European Journal of Medicinal Chemistry 255 (2023) 115401

Contents lists available at ScienceDirect

European Journal of Medicinal Chemistry

journal homepage: www.elsevier.com/locate/ejmech

Research paper

DeepCancerMap: A versatile deep learning platform for target- and cell-based anticancer drug discovery

Jingxing Wu [a,1], Yi Xiao [a,1], Mujie Lin [a,1], Hanxuan Cai [a], Duancheng Zhao [a], Yirui Li [b], Hailin Luo [b], Chuanqi Tang [c], Ling Wang [a,*]

[a] School of Biology and Biological Engineering, South China University of Technology, Guangzhou, 510006, China
[b] School of Software Engineering, South China University of Technology, Guangzhou, 510006, China
[c] School of Design, South China University of Technology, Guangzhou, 510006, China

ARTICLE INFO

Keywords:
Anticancer drug
Deep learning
Phenotypic-based drug discovery
Target-based drug discovery
Webserver

ABSTRACT

Discovering new anticancer drugs has been widely concerned and remains an open challenge. Target- and phenotypic-based experimental screening represent two mainstream anticancer drug discovery methods, which suffer from time-consuming, labor-intensive, and high experimental costs. In this study, we collected 485,900 compounds involving in 3,919,974 bioactivity records against 426 anticancer targets and 346 cancer cell lines from academic literature, as well as 60 tumor cell lines from NCI-60 panel. A total of 832 classification models (426 target- and 406 cell-based predictive models) were then constructed to predict the inhibitory activity of compounds against targets and tumor cell lines using FP-GNN deep learning method. Compared to the classical machine learning and deep learning methods, the FP-GNN models achieve considerable overall predictive performance, with the highest AUC values of 0.91, 0.88, 0.91 for the test sets of targets, academia-sourced and NCI-60 cancer cell lines, respectively. A user-friendly webserver called DeepCancerMap and its local version were developed based on these high-quality models, enabling users to perform anticancer drug discovery-related tasks including large-scale virtual screening, profiling prediction of anticancer agents, target fishing, and drug repositioning. We anticipate this platform to accelerate the discovery of anticancer drugs in the field. DeepCancerMap is freely available at https://deepcancermap.idruglab.cn.

1. Introduction

Cancer is a severe public health problem that imposes a heavy burden to humans worldwide [1]. Chemotherapeutics, one of the most commonly used strategies in cancer treatment, aim to interfere with the rapidly and disorderly dividing tumor cells out of normal mechanisms of growth suppression control [2]. However, many anticancer drugs exhibit various typical weaknesses and unsatisfactory therapeutic effects. For one thing, the low specificity of existing chemotherapeutic drugs for cancer cells leads to killing normal cells, especially those in the rapidly proliferating stage, resulting in serious adverse effects [3,4]. For another, the resistance of tumor cells to conventional chemotherapeutic drugs further limits their clinical efficacies and applications. For example, one common mechanism for cancer cells to acquire resistance is through the active efflux function of ATP-binding cassette (ABC) transporter proteins, which can cause some drugs to be ineffective, such as Vincristine and Adriamycin [5]. In addition, the ATP-independent manners, such as chromosomal genes mutation and altered drug permeation, have also been reported in recent years [6,7]. Accordingly, there is an urgent need to identify novel anticancer drugs to overcome these shortcomings.

Since the 1990s, with the influential progress of combinatorial chemistry and high-throughput screening (HTS), the explosive growth of biological data has brought a promising impact on drug discovery. For instance, ChEMBL is a large, open-access, and manually curated database collecting medicinal chemistry data and bioactivity, containing more than 5.4 million bioactivity data released in 2011 [8]. Currently, more than 20 million compound-target pairs obtained from over 2.3 million compounds tested against over 15 thousand targets is accessible on the ChEMBL database (released Jan 26, 2023). Similar medicinal chemistry databases are DrugBank and Scifinder, which collect chemical structures, pharmacological actions, acting protein targets and

* Corresponding author.
E-mail address: lingwang@scut.edu.cn (L. Wang).
[1] These authors made equal contribution to this work.

https://doi.org/10.1016/j.ejmech.2023.115401
Received 7 December 2022; Received in revised form 29 March 2023; Accepted 18 April 2023
Available online 23 April 2023

physiological pathways of drugs. Based on the rapidly growing and aggregated data, the computational-based approaches to drug development began to show prospective research value and application directions. In traditional research channels, target-based drug discovery (TDD) and phenotypic-based drug discovery (PDD) represent two mainstream approaches for anticancer drug discovery. With the advantages of high screening capacity, relatively lower cost and easily to achieving efficient structure-activity relationship (SAR) for subsequent lead optimization [9,10], TDD has been widely applied in the pharmaceutical industry as the mainstream method in the past decades [11–16]. However, a major shortcoming of TDD is that identification and validation of druggable anticancer targets is extremely difficult, as it requires proof that drug candidates acting on potential target have definite clinical therapeutic effects in cancer treatment, without unacceptable side effects, which is nearly impossible for implementation in the early stage of TDD [17]. This fatal lag in target validation has been recently identified as one of the possible reasons for the low likelihood of clinical translation [18]. PDD is an original but significant method that has revived interest in recent years, allowing direct observation of molecules modulating the behavior (phenotype) of biological systems. PDD is referred to an alternative to TDD due to its higher clinical-translatability [19]. For example, cancer cell-based screening model as a typical biological system, has been successfully applied to many cases of anticancer drug discovery [20–22]. However, the main difficulty of PDD is how to quickly fish the potential targets for a given drug candidate. Although there are many experimental and computational methods for target fishing [23,24], undeniably, it is not easy to find potential targets for a given drug from many proteins in the human body. It is clear that PDD and TDD can complement each other for anticancer drug discovery.

With increasing of the target- and phenotypic-based screening data in the field of anticancer drug discovery, machine learning (ML), and deep learning (DL) algorithms are used to develop predictive and/or generative models based on various molecular representations such as molecular descriptors, fingerprints, and graphs for the discovery of new anticancer agents. For example, in the aspects of utilizing target-based screening data, Chen et al. used support vector machine (SVM) method to identify a nanomolar dual inhibitor against FGFR4 and EGFR [25]. Yang et al. reported extreme gradient boosting (XGBoost)-based predictive models for the identification of six active compounds against JAK2 with IC_{50} values less than 100 nM [26]. Moreover, ML and DL methods have been used for the inhibitory activity prediction [27,28] of various kinases inhibitors and profiling prediction of kinase inhibitors [29,30] in the field of anticancer drug discovery. Cell-based screening represents the most commonly used phenotypical-based screening method for the determination of inhibitory activity of tested compounds against tumor cell lines, which has substantially improved the ability to discover new anticancer agents. Luo et al. [31,32] constructed naïve Bayesian (NB)-based predictive models for the cellular inhibitory activity prediction of anticancer molecules and subsequently discovered novel and diverse anticancer agents with significant inhibitory activity against various tumor cell lines. Recently, He et al. reported interpretable ML models for the accurate prediction of active molecules against breast cancer cells [33]. In addition, DL-based molecular generative models have also been reported to discover inhibitors of DDR1 [34–36], AKT [37], and CDK9 [38] kinases for cancer treatment. Collectively, these models have been successfully used in the discovery of new anticancer active molecules in real-world scenarios, demonstrating that ML and DL methods play an important role in accelerating anticancer drug discovery.

