1.Gold dataSet

The gold standard dataset is divided into four datasets; In other words, the four datasets of enzymes, ion channels, G protein-coupled receptors (GPCRs), and nuclear receptors are actually classes of proteins in the human body. This gold standard data was introduced in 2008 by Yamanishi et al[1]. The statistical quantities of existing drugs are 445, 223, 210, and 54, respectively. The numbers of known proteins are 664, 95, 204, and 26, respectively. The counts of the DTIs which have been proven are 2926, 635, 1476, and 90, respectively. Most new methods of predicting drug-target interactions have evaluated their method and reported results using this dataset [1-5]. Each data set in this gold data set includes three txt files of drug-target interactions, drug properties, and target properties, and the data in this data set are numeric.

The golden data set is explained 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

All data sets originate from the databases of DrugBank [6], SuperTarget [7], BRENDA [8], and KEGG BRITE [9]. 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.

1-2. Drugs and targets information

Note that in the enzyme class focused on the regulatory interactions between enzymes and compounds rather than the metabolic interactions, so all the ligands in the enzyme data are inhibitors or activators rather than substrates or products. Cofactors such as adenosine triphosphate (ATP) and nicotinamide adenine dinucleotide phosphate (NADPH) are also not included except when they are annotated as regulators in the BRENDA database. Also, do not use compounds whose molecular weights are <100.

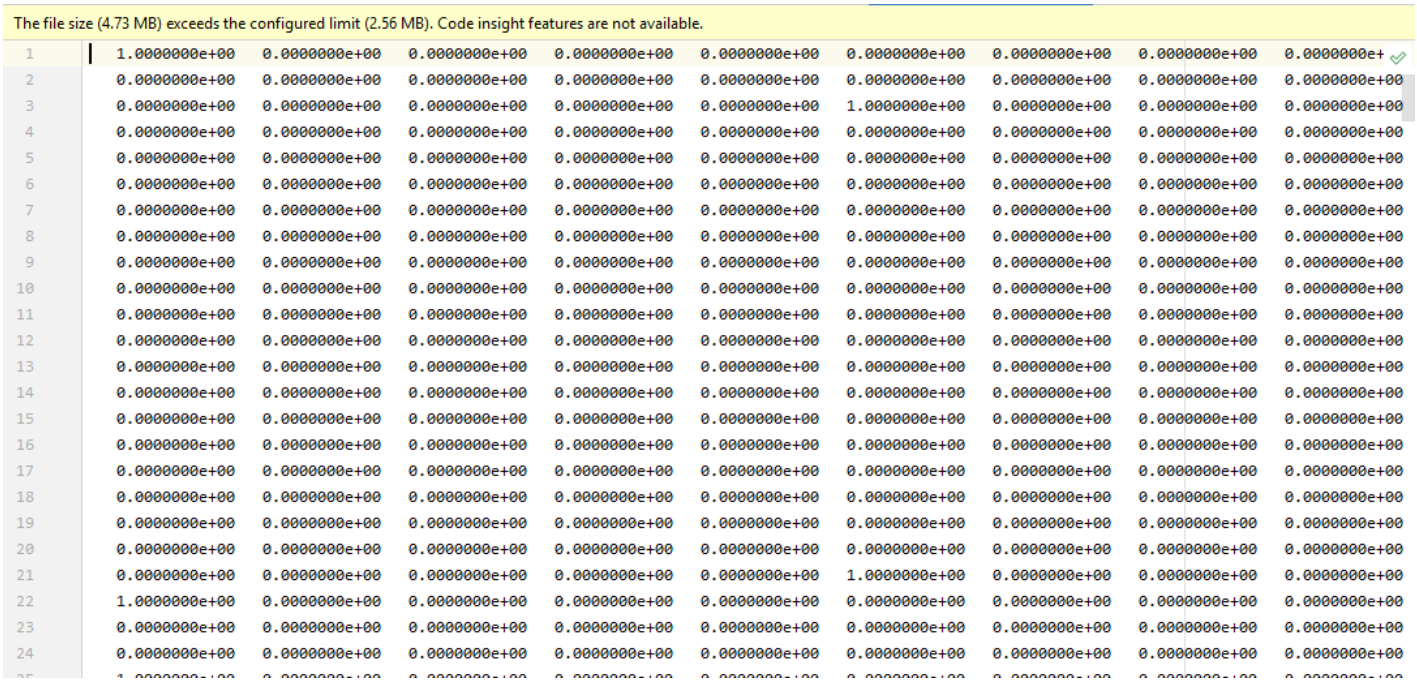
Chemical structures of the drugs were obtained from the DRUG and COMPOUND Sections in the KEGG LIGAND database. Amino acid sequences of the target proteins were obtained from the KEGG GENES database.

