   1   0   0   0   0   0   0   0   0   1   0   0   0
1003    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   1   1   1   0   0   0   0   0   0   0   0   0   1   0   0   0
1005    1   0   0   0   0   0   0   0   0   1   0   0   0   0   0   0   0   0   1   1   1   0   0   0   0   0   0   0   0   0   1   0   0   0
10062751    1   1   0   0   0   0   0   0   0   1   1   1   0   0   1   0   0   0   1   1   1   0   0   0   0   0   0   0   0   0   0   0   0
10132   0   0   0   0   0   0   0   0   0   1   0   0   0   0   0   0   0   0   1   1   1   0   0   0   0   0   0   0   0   0   0   0   0   0
1014    1   1   0   0   0   0   0   0   0   1   1   0   0   0   1   0   0   0   1   1   1   0   0   0   0   0   0   0   0   0   1   0   0   0
10180   1   1   1   0   0   0   0   0   0   1   1   1   1   0   1   1   0   0   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0   1
1021    1   1   0   0   0   0   0   0   0   1   1   1   0   0   1   1   0   0   1   1   1   0   0   0   0   0   0   0   0   0   0   0   0   0
10214   1   1   0   0   0   0   0   0   0   1   1   1   0   0   1   1   1   0   1   1   1   1   0   0   0   0   0   0   0   0   1   1   0   1
10267   1   1   0   0   0   0   0   0   0   1   1   0   0   0   0   0   0   0   1   1   1   1   0   0   0   0   0   0   0   0   1   1   0   0
1030    1   1   0   0   0   0   0   0   0   1   0   0   0   0   0   0   0   0   1   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0
1045    1   1   0   0   0   0   0   0   0   1   1   0   0   0   1   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
1046    1   0   0   0   0   0   0   0   0   1   1   0   0   0   1   1   0   0   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
104741  1   1   1   1   0   0   0   0   0   1   1   1   1   1   0   0   0   0   1   1   0   0   0   1   1   1   0   0   0   0   0   0   0   1
104758  1   1   1   0   0   0   0   0   0   1   1   1   1   0   1   1   1   0   1   1   1   0   0   0   0   0   0   0   0   0   0   0   0   1
104799  1   1   1   0   0   0   0   0   0   1   1   1   0   0   1   1   0   0   1   1   1   0   0   0   0   0   0   0   0   0   1   0   0   0
104850  1   1   1   0   0   0   0   0   0   1   1   1   1   0   1   1   1   0   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
104865  1   1   1   0   0   0   0   0   0   1   1   1   1   0   1   1   1   0   1   1   1   0   0   0   0   0   0   0   0   0   0   0   0   1
1050    1   1   0   0   0   0   0   0   0   1   1   1   0   0   1   0   0   0   1   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0
1051    1   1   0   0   0   0   0   0   0   1   1   1   0   0   1   0   0   0   1   1   1   0   0   0   0   0   0   0   0   0   1   0   0   0
10517   1   1   1   0   0   0   0   0   0   1   1   1   1   0   1   0   0   0   1   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0
1053    1   1   0   0   0   0   0   0   0   1   1   1   0   0   1   1   0   0   1   1   1   0   0   0   0   0   0   0   0   0   1   0   0   0
10531   1   1   1   1   0   0   0   0   0   1   1   1   1   1   1   1   1   0   1   1   1   0   0   0   0   0   0   0   0   0   0   0   0   0
1054    1   1   0   0   0   0   0   0   0   1   1   1   0   0   1   0   0   0   1   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0
```

Figure 2. view of the drug features file

References:

[١] Y. Yamanishi, E. Pauwels, H. Saigo, and V. Stoven, "Extracting sets of chemical substructures and protein domains governing drug-target interactions," *Journal of chemical information and modeling,* vol. 51, no. 5, pp. 1183-1194, 2011.

[٢] C. Knox *et al*" ,.DrugBank 3.0: a comprehensive resource for ‘omics’ research on drugs," *Nucleic acids research,* vol. 39, no. suppl_1, pp. D1035-D1041, 2010.

[٣] D. S. Wishart *et al.*, "DrugBank: a knowledgebase for drugs, drug actions and drug targets," *Nucleic acids research,* vol. 36, no. suppl_1, pp. D901-D906, 2007.

[٤] R. D. Finn *et al.*, "The Pfam protein families database: towards a more sustainable future," *Nucleic acids research,* vol. 44, no. D1, pp. D279-D285, 2016.