CandidateDrug4Cancer: An Open Molecular Graph Learning Benchmark on Drug Discovery for Cancer

Xianbin Ye*
Ziliang Li*
Ping An Healthcare Technology
Beijing, China

Fei Ma
Zongbi Yi
Chinese Academy of Medical Sciences
National Cancer Center
Beijing, China

Jun Wang†
Ping An Healthcare Technology
Beijing, China
junwang.deeplearning@gmail.com

Pengyong Li
Tsinghua University
Department of Biomedical
Engineering, Beijing, China

Peng Gao
Ping An Healthcare Technology
Beijing, China
gaopeng712@pingan.com.cn

Guotong Xie
Ping An Healthcare Technology
Beijing, China
xieguotong@pingan.com.cn

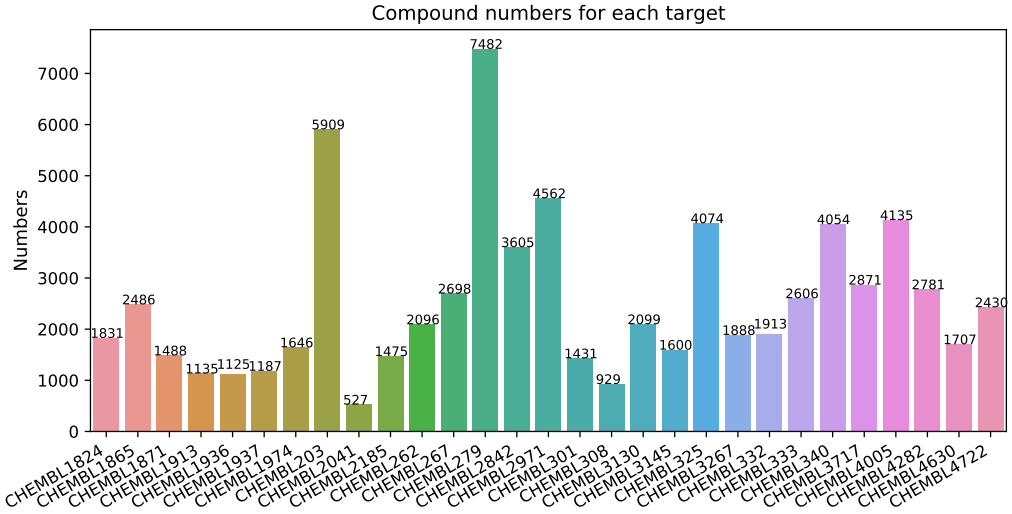

Figure 1: The CandidateDrug4Cancer benchmark encompasses the most-mentioned 29 targets for cancer, covering 54869 cancer-related drug molecules which is ranged from pre-clinical, clinical and FDA-approved.

ABSTRACT

Anti-cancer drug discoveries have been serendipitous, we sought to present the Open Molecular Graph Learning Benchmark, named CandidateDrug4Cancer, a challenging and realistic benchmark dataset to facilitate scalable, robust, and reproducible graph machine learning research for anti-cancer drug discovery. CandidateDrug4Cancer dataset encompasses multiple most-mentioned 29 targets for cancer, covering 54869 cancer-related drug molecules which are ranged from pre-clinical, clinical and FDA-approved. Besides building the datasets, we also perform benchmark experiments with effective Drug Target Interaction (DTI) prediction baselines using descriptors and expressive graph neural networks. Experimental results suggest that CandidateDrug4Cancer presents significant challenges for learning molecular graphs and targets in practical application, indicating opportunities for future researches on developing candidate drugs for treating cancers.

KEYWORDS

molecular graph, cancer, drug target interaction, benchmark

ACM Reference Format:
Xianbin Ye, Ziliang Li, Fei Ma, Zongbi Yi, Jun Wang, Pengyong Li, Peng Gao, and Guotong Xie. 2018. CandidateDrug4Cancer: An Open Molecular Graph Learning Benchmark on Drug Discovery for Cancer. In ,. GLB, NY, USA, 5 pages.

*These authors contribute equally to this work.
†Corresponding author.

