۱.مجموعه داده طلایی

مجموعه داده طلایی به چهار مجموعه داده تقسیم می‌شود؛ به عبارت دیگر چهار مجموعه داده آنزیم‌ها[1]، کانال‌های یونی[2]، GPCRها[3] و گیرنده‌های هسته‌ای[4] در حقیقت کلاس‌های پروتئین در داخل بدن انسان هستند. این مجموعه داده طلایی در سال ۲۰۰۸ توسط یامانیشی[5] و همکاران [۱] معرفی شده است. این مجموعه داده به ترتیب شامل ۹۰، ۶۳۵ ، ۱۴۷۶ و ۲۰۲۹ تعامل شناخته‌شده میان ۵۴، ۲۲۳، ۲۱۰ و ۴۴۵ دارو و ۲۶، ۹۵، ۲۰۴ و ۶۶۴ هدف است. عموم روش‌های پیش‌بینی تعاملات دارو-هدف، روش خود را بر پایه این مجموعه داده‌ها پیاده‌سازی و نتایج را گزارش کرده‌اند[۴-۲]. هر مجموعه داده موجود در این مجموعه داده طلایی شامل سه فایل متنی تعاملات دارو- هدف، ویژگی‌های داروها و ویژگی‌های اهداف می‌باشد و داده‌های موجود در این مجموعه داده، از نوع عددی هستند.

مجموعه داده طلایی در دو قسمت توضیح داده می‌شود؛ در قسمت اول، اطلاعاتی درباره تعاملات دارو-هدف و در قسمت دوم اطلاعات دارو و هدفی که مورد استفاده قرار گرفته شده است و نحوه نمایش آن‌ها به صورت بردارهای ویژگی، تشریح می‌شود.

۱-۱.اطلاعات تعاملات دارو-هدف

تعاملات موجود در این چهار مجموعه داده با استفاده از چند منبع معتبر KEGG BRITE [۵]، BRENDA [۶]، SuperTarget [۷] و DrugBank [۸] جمع‌آوری شده است. عناصر فایل‌های تعاملات دارو- هدف از اعداد صفر و یک تشکیل شده است؛ این فایل‌ها به صورت یک ماتریس دوبعدی $Y \in \mathbb{R}^{n \times m}$ دارای n سطر دارو و m ستون پروتئین هدف می‌باشد؛ به طوریکه اگر $Y_{ij} = 1$ باشد، دارو d_i با هدف t_j تعامل دارد و در غیر اینصورت $Y_{ij} = 0$ می‌باشد. به عبارت دیگر عدد یک نشان‌دهنده وجود تعامل و عدد صفر نشان‌دهنده عدم تعامل بین دارو d_i که در سطر ماتریس و هدف t_j که در ستون ماتریس قرار گرفته است، می‌باشد.

[1] Enzymes (E)
[2] Ion channels (IC)
[3] G protein-coupled receptors (GPCRs)
[4] Nuclear receptors (NR)
[5] Yamanishi

۲-۱.اطلاعات داروها و اهداف

مورد توجه است که در مجموعه داده آنزیم‌ها بر روی فعل و انفعالات بین آنزیم ها و ترکیبات به جای فعل و انفعالات متابولیکی، بنابراین تمام لیگاندهای موجود در داده های آنزیمی به جای سوبستراها یا محصولات، بازدارنده یا فعال کننده هستند. کوفاکتورهایی مانند آدنوزین تری فسفات (ATP) و نیکوتین آمید آدنین دی نوکلئوتید فسفات (NADPH) نیز شامل نمی شوند، مگر زمانی که به عنوان تنظیم کننده در پایگاه داده BRENDA مشروح شده باشند. همچنین، از ترکیباتی که وزن مولکولی آنها کمتر از ۱۰۰ است استفاده نمی‌شود.

