Seven attributes, i.e., molecular structure, target, pathway, side effect, phenotype, gene, and disease are employed in this work for ADDI modeling. In the experiments, there are overall 25,495 adverse drug pairs extracted from Twosides associated with 752 drugs for ADDI modeling. This folder contains the seven attribute information of 752 drugs. The details of the seven subfolders included in this folder are in the following. We recommend to open these files by software Emeditor (available at https://www.emeditor.com/) that can better show the rows and columns of the files.

**Catalogue**

|  |  |
| --- | --- |
| **Folder/File Name** | **Page** |
| **1. MolecularStructure** | **3** |
| **DrugSet.txt** | **3** |
| **DrugDBSMILES.txt** | **3** |
| **DrugMolecularStructureMatrix.txt** | **3** |
| **2. Target** | **4** |
| **DrugSet.txt** | **4** |
| **TargetSet.txt** | **4** |
| **DrugDBTargetRelation.txt** | **4** |
| **DrugTargetMatrix.txt** | **4** |
| **3. Pathway** | **5** |
| **DrugSet.txt** | **5** |
| **PathwaySet.txt** | **5** |
| **DrugDBPathwayRelation.txt** | **5** |
| **DrugPathwayMatrix.txt** | **5** |
| **4. SideEffect** | **6** |
| **DrugSet.txt** | **6** |
| **SideEffectSet.txt** | **6** |
| **DrugDBSideEffectRelation.txt** | **6** |
| **DrugSideEffectMatrix.txt** | **6** |
| **5. Phenotype** | **7** |
| **DrugSet.txt** | **7** |
| **PhenotypeSet.txt** | **7** |
| **DrugDBPhenotypeidRelation.txt** | **7** |
| **DrugPhenotypeMatrix.txt** | **7** |
| **6. Gene** | **8** |
| **DrugSet.txt** | **8** |
| **GeneSet.txt** | **8** |
| **DrugDBGeneRelation.txt** | **8** |
| **DrugGeneMatrix.txt** | **8** |
| **7. Disease** | **8** |
| **DrugSet.txt** | **9** |
| **DiseaseSet.txt** | **9** |
| **DrugDBDiseaseRelation.txt** | **9** |
| **DrugDiseaseMatrix.txt** | **9** |

**1. MolecularStructure**

This subfolder consists of three files: **DrugSet.txt**, **DrugDBSMILES.txt**, and **DrugMolecularStructureMatrix.txt**. The details of these three files are as follows.

**DrugSet.txt** records the index of drugs (0-751), the drug identifier with prefix 'DB', and the drug name, in which the first column is the index of 752 drugs, the second column is the drug identifier with prefix 'DB', and the third column represents the drug name.

**DrugDBSMILES.txt** contains the molecular structure information of 752 drugs, where the first column is the drug identifier with prefix 'DB' and the second column represents the canonical SMILES of 752 drugs from DrugBank (https://go.drugbank.com/) that can be converted into their 2D molecular structures. For example, the canonical SMILES of drug Leuprolide with identifier DB00007 (the second row in this file) is CCNC(=O)[C@@H]1CCCN1C(=O)[C@H] (CCCNC(N)=N)NC(=O)[C@H](CC(C)C)NC(=O)[C@@H](CC(C)C)NC(=O)[C@H](CC1=CC=C(O)C=C1)NC(=O)[C@H](CO)NC(=O)[C@H](CC1=CNC2=C1C=CC=C2)NC(=O)[C@H](CC1=CNC=N1)NC(=O)[C@@H]1CCC(=O)N1, and its 2D molecular structure is given in Fig. 1. Based on the canonical SMILES of drugs, we adopt ‘rcdk’ library from R language with PubChem substructure fingerprints to transform the canonical SMILES of drugs into their integer vectors ranging from 1 to 881, which are presented in the second column of DrugMolecularStructureMatrix.txt, where integer *i* included in a vector indicates the drug owns the i-th molecular substructure (*i* is from 1 to 881).

|  |
| --- |
|  |
| Fig. 1. The 2D molecular structure of drug Leuprolide  with drug identifier DB00007. |

**DrugMolecularStructureMatrix.txt** contains the molecular structure vectors of 752 drugs, where the first column represents the identifiers of all drugs and the second column denotes the 881-dimensional binary vectors, in which if a drug owns the *i*-th molecular substructure, then the *i*-th entry in its binary vector in the second column is 1; otherwise the *i*-th entry in its binary vector is 0 (*i* is from 0 to 880). The second column in this file can be regarded as the original molecular structure matrix of drugs for multi-attribute representation learning.

