Compound2Drug – a machine/deep learning tool for predicting the bio-activity of PubChem compounds

A S Ben Geoffrey[a], Pavan Preetham Valluri[b], Akhil Sanker[c], Rafal Madaj[d] , Host Antony David[e], Beutline Malgija[e], Konka Dinesh[f], Suyash Pant[g], Shweta Chakrabarti[h], Sharvani Togata[i], Bharti Mittal[i], Manish Upadhyay[i], Judith Gracia[a], Adarsh VK[j], Varun TK[k]

[a]University of Madras, Chepauk, Chennai 600 005, India
[b]PSG College of Technology, Coimbatore, Tamil Nadu 641004, India
[c]SRM University, Tamil Nadu 603203, India
[d]Centre of Molecular and Macromolecular Studies, Polish Academy of Sciences, Poland
[e]Madras Christian College, East Tambaram, Chennai 600 059, India
[f]Ahmedabad University, Ahmedabad , Gujarat, 380009, India
[g]National Institute Of Pharmaceutical Education And Research, Kolkata, West Bengal, 700032, India
[h]Manipal School of Life Sciences, MAHE, Manipal, Karnataka, 576104, India
[i] Independent Researcher
[j]National Institute of Technology , Calicut, 673601, India.
[k] Kannur University, Kerala, 670 002, India
Corresponding email : bengeof@gmail.com

Abstract

Network data is composed of nodes and edges. Successful application of machine learning/deep learning algorithms on network data to make node classification and link prediction has been shown in the area of social networks through which highly customized suggestions are offered to social network users. Similarly one can attempt the use of machine learning/deep learning algorithms on biological network data to generate predictions of scientific usefulness. In the present work, compound-drug target interaction data set from bindingDB has been used to train machine learning/deep learning algorithms which are used to predict the drug targets for any PubChem compound queried by the user. The user is required to input the PubChem Compound ID (CID) of the compound the user wishes to gain information about its predicted biological activity and the tool outputs the RCSB PDB IDs of the predicted drug target. The tool also incorporates a feature to perform automated *In Silico* modelling for the compounds and the predicted drug targets to uncover their protein-ligand interaction profiles. The programs fetches the structures of the compound and the predicted drug targets, prepares them for molecular docking using standard AutoDock Scripts that are part of MGLtools and performs molecular docking, protein-ligand interaction profiling of the targets and the compound and stores the visualized results in the working folder of the user. The program is hosted, supported and maintained at the following GitHub repository https://github.com/bengeof/Compound2Drug

Introduction

A network data is composed of nodes and edges[1]. An example of such network data would be social network data where nodes are people and their interests and edges are interconnections between them[2-5]. Many useful applications such as customized suggestions for social media users have been developed through the use of Machine/Deep learning algorithms which accomplish this through node classification and link prediction protocols[5-10]. Similar techniques are transferable to gain insights and predictions from biological network data. Biological network data include, protein-protein interaction networks, differential gene expression and regulatory networks, metabolic pathways and cell signalling networks, etc [11,12]. Using these techniques Vazquez, Alexei, et al have developed a tool for protein function prediction from protein-protein interaction networks [13]. Similarly Hashemifar, Somaye, et al and other groups have developed a tool for predicting protein-protein interaction using deep learning algorithms [14,15]. From gene expression network data different groups have developed tools that use deep learning algorithms to classify cancer types [16-18]. Similarly advances in understanding differential gene expression from gene expression networks have also been carried out using Deep Learning techniques by different groups [19,20]. The previous works of our research group has involved incorporating machine/deep learning techniques for automation in screening PubChem compound library and identifying the best small drug molecules for a particular drug target [21-23]. In keeping with our research focus, the present work presents a complimentary approach to drug screening, wherein, given a particular PubChem compound ID for a particular compound, the developed tool predicts the most likely pharmaceutical activity of the compound and followingly performs an automated *In Silico* modelling to uncover the molecular details of its pharmaceutical activity. To accomplish the task mentioned above we have used different Machine and Deep Learning algorithm which predict the bio-activity of a given PubChem compound from the their prior training knowledge on a training dataset on protein-compound interaction network data downloaded from BindingDB [24,25]. To automate the discovery of the molecular basis of the predicted pharmaceutical activity of the compound, an automated *In Silico* modelling was carried out against the predicted drug targets. This has been carried out by programmatic access of AutoDock Vina and MGLtools from the main program [26].