Although the reported computational models provided valuable insight into the discovery of anticancer agents, there is still no comprehensive platform for anticancer drug discovery. To this end, in this study, we developed a DL-based webserver called DeepCancerMap, which, to the best of our knowledge, is the first comprehensive online platform dedicated to the discovery of anticancer drugs. The schematic diagram of modelling datasets, model construction, function, and organization involved in DeepCancerMap is shown in Fig. 1.

2. Methods

2.1. Data collection and processing

As shown in Fig. 1, target- and cell-based screening data were collected and subsequently used to build predictive models, each model is obtained from the original algorithm framework by fitting and optimizing parameters on a dataset. Firstly, anticancer targets were collected from the therapeutic target database (TTD: http://db.idrblab.net/ttd/) [39], and small-molecules and their bioassay records for these targets were acquired from ChEMBL database (version ChEMBL29) [8] using the *chembl_webresource_client* Python module. To gather high-quality bioactivity datasets, each dataset from target-based screening data was processed as follows: 1) activity records with assay type B were kept; 2) compounds with biological activity reported as IC_{50}, EC_{50}, K_i or K_d were retained, whereas records with no bioactivity value or labeled indefinitely (e.g., activity $<$50 μM or activity $>$0.1 μM) were deleted, and the diverse units of bioactivity values (i.e., ug/mL, nM, and μM) were converted to a unique unit in μM; 3) molecules with molecular weight greater than 1000 were discarded; and 4) if a compound has multiple activity records, the average of the activity values for the compound was used as its final bioassay activity. All chemical structures were standardized using the Python module standardizer (https://github.com/flatkinson/standardiser), including removing salt or solvent fraction, neutralizing charge by adding or subtracting atoms, and standardizing structures to a common representation. We excluded the compounds containing metal ions after standardization. Finally, compounds with biological activity values (e.g., IC_{50}, EC_{50}, K_i, and K_d) $\leq$ 1 μM were regarded as active, while $>$1 μM were regarded as inactive [40,41].

Secondly, phenotypic cell-based screening data were collected from the ChEMBL database (accessed by September 2021) and NCI-60 DTP project. Molecules with their activities records from ChEMBL database are referred to as academic-sourced phenotypic cell-based modelling dataset, while compounds and their activities from NCI-60 DTP are termed as industry-standard phenotypic cell-based modelling dataset. For the academia-sourced cell-based modelling dataset, all cancer cell lines were obtained from the ChEMBL and were then manually annotated and classified into 32 categories according to source issue and description information. The full activity records of tested compounds against these cancer cell lines were obtain by matching the cell ChEMBL ID. Meanwhile, the industry-standard cell-based modelling datasets including structures and their activity records were collected from the NCI DTP project (https://dtp.cancer.gov/), a high-throughput screening platform for anticancer drug based on 60 human tumor cell lines. For academia-sourced cell-based screening data from ChEMBL, each cancer cell line dataset was refined using the following criteria: 1) only keeping activity records with assay type F; 2) molecules with definitely bioassay values, specifically reported as IC_{50}, EC_{50}, GI_{50} or ED_{50} were retained, and the different units of bioassay values, such as ug/mL, nM, μM, were also converted to a unique unit in μM; 3) similar to the processing of target-based screening data, the average of multiple testing values represented the final test value for a given compound, and all structures were also standardized using the same protocol. Molecules with biological activity values (i.e., IC_{50}, EC_{50}, GI_{50}, and ED_{50}) $\leq$ 10 μM were labeled as actives and vice versa [42,43].

In addition, for the industry-standard NCI-60 cell-based screening data, the dataset for each tumor cell line was processed as follows: 1) keeping activity records with the $\log_{10}$ of the highest concentration of −5 and removing data with concentration units other than 'M'; 2) the GIPRCNT value represents the bioactivity value in the NCI-60 screening data, whereas the average GIPRCNT value was used as the final bioactivity value for each compound with multiple records [30,44]; 3) all molecules were standardized using the above protocol; 4) compounds

