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

Pengyong Li
Xidian University
School of Computer Science and Technology
Xian, China

Jun Wang†
Ping An Healthcare Technology
Beijing, China
deeplearning.pku@qq.com

Peng Gao
Yixuan Qiao
Ping An Healthcare Technology
Beijing, China

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

Figure 1: The CandidateDrug4Cancer benchmark encompasses the most-mentioned 29 targets for cancer, covering 54869 cancer-related drug molecules which are 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

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-XXXX-X/21/06 . $15.00
https://doi.org/10.1145/1122445.1122456

arXiv:2203.00836v2  [cs.LG]  22 Aug 2022
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 [23, 33]. 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 [24] 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 [32]. 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].

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 [31]. In the early years, molecular representations are based on hand-crafted features such as molecular descriptors or fingerprints [30]. 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, 22]. 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, 20, 27]. Generally, the deep learning algorithms for DTI prediction comprise of a compound encoder and a protein encoder. Recently, Tsubaki et al. [25] 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 [19, 29]. 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, 17, 22].

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 [21]. 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 (GDS) [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 cancers.

Table 1: The most frequent target proteins of cancers collected from DepMap[6], DrugBank[28] 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      |
| chembl1956| Stem cell growth factor receptor                  |
| chembl1957| 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         |
| chembl2262| Glycogen synthase kinase-3 beta                  |
| chembl2267| Glycogen synthase kinase-3 beta                  |
| chembl2279| 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 [28], 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 represents comparable measures of concentrations to reach half-maximal response transformed to a negative logarithmic scale. pChEMBL is defined as:

\[
pChEMBL = -\log(molarIC50(M))
\]

For example, an IC50 measurement of 10 μM would have a pChEMBL value of 5. However, in order to efficiently verify anti-cancer 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  | NumH                        | Number of connected hydrogen[unk, 1, 2, 3, 4, 5]                             |
|               | Implicit valence            | [unk, 0, 1, 2, 3, 4, 5]                                                     |
|               | Degree                      | Number of covalent bonds[unk, 0, 1, 2, 3, 4, 5, 6, 7, 8]                   |
|               | Aromatic                    | Whether the atom is part of an aromatic system[0,1]                         |
| Edge feature  | Bond direction              | [None, enddiproline, 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, 20, 25, 27]. 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 [18], to evaluate the effectiveness of pre-trained GNN on DTI prediction for cancer 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. Weakly supervised learning. It is also worthwhile to utilize semi-supervised learning and active learning techniques [26] to further boost prediction performance. 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/1gXpGe5UhAYB9zVYnSl6E2MEaWfl5w98O/view?usp=sharing.