GLB 2021, The Web Conference 2021

ACM ISBN 978-1-4503-XXXX-X/21/06…$15.00

1 INTRODUCTION

Cancer is one of the leading causes of mortality worldwide. Opportunities to help reduce the death rate from cancer through the discovery of new drugs are benefiting from the increasing advances in technology and enhanced knowledge of human neoplastic disease[21, 30]. However, drug discovery is a time-consuming, labor intensive, and expensive process, far slower than expected. The entire process from discovery to the regulatory approval of a new drug can take as much as 12 years and cost estimated at US 3 billion [7]. It has been usually hampered by experimental discovery of molecules and targets, and following with validation with in vitro experiments on cell lines and animals before moving to clinical testing [12]. Furthermore, stagnant success rate (1:5000) is associated with each drug development stage.

Fortunately, many researchers have proposed various effective computer-aided drug discovery (CADD) methods [22] to decrease the costs and speed up projects [14]. As an feasible assistant technique, CADD has made a significant contribution to drug discovery and has successfully developed dozens of drugs to market in recent decades[29]. Generally, CADD can be categorized into receptor-based methods and ligand-based methods upon the availability of target proteins and molecules. Receptor-based CADD relies on the target protein structure to calculate interaction for compounds, while ligand-based CADD exploits the information of compounds with diverse structures and known activity and inactivity, for construction of predictive, quantitative structure-activity relation (QSAR)[4].

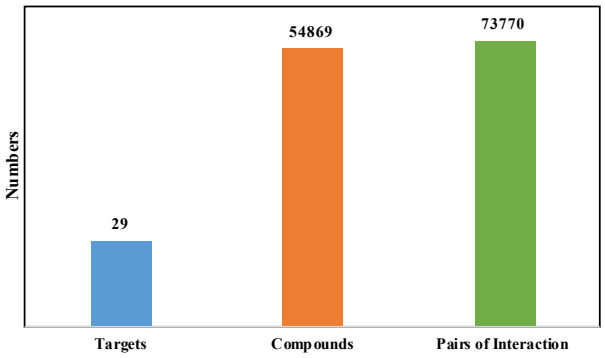

Figure 2: CandidateDrug4Cancer dataset encompasses multiple most-mentioned 29 targets for cancer, with 73770 drug-target pairs, covering a diverse range of 54869 related drug molecules which is ranged from pre-clinical, clinical and FDA-approved.

At present, the interdisciplinary studies between advanced Artificial Intelligence (AI) and drug discovery have received increasing attention due to the superior speed and performance. Many AI technologies have been successfully applied in a variety of tasks for CADD, such as receptor-based Drug Target Interaction (DTI) prediction [1, 8]. Due to the large size of the chemical space, one of the fundamental challenges for these studies is how to learn expressive representations from molecular structures [28]. In the early years, molecular representations are based on hand-crafted features such as molecular descriptors or fingerprints [27]. In contrast, there has been a surge of interests in end-to-end molecular representation learned by graph neural networks. However, challenges for deep learning in molecular representation mainly arise from the scarcity of labeled data, resulting in models that lack practicability [13, 20]. In particular, so far there is limited published information on molecular graph learning for cancer to our knowledge.

Altogether, these issues make it difficult producing favorable benchmark assembling proper information of cancer targets and potential molecules, for assessing and comparing the performance of different methods on drug discovery for cancer. To address these issues, we introduced an open molecular graph learning benchmark on drug discovery for cancer, named CandidateDrug4Cancer. Our main contributions can be summarized as follows:

- We introduce a novel graph learning task which has the potential to help understand the performance and limitations of graph representation facilitated models on candidate anti-cancer drug discovery problems.
- We provide CandidateDrug4Cancer benchmark dataset, which encompasses multiple most-mentioned 29 targets for cancer, covering 54869 cancer-related drug molecules which are ranged from pre-clinical, clinical and FDA-approved.
- We conduct benchmark evaluations for drug target interaction with baseline encoders including powerful graph neural network for drug-like molecules.