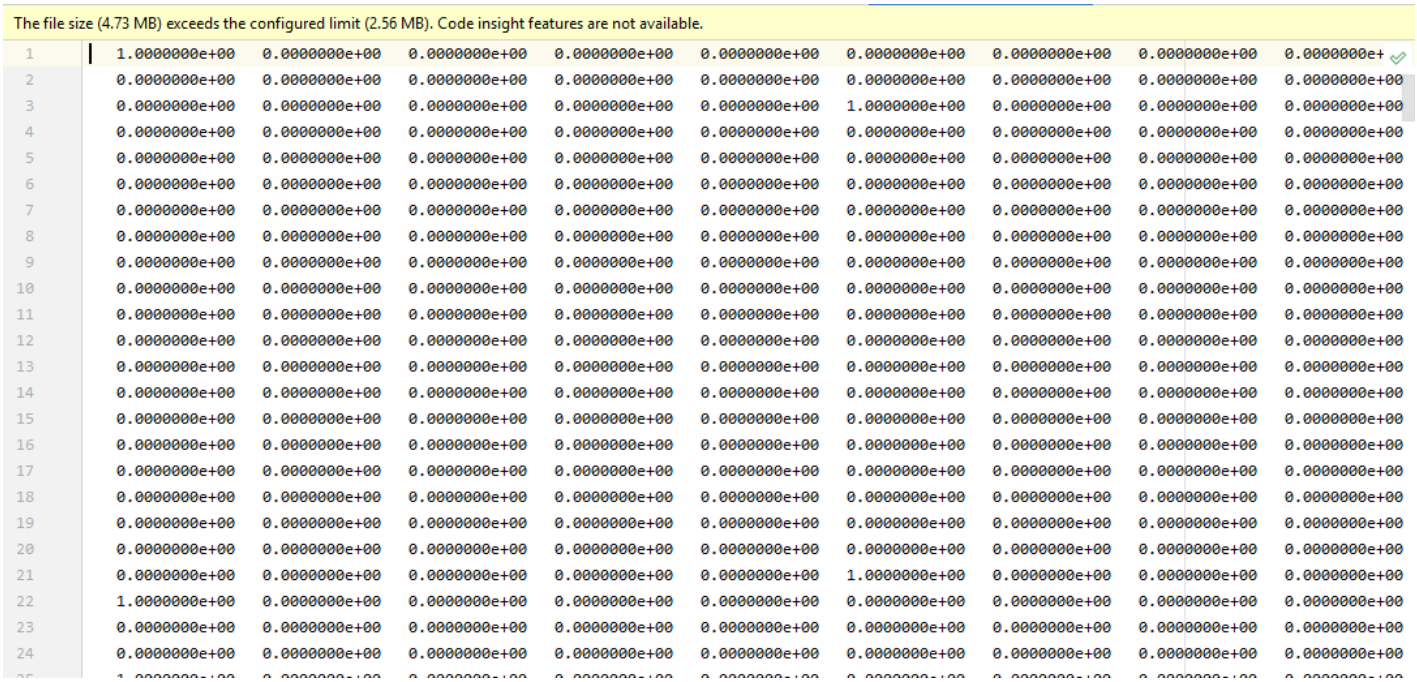
ساختارهای شیمیایی داروها از بخش DRUG و COMPOUND در پایگاه داده KEGG LIGAND و توالی های اسید آمینه پروتئین های مورد نظر از پایگاه داده KEGG GENES به دست آمده است؛ در واقع در این مطالعه بر روی پروتئین های موجود در بدن انسان تمرکز شده است.

فایل‌های ویژگی‌های داروها به صورت ماتریس $D \in \mathbb{R}^{n\times f}$ و ویژگی‌های اهداف به شکل ماتریس $T \in \mathbb{R}^{m\times g}$ می‌باشد که f و g تعداد ویژگی‌های داروها و اهداف است؛ به این صورت که هر دارو به شکل $d = [d_1, d_2, \dots, d_p]$ و هر هدف به صورت $t = [t_1, t_2, \dots, t_q]$ نمایش داده می‌شود. جزئیات این مجموعه داده در جدول ۱ آورده شده است.

جدول ۱.جزئیات مجموعه داده طلایی یامانیشی و همکاران

مجموعه داده	تعداد داروها	تعداد اهداف	تعداد تعاملات
آنزیم‌ها	۴۴۵	۶۶۴	۲۹۲۶
کانال‌های یونی	۲۱۰	۲۰۴	۱۴۷۶
GPCRs	۲۲۳	۹۵	۶۳۵
گیرنده‌های هسته‌ای	۵۴	۲۶	۹۰

تشکیل بردارهای دارو-هدف برای استفاده در مدل یادگیری ماشین در هر مجموعه داده به صورت $[d_1, d_2, \dots, d_p, t_1, t_2, \dots, t_q, I]$ می‌باشد که برچسب I یک عدد باینری است و نشان دهنده وجود یا عدم وجود تعامل بین دارو و هدف موجود در بردار می‌باشد. برچسب I از طریق ماتریس تعاملات دارو-هدف مشخص می‌شود. شکل ۱ و ۲ به ترتیب نمایی از فایل تعاملات دارو-هدف آنزیم‌ها و فایل ویژگی‌های داروهای موجود در آن می‌باشد.



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
```

شکل ۱ نمایی از فایل تعاملات دارو-هدف آنزیم‌ها

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

1	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-
2	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	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
4	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
5	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
6	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
7	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
8	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
9	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
10	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
11	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
12	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
13	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
14	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
15	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
16	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
17	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
18	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
19	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
20	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
21	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
22	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
23	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
24	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

شکل ۲ نمایی از فایل ویژگی‌های دارو های موجود در مجموعه داده آنزیم

فهرست منابع

[۱] 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.

[۲] 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.

[۳] Y. Ding, J. Tang, and F. Guo, "Identification of drug–target interactions via fuzzy bipartite local model," *Neural Computing and Applications,* vol. 32, no. 14, pp. 10303-10319, 2020.

[٤] C. Chen *et al.*, "DNN-DTIs: Improved drug-target interactions prediction using XGBoost feature selection and deep neural network," *Computers in Biology and Medicine,* vol. 136, p. 104676, 2021.

[٥] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima, "KEGG: new perspectives on genomes, pathways, diseases and drugs," *Nucleic acids research,* vol. 45, no. D1, pp. D353-D361, 2017.

[٦] 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.

[۷] 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.

[۸] 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, 2008.