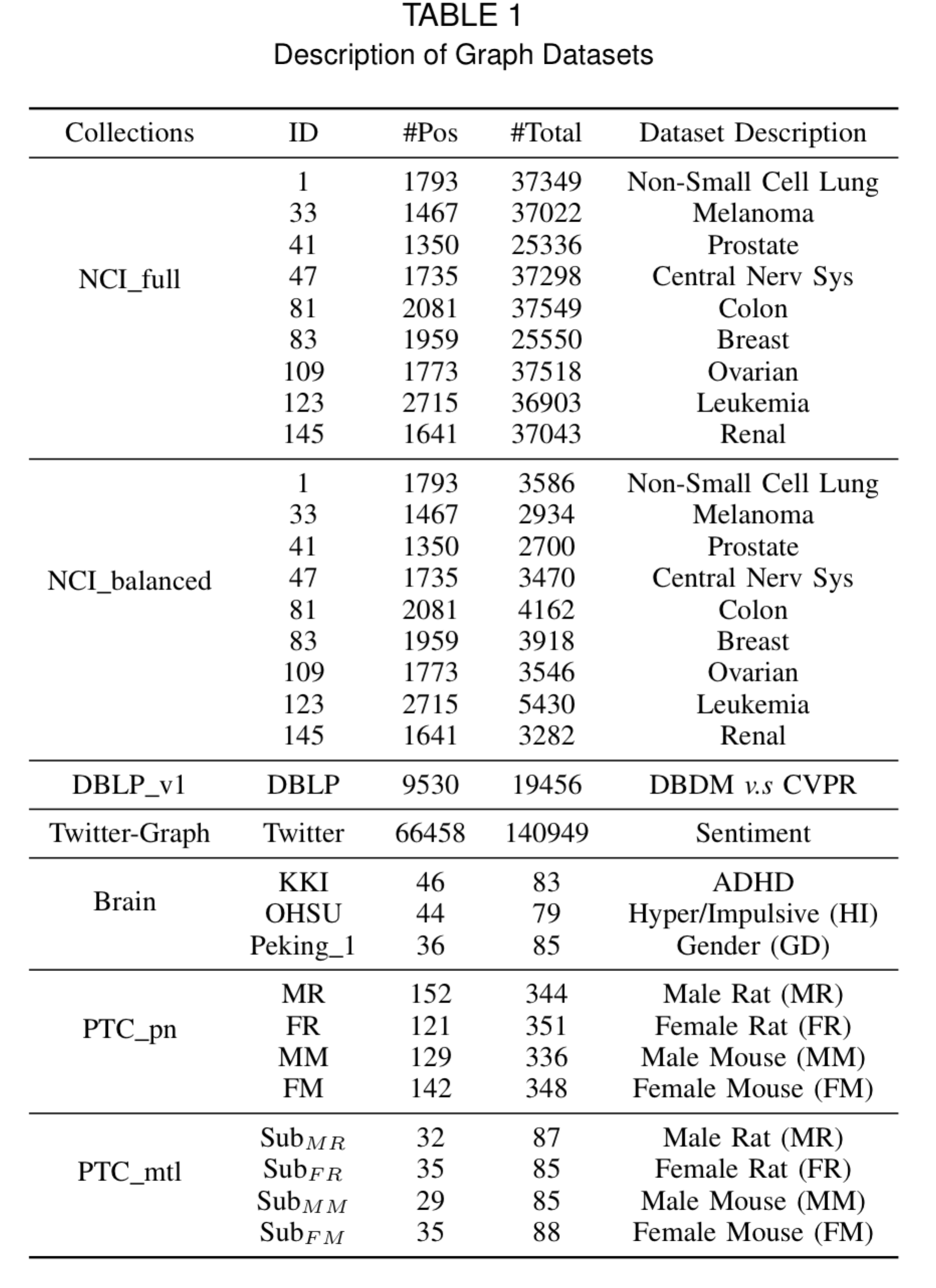
**A Repository of Benchmark Graph Dataset**

This repository maintains several benchmark graph datasets, which are widely used for graph classification. The graph datasets ranges from chemical compounds, scientific publications, to twitter sentiment graphs. The chemical compound graph datasets are in “.sdf” or “.smi” format, and other graph dataset are represented as “.nel” format. All these graph datasets can be handle by frequent subgraph miner packages such as [**Moss**](http://www.borgelt.net/moss.html) **[1]**. A summarization of our graph datasets is given in Table 1.