The files 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. Each drug is displayed as $d = [d_1, d_2, \dots, d_p]$ and each target as $t = [t_1, t_2, \dots, t_q]$. Details of this dataset are given in Table 1.

Table 1.Gold Dataset of Yamanishi et al

Dataset	Drugs	Targets	Interactions
Enzyme	445	664	2926
Ion channel	210	204	1476
GPCRs	223	95	635
Nuclear receptor	54	26	90

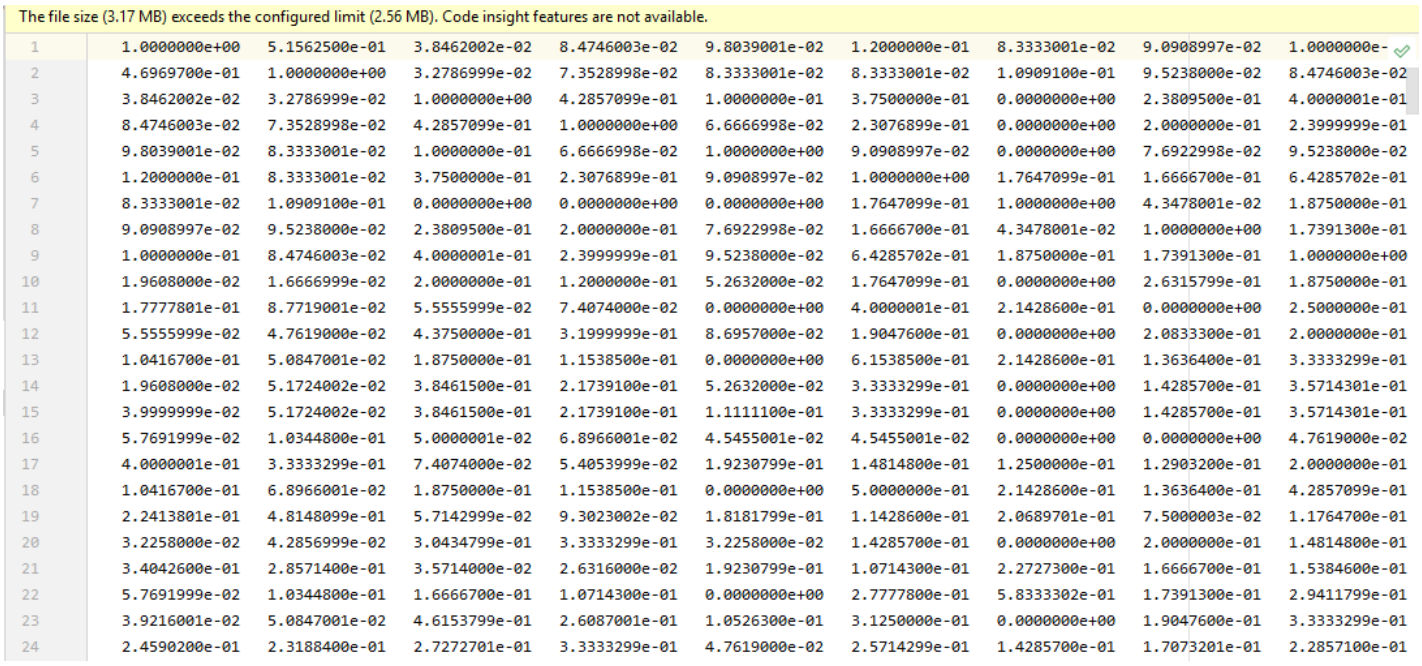
Formation of drug-target vectors for use in machine learning model as $[d_1, d_2, \dots, d_p, t_1, t_2, \dots, t_q, I]$, which label I is a binary number and indicates the presence or absence of interaction Is between the drug and the target in the vector. Label I is identified by the drug-target interaction matrix. Figures 1 and 2 show a view of the drug-target interaction file and the drug features file, respectively.