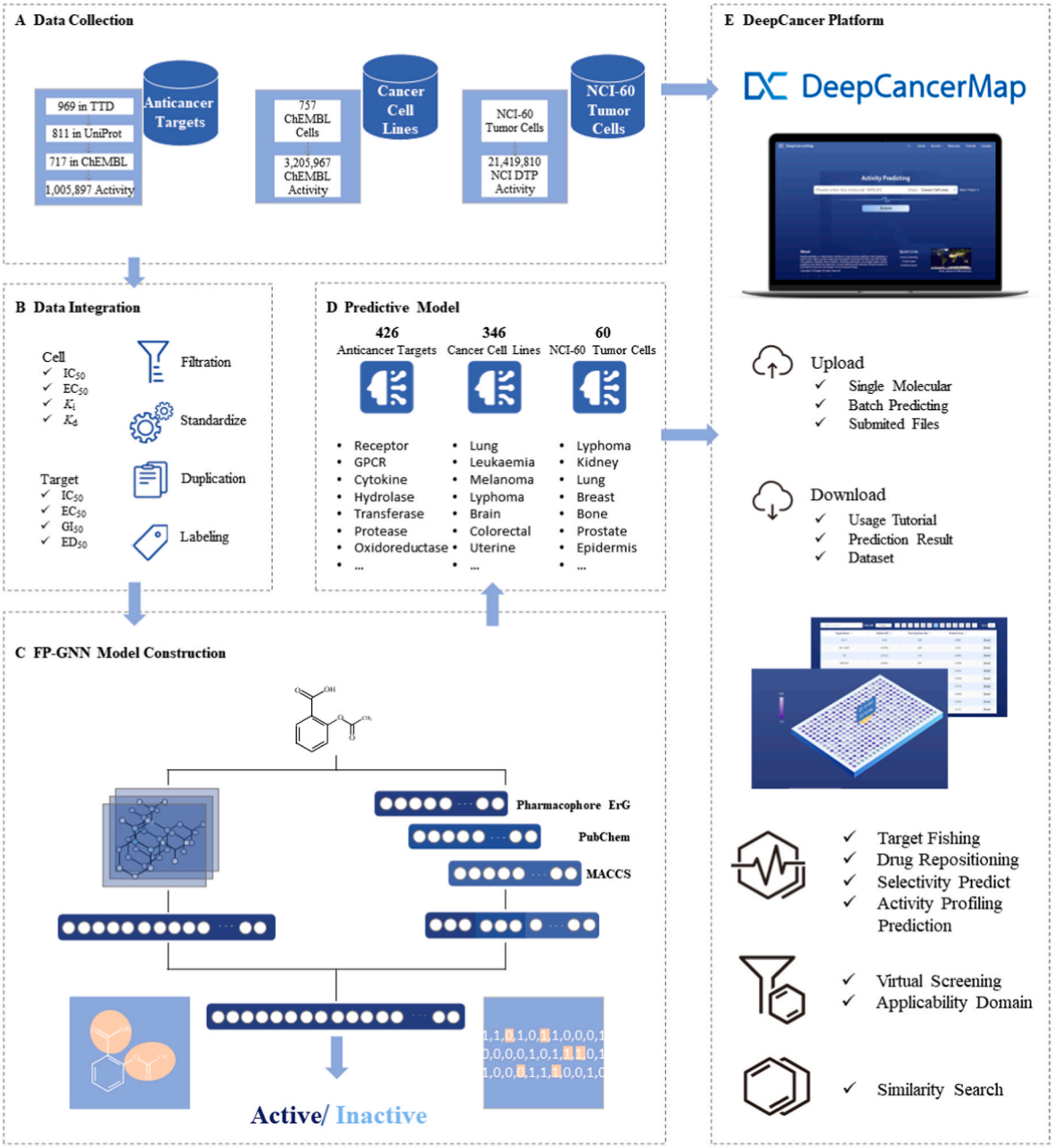


Fig. 1. DeepCancerMap workflow.

with bioactivity values $\leq$ 50 were labeled as actives and those with values $\geq$ 90 were labeled as inactives [30,44].

The processed dataset for each target or cell line was randomly split into three sub-datasets of training, validation, and test sets by the ratio of 8:1:1. All these modelling datasets are freely available at https://deepcancer.idruglab.cn.

2.2. FP-GNN DL architecture and model training protocol

In this study, the FP-GNN DL algorithm [45] developed by our group was applied to construct classification models to estimate the bioactivity of small molecules against various targets and tumor cell lines. Briefing, FP-GNN simultaneously learns and explores the effective information of molecules in both one-dimensional fingerprints and bidimensional graphs. As shown in Fig. 1, it is divided into two parts: graph neural network (GNN) and fingerprints network (FPN).

The spatial GNN was used as the GNN part of FP-GNN model. Molecules were individually transformed into sets of nodes representing for atomic features. The attention mechanism in graph convolutional network was used to update each node by aggregating information from its neighbor nodes. After all nodes were updated, the full graph information was aggregated as the eventual output of the graph proportion. Meanwhile, three fingerprints including PubChem fingerprint, MACCS fingerprint, and Pharmacophore ErG fingerprint were applied in the FPN module. These three fingerprints were tandemly combined for their respective advantages, such as the atomic properties, bond properties in MACCS fingerprint, a broad spectrum of chemical substructures covered in PubChem fingerprint, and pharmacophore features identified in Pharmacophore ErG fingerprint. Finally, fully connected layers were used to combine the outcomes from both GNN and FPN to produce the ultimate output. Accordingly, FP-GNN architecture has the outstanding advantage of integrating molecular fingerprints and structural graphs,

aiming to improve the performance of target- and cell-based anticancer activity predictions.

To obtain the predictive models with the best robustness and precision, Bayesian optimization strategy was used to perform hyperparameters optimization during the training of the FP-GNN models. The optimized parameters included the hidden size of attentions, the hidden size and dropout rate of FPN, the dropout rate of GNN, the number of multi-head attentions, and the ratio of GNN in the FP-GNN. Among them, the ratio of GNN in the FP-GNN is a necessary parameter for optimization due to the large difference in the amount of each target- or cell-based modelling dataset. Additionally, the training epochs for each parameter combination were limited to 25, while the number of explorations of hyperparameters choices were restricted to 20. FP-GNN models were trained using the PyTorch framework (GPU version 1.5.0).

In addition, random forest (RF), XGBoost, SVM, K-nearest neighbor (KNN), NB, deep neural network (DNN), graph convolutional network (GCN) and graph attention network (GAT) models were built as baseline models for comparison. A brief introduction to these methods can be found elsewhere [32,46,47]. The RF, SVM, KNN, NB, and DNN models were developed using the Scikit-learn python package (https://github.com/scikit-learn/scikit-learn, version: 0.24.1); the XGBoost models were developed using the XGBoost python package (https://github.com/dmlc/xgboost, version: 1.3.3) [48]; and the GCN and GAT models were established using the DeepChem python package (https://deepchem.io/). All models were trained on CPU [Intel(R) Xeon (R) Silver 4216 CPU @ 2.10 GHz] and GPU [NVIDIA Corporation GV100GL (Tesla V100 PCIe 32 GB)], respectively. In this study, all models were finally obtained through the hyperparameters optimization, and detailed hyperparameters are provided in Tables S1–S6.

2.3. Model evaluation

We established classification models for predicting active molecules against targets or tumor cell lines. Five parameters were selected for evaluation of the predictive performance of the models: the area under the ROC curve (AUC), specificity (SP/TNR), sensitivity (SE/TPR/Recall), accuracy (ACC), and Matthews correlation coefficient (MCC) [49]. They are defined as follows:

$$\mathrm{SE} = \frac{TP}{TP + FN} \tag{1}$$

$$\mathrm{SP} = \frac{TN}{TN + FP} \tag{2}$$

$$\mathrm{ACC} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$\mathrm{MCC} = \frac{TP \times TN - FN \times FP}{\sqrt{(TP + FN) \times (TP + FP) \times (TN + FN) \times (TN + FP)}} \tag{4}$$

All models were optimized and selected based on the AUC. The best-performing models were saved and used for the development of the DeepCancerMap platform.