If you used the dataset, please cite the related papers properly.

1. **NIC Anti-cancer activity prediction data**

*Description:*

The NCI graph datasets are commonly used as the benchmark for graph classification. Each NCI dataset belongs to a bioassay task for anticancer activity prediction, where each chemical compound is represented as a graph, with atoms representing nodes and bonds as edges. A chemical compound is positive if it is active against the corresponding cancer, or negative otherwise. Table 1 summarizes the NCI graph data we download from [PubChem](http://pubchem.ncbi.nlm.nih.gov). We have removed disconnected graphs and graphs with unexpected atoms (some graphs have atoms represented as `\*`) in the original graphs. Columns 2-3 show the number of positive and total number of graphs in each dataset, and Columns 4-5 indicate the average number of nodes and edges in each dataset, respectively.

*Full Dataset:*

The full datasets of NCI graphs can be downloaded here (**NCI\_full.zip**), which is naturally imbalanced and ideal benchmark for imbalanced or cost-sensitive graph classification. We have considered cost-sensitive graph classification in [2], and graph stream classification in [3][4].

*Partial Dataset:*

We randomly select #Pos number of negative graphs from each original graph set to create balanced graph datasets, which are available here (**NCI\_balanced.zip**). This dataset was used in [5] for multi-task graph classification.

*Citations:*

If you used this dataset, please cite 1-2 following papers:

Shirui Pan, Jia Wu, and Xingquan Zhu “CogBoost: Boosting for Fast Cost-sensitive Graph Classification", IEEE Transactions on Knowledge and Data Engineering (**TKDE**), 27(11): 2933-2946 (2015)

Shirui Pan, Jia Wu, Xingquan Zhu, and Chengqi Zhang, “Graph Ensemble Boosting for Imbalanced Noisy Graph Stream Classification", IEEE Transactions on Cybernetics (**TCYB**), 45(5): 940-954 (2015).

Shirui Pan, Xingquan Zhu, Chengqi Zhang, and Philip S. Yu. "Graph Stream Classification using Labeled and Unlabeled Graphs", International Conference on Data Engineering (**ICDE**), pages 398-409, 2013

Shirui Pan, Jia Wu, Xingquan Zhu, Guodong Long, and Chengqi Zhang. " Task Sensitive Feature Exploration and Learning for Multi-Task Graph Classification." IEEE Trans. Cybernetics (**TCYB**) 47(3): 744-758 (2017).

1. **PTC Predictive Toxicology Challenge Data**

*Description*:

This [PTC](http://www.predictive-toxicology.org/ptc/) graph dataset include a number of carcinogenicity tasks for toxicology prediction of chemical compounds.

The dataset we selected contains 417 compounds from four types of test animals: MM (male mouse), FM (female mouse), MR (male rat), and FR (female rat). Each compound is with one label selected from {CE, SE, P, E, EE, IS, NE, N}, which stands for Clear Evidence of Carcinogenic Activity (CE), Some Evidence of Carcinogenic Activity (SE), Positive (P), Equivocal (E), Equivocal Evidence of Carcinogenic Activity (EE), Inadequate Study of Carcinogenic Activity (IS), No Evidence of Carcinogenic Activity (NE), and Negative (N).

*Full Dataset:*

By setting {CE, SE, P} as positive label, {NE, N} as negative label, and remove the data with {E, EE, IS} we can obtain four graph datasets **PTC\_pn.zip.**

*Sub-dataset:*

The data can be formulated as a multi-task problem. We can randomly split 417 compounds into four equal non-overlapping subsets. For each subset, we only consider one type of carcinogenicity test as its learning task. The multi-task graph dataset can be downloaded here (**PTC\_mtl.zip**).