Methods

The bindingDB database [27] was downloaded and a network was constructed using NetworkX [28] wherein the nodes where compounds and proteins and edges where the interactions between them. Lower the IC_{50} value for a compound to inhibit a particular protein, the shorter the edges were that link them together. Each compound is identified using the PubChem Compound ID (CID) and proteins are identified with the Protein Data Bank ID (PDB ID). The dataset visualized using NetworkX and select visualization is shown in Fig.1. The Dataset consists of 536435 unique CIDs and 2707 unique PDB IDs. To generate 2D embeddings of the network, the node2vec [29] python package was used. The module learnt the embeddings of 65 graphs and they were used to perform a machine learning/deep learning based multi-class classification [30-35]. To address the problem of multi-class classification for graphs with large data a fully connected deep neural network was constructed, which consisted of an input layer, three hidden layers which were activated by a RELU activation function and a output layer which uses a sigmoid activation function to perform the multi-label classification. The categorical labels were vectorized using OneHotEncoder method. Given an input node which is a PubChem compound ID (CID), the program generates a sub-network of structurally related CIDs to the input CID and performs a multi-class classification using the Deep Neural Network the classify CID into the PDB ID class it belongs to or to say it otherwise, predict the PDB ID of the protein the compound with a given input CID is likely to interact with. Dropouts were used as regularization technique to overcome over-fitting and the neural network performed prediction with a accuracy of over 80%. The multi-label classification for smaller graphs were handled with a machine learning based approach with the logistic regression algorithm. Therefore this Machine Learning/Deep Learning based programmatic tool is useful to predict the bio-activity of a PubChem compound. The workflow of the program is shown in Fig.2.

The program is required to be run in python3 environment with following dependencies, code files and models kept in the working folder of the user which are downloadable from the links given below.