2.4. Model applicability domain

To apply predictive DL models in drug discovery, the suitable and practical applicability domain (AD) should be determined for each model. The size of chemical space of training compounds was commonly employed. In this study, the similarity distance strategy was therefore used to define and perform the AD analysis [50]. Euclidean distance was calculated based Morgan fingerprint using RDKit software (http://www.rdkit.org/) for the measurment of structural similarity. The AD of the model was computed and described as a threshold value of distance (D_c)

$$D_c = d_{ave} + Z * d_{std}$$

To calculate d_{ave} and d_{std}, all the molecules in training set were transferred into Morgan fingerprint in bit string, and the Euclidean distances between every molecule and the other molecules in training set were calculated. An empirical parameter k was defined and the avearage of k smallest Euclidean distance for each molecules in training set was obtained, then took the average of these mean values as d_{ave}, and d_{std} is the standard deviation of these mean values. With an empirical parameter Z, the D_c was computable as the threshold of AD. To evaluate an external compound as outside of domain (OD) or inside of domain (ID), the nearest distance between this compound and molecules in training set (also calculated by Euclidean distance with Morgan fingerprint string) was compared with the D_c, where surpassing the D_c means OD and vice versa.

In this study, the 832 D_c of each training set was calculated and the molecules in corresponding test sets were evaluated as OD and ID compounds. For each model, the predictive ability to ID and OD sets were evaluated, respectively. Different combinations of k and Z were conducted to determine the most suitable parameters.

2.5. Software implementation

DeepCancerMap was built based on the python web framework Django and deployed on an elastic compute service (ECS) of Tencent-Cloud to run a CentOS Linux system. MySQL was adopted as a data management tool for storing predictive models and anticancer molecules with related information. The data query, browser, prediction, network display, web interfaces and access were implemented using HTML, PHP, JavaScript and Nginx HTTP servers. The trained FP-GNN models and their associated files (e.g., model information, predictive results, and local python software) were stored, displayed, and available in the DeepCancerMap webserver. JSME was used as the online molecular structure editor [51]. The similarity searching was performed based on the MACCS, AtomPairs, and Morgen fingerprints implemented in RDKit software. To manage the multiple and simultaneous computational tasks submitted by users, we implemented the Celery system (version 5.2.5: https://github.com/celery/celery/) in DeepCancerMap to process distributed task queues. Consequently, the email notification function is used to remind the users that the task calculation is complete, and the entire tutorial is available on the tutorial page. DeepCancerMap can be accessed through Firefox, Google Chrome, and Apple Safari. Details of the methods involved in the development of DeepCancerMap are listed in Table S7.

3. Results and discussion

3.1. Datasets statistics and analysis in DeepCancerMap

After the data collection and optimization mentioned above, we obtained 485,900 unique and experimentally-validated molecules involving in 3,919,974 pharmacological data points from the original target- and cell-based anticancer screening data (Table 1).

Specifically, 426 potential therapeutic targets for cancer treatment were eventually obtained, which can be classified into 20 categories (Fig. 2A). These target-based modelling datasets included 307,470 unique compounds involving in 447,050 target-compound associations. Among them, 315,269 associations were annotated as actives and 131,781 were labeled as inactives, both of which were then used to construct target-based prediction models. Table S8 summarizes detailed

Table 1
Overview of data scale collected and used in DeepCancerMap.

Dataset	Num. of task	Num. of categories	Num. of activities
Anticancer targets	426	20	447,050
Cancer cell lines	354	32	343,940
NCI-60 tumor cells	60	9	3,128,984

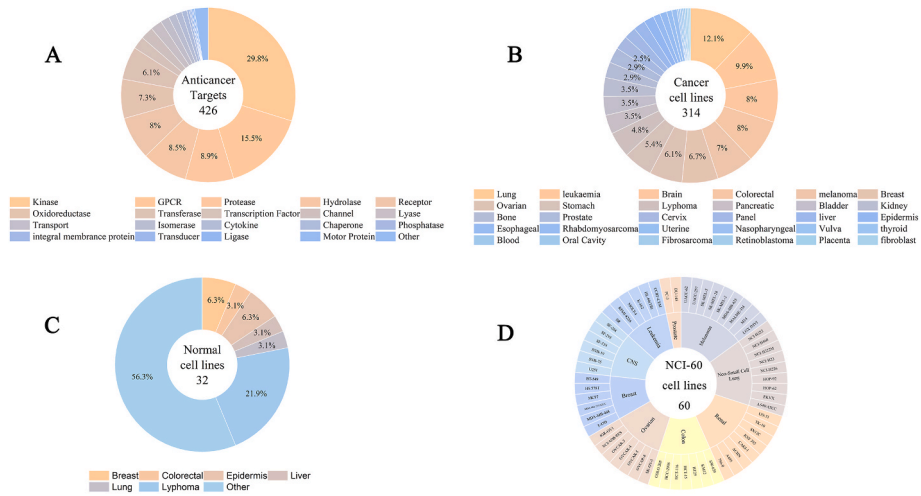


Fig. 2. Distribution of targets and cell lines for modelling datasets construction in DeepCancerMap. Statistical categories distribution of anticancer targets (A), academia-sourced cancer cell lines (B) and normal cell lines (C) from ChEMBL, and NCI-60 tumor cancer cell lines (D).

information on these targets and their related compounds. For academia-sourced cell-based screening data, a total of 314 cancer cell lines and 32 normal cell lines from the ChEMBL were finally obtained, which are artificially categorized into 32 groups according to source tissue (Fig. 2B and C). For these cell lines, a total of 118,828 compounds involving in 343,940 cellular assay data points were utilized to build cell-based predictive models. Notably, we included 32 normal cell lines to establish their predictive models to predict the inhibitory activity of molecules against these normal cell lines, which can be used to assess the cytotoxicity of anticancer agents to normal cells, i.e. selectivity. The detailed information of these cell lines and their related compounds are provided in Tables S9 and S10. In addition, NCI-60 human tumor cell lines from the industry-standard NCI-60 cell-based screening platform cover nine panels (Fig. 2D), including melanoma, non-small cell lung, colon, ovarian, Leukemia, CNS, renal, prostate, and breast. In the present study, a total of 84,552 compounds and their 3,128,984 testing data points against these nine panels of tumor cell lines are obtained to build the predictive models in the DeepCancerMap. The detailed information on these NCI-60 tumor cell lines and the corresponding compounds are listed in Tables S11–S12.

In addition, Bemis–Murcko scaffold analysis was used to analyze the structural diversity of each modelling dataset. As shown in Tables S13–S16, the proportion of the scaffolds for each target or cell line modelling dataset was between 8.24% and 93.98%, with an average value of 33.92%. The proportion of scaffolds in the modelling datasets of anticancer targets, academically-derived cancer cell lines and normal cell lines, as well as industrial-derived tumor cell lines were 28.41%, 44.92%, 36.26%, and 14.17% (Fig. 3), respectively, indicating that the compounds of each target or cell line modelling dataset were structurally more diverse. All these results suggest that the predictive models based on the collected modelling datasets in this study will have good robustness and scalability. All datasets mentioned in the present study are freely available at https://github.com/idrugLab/DeepCancerMap.