2 RELATED WORKS

Drug Target Interaction One of the initial steps of drug discovery is the identification of novel drug-like compounds that interact with the predefined target proteins. Various deep learning methods have been developed and achieved excellent performance for drug-target interaction (DTI) prediction [5, 18, 24]. Generally, the deep learning algorithms for DTI prediction comprise of a compound encoder and a protein encoder. Recently, Tsubaki et al. [23] proposed a new DTI framework by combining GNN for compounds and a CNN for proteins, which significantly outperformed existing methods.

Molecular Graph Learning Recently, among the promising deep learning architectures, graph neural network (GNN), such as message passing neural network (MPNN) [11] has gradually emerged as a powerful candidate for modeling molecular data. Because a molecule is naturally a graph that consists of atoms (nodes) connected through chemical bonds (edges), it is ideally suited for GNN. Up to now, various GNN architectures have achieved great progress in drug discovery [26]. However, there are some limits that need to be addressed. Challenges for deep learning in molecular representation mainly arise from the scarcity of labeled data, as lab experiments are expensive and time-consuming. Thus, training datasets in drug discovery are usually limited in size, and GNNs tend to overfit them, resulting in learned representations that lack practicability [13, 20].

Anti-Cancer Candidate Drug Datasets Unfortunately, the current number of drugs (FDA approved or at the experimental stage) is only around 10,000, the current knowledge about the drug–target space is limited, especially for severe life-threatening cancers, and novel approaches are required to widen our knowledge[19]. However, most cancer drug discoveries have been serendipitous, the extent to which non-cancer drugs have potential as future cancer

therapeutics is unknown[6]. Recent efforts have demonstrated the power of cancer cell line screening—testing either many compounds across a limited number of cell lines (for example, NCI-60 [2], Genomics of Drug Sensitivity in Cancer (GDSC) [9], the Cancer Target Discovery and Development (CTD2)[3] and DepMap[6]. The ideal study would involve screening larger amount of drug candidates (most of which are non-oncology drugs) across most target proteins to capture the molecular diversity of human cancer.

Table 1: The most frequent target proteins of cancers collected from DepMap[6], DrugBank[25] and ChEMBL[10].

Target_id	Target_name
chembl1824	Receptor protein-tyrosine kinase erbB-2
chembl1865	Histone deacetylase 6
chembl1871	Androgen Receptor
chembl1913	Platelet-derived growth factor receptor beta
chembl1936	Stem cell growth factor receptor
chembl1937	Histone deacetylase 2
chembl1974	Tyrosine-protein kinase receptor FLT3
chembl203	Epidermal growth factor receptor erbB1
chembl2041	Tyrosine-protein kinase receptor RET
chembl2185	Serine/threonine-protein kinase Aurora-B
chembl262	Glycogen synthase kinase-3 beta
chembl267	Glycogen synthase kinase-3 beta
chembl279	Vascular endothelial growth factor receptor 2
chembl2842	Serine/threonine-protein kinase mTOR
chembl2971	Tyrosine-protein kinase JAK2
chembl301	Cyclin-dependent kinase 2
chembl308	Cyclin-dependent kinase 1
chembl3130	PI3-kinase p110-delta subunit
chembl3145	PI3-kinase p110-beta subunit
chembl325	Histone deacetylase 1
chembl3267	PI3-kinase p110-gamma subunit
chembl332	Matrix metalloproteinase-1
chembl333	Matrix metalloproteinase-2
chembl340	Cytochrome P450 3A4
chembl3717	Hepatocyte growth factor receptor
chembl4005	PI3-kinase p110-alpha subunit
chembl4282	Serine/threonine-protein kinase AKT
chembl4630	Serine/threonine-protein kinase Chk1
chembl4722	Serine/threonine-protein kinase Aurora-A