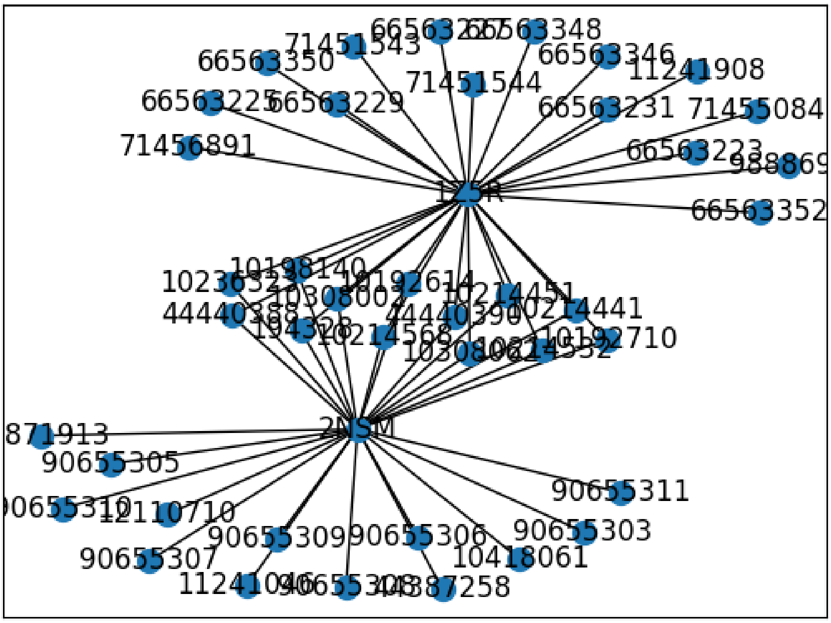


Fig.1 NetworkX visualization of compound-drug target interaction network

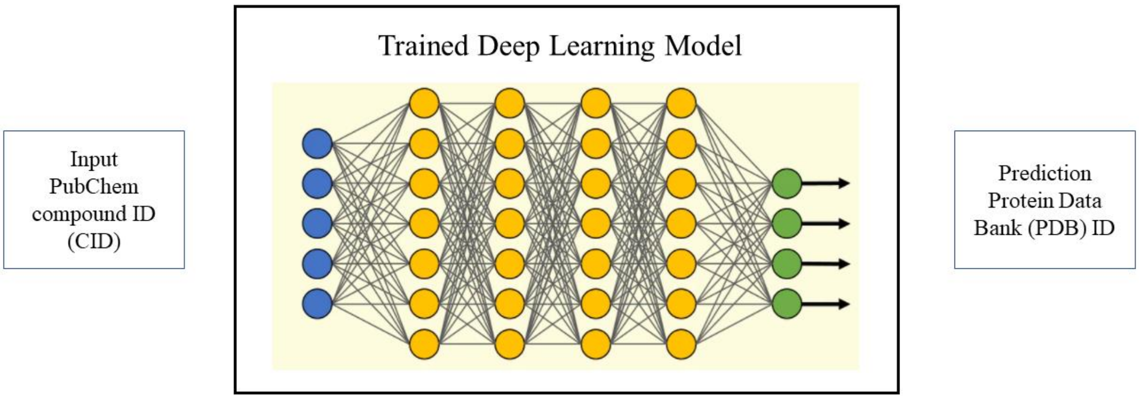


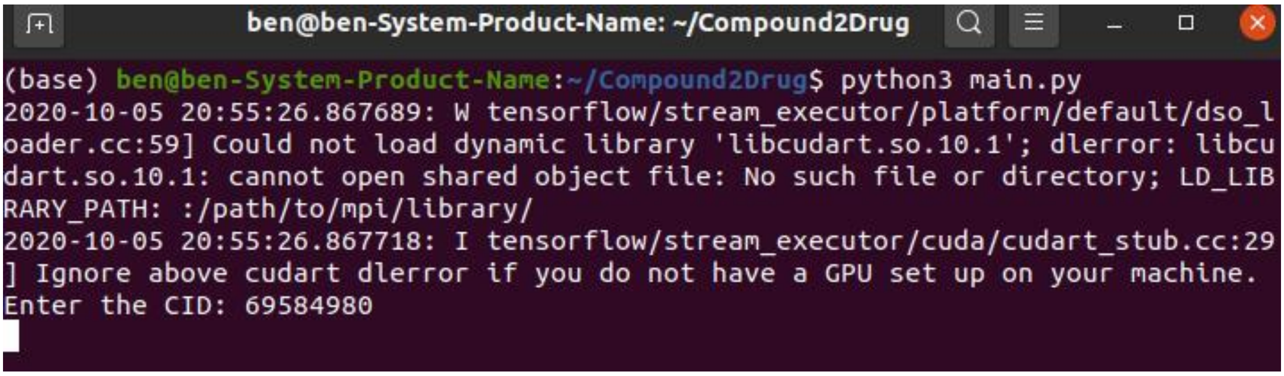
Fig. 2 Overall algorithmic workflow

Dependencies

gensim==3.8.3
gunicorn==20.0.4
Keras-Preprocessing==1.1.2
kfac==0.2.0
matplotlib==3.3.0
networkx==2.4
node2vec==0.3.2
nodevectors==0.1.22
numpy==1.19.1
pandas==1.1.1
scikit-learn==0.23.2
scipy==1.5.2
seaborn==0.10.1
mgltools==1.5.6

tensor2tensor==1.15.7
tensorboard==2.3.0
tensorboard-plugin-wit==1.7.0
tensorflow==2.3.0
tensorflow-addons==0.10.0
tensorflow-datasets==3.2.1
tensorflow-estimator==2.3.0
tensorflow-gan==2.0.0
tensorflow-hub==0.8.0
tensorflow-metadata==0.22.2
tensorflow-probability==0.7.0
tensorflow-text==2.3.0
xgboost==1.1.1
autoDock vina==4.2.6

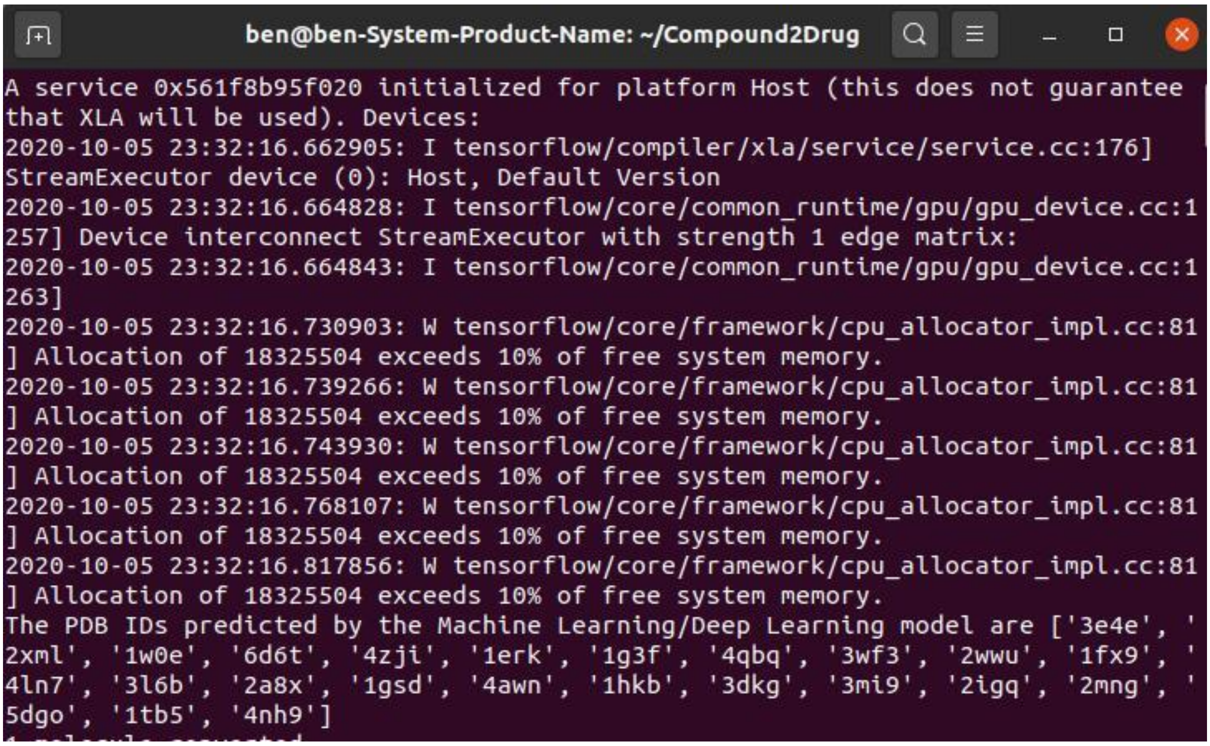
The command line user interface of the tool is shown below and the usefulness of the tool is demonstrated by performing a few select examples using a randomly selected CID input. When the user runs the main program he is prompted to enter the CID of the compound for which he requires prediction of drug targets.



```
(base) ben@ben-System-Product-Name:~/Compound2Drug$ python3 main.py
2020-10-05 20:55:26.867689: W tensorflow/stream_executor/platform/default/dso_l
oader.cc:59] Could not load dynamic library 'libcudart.so.10.1'; dlerror: libcu
dart.so.10.1: cannot open shared object file: No such file or directory; LD_LIB
RARY_PATH: :/path/to/mpi/library/
2020-10-05 20:55:26.867718: I tensorflow/stream_executor/cuda/cudart_stub.cc:29
] Ignore above cudart dlerror if you do not have a GPU set up on your machine.
Enter the CID: 69584980
```