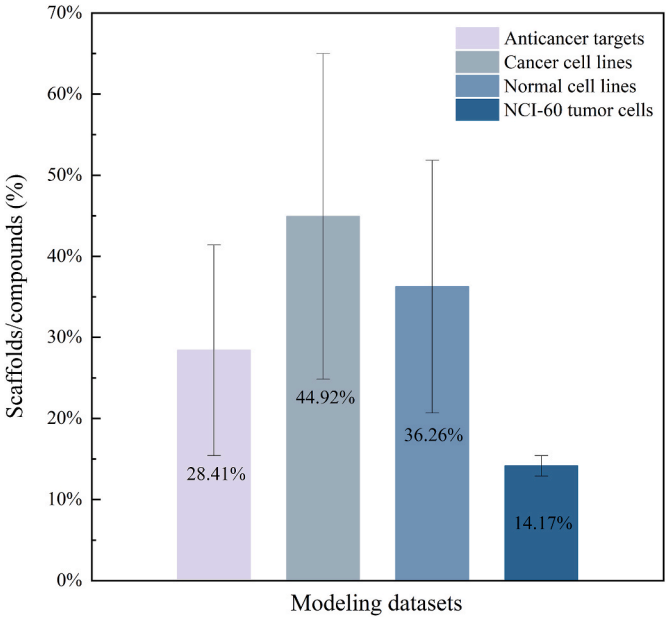


Fig. 3. Average ratio of the number of scaffolds to the number of compounds in the modelling datasets of anticancer targets, cancer cell lines, normal cell lines, and NCI-60 tumor cells.

3.2. Performance of target-based predictive models in DeepCancerMap

Each target-based modelling dataset was randomly split into training (80%), validation (10%), and test (10%) sets. The training set was used to build the model, and the validation set was employed for optimizing hyperparameters, and the test set was used to estimate the performance of the model. FP-GNN DL algorithm was used to construct classification models for predicting active molecules against these anticancer targets. The detailed evaluation results of FP-GNN model for each target are provided in Table S17. As shown in Table S17, FP-GNN models performed well for these 426 targets, with an average AUC value of 0.905 for the test sets. Further analysis found that the performance results of the FP-GNN models varied by protein family (Table 2), with average AUC values ranging between 0.714 and 0.973 for the 20 target classes. The interval distribution analysis of the AUC values of the test sets of 426 targets found that the AUC values of the majority of FP-GNN models (98.59%) were greater than 0.7. For example, the number of high quality (HQ, AUC >0.8) FP-GNN models was 387. In addition, over half of the models (264 anticancer targets) achieved remarkable performance with AUC value greater than 0.9, showing superior accuracy covering the entire anticancer targets.

As shown in Fig. S1A and Table S18, compared to other commonly used CML (i.e., RF, XGBoost, KNN, NB, and SVM) and DL (i.e., GAT, GCN, and DNN) algorithms, FP-GNN method achieved the best overall predictive performance for these 426 targets, with not only the highest average AUC value, but also the highest average MCC value of 0.573. In addition, the comparative analysis of the AUC value of each target of the

Table 2
The average predictive performance of FP-GNN models for the test sets against 20 target classes covering 426 anticancer targets.

Target class	Num. of targets	SE[a]	SP[b]	ACC[c]	MCC[d]	AUC[e]
Channel	10	0.469	0.822	0.712	0.404	0.879
Chaperone	4	0.546	0.717	0.739	0.574	0.914
Cytokine	5	0.700	0.369	0.693	0.294	0.985
GPCR	66	0.569	0.852	0.802	0.537	0.897
Hydrolase	36	0.705	0.716	0.817	0.612	0.909
Integral membrance protein	1	1.000	0.000	0.778	–	0.714
Isomerase	6	0.792	0.457	0.783	0.726	0.912
Kinase	127	0.565	0.876	0.818	0.569	0.895
Ligase	1	0.780	0.868	0.827	0.652	0.902
Lyase	9	0.697	0.734	0.834	0.673	0.937
Motor protein	1	0.690	0.787	0.756	0.462	0.861
Oxidoreductase	31	0.712	0.730	0.772	0.589	0.898
Phosphatase	2	0.167	1.000	0.778	0.516	0.875
Protease	38	0.669	0.820	0.534	0.627	0.908
Receptor	34	0.640	0.732	0.816	0.602	0.928
Transcription factor	10	0.758	0.581	0.760	0.616	0.901
Transducer	1	1.000	0.000	0.771	–	1.000
Transferase	26	0.681	0.622	0.917	0.593	0.914
Transport	7	0.618	0.699	0.764	0.555	0.936
Others	11	0.605	0.773	0.680	0.714	0.942
Average		**0.668**	**0.658**	**0.794**	**0.573**	**0.905**

-, not computable.
[a] SE: Sensitivity.
[b] SP: Specificity.
[c] ACC: Accuracy.
[d] MCC: Matthews correlation coefficient.
[e] AUC: The area under receiver operating characteristic.

FP-GNN model and the baseline models (Fig. 4A) also found that the models based on FP-GNN showed satisfactory predictive performance, because most of the FP-GNN models outperformed the baseline models (Fig. 4 and Fig. S2). A similar trend can be observed from the analysis of the HQ (AUC >0.8) models between FP-GNN and baseline methods (Fig. 5A). All these results indicate that the target-based predictive models constructed by the FP-GNN method in this study show a powerful ability to predict active molecules against these 426 targets, which can be used for target-based anticancer drug design and discovery.

3.3. Performance of cell-based predictive models in DeepCancerMap

A total of 406 FP-GNN-based predictive models were constructed for predicting active molecules against various tumor cells. Detailed performance results of these models are summarized in Tables S19–S21. As shown in Table 3, FP-GNN algorithm achieved considerable predictive performance on these 314 academia-sourced tumor cell lines, with an average AUC value of 0.877 for the test sets (Table 3 and Fig. 4B). Importantly, the FP-GNN model exhibited remarkable predictive accuracy on the industry-standard NCI-60 modelling datasets for 60 tumor cell lines, with the higher average AUC value of 0.912 for the test sets (Table 4 and Supporting Information Table S20). It is obvious that the overall performance of models (AUC = 0.912) based on the industry-standard NCI-60 cellular-level screening data outperforms the models (AUC = 0.881) based on academic-sourced modelling data, despite both use the FP-GNN modelling method. This may be due to the higher quality and larger in size of the NCI-60 modelling datasets than those from academic sources.