3 CANDIDATEDRUG4CANCER DATASETS

Inspired by DepMap [6] and DrugBank [25], we sought to create a public resource containing the candidate compounds for the most-used targets for cancers, as shown in Table 1. Firstly, the most commonly researched and cancer-related targets are collected based on DepMap and DrugBank, subsequently, the corresponding compounds and drug-target pairs are collected to establish CandidateDrug4Cancer from the ChEMBL database [10] (https://www.ebi.ac.uk/chembl/g/, ChEMBL is a manually curated database of bioactive molecules with drug-like properties. It brings together chemical, bioactivity and genomic data to aid the translation of genomic information into effective new drugs). In addition to the published IC50 measurement, ChEMBL have added an additional field called pChEMBL to the activities table. This value represent comparable measures of concentrations to reach half-maximal response transformed to a negative logarithmic scale. pChEMBL is defined as:

$$pChEMBL = -\lg(molarIC50(M)) \quad (1)$$

For example, an IC50 measurement of 10 uM would have a pChEMBL value of 5. However, in order to efficiently verify anticancer inhibitors and reduce the cost of a large number of subsequent experiments, we set decision boundary at pChEMBL is 7(approximately 100 nM) more strictly, defining pChEMBL larger than or equal to 7 as positive interaction [16] (see Figure 3).

Table 2: The features used in molecular graph. These features are obtained by RDKit [15].

Type	Name	Description
Node feature	Atom type	Atomic number (0-122)
	Formal charge	[unk,-5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5]
	Chirality type	[unk, unspecifie, tetrahedral-CW, tetrahedralL-CCW, other]
	Hybridization	[unk, sp, sp2, sp3, sp3d, sp3d2, unspecified]
	NumH	Number of connected hydrogens[unk,0, 1, 2, 3, 4, 5, 6, 7, 8]
	Implicit valence	[unk, 0, 1, 2, 3, 4, 5, 6]
	Degree	Number of covalent bonds [unk, 0, 1, 2, 3, 4, 5,6,7,8,9,10]
	Aromatic	Whether the atom is part of an aromatic system [0,1]
Edge feature	Bond direction	[None,endupright, enddownright]
	Bond type	[Single, double, triple, aromatic]
	Conjugation	Whether the bond is conjugated [0,1]
	Ring	Whether the bond is in Ring [0,1]
	Stereo	[StereoNone, StereoAny, StereoZ, StereoE]

4 EXPERIMENTAL SETUP

Baseline models. Various deep learning methods have been developed and achieved promising performance for DTI prediction [5, 18, 23, 24].Generally, the common DTI prediction models comprise of a compound encoder and a protein encoder. We have compared baselines using different combinations of compound and protein encoders including: MPNN-CNN, Daylight-AAC, Morgan-AAC and MolGNet-CNN (For compounds: Morgan fingerprints, Daylight-type fingerprints, MPNN on molecular graph. For proteins: Amino Acid Composition up to 3-mers, Convolutional Neural Network (CNN) on target sequence). In particular, we further adopt our pre-trained GNN named MolGNet [17], to evaluate the effectiveness of pre-trained GNN on DTI prediction for caner targets.

Implementation details. We conduct leave-one-target-out evaluation, the drug-target interaction pairs of the rest 28 targets are used for training, and the performances are calculated on each unseen target proteins. So as to evaluate whether models are capable of prediction on unseen targets. Due to hardware limitations and multiple baseline comparisons, we evaluate the baselines on 20% of total datasets (the sub-datasets are also provided in our link). Each model is trained for 50 epochs. For graph learning, all drug molecules are pre-processed into hydrogen-depleted molecular graphs with nodes features, edge features, and adjacency matrix with RDKit [15]. The detailed information about nodes features and edge features can be referred to Table 2.

5 RESULTS AND CONCLUSIONS

As shown in Figure 4, the average AUC scores of MPNN-CNN, Daylight-AAC, Morgan-AAC and MolGNet-CNN are 0.57, 0.62, 0.63,