Fig.3a Tool Interface

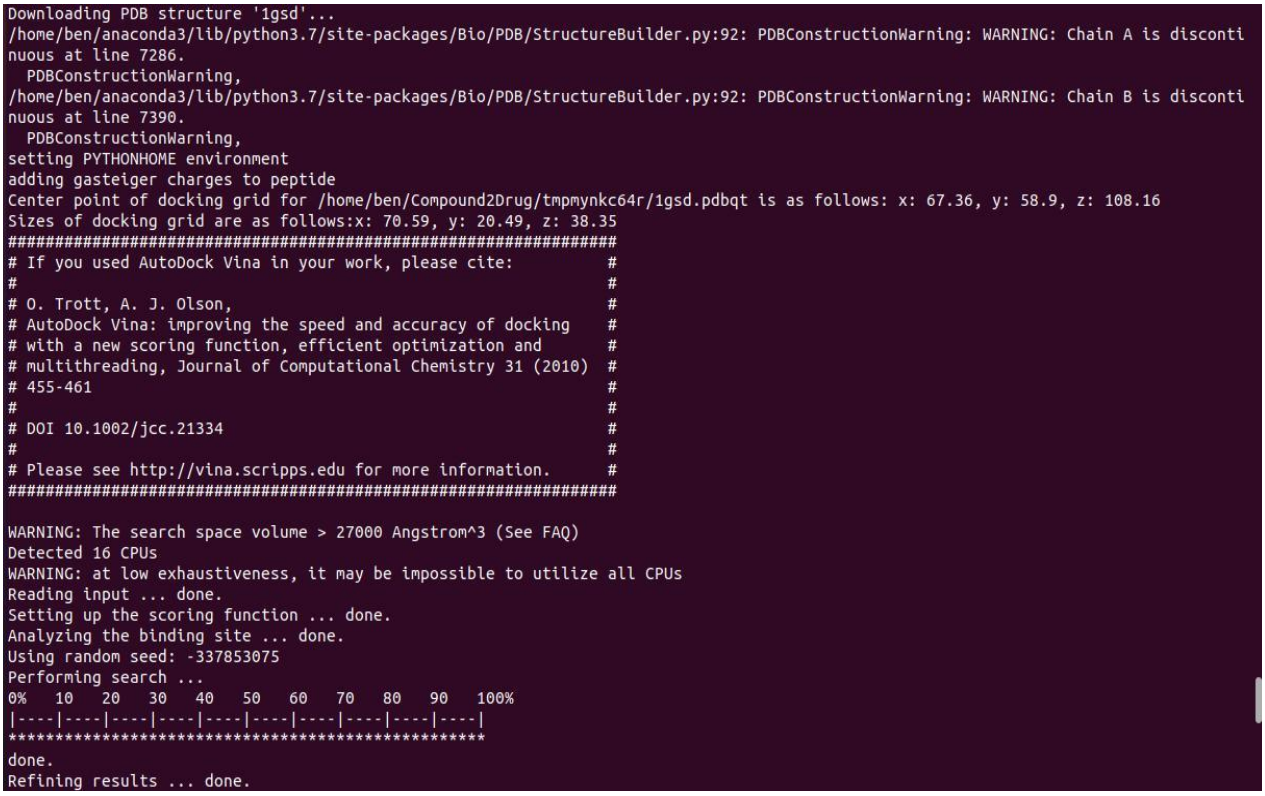
Following this, the tool carries out the prediction task and prints out the predicted target PDB IDs as follows



```
A service 0x561f8b95f020 initialized for platform Host (this does not guarantee
that XLA will be used). Devices:
2020-10-05 23:32:16.662905: I tensorflow/compiler/xla/service/service.cc:176]
StreamExecutor device (0): Host, Default Version
2020-10-05 23:32:16.664828: I tensorflow/core/common_runtime/gpu/gpu_device.cc:1
257] Device interconnect StreamExecutor with strength 1 edge matrix:
2020-10-05 23:32:16.664843: I tensorflow/core/common_runtime/gpu/gpu_device.cc:1
263]
2020-10-05 23:32:16.730903: W tensorflow/core/framework/cpu_allocator_impl.cc:81
] Allocation of 18325504 exceeds 10% of free system memory.
2020-10-05 23:32:16.739266: W tensorflow/core/framework/cpu_allocator_impl.cc:81
] Allocation of 18325504 exceeds 10% of free system memory.
2020-10-05 23:32:16.743930: W tensorflow/core/framework/cpu_allocator_impl.cc:81
] Allocation of 18325504 exceeds 10% of free system memory.
2020-10-05 23:32:16.768107: W tensorflow/core/framework/cpu_allocator_impl.cc:81
] Allocation of 18325504 exceeds 10% of free system memory.
2020-10-05 23:32:16.817856: W tensorflow/core/framework/cpu_allocator_impl.cc:81
] Allocation of 18325504 exceeds 10% of free system memory.
The PDB IDs predicted by the Machine Learning/Deep Learning model are ['3e4e', '
2xml', '1w0e', '6d6t', '4zji', '1erk', '1g3f', '4qbq', '3wf3', '2wwu', '1fx9', '
4ln7', '3l6b', '2a8x', '1gsd', '4awn', '1hkb', '3dkg', '3mi9', '2igq', '2mng', '
5dgo', '1tb5', '4nh9']
```