In addition, updated NCI-60 screening data (2016–2021) were collected and used as time-split external test sets to further measure the accuracy of the established FP-GNN models based on the industry-standard NCI-60 cellular-level screening data (Table 5 and Fig. 4C). The average AUC value for these time-split external test sets is 0.912, indicating that the FP-GNN models have excellent robustness and scalability for the discovery of new active molecules against various tumor cell lines in real-world scenarios of drug discovery. Considering the intrinsic correlation between the screening data of each tumor cell line in the NCI-60 panel, we further constructed a multi-task FP-GNN model based on the NCI-60 tumor cell line screening datasets [52], achieving the highest AUC of 0.913 and 0.921 on test set and external test set (Table S22) compared to the single-task model. Accordingly, such multi-task FP-GNN model was also implemented on DeepCancerMap, enabling users to utilize the advanced multiple-task model to predict the inhibitory activity of molecules against 60 tumor cell lines.

As shown in Figs. S1B and S1C, FP-GNN models performed better than CML (e.g., RF, XGBoost, KNN, NB, and SVM) and DL (e.g., GAT, GCN, and DNN) models for tumor cell lines from both academia-sourced and industry-standard NCI-60 screening data, the detail evaluation information was provided in Tables S23 and S24. Fig. 4B and C indicate the superiority of FP-GNN models against the optimal baseline models on academia-sourced cancer cell lines and external test sets as well as industrial tumor cell lines through the per-cell line comparison. In addition, Fig. 5 and Figs. S3–S4 show that the FP-GNN models contribute the most unique HQ models (AUC ≥0.8) for these tumor cell lines.

It is well known that anticancer drugs lack selective toxicity to human normal cells, leading to severe side effects which limits their clinical applications. Therefore, predicting the cytotoxicity of compounds against normal cell lines is essential in the early stages of anticancer drug discovery. Accordingly, we established predictive models for 32 normal cell lines, involving in 9668 compounds and 10,861 inhibitory activity records. These models based on FP-GNN method performed well against 32 normal cell lines (Table 5 and Table S19), with an average AUC value of 0.914 for the test sets. Such models can be used to predict whether anticancer agents are toxic to normal cells. Collectively, cell-based predictive models in DeepCancerMap can be used not only to predict active molecules against various cancer cell lines, but also to detect the potential cytotoxicity of compounds to normal cell lines, which can be used for cell-based anticancer drug discovery.

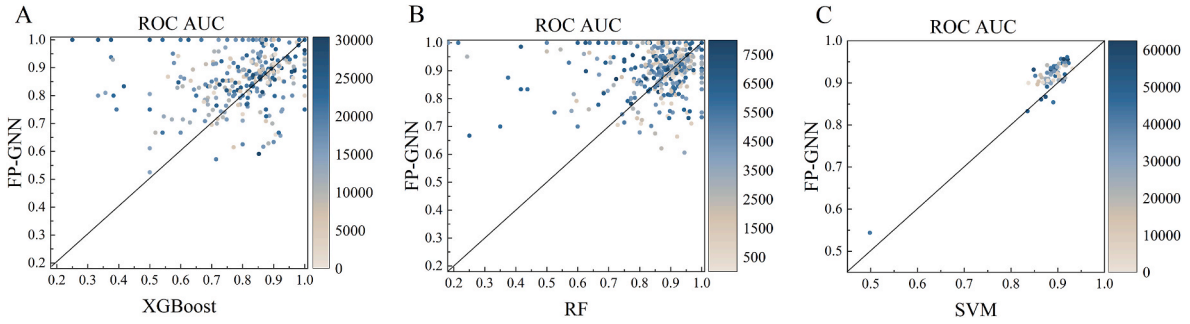


Fig. 4. Performance comparison of FP-GNN model and the best baseline algorithm for test set of each anticancer target (A), academia-sourced cancer cell line (B), and time-split NCI-60 external data (C).

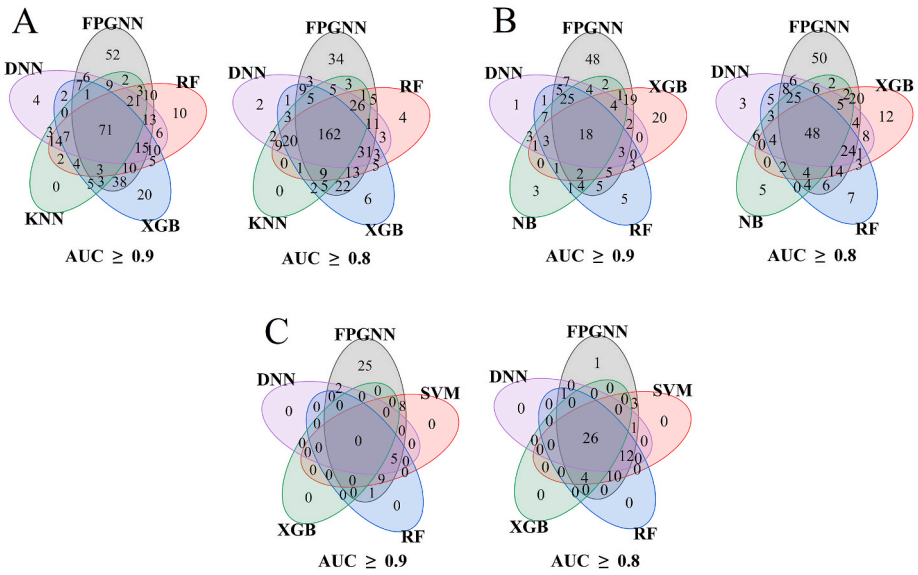


Fig. 5. Overlap analyses of high-quality (HQ) FP-GNN models and baseline algorithm models with the average AUC higher than 0.8 and 0.9. (A), (B), and (C) represent analysis results for anticancer targets, academia-sourced cancer cell lines, and time-split NCI-60 external data (C).

Table 3
Performance of FP-GNN models for the test sets against 314 academia-sourced cancer cell lines.