**2. Target**

This subfolder consists of four files: **DrugSet.txt**, **TargetSet.txt, DrugDBTargetRelation.txt,** and **DrugTargetMatrix.txt**. The details of these four files are as follows.

**DrugSet.txt** records the index of drugs (0-751), the drug identifier with prefix 'DB', and the drug name, in which the first column is the index of 752 drugs, the second column is the drug identifier with prefix 'DB', and the third column represents the drug name.

**TargetSet.txt** records the detailed information of targets. There are totally 1431 targets in association with 752 drugs. In this file, the first column represents the indexes of all targets from 0 to 1430 and the second column indicates the names of all targets from DrugBank (https://go.drugbank.com/).

**DrugDBTargetRelation.txt** records the relations between targets and drugs, where **the first column is the identifiers of 752 drugs, and the second column represents the identifiers of the related targets for each drug.** For example, the drug with identifier DB00007 (i.e., the second row in this file) is associated with targets P30968 and P08684, and the drug with identifier DB00035 (i.e., the fourth row in this file) is related to targets P30518, P37288, P47901, P23219, and P35354. The related targets w.r.t. each drug are split by spaces. Each drug is in association with a number of targets shown in this file. In particular, there are overall 38 drugs in correlation with no target (e.g., DB00080, DB00225), which means that the target information of the corresponding drugs are not recorded in DrugBank database, leading to their absence of target information. The absent target information of 38 drugs are to be recovered by multi-attribute representation learning.

**DrugTargetMatrix.txt** presents the target vectors of 752 drugs, where the first column represents the identifiers of all drugs and the second column indicates the 1431-dimensional binary target vectors of drugs (the entry index for the binary vector is from 0 to 1430). For each drug (row), if it is associated with the *i*-th target, then the *i*-th entry of its binary vector in the second column is 1; otherwise, the *i*-th entry of its binary vector is 0. Due to that 38 drugs suffer from the absent target information from DrugBank database, the second column for these drugs are empty, which is in correspondence to the second column in **DrugDBTargetRelation.txt.**

**3. Pathway**

This subfolder consists of four files: **DrugSet.txt**, **PathwaySet.txt**, **DrugDBPathwayRelation.txt**, and **DrugPathwayMatrix.txt**. The details of these four files are as follows.

**DrugSet.txt** records the index of drugs (0-751), the drug identifier with prefix 'DB', and the drug name, in which the first column is the index of 752 drugs, the second column is the drug identifier with prefix 'DB', and the third column represents the drug name.

**PathwaySet.txt** records the detailed information of pathways. There are totally 427 pathways in association with 752 drugs. In this file, the first column represents the indexes of all pathways from 0 to 426 and the second column indicates the names of all pathways from KEGG (https://www.kegg.jp/).

**DrugDBPathwayRelation.txt** records the relations between pathways and drugs, where **the first column is the identifiers of 752 drugs, and the second column represents the identifiers of the related pathways for each drug.** For example, the drug with identifier DB00007 (i.e., the second row in this file) is associated with pathways hsa04080, hsa04912, hsa00140, hsa00591, hsa00830, hsa00980, hsa00982, hsa00983, hsa01100, hsa04976, and hsa05204. The related pathways w.r.t. each drug are split by spaces. Each drug is in association with a number of pathways shown in this file. There are overall 47 drugs in correlation with no pathway (e.g., DB00080, DB00225, DB00258), which means that the pathway information of the corresponding drugs are not recorded in KEGG database, leading to their absence of pathway information. The absent pathway information of 47 drugs are to be recovered by multi-attribute representation learning.