Fig.3b – Drug target prediction by the tool

For each given input CID, the program also performs automated *In Silico* modelling and stores the visualized results of protein-ligand interaction in the working folder of the user. The structures of the ligand(compound) and the protein are automatically downloaded from PubChem and RCSB Protein Data Bank and they are prepared for molecular docking using the standard AutoDock scripts available through MGLTools. The program uses Web API to perform PLIP protein-ligand interaction profile and stores the results of the protein-ligand interaction profile in the working folder of the user.



```
Downloading PDB structure '1gsd'...
/home/ben/anaconda3/lib/python3.7/site-packages/Bio/PDB/StructureBuilder.py:92: PDBConstructionWarning: WARNING: Chain A is disconti
nuous at line 7286.
  PDBConstructionWarning,
/home/ben/anaconda3/lib/python3.7/site-packages/Bio/PDB/StructureBuilder.py:92: PDBConstructionWarning: WARNING: Chain B is disconti
nuous at line 7390.
  PDBConstructionWarning,
setting PYTHONHOME environment
adding gasteiger charges to peptide
Center point of docking grid for /home/ben/Compound2Drug/tmpmynkc64r/1gsd.pdbqt is as follows: x: 67.36, y: 58.9, z: 108.16
Sizes of docking grid are as follows:x: 70.59, y: 20.49, z: 38.35
#################################################################
# If you used AutoDock Vina in your work, please cite:          #
#                                                               #
# O. Trott, A. J. Olson,                                        #
# AutoDock Vina: improving the speed and accuracy of docking    #
# with a new scoring function, efficient optimization and       #
# multithreading, Journal of Computational Chemistry 31 (2010)  #
# 455-461                                                       #
#                                                               #
# DOI 10.1002/jcc.21334                                         #
#                                                               #
# Please see http://vina.scripps.edu for more information.      #
#################################################################

WARNING: The search space volume > 27000 Angstrom^3 (See FAQ)
Detected 16 CPUs
WARNING: at low exhaustiveness, it may be impossible to utilize all CPUs
Reading input ... done.
Setting up the scoring function ... done.
Analyzing the binding site ... done.
Using random seed: -337853075
Performing search ...
0%   10   20   30   40   50   60   70   80   90   100%
|----|----|----|----|----|----|----|----|----|----|
***************************************************
done.
Refining results ... done.
```

Fig.4 – Automated *In Silico* modelling and protein-ligand interaction profiling

The tool is required to be run with the following files as shown in the working folder. They are downloadable from the links given below.