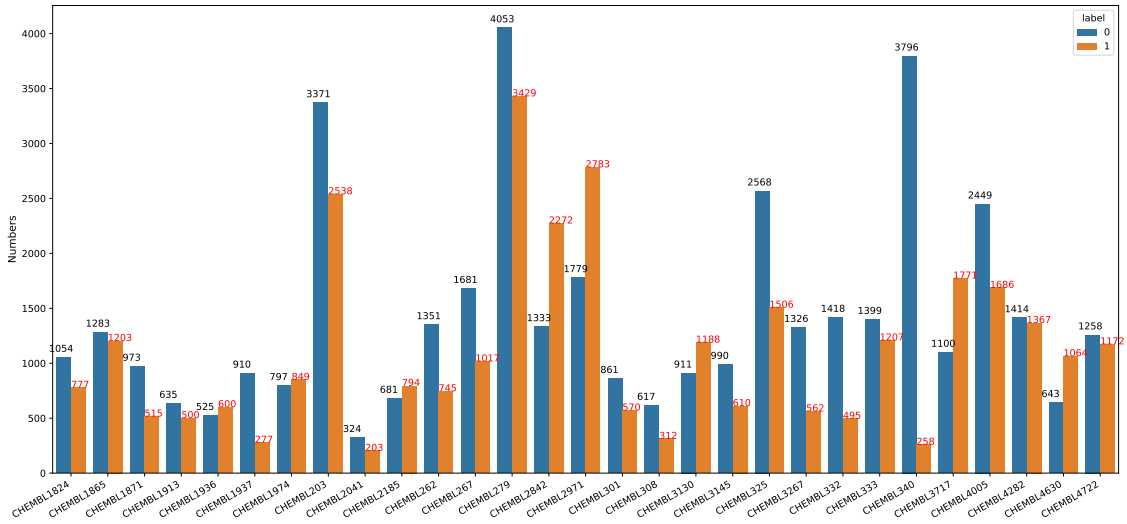

Figure 3: Label distribution of drug-target interaction pairs (boundary with pChEMBL at 7.0) across most-mentioned 29 targets for cancer in CandidateDrug4Cancer.

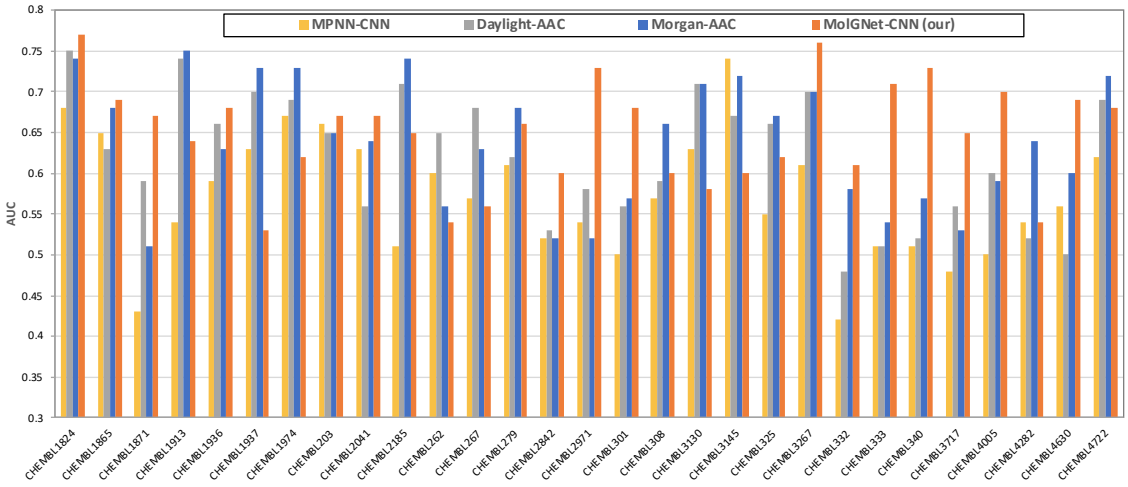

Figure 4: Baseline Comparison of the leave-one-out validation with DTI prediction AUC scores across 29 targets for cancer.

0.65, respectively. The leave-one-out validation demonstrated that, the DTI model with expressive MolGNet encoder yielded superior performances on drug target interaction prediction , compared with other baselines. It re-confirmed graph learning as fundamental and powerful tools for modeling molecules.