*Citation:*

If you used this dataset, please cite the following paper:

Shirui Pan, Jia Wu, Xingquan Zhu, Guodong Long, and Chengqi Zhang. " Task Sensitive Feature Exploration and Learning for Multi-Task Graph Classification." IEEE Trans. Cybernetics (**TCYB**) 47(3): 744-758 (2017).

1. **DBLP Graph Datasets**

*Description:*

The DBLP dataset consists of bibliography data in computer science. Each record in DBLP is associated with a number of attributes such as abstract, authors, year, venue, title, and reference ID. To build a graph stream, we select a list of conferences (as shown in Table I) and use the papers published in these conferences (in chronological order) to form a binary-class graph stream. The classification task is to predict whether a paper belongs to DBDM (database and data mining) or CVPR (computer vision and pattern recognition) field, by using the references and the title of each paper.

*Version 1:*

Each paper in DBLP is represented as a graph, where each node denotes a Paper ID or a keyword and each edge denotes the citation relationship between papers or keyword relations in the title. More specifically, we denote that (1) each paper ID is a node; (2) if a paper P.A cites another paper P.B, there is an edge between P.A and P.B; (3) each keyword in the title is also a node; (4) each paper ID node is connected to the keyword nodes of the paper; and (5) for each paper, its keyword nodes are fully connected with each other. An example of DBLP graph data is shown in Fig. 4.

The dataset can be downloaded here (**DBLP\_v1.zip**).

*Citation:*

If you used this dataset, please cite the following paper:

Shirui Pan, Xingquan Zhu, Chengqi Zhang, and Philip S. Yu. "Graph Stream Classification using Labeled and Unlabeled Graphs", International Conference on Data Engineering (**ICDE**), pages 398-409, 2013

1. **Twitter Sentiment Graph Data**

*Description:*

This dataset is extracted from [twitter sentiment](file://localhost/%7Bhttp/::jmgomezhidalgo.blogspot.com.au:2013:01:a-list-of-datasets-for-opinion-mining.html) classification. Because of the inherently short and sparse nature, twitter sentiment analysis (i.e., predicting whether a tweet reflects a positive or a negative feeling) is a difficult task. To build a graph dataset, we represent each tweet as a graph by using tweet content, with nodes in each graph denoting the terms and/or smiley symbols (e.g, :-D and :-P) and edges indicating the co-occurrence relationship between two words or symbols in each tweet. To ensure the quality of the graph, we only use tweets containing 20 or more words. We select the tweets from April 6 to June 16 to generate 140,949 graphs (in a chronological order). This dataset has been used for graph stream classification in [3] and cost-sensitive learning in [2].

*Dataset*:

The data set is available here (**Twitter-Graph.zip**)

*Citations:*

If you used this dataset, please cite the following papers:

Shirui Pan, Jia Wu, and Xingquan Zhu “CogBoost: Boosting for Fast Cost-sensitive Graph Classification", IEEE Transactions on Knowledge and Data Engineering (**TKDE**), 27(11): 2933-2946 (2015)

Shirui Pan, Jia Wu, Xingquan Zhu, and Chengqi Zhang, “Graph Ensemble Boosting for Imbalanced Noisy Graph Stream Classification", IEEE Transactions on Cybernetics (**TCYB**), 45(5): 940-954 (2015).

1. **Functional Brain Network Analysis Data**

*Description*:

*BrainNet Functional Brain Network Analysis Data* are constructed from the whole brain functional magnetic res- onance image (fMRI) atlas [6]. The purpose of the study is to map brain as a network (or a graph) where each node corresponds to a region of Interest (ROI) and the edge indicates correlations between two ROIs. In our experiments, we use functional parcellation results, CC200, from [6], which parcellate each brain into 200 regions of interest. In order to discover relationships between ROIs, the mean values of each ROI are recorded with respect to certain voxel time courses. By using Pearson correlations between two time courses, we can calculate correlation between two ROIs, and a graph is constructed by connecting ROIs whose correlations is higher than a threshold value. For ADHD and HI tasks, the functional response is real values, so we discretize the functional response to binary values by using a simple threshold.