**DrugPathwayMatrix.txt** presents the pathway vectors of 752 drugs, where the first column represents the identifiers of all drugs and the second column indicates the 427-dimensional binary pathway vectors of drugs (the entry index for the binary vector is from 0 to 426). For each drug (row), if it is associated with the *i*-th pathway, then the *i*-th entry of its binary vector in the second column is 1; otherwise, the *i*-th entry of its binary vector is 0. Due to that 47 drugs suffer from the absent pathway information from KEGG database, the second column for these drugs are empty, which is in correspondence to the second column in **DrugDBPathwayRelation.txt.**

**4. SideEffect**

This subfolder consists of four files: **DrugSet.txt**, **SideEffectSet.txt, DrugDBSideEffectRelation.txt,** and **DrugSideEffectMatrix.txt**. The details of these four files are as follows.

**DrugSet.txt** records the index of drugs (0-751), the drug identifier with prefix 'DB', and the drug name, in which the first column is the index of 752 drugs, the second column is the drug identifier with prefix 'DB', and the third column represents the drug name.

**SideEffectSet.txt** records the detailed information of side effects. There are totally 3750 side effects in association with 752 drugs. In this file, the first column represents the indexes of all side effects from 0 to 3749 and the second column indicates the names of all side effects from SIDER (http://sideeffects.embl.de/).

**DrugDBSideEffectRelation.txt** records the relations between side effects and drugs, where **the first column is the identifiers of 752 drugs, and the second column represents the identifiers of the related side effects for each drug.** For example, the drug with identifier DB00165 (i.e., the 18th row in this file) is associated with side effects Headache, Nausea, Paraesthesia, Neuropathy peripheral, Somnolence. The related side effects w.r.t. each drug are split by symbol ‘|’. Each drug is in association with a number of side effects shown in this file. There are overall 148 drugs in correlation with no side effect (e.g., DB00121, DB00137, DB00147, DB00163), which means that the side effect information of the corresponding drugs are not recorded in SIDER database, leading to their absence of side effect information. The absent side effect information of 148 drugs are to be recovered by multi-attribute representation learning.

**DrugSideEffectMatrix.txt** presents the side effect vectors of 752 drugs, where the first column represents the identifiers of all drugs and the second column indicates the 3750-dimensional binary side effect vectors of drugs (the entry index for the binary vector is from 0 to 3749). For each drug (row), if it is associated with the *i*-th side effect, then the *i*-th entry of its binary vector in the second column is 1; otherwise, the *i*-th entry of its binary vector is 0. Due to that 148 drugs suffer from the absent side effect information from SIDER database, the second column for these drugs are empty, which is in correspondence to the second column in **DrugDBSideEffectRelation.txt.**

**5. Phenotype**

This subfolder consists of four files: **DrugSet.txt**, **PhenotypeSet.txt, DrugDBPhenotypeidRelation.txt,** and **DrugPhenotypeMatrix.txt**. The details of these four files are as follows.

**DrugSet.txt** records the index of drugs (0-751), the drug identifier with prefix 'DB', and the drug name, in which the first column is the index of 752 drugs, the second column is the drug identifier with prefix 'DB', and the third column represents the drug name.

**PhenotypeSet.txt** records the detailed information of phenotypes. There are totally 2656 phenotypes in association with 752 drugs. In this file, the first column represents the indexes of all phenotypes from 0 to 2655 and the second column indicates the names of all phenotypes from CTD database (http://ctdbase.org/).

**DrugDBPhenotypeidRelation.txt** records the relations between phenotypes and drugs, where **the first column is the identifiers of 752 drugs, and the second column represents the identifiers of the related phenotypes for each drug.** For example, the drug with identifier DB00007 (i.e., the second row in this file) is associated with phenotypes GO:0006915 and GO:0007283. The related phenotypes w.r.t. each drug are split by spaces. Each drug is in association with a number of phenotypes shown in this file. There are overall 154 drugs in correlation with no phenotype (e.g., DB00006, DB00080, DB00163, DB00209), which means that the phenotype information of the corresponding drugs are not recorded in CTD database, leading to their absence of phenotype information. The absent phenotype information of 154 drugs are to be recovered by multi-attribute representation learning.