Cell class	Num. of cell lines	SE[a]	SP[b]	ACC[c]	MCC[d]	AUC[e]
Bladder	11	0.739	0.483	0.580	0.637	0.886
Blood	1	1.000	0.000	0.333	–	1.000
Bone	9	0.826	0.508	0.705	0.456	0.867
Brain	25	0.792	0.623	0.750	0.611	0.865
Breast	21	0.835	0.495	0.772	0.464	0.889
Cervix	8	0.700	0.510	0.613	0.589	0.929
Colorectal	25	0.830	0.471	0.738	0.487	0.847
Epidermis	6	0.797	0.546	0.718	0.704	0.911
Esophageal	4	0.647	0.638	0.610	0.569	0.925
Fibroblast	1	0.000	1.000	0.200	–	1.000
Fibrosarcoma	1	0.902	0.595	0.786	0.533	0.891
Kidney	11	0.733	0.547	0.697	0.414	0.807
Leukemia	31	0.604	0.625	0.689	0.533	0.918
Liver	7	0.929	0.630	0.819	0.680	0.924
Lung	38	0.643	0.656	0.654	0.446	0.851
Lyphoma	15	0.773	0.354	0.670	0.524	0.897
Melanoma	22	0.769	0.568	0.731	0.479	0.862
Nasopharyngeal	3	0.513	0.861	0.800	0.562	0.917
Oral Cavity	1	0.800	0.800	0.800	0.600	0.880
Ovarian	19	0.851	0.449	0.738	0.485	0.883
Pancreatic	11	0.767	0.547	0.697	0.504	0.852
Panel	7	0.420	0.679	0.569	0.713	0.960
Placenta	1	0.750	0.667	0.692	0.386	0.806
Prostate	9	0.761	0.568	0.756	0.605	0.931
Retinoblastoma	1	0.000	1.000	0.800	–	1.000
Rhabdomyosarcoma	4	0.542	0.556	0.601	0.255	0.830
Stomach	17	0.674	0.622	0.659	0.459	0.840
Thyroid	1	1.000	0.000	0.609	–	0.778
Uterine	3	0.780	0.619	0.798	0.600	0.950
Vulva	1	0.750	0.200	0.538	−0.058	0.525
Average		**0.704**	**0.560**	**0.671**	**0.509**	**0.881**

-, not computable.
[a] SE: Sensitivity.
[b] SP: Specificity.
[c] ACC: Accuracy.
[d] MCC: Matthews correlation coefficient.
[e] AUC: The area under receiver operating characteristic.

Table 4
Performance of FP-GNN models for the test sets and time-split external test sets against the industry-standard NCI-60 cancer cell lines.

Tumor cell class	Num. of cell lines	Test set			External test set		
		ACC	MCC	AUC	ACC	MCC	AUC
Breast	6	0.890	0.632	0.919	0.950	0.532	0.923
CNS	6	0.896	0.634	0.913	0.958	0.496	0.917
Colon	7	0.898	0.636	0.916	0.957	0.504	0.920
Leukemia	6	0.848	0.630	0.901	0.918	0.546	0.887
Melanoma	9	0.898	0.626	0.913	0.962	0.519	0.928
Non-Small Cell Lung	9	0.888	0.622	0.910	0.942	0.507	0.926
Ovarian	7	0.905	0.615	0.912	0.963	0.465	0.926
Prostate	2	0.892	0.637	0.921	0.954	0.553	0.931
Renal	8	0.889	0.615	0.907	0.890	0.431	0.862
Average		**0.889**	**0.627**	**0.912**	**0.944**	**0.506**	**0.913**

Table 5
Performance of FP-GNN models for the test sets against 32 academic normal cell lines.

Normal cell class	Num. of cell lines	SE[a]	SP[b]	ACC[c]	MCC[d]	AUC[e]
Breast	2	0.921	0.423	0.622	0.681	0.864
Colorectal	1	1.000	0.000	0.867	–	0.923
Epidermis	2	1.000	0.000	0.750	–	1.000
Liver	1	0.000	1.000	0.750	–	1.000
Lung	1	1.000	0.000	0.968	–	0.800
Lyphoma	7	0.669	0.652	0.694	0.594	0.927
Others	18	0.862	0.396	0.717	0.499	0.905
Average		**0.638**	**0.947**	**0.889**	**0.627**	**0.912**

-, not computable.
[a] SE: Sensitivity.
[b] SP: Specificity.
[c] ACC: Accuracy.
[d] MCC: Matthews correlation coefficient.
[e] AUC: The area under receiver operating characteristic.

3.4. Analysis of applicability domain of the predictive models in DeepCancerMap

The determination of AD help to judge whether the input molecule

falls into chemical space of training sets, thereby determining the reliability of the predicted results. According to the definition of AD, the d_{ave} and d_{std} of each dataset were generated by calculating the distance matrix. Next, to suit the different characteristics among the modelling data from anticancer targets, academia-sourced cancer cell lines, and NCI-60 cancer cell lines (e.g., the gap of several order of magnitude between some target-based datasets and NCI-60 cancer cell datasets), the empirical parameter Z was explanatively attempted for the suitable value to control the significance level. According to the different Z and D_c values, all test sets were divided into ID and OD sets. Then the trained models were used to perform prediction on the ID and OD compounds defined by different parameter strategies in the corresponding test set. As shown in Table S25, the prediction result illustrates that the evaluation AUC and ACC indicators in ID were improved and those in OD were worse when Z was set as 0.2 and k was set as 3 for both cancer cells and anticancer targets, suggesting Z parameter combination could distinguish the boundaries of AD appropriately (notice that those datasets with few compounds for evaluation were excluded for results statistics). The results also indicate that the definition of AD in this study is suitable for the FP-GNN models. The D_c value of each model was summarized in Tables S26–S28, which could provide judgment on whether it is an ID compound or not together with the prediction result.

3.5. Web interfaces and usage of DeepCancerMap

DeepCancerMap offers a user-friendly and practical interface with a fully functional backend. The front-end interface is implemented using HTML, JavaScript, CSS and Vue, and the tools including Celery, Nginx, and Elastic Search are applied to achieve superior back-end performance. Three modules are designed to support various anticancer drug discovery tasks, including activity profiling predicting, virtual screening, and similarity search, whose brief introduction and usage are shown on the home page and explained below.

3.5.1. Activity profiling predicting

The activity profiling predicting module enables rapid and accurate bioactivity prediction of the submitted compound against 346 academic cancer cell lines (including 32 normal cell lines) or 426 anticancer targets or 60 NCI tumor cancer cell lines of interest to user at once. Taking 1-(1H-benzo[*d*]imidazole-2-yl)-4-(4-hydroxy-3,5-dimethoxyphenyl)-1,4,5,7-tetrahydro-6H-pyrazolo[3,4-*b*]pyridin-6-one (compound **1**, not included in the NCI-60 modelling dataset) as an example [53], user can easily upload/paste the SMILES or draw the structure online with JSME editor to predict its inhibitory activity against NCI-60 tumor cell line panel (Fig. 6A). Once the computational task is completed, the results are visually and clearly displayed in table format and well plate-liked illustration (Fig. 6B and C), respectively. On the visualized result, each hole represents a prediction score from the constructed FP-GNN model for the corresponding tumor cell line, where lighter color stands for a higher predicted score, meaning it may show better inhibitory activity against the tumor cell, and vice versa. The predicted anticancer profiling results of compound **1** by DeepCancerMap were overall consistent with the experimental inhibition results from the NCI-60 screening platform (Table S29) [53], with an AUC value of 0.711 (Fig. 7A). Similar trends and prediction results (Fig. 7A and Table S29) can also be observed on compounds **2** (AUC = 0.696) and **3** (AUC = 0.687), which are also not included in the modelling dataset [53,54]. All such results indicate the accuracy and usability of the DeepCancerMap platform.