In summary, the CandidateDrug4Cancer is a starting point to develop new oncology therapeutics, and more rarely, for potential de novo drug design with powerful graph learning. Despite the successes, there is still future improvement directions in the following aspects: **Larger models and larger datasets**. It is interesting to employ even larger models and larger datasets for drug discovery on cancer. Larger models can potentially better handle more complicated learning. **More suitable protein encoder**. It is desirable to explore more powerful protein encoders, so as to handle more complicated drug-target interaction. **Explainability of graph learning**. It is desirable to explore what useful insights or representations were learned in the graph embedding. Moreover, noting that DTI for cancer is still in the early stages but seen rapid growth/interest. Further analysis both theoretically and empirically is desired to better understand when/why/how graph learning can better work in drug discovery.

6 ONLINE RESOURCES

The CandidateDrug4Cancer datasets and details are available at https://drive.google.com/file/d/1gXpGc5UhAYB9zVYnSl6E2MEaWfl5w98O/view?usp=sharing.

REFERENCES

[1] Karim Abbasi, Parvin Razzaghi, Antti Poso, Saber Ghanbari-Ara, and Ali Masoudi-Nejad. 2020. Deep Learning in Drug Target Interaction Prediction: Current and Future Perspective. *Current Medicinal Chemistry* (2020).

[2] Michael C Alley, Dominic A Scudiero, Anne Monks, Miriam L Hursey, Maciej J Czerwinski, Donald L Fine, Betty J Abbott, Joseph G Mayo, Robert H Shoemaker, and Michael R Boyd. 1988. Feasibility of drug screening with panels of human tumor cell lines using a microculture tetrazolium assay. *Cancer research* 48, 3 (1988), 589–601.

[3] Amrita Basu, Nicole E Bodycombe, Jaime H Cheah, Edmund V Price, Ke Liu, Giannina I Schaefer, Richard Y Ebright, Michelle L Stewart, Daisuke Ito, Stephanie Wang, et al. 2013. An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. *Cell* 154, 5 (2013), 1151–1161.

[4] Adrià Cereto-Massagué, María José Ojeda, Cristina Valls, Miquel Mulero, Santiago Garcia-Vallvé, and Gerard Pujadas. 2015. Molecular fingerprint similarity search in virtual screening. *Methods* 71 (2015), 58–63.

[5] Ruolan Chen, Xiangrong Liu, Shuting Jin, Jiawei Lin, and Juan Liu. 2018. Machine learning for drug-target interaction prediction. *Molecules* 23, 9 (2018), 2208.

[6] Steven M Corsello, Rohith T Nagari, Ryan D Spangler, Jordan Rossen, Mustafa Kocak, Jordan G Bryan, Ranad Humeidi, David Peck, Xiaoyun Wu, Andrew A Tang, et al. 2020. Discovering the anticancer potential of non-oncology drugs by systematic viability profiling. *Nature cancer* 1, 2 (2020), 235–248.

[7] Joseph A DiMasi, Henry G Grabowski, and Ronald W Hansen. 2016. Innovation in the pharmaceutical industry: new estimates of R&D costs. *Journal of health economics* 47 (2016), 20–33.

[8] Sofia D'Souza, KV Prema, and Seetharaman Balaji. 2020. Machine learning models for drug–target interactions: current knowledge and future directions. *Drug Discovery Today* 25, 4 (2020), 748–756.

[9] Mathew J Garnett, Elena J Edelman, Sonja J Heidorn, Chris D Greenman, Anahita Dastur, King Wai Lau, Patricia Greninger, I Richard Thompson, Xi Luo, Jorge Soares, et al. 2012. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* 483, 7391 (2012), 570–575.

[10] Anna Gaulton, Louisa J Bellis, A Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, et al. 2012. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic acids research* 40, D1 (2012), D1100–D1107.

[11] Justin Gilmer, Samuel S Schoenholz, Patrick Riley, Oriol Vinyals, and George E Dahl. 2017. Neural Message Passing for Quantum Chemistry. *international conference on machine learning* (2017), 1263–1272.

[12] Raymond G Hill. 2012. *Drug discovery and development-E-book: technology in transition.* Elsevier Health Sciences.