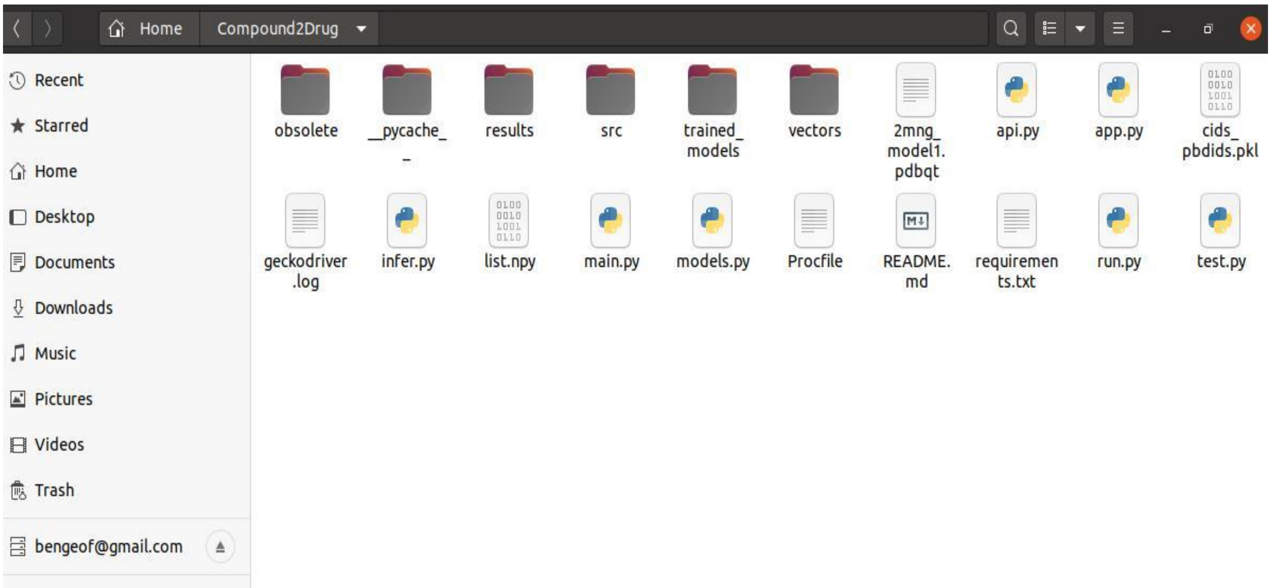


Fig.5 – Working folder

The trained models, vectors, pickle file can be downloaded from the drive link given below
https://drive.google.com/drive/folders/1wwgrS6EWCnUFnPRohDFmzzShjZDb0GFe?usp=sharing
https://drive.google.com/drive/folders/1JOpIdckxhCVz1A5R67YzXPxBWOlkFLJs?usp=sharing
https://drive.google.com/file/d/1ENt5pb7liNctR_8CE54g35hBU1WQ1TPx/view?usp=sharing

The code is downloadable from the GitHub repository link given below
https://github.com/bengeof/Compound2Drug

Results and Discussion

To demonstrate the use of the tool with a randomly selected user input, the tool was run as described in the methodology section with a randomly chosen PubChem CID : 69584980. The tool generated a list of predicted targets and automatically estimated the strength of interaction of the compound with the predicted targets and the results are given below in Table 1. The strongest interaction was found to be with the target identified with PDB ID : 1gsd which is identified to be the enzyme Glutathione Transferase. Glutathione Transferase inhibitors increase the sensitivity of cancer cells to anti-cancer drugs and also possess several other therapeutic applications [36]. The protein-ligand interaction profile generated by the tool is shown in Fig. 6 below