The file size (4.73 MB) exceeds the configured limit (2.56 MB). Code insight features are not available.

```
1.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+
0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00
0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   1.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00
0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00
0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00
0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00
0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00
0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00
0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00
0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00
0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00
0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00
0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00
0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00
0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00
0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00
0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00
0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00
0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00
0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00
0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   1.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00
1.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00
0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00
0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00   0.0000000e+00
```

Figure 1. view of the drug-target interaction file



The file size (3.17 MB) exceeds the configured limit (2.56 MB). Code insight features are not available.

```
1.0000000e+00   5.1562500e-01   3.8462002e-02   8.4746003e-02   9.8039001e-02   1.2000000e-01   8.3333001e-02   9.0908997e-02   1.0000000e-
4.6969700e-01   1.0000000e+00   3.2786999e-02   7.3528998e-02   8.3333001e-02   8.3333001e-02   1.0909100e-01   9.5238000e-02   8.4746003e-02
3.8462002e-02   3.2786999e-02   1.0000000e+00   4.2857099e-01   1.0000000e-01   3.7500000e-01   0.0000000e+00   2.3809500e-01   4.0000001e-01
8.4746003e-02   7.3528998e-02   4.2857099e-01   1.0000000e+00   6.6666998e-02   2.3076899e-01   0.0000000e+00   2.0000000e-01   2.3999999e-01
9.8039001e-02   8.3333001e-02   1.0000000e-01   6.6666998e-02   1.0000000e+00   9.0908997e-02   0.0000000e+00   7.6922998e-02   9.5238000e-02
1.2000000e-01   8.3333001e-02   3.7500000e-01   2.3076899e-01   9.0908997e-02   1.0000000e+00   1.7647099e-01   1.6666700e-01   6.4285702e-01
8.3333001e-02   1.0909100e-01   0.0000000e+00   0.0000000e+00   0.0000000e+00   1.7647099e-01   1.0000000e+00   4.3478001e-02   1.8750000e-01
9.0908997e-02   9.5238000e-02   2.3809500e-01   2.0000000e-01   7.6922998e-02   1.6666700e-01   4.3478001e-02   1.0000000e+00   1.7391300e-01
1.0000000e-01   8.4746003e-02   4.0000001e-01   2.3999999e-01   9.5238000e-02   6.4285702e-01   1.8750000e-01   1.7391300e-01   1.0000000e+00
1.9608000e-02   1.6666999e-02   2.0000000e-01   1.2000000e-01   5.2632000e-02   1.7647099e-01   0.0000000e+00   2.6315799e-01   1.8750000e-01
1.7777801e-01   8.7719001e-02   5.5555999e-02   7.4074000e-02   0.0000000e+00   4.0000001e-01   2.1428600e-01   0.0000000e+00   2.5000000e-01
5.5555999e-02   4.7619000e-02   4.3750000e-01   3.1999999e-01   8.6957000e-02   1.9047600e-01   0.0000000e+00   2.0833300e-01   2.0000000e-01
1.0416700e-01   5.0847001e-02   1.8750000e-01   1.1538500e-01   0.0000000e+00   6.1538500e-01   2.1428600e-01   1.3636400e-01   3.3333299e-01
1.9608000e-02   5.1724002e-02   3.8461500e-01   2.1739100e-01   5.2632000e-02   3.3333299e-01   0.0000000e+00   1.4285700e-01   3.5714301e-01
3.9999999e-02   5.1724002e-02   3.8461500e-01   2.1739100e-01   1.1111100e-01   3.3333299e-01   0.0000000e+00   1.4285700e-01   3.5714301e-01
5.7691999e-02   1.0344800e-01   5.0000001e-02   6.8966001e-02   4.5455001e-02   4.5455001e-02   0.0000000e+00   0.0000000e+00   4.7619000e-02
4.0000001e-01   3.3333299e-01   7.4074000e-02   5.4053999e-02   1.9230799e-01   1.4814800e-01   1.2500000e-01   1.2903200e-01   2.0000000e-01
1.0416700e-01   6.8966001e-02   1.8750000e-01   1.1538500e-01   0.0000000e+00   5.0000000e-01   2.1428600e-01   1.3636400e-01   4.2857099e-01
2.2413801e-01   4.8148099e-01   5.7142999e-02   9.3023002e-02   1.8181799e-01   1.1428600e-01   2.0689701e-01   7.5000003e-02   1.1764700e-01
3.2258000e-02   4.2856999e-02   3.0434799e-01   3.3333299e-01   3.2258000e-02   1.4285700e-01   0.0000000e+00   2.0000000e-01   1.4814800e-01
3.4042600e-01   2.8571400e-01   3.5714000e-02   2.6316000e-02   1.9230799e-01   1.0714300e-01   2.2727300e-01   1.6666700e-01   1.5384600e-01
5.7691999e-02   1.0344800e-01   1.6666700e-01   1.0714300e-01   0.0000000e+00   2.7777800e-01   5.8333302e-01   1.7391300e-01   2.9411799e-01
3.9216001e-02   5.0847001e-02   4.6153799e-01   2.6087001e-01   1.0526300e-01   3.1250000e-01   0.0000000e+00   1.9047600e-01   3.3333299e-01
2.4590200e-01   2.3188400e-01   2.7272701e-01   3.3333299e-01   4.7619000e-02   2.5714299e-01   1.4285700e-01   1.7073201e-01   2.2857100e-01
```