*Dataset*:

The data set is available here (**Brain.zip**)

*Citations:*

If you used this dataset, please cite the following papers:

Shirui Pan, Jia Wu, Xingquan Zhu, Guodong Long, and Chengqi Zhang. " Task Sensitive Feature Exploration and Learning for Multi-Task Graph Classification." IEEE Trans. Cybernetics (**TCYB**) 47(3): 744-758 (2017).

**Reference:**

1. C. Borgelt and M. R. Berthold, “Mining molecular fragments: Finding relevant substructures of molecules,” 2002 IEEE International Conference on Data Mining. IEEE, 2002, pp. 51–58.
2. S. Pan, J. Wu, and X. Zhu, “Cogboost: Boosting for fast cost-sensitive graph classification,” *IEEE Transactions on Knowledge and Data Engineering*, 2015.
3. S. Pan, J. Wu, X. Zhu, and C. Zhang, “Graph ensemble boosting for imbalanced noisy graph stream classification,” *IEEE*  *Transactions on Cybernetic*, 2015.
4. S. Pan, X. Zhu, C. Zhang, and P. S. Yu, “Graph stream classification using labeled and unlabeled graphs,” in *Proc. of ICDE*. IEEE, 2013.
5. Shirui Pan, Jia Wu, Xingquan Zhu, and Chengqi Zhang, “Graph Ensemble Boosting for Imbalanced Noisy Graph Stream Classification", IEEE Transactions on Cybernetics (TCYB), 45(5): 940-954 (2015).
6. R. Craddock, C. James, P. Holtzheimer, X. Hu, and H. Mayberg, “A whole brain fmri atlas generated via spatially constrained spectral clustering,” *Human Brain Mapping*, vol. 33, 2012.

**Appendix:**

About the File Format:

1. Molecular Graphs:

**SDF file:**

SDF is one of a family of chemical-data file formats developed by MDL; it is intended especially for structural information. "SDF" stands for structure-data file, and SDF files actually wrap the molfile (MDL Molfile) format. Multiple compounds are delimited by lines consisting of four dollar signs ($$$$). A feature of the SDF format is its ability to include **associated data**. An example of SDF file is available here (**example.sdf**):

Note that the associated data “> <value>” in the file indicates the class label of a chemical compound, -1.0 means it is a negative example while 1.0 means it is a positive example.

**SMI file:**

The Simplified Molecular Input Line Entry Specification (**SMILES**) is a line notation for molecules. SMILES strings include connectivity but do not include 2D or 3D coordinates. Example of our smi format is:

TR155,-1,c1c(Cl)cc(Cl)c(O)c1Cl

TR174,-1,c1cc(N)ccc1N

TR366,1,c1cc(O)ccc1O

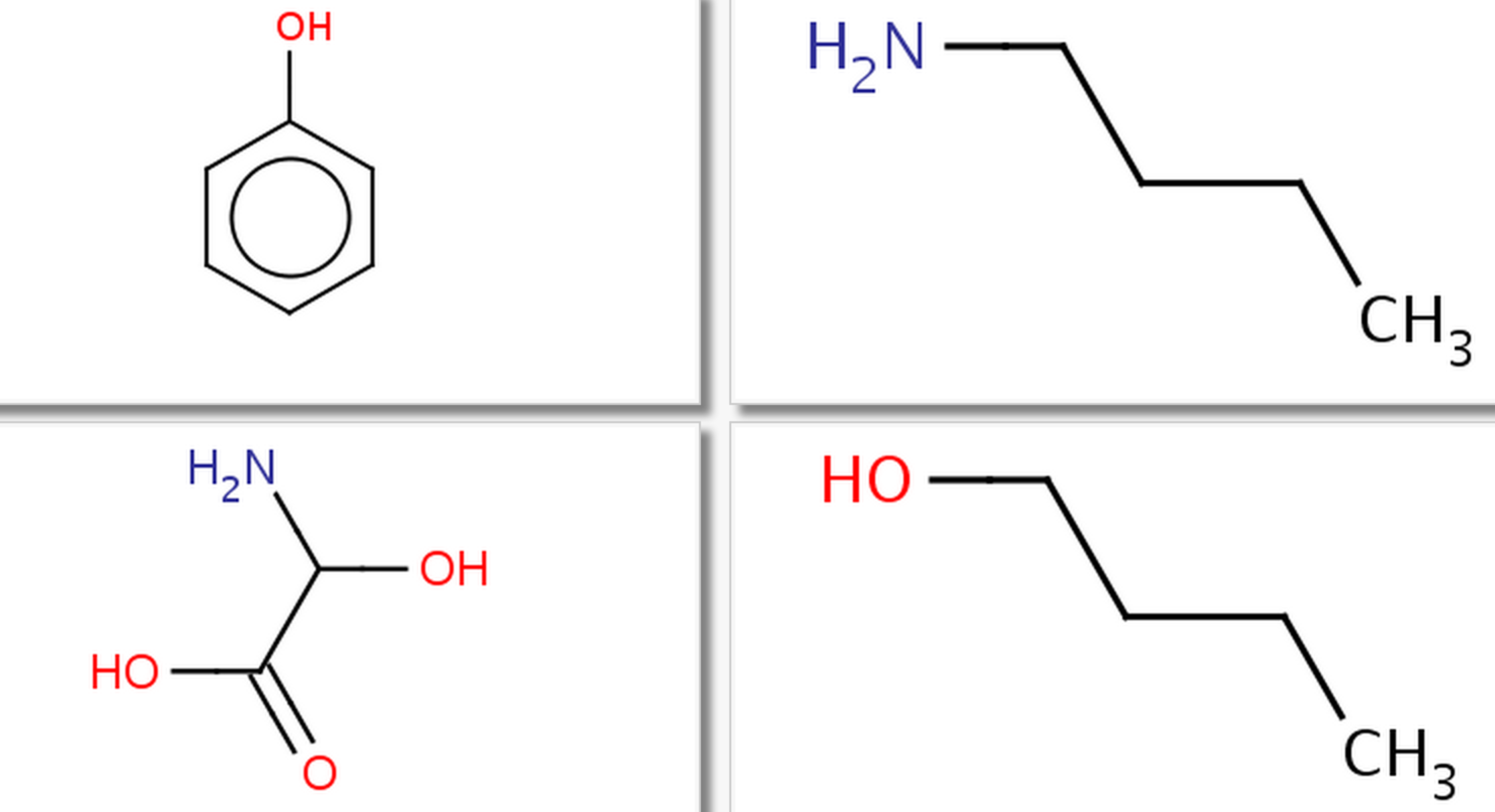
Where each line indicates a chemical compounds in the format as follows:

ID, Class Label, Smiles String

**Depict a chemical compound:**

The structure of chemical compounds can be depicted in a number of online toolboxes:

Here is a **link (http://cdb.ics.uci.edu/cgibin/Smi2DepictWeb.py)** you can have a try.



1. General Graphs:

**NEL file**

The NEL file is a general representation of graph objects, which explicitly shows the node edges information. An example of NEL file is as follows:

n 1 a

n 2 b

n 3 c

e 1 2 A

e 1 3 B

g graph\_1

x 1.0

In this example, the first 3 lines define 3 nodes with node label ‘a’, ‘b’, and ‘c’.

‘e 1 2 A’ means there is an edge with label A between the first and second nodes.

‘g graph\_1’ defines the name of this graph.

‘x 1.0’ indicates the class label of this graph. For binary classification, 1.0 means positive, -1.0 means negative.