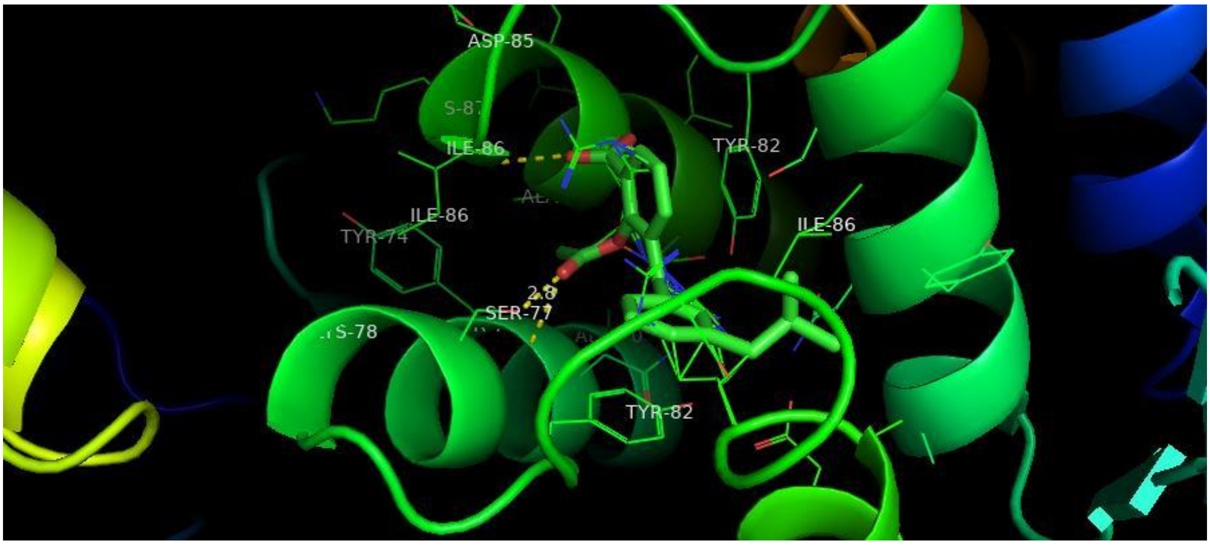


Fig. 6 – Protein-ligand interaction

Table 1 – Results of protein-ligand interaction prediction and modelling by the tool

Compound Information	Target Information	Interaction Strength
PubChem CID	RCSB PDB ID	Binding Affinity (Kcal/mol)
69584980	3e4e	-7.6
	2xml	-9.2
	1w0e	-9.1
	6d6t	-7
	4zji	-7.1
	1erk	-7
	1g3f	-6.4
	4qbq	-7
	3wf3	-8.6
	2wwu	-7.5
	1fx9	-8.4
	4ln7	-7.4
	3l6b	-7
	2a8x	-9.1
	1gsd	**-9.9**
	4awn	-6.7
	1hkb	-7.3
	3dkg	-7.1
	3mi9	-7.9
	2igq	-7.9
	5dgo	-7.3
	1tb5	-9.3
	4nh9	-8.6

Conclusion

In the present work, the compound-drug target interaction data set from bindingDB has been used to train machine learning/deep learning algorithms which were used to predict the drug targets for any PubChem compound. The user is required to input the PubChem Compound ID (CID) of the compound the user wishes to gain information about its predicted biological activity and the tool outputs the RCSB PDB IDs of the predicted drug targets for the compound. The tool also incorporates a feature to perform automated *In Silico* modelling for the compounds and the predicted drug targets to uncover their protein-ligand interaction profiles. To demonstrate the use of the tool a randomly selected PubChem Compound ID (CID) was given as input to the program and the use of the tool in identifying the bio-activity of the compound was demonstrated.

References

1. Trinajstic, Nenad. *Chemical graph theory*. Routledge, 2018.
2. Lee, Yofay Kari, Peter Xiu Deng, and Luke Andrew DeLorme. "Automatically generating nodes and edges in an integrated social graph." U.S. Patent No. 8,572,129. 29 Oct. 2013.
3. Al Hasan, Mohammad, and Mohammed J. Zaki. "A survey of link prediction in social networks." *Social network data analytics*. Springer, Boston, MA, 2011. 243-275.
4. da Silva Soares, Paulo Ricardo, and Ricardo Bastos Cavalcante Prudêncio. "Time series based link prediction." *The 2012 international joint conference on neural networks (IJCNN)*. IEEE, 2012.
5. Zheleva, Elena, et al. "Using friendship ties and family circles for link prediction." *International Workshop on Social Network Mining and Analysis*. Springer, Berlin, Heidelberg, 2008.
6. Aggarwal, Charu C. "An introduction to social network data analytics." *Social network data analytics*. Springer, Boston, MA, 2011. 1-15.
7. Murata, Tsuyoshi, and Sakiko Moriyasu. "Link prediction based on structural properties of online social networks." *New Generation Computing* 26.3 (2008): 245-257.
8. Aggarwal, Charu C. "An introduction to social network data analytics." *Social network data analytics*. Springer, Boston, MA, 2011. 1-15.
9. Jalili, Mahdi, et al. "Link prediction in multiplex online social networks." *Royal Society open science* 4.2 (2017): 160863.
10. Tang, Jiliang, Charu Aggarwal, and Huan Liu. "Node classification in signed social networks." *Proceedings of the 2016 SIAM international conference on data mining*. Society for Industrial and Applied Mathematics, 2016.
11. Almansoori, Wadhah, et al. "Link prediction and classification in social networks and its application in healthcare and systems biology." *Network Modeling Analysis in Health Informatics and Bioinformatics* 1.1-2 (2012): 27-36.
12. Pavlopoulos, Georgios A., et al. "Using graph theory to analyze biological networks." *BioData mining* 4.1 (2011): 10.
13. Vazquez, Alexei, et al. "Global protein function prediction from protein-protein interaction networks." *Nature biotechnology* 21.6 (2003): 697-700.
14. Hashemifar, Somaye, et al. "Predicting protein–protein interactions through sequence-based deep learning." *Bioinformatics* 34.17 (2018): i802-i810.

