Online Appendix for Submission 1706

Anonymous Author(s)

1 DATASETS

Molecular Interaction Prediction. For the datasets used in the molecular interaction prediction task, we convert the SMILES string into graph structure by using the Github code of CIGIN [7]. Moreover, for the datasets that are related to solvation free energies, i.e., MNSol, FreeSolv, CompSol, Abraham, and CombiSolv, we use the SMILES-based datasets provided in the previous work [8]. Only solvation free energies at temperatures of 298 K (± 2) are considered and ionic liquids and ionic solutes are removed [8].

- **Chromophore** [3] contains 20,236 combinations of 7,016 chromophores and 365 solvents which are given in the SMILES string format. All optical properties are based on scientific publications and unreliable experimental results are excluded after examination of absorption and emission spectra. In this dataset, we measure our model performance on predicting **maximum absorption wavelength (Absorption)**, **maximum emission wavelength (Emission)** and **excited state lifetime (Lifetime)** properties which are important parameters for the design of chromophores for specific applications. We delete the NaN values to create each dataset which is not reported in the original scientific publications. Moreover, for Lifetime data, we use log normalized target value since the target value of the dataset is highly skewed inducing training instability.
- **MNSol** [4] contains 3,037 experimental free energies of solvation or transfer energies of 790 unique solutes and 92 solvents. In this work, we consider 2,275 combinations of 372 unique solutes and 86 solvents following previous work [8].
- **FreeSolv** [5] provides 643 experimental and calculated hydration free energy of small molecules in water. In this work, we consider 560 experimental results following previous work [8].
- **CompSol** [6] dataset is proposed to show how solvation energies are influenced by hydrogen-bonding association effects. We consider 3,548 combinations of 442 unique solutes and 259 solvents in the dataset following previous work [8].
- **Abraham** [2] dataset is a collection of data published by the Abraham research group at College London. We consider 6,091 combinations of 1,038 unique solutes and 122 solvents following previous work [8].
- **CombiSolv** [8] contains all the data of MNSol, FreeSolv, CompSol, and Abraham, resulting in 10,145 combinations of 1,368 solutes and 291 solvents.

Drug-Drug Interaction Prediction. For the datasets used in the drug-drug interaction prediction task, we use the positive drug pairs given in MIRACLE Github link[1], which removed the data instances that cannot be converted into graphs from SMILES strings. Then, we generate negative counterparts by sampling a complement set of positive drug pairs as the negative set for both datasets. We also follow the graph converting process of MIRACLE [9] for classification task.

- **ZhangDDI** [11] contains 548 drugs and 48,548 pairwise interaction data and multiple types of similarity information about these drug pairs.
- **ChChMiner** [13] contains 1,322 drugs and 48,514 labeled DDIs, obtained through drug labels and scientific publications.

Although ChChMiner dataset has much more drug instances than ZhangDDI dataset, the number of labeled DDI is almost the same. This indicates that ChChMiner dataset has much more sparse relationship between the drugs.

Graph Similarity Learning. For graph similarity learning task, we use three commonly used datasets, i.e., AIDS, IMDB [1], and OpenSSL [10].

- **AIDS** [1] contains 700 antivirus screen chemical compounds and the labels that are related to the similarity information of all pair combinations, i.e., 490K labels. The labels are Graph Edit Distance (GED) scores which are computed with A^* algorithm.
- **IMDB** [1] contains 1,500 ego-networks of movie actors/actresses, where there is an edge if the two people appear in the same movie. Labels are related to the similarity information of all pair combinations, i.e., 2.25M labels. The labels are Graph Edit Distance (GED) scores which are computed with A^* algorithm.
- **OpenSSL** [10] dataset is generated from popular open-source software OpenSSL[2], whose graphs denote the binary function's control flow graph. Labels are related to whether two binary functions are compiled from the same source code or not, since the binary functions that are compiled from the same source code are semantically similar to each other. In this work, we only consider the graphs that contain more than 50 nodes, i.e., OpenSSL [50, 200] setting in previous work [12].

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-XXXX-X/18/06...$15.00
https://doi.org/XXXXXXX.XXXXXXX

[1] https://github.com/isjakewong/MIRACLE/tree/main/MIRACLE/datachem
[2] https://www.openssl.org/

REFERENCES

[1] Yunsheng Bai, Hao Ding, Song Bian, Ting Chen, Yizhou Sun, and Wei Wang. 2019. Simgnn: A neural network approach to fast graph similarity computation. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 384–392.

[2] Laura M Grubbs, Mariam Saifullah, E Nohelli, Shulin Ye, Sai S Achi, William E Acree Jr, and Michael H Abraham. 2010. Mathematical correlations for describing solute transfer into functionalized alkane solvents containing hydroxyl, ether, ester or ketone solvents. *Fluid phase equilibria* 298, 1 (2010), 48–53.

[3] Joonyoung F Joung, Minhi Han, Minseok Jeong, and Sungnam Park. 2020. Experimental database of optical properties of organic compounds. *Scientific data* 7, 1 (2020), 1–6.

[4] Aleksandr V Marenich, Casey P Kelly, Jason D Thompson, Gregory D Hawkins, Candee C Chambers, David J Giesen, Paul Winget, Christopher J Cramer, and Donald G Truhlar. 2020. Minnesota solvation database (MNSOL) version 2012. (2020).

[5] David L Mobley and J Peter Guthrie. 2014. FreeSolv: a database of experimental and calculated hydration free energies, with input files. *Journal of computer-aided molecular design* 28, 7 (2014), 711–720.

[6] Edouard Moine, Romain Privat, Baptiste Sirjean, and Jean-Noël Jaubert. 2017. Estimation of solvation quantities from experimental thermodynamic data: Development of the comprehensive CompSol databank for pure and mixed solutes. *Journal of Physical and Chemical Reference Data* 46, 3 (2017), 033102.

[7] Yashaswi Pathak, Siddhartha Laghuvarapu, Sarvesh Mehta, and U Deva Priyakumar. 2020. Chemically interpretable graph interaction network for prediction of pharmacokinetic properties of drug-like molecules. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 873–880.

[8] Florence H Vermeire and William H Green. 2021. Transfer learning for solvation free energies: From quantum chemistry to experiments. *Chemical Engineering Journal* 418 (2021), 129307.

[9] Yingheng Wang, Yaosen Min, Xin Chen, and Ji Wu. 2021. Multi-view graph contrastive representation learning for drug-drug interaction prediction. In *Proceedings of the Web Conference 2021*. 2921–2933.

[10] Xiaojun Xu, Chang Liu, Qian Feng, Heng Yin, Le Song, and Dawn Song. 2017. Neural network-based graph embedding for cross-platform binary code similarity detection. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. 363–376.

[11] Wen Zhang, Yanlin Chen, Feng Liu, Fei Luo, Gang Tian, and Xiaohong Li. 2017. Predicting potential drug-drug interactions by integrating chemical, biological, phenotypic and network data. *BMC bioinformatics* 18, 1 (2017), 1–12.

[12] Zhen Zhang, Jiajun Bu, Martin Ester, Zhao Li, Chengwei Yao, Zhi Yu, and Can Wang. 2021. H2mn: Graph similarity learning with hierarchical hypergraph matching networks. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 2274–2284.

[13] Marinka Zitnik, Rok Sosic, and Jure Leskovec. 2018. BioSNAP Datasets: Stanford biomedical network dataset collection. *Note: http://snap. stanford. edu/biodata Cited by* 5, 1 (2018).