Meanwhile, to help users analyze the interpretation of result, the FP-GNN model related information, including the evaluation indicators, the

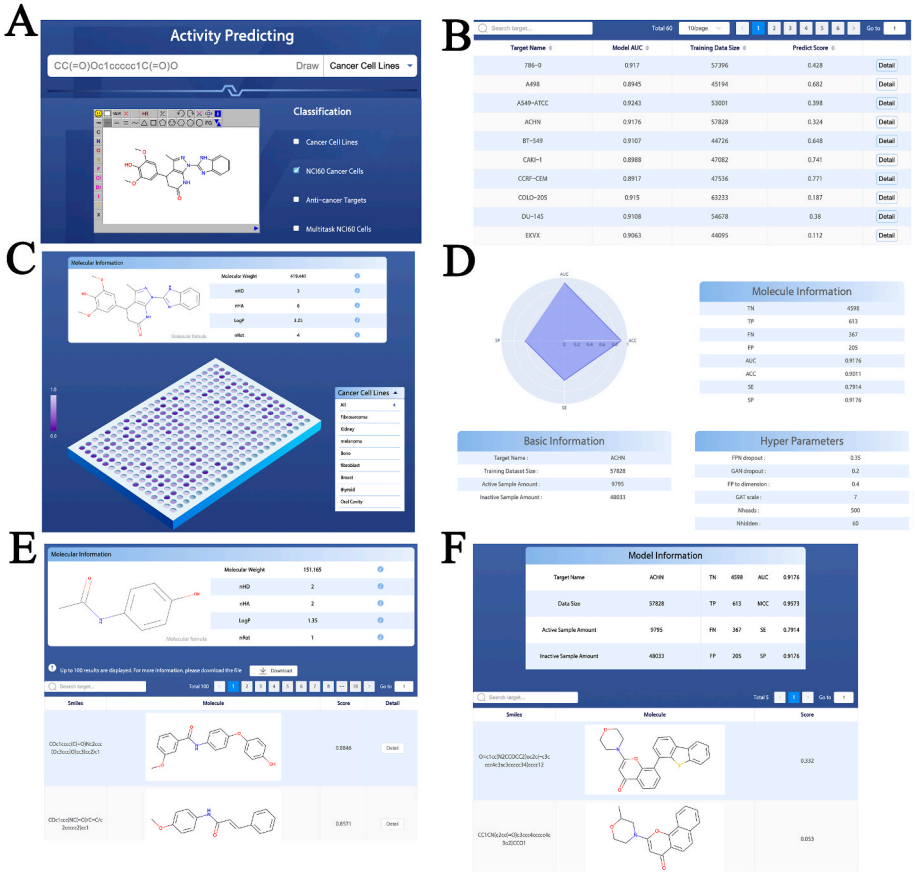
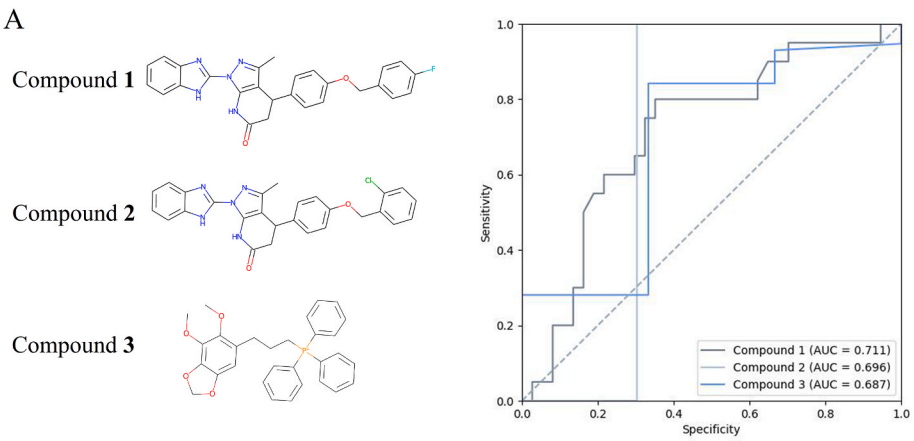


Fig. 6. Website interface schematic diagram of DeepCancerMap. In activity predicting module, compound could be submitted in diversified forms (A), and the detailed prediction results (B) with model performance information (C) would be displayed. (D) Represents the detail information of specific models that could be accessed. (E) and (F) represent the visualized result of virtual screening and similarity search, respectively.



B Compound 4

Cell Line	IC_{50} (μM) ± SD	Predicted score
CNE2	2.99 ± 0.30	0.691
A549	>50	0.231

C Compound 5

Cell Line	IC_{50} (μM) ± SD	Predicted score
CNE2	2.03 ± 0.33	0.707
A549	>50	0.116

Fig. 7. Experimental-based case studies for DeepCancerMap. (A) Represents the comparison analysis of profiling predicted results for compounds **1**, **2**, and **3** against 60 NCI tumor cell lines by DeepCancerMap with actual experimental test results through receiver operating characteristic (ROC) plot. (B) and (C) represent the comparison of experimental test and DeepCancerMap prediction results for compounds **4** and **5**, respectively. All these compounds are not included in the modelling datasets.

size and balance degree of training data, and the hyperparameters settings, is provided in a separate page. Users can access it by clicking the *DETAIL* button beside the score in the table below (Fig. 6C). If users are only interested in one class of cancer cell lines or targets, they can utilize the clustering display filter to show wells of the specific category only (Fig. 6B).

3.5.2. Virtual screening

In this module, users can select a cancer cell line or target of interest to screen their own compounds library to perform virtual screening online using DeepCancerMap platform. Similar to batch activity profiling predicting, the compound library (<2000 compounds) could be uploaded as a CSV formatted file. The screening results will be displayed on the web page in the form of a table (Fig. 6E). In addition, users can combine target- and cell phenotype-based virtual screening results to ensure that selected compounds show inhibitory activity not only at the target level, but also at the cancer cell level. This is critical for subsequent lead compound optimization and mechanism exploration. More importantly, all of FP-GNN DL models for 426 targets and 346 cancer cell lines (including normal cell lines) are packed into locally executable Python software for large-scale virtual screening campaigns. A fully functional and annotated program are freely available at https://deepcancer.idruglab.cn.

3.5.3. Similarity search

Based on the collected comprehensive data related to anticancer targets and cancer cell