[13] Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. 2019. Strategies for Pre-training Graph Neural Networks. In *International Conference on Learning Representations.*

[14] IM Kapetanovic. 2008. Computer-aided drug discovery and development (CADDD): in silico-chemico-biological approach. *Chemico-biological interactions* 171, 2 (2008), 165–176.

[15] Greg Landrum. 2006. *RDKit: Open-source cheminformatics.* http://www.rdkit.org

[16] Eelke B Lenselink, Niels Ten Dijke, Brandon Bongers, George Papadatos, Herman WT Van Vlijmen, Wojtek Kowalczyk, Adriaan P IJzerman, and Gerard JP Van Westen. 2017. Beyond the hype: deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set. *Journal of cheminformatics* 9, 1 (2017), 1–14.

[17] Pengyong Li, Jun Wang, Yixuan Qiao, Hao Chen, Yihuan Yu, Xiaojun Yao, Peng Gao, Guotong Xie, and Sen Song. 2020. Learn molecular representations from large-scale unlabeled molecules for drug discovery. *arXiv preprint arXiv:2012.11175* (2020).

[18] Zaynab Mousavian and Ali Masoudi-Nejad. 2014. Drug–target interaction prediction via chemogenomic space: learning-based methods. *Expert opinion on drug metabolism & toxicology* 10, 9 (2014), 1273–1287.

[19] Ahmet Sureyya Rifaioglu, Esra Nalbat, Volkan Atalay, Maria Jesus Martin, Rengul Cetin-Atalay, and Tunca Doğan. 2020. DEEPScreen: high performance drug–target interaction prediction with convolutional neural networks using 2-D structural compound representations. *Chemical science* 11, 9 (2020), 2531–2557.

[20] Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. 2020. Self-Supervised Graph Transformer on Large-Scale Molecular Data. *Advances in Neural Information Processing Systems* 33 (2020).

[21] Yuval Shaked. 2019. The pro-tumorigenic host response to cancer therapies. *Nature Reviews Cancer* 19, 12 (2019), 667–685.

[22] Gregory Sliwoski, Sandeepkumar Kothiwale, Jens Meiler, and Edward W Lowe. 2014. Computational methods in drug discovery. *Pharmacological reviews* 66, 1 (2014), 334–395.

[23] Masashi Tsubaki, Kentaro Tomii, and Jun Sese. 2019. Compound-protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. *Bioinformatics* 35, 2 (2019), 309–318. https://doi.org/10.1093/bioinformatics/bty535

[24] Ming Wen, Zhimin Zhang, Shaoyu Niu, Haozhi Sha, Ruihan Yang, Yonghuan Yun, and Hongmei Lu. 2017. Deep-learning-based drug–target interaction prediction. *Journal of proteome research* 16, 4 (2017), 1401–1409.

[25] David S Wishart, Craig Knox, An Chi Guo, Dean Cheng, Savita Shrivastava, Dan Tzur, Bijaya Gautam, and Murtaza Hassanali. 2008. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic acids research* 36, suppl_1 (2008), D901–D906.

[26] Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. 2018. MoleculeNet: A benchmark for molecular machine learning. *Chemical Science* 9, 2 (2018), 513–530. https://doi.org/10.1039/c7sc02664a arXiv:1703.00564

[27] Ling Xue and Jurgen Bajorath. 2000. Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening. *Combinatorial chemistry & high throughput screening* 3, 5 (2000), 363–372.

[28] Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, et al. 2019. Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling* 59, 8 (2019), 3370–3388.

[29] Xin Yang, Yifei Wang, Ryan Byrne, Gisbert Schneider, and Shengyong Yang. 2019. Concepts of artificial intelligence for computer-assisted drug discovery. *Chemical reviews* 119, 18 (2019), 10520–10594.

[30] Zhe Zhang, Li Zhou, Na Xie, Edouard C Nice, Tao Zhang, Yongping Cui, and Canhua Huang. 2020. Overcoming cancer therapeutic bottleneck by drug repurposing. *Signal transduction and targeted therapy* 5, 1 (2020), 1–25.