**DrugPhenotypeMatrix.txt** presents the phenotype vectors of 752 drugs, where the first column represents the identifiers of all drugs and the second column indicates the 2656-dimensional binary phenotype vectors of drugs (the entry index for the binary vector is from 0 to 2655). For each drug (row), if it is associated with the *i*-th phenotype, then the *i*-th entry of its binary vector in the second column is 1; otherwise, the *i*-th entry of its binary vector is 0. Due to that 154 drugs suffer from the absent phenotype information from CTD database, the second column for these drugs are empty, which is in correspondence to the second column in **DrugDBPhenotypeidRelation.txt.**

**6. Gene**

This subfolder consists of four files: **DrugSet.txt**, **GeneSet.txt, DrugDBGeneRelation.txt,** and **DrugGeneMatrix.txt**. The details of these four files are as follows.

**DrugSet.txt** records the index of drugs (0-751), the drug identifier with prefix 'DB', and the drug name, in which the first column is the index of 752 drugs, the second column is the drug identifier with prefix 'DB', and the third column represents the drug name.

**GeneSet.txt** records the detailed information of genes. There are totally 6267 genes in association with 752 drugs. In this file, the first column represents the indexes of all genes from 0 to 6266 and the second column indicates the names of all genes from KEGG database (https://www.kegg.jp/).

**DrugDBGeneRelation.txt** records the relations between genes and drugs, where **the first column is the identifiers of 752 drugs, and the second column represents the identifiers of the related genes for each drug.** For example, the drug with identifier DB00006 (i.e., the first row in this file) is associated with genes F2, PT, RPRGL2, THPH1, and MPO. The related genes w.r.t. each drug are split by spaces. Each drug is in association with a number of genes shown in this file. There are overall 40 drugs in correlation with no gene (e.g., DB00080, DB00258, DB00410), which means that the gene information of the corresponding drugs are not recorded in KEGG database, leading to their absence of gene information. The absent gene information of 40 drugs are to be recovered by multi-attribute representation learning.

**DrugGeneMatrix.txt** presents the gene vectors of 752 drugs, where the first column represents the identifiers of all drugs and the second column indicates the 6267-dimensional binary gene vectors of drugs (the entry index for the binary vector is from 0 to 6266). For each drug (row), if it is associated with the *i*-th gene, then the *i*-th entry of its binary vector in the second column is 1; otherwise, the *i*-th entry of its binary vector is 0. Due to that 40 drugs suffer from the absent gene information from KEGG database, the second column for these drugs are empty, which is in correspondence to the second column in **DrugDBGeneRelation.txt.**

**7. Disease**

This subfolder consists of four files: **DrugSet.txt**, **DiseaseSet.txt, DrugDBDiseaseRelation.txt,** and **DrugDiseaseMatrix.txt**. The details of these four files are as follows.

**DrugSet.txt** records the index of drugs (0-751), the drug identifier with prefix 'DB', and the drug name, in which the first column is the index of 752 drugs, the second column is the drug identifier with prefix 'DB', and the third column represents the drug name.

**DiseaseSet.txt** records the detailed information of diseases. There are totally 613 diseases in association with 752 drugs. In this file, the first column represents the indexes of all diseases from 0 to 612 and the second column indicates the names of all diseases from KEGG (https://www.kegg.jp/).

**DrugDBDiseaseRelation.txt** records the relations between diseases and drugs, where **the first column is the identifiers of 752 drugs, and the second column represents the identifiers of the related diseases for each drug.** For example, the drug with identifier DB00006 (i.e., the first row in this file) is associated with diseases H00223, H01254, H00101, and H02025. The related diseases w.r.t. each drug are split by spaces. Each drug is in association with a number of diseases shown in this file. There are overall 117 drugs in correlation with no disease (e.g., DB00080, DB00137, DB00147), which means that the disease information of the corresponding drugs are not recorded in KEGG database, leading to their absence of disease information. The absent disease information of 117 drugs are to be recovered by multi-attribute representation learning.

**DrugDiseaseMatrix.txt** presents the disease vectors of 752 drugs, where the first column represents the identifiers of all drugs and the second column indicates the 613-dimensional binary disease vectors of drugs (the entry index for the binary vector is from 0 to 612). For each drug (row), if it is associated with the *i*-th disease, then the *i*-th entry of its binary vector in the second column is 1; otherwise, the *i*-th entry of its binary vector is 0. Due to that 117 drugs suffer from the absent disease information from KEGG database, the second column for these drugs are empty, which is in correspondence to the second column in **DrugDBDiseaseRelation.txt.**