15. Zhang, Buzhong, et al. "Sequence-based prediction of protein-protein interaction sites by simplified long short-term memory network." *Neurocomputing* 357 (2019): 86-100.
16. Lyu, Boyu, and Anamul Haque. "Deep learning based tumor type classification using gene expression data." *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*. 2018.
17. Fakoor, Rasool, et al. "Using deep learning to enhance cancer diagnosis and classification." *Proceedings of the international conference on machine learning*. Vol. 28. New York, USA: ACM, 2013.
18. Xiao, Yawen, et al. "A deep learning-based multi-model ensemble method for cancer prediction." *Computer methods and programs in biomedicine* 153 (2018): 1-9.
19. Tasaki, Shinya, et al. "Deep learning decodes the principles of differential gene expression." *Nature Machine Intelligence* (2020): 1-11.
20. Sekhon, Arshdeep, Ritambhara Singh, and Yanjun Qi. "DeepDiff: DEEP-learning for predicting DIFFerential gene expression from histone modifications." *Bioinformatics* 34.17 (2018): i891-i900.
21. Geoffrey A S, Ben; Madaj, Rafal; Sanker, Akhil; Tresanco, Mario Sergio Valdés; Davidd, Host Antony; Roy, Gitanjali (2020): A programmatic tool for automatic ease in coronavirus drug discovery through programmatically automated data mining, QSAR and In Silico modelling. ChemRxiv. Preprint.
22. AS, B. G., Sanker, A., Madaj, R., Tresanco, M. S., Upadhyay, M., & Gracia, J. (2020). A program to automate the discovery of drugs for West Nile and Dengue virus—programmatic screening of over a billion compounds on PubChem, generation of drug leads and automated In Silico modelling. BioRxiv.
23. Madaj, R., AS, B. G., Sanker, A., David, H. A., Faletif, A. I., Gracia, J., ... & Verma, S. (2020). Automated identification of small drug molecules for Hepatitis C virus through a novel programmatic tool and extensive Molecular Dynamics studies of select drug candidates. *bioRxiv*.
24. Liu, Tiqing, et al. "BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities." *Nucleic acids research* 35.suppl_1 (2007): D198-D201.
25. Gilson, Michael K., et al. "BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology." *Nucleic acids research* 44.D1 (2016): D1045-D1053.

26. Jaghoori, Mohammad Mahdi, Boris Bleijlevens, and Silvia D. Olabarriaga. "1001 ways to run AutoDock Vina for virtual screening." *Journal of computer-aided molecular design* 30.3 (2016): 237-249.
27. Gilson, Michael K., et al. "BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology." *Nucleic acids research* 44.D1 (2016): D1045-D1053.
28. Hagberg, Aric, Pieter Swart, and Daniel S Chult. *Exploring network structure, dynamics, and function using NetworkX*. No. LA-UR-08-05495; LA-UR-08-5495. Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.
29. Grover, Aditya, and Jure Leskovec. "node2vec: Scalable feature learning for networks." *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. 2016.
30. Farooq, Ammarah, et al. "A deep CNN based multi-class classification of Alzheimer's disease using MRI." *2017 IEEE International Conference on Imaging systems and techniques (IST)*. IEEE, 2017.
31. Maxwell, Andrew, et al. "Deep learning architectures for multi-label classification of intelligent health risk prediction." *BMC bioinformatics* 18.14 (2017): 523.
32. Wang, Sheng, Ashwin Raju, and Junzhou Huang. "Deep learning based multi-label classification for surgical tool presence detection in laparoscopic videos." *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*. IEEE, 2017.
33. Murthy, Venkatesh N., et al. "Deep decision network for multi-class image classification." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
34. Szymański, Piotr, and Tomasz Kajdanowicz. "A scikit-based Python environment for performing multi-label classification." *arXiv preprint arXiv:1702.01460* (2017).
35. Xu, Youjun, Jianfeng Pei, and Luhua Lai. "Deep learning based regression and multiclass models for acute oral toxicity prediction with automatic chemical feature extraction." *Journal of chemical information and modeling* 57.11 (2017): 2672-2685.
36. Allocati, Nerino, et al. "Glutathione transferases: substrates, inihibitors and pro-drugs in cancer and neurodegenerative diseases." *Oncogenesis* 7.1 (2018): 1-15.