Figure 2. view of the drug features file

References:

[١] Y. Yamanishi, M. Araki, A. Gutteridge, W. Honda, and M. Kanehisa, "Prediction of drug–target interaction networks from the integration of chemical and genomic spaces," *Bioinformatics,* vol. 24, no. 13, pp. i232-i240, 2008.

[٢] Z.-Y. Zhao *et al.*, "An Ensemble Learning-Based Method for Inferring Drug-Target Interactions Combining Protein Sequences and Drug Fingerprints," *BioMed Research International,* vol. 2021, 2021.

[٣] Y. Tabei, E. Pauwels, V. Stoven, K. Takemoto, and Y. Yamanishi, "Identification of chemogenomic features from drug–target interaction networks using interpretable classifiers," *Bioinformatics,* vol. 28, no. 18, pp. i487-i494, 2012.

[٤] A. Sharma and R. Rani, "BE-DTI': ensemble framework for drug target interaction prediction using dimensionality reduction and active learning," *Computer methods and programs in biomedicine,* vol. 165, pp. 151-162, 2018.

[٥] C. Réda, E. Kaufmann, and A. Delahaye-Duriez, "Machine learning applications in drug development," *Computational and structural biotechnology journal,* vol. 18, pp. 241-252, 2020.

[٦] D. S. Wishart *et al.*, "DrugBank: a comprehensive resource for in silico drug discovery and exploration," *Nucleic acids research,* vol. 34, no. suppl_1, pp. D668-D672, 2006.

[٧] S. Günther *et al.*, "SuperTarget and Matador: resources for exploring drug-target relationships," *Nucleic acids research,* vol. 36, no. suppl_1, pp. D919-D922, 2007.

[٨] I. Schomburg *et al.*, "BRENDA, the enzyme database: updates and major new developments," *Nucleic acids research,* vol. 32, no. suppl_1, pp. D431-D433, 2004.

[٩] M. Kanehisa *et al.*, "KEGG for linking genomes to life and the environment," *Nucleic acids research,* vol. 36, no. suppl_1, pp. D480-D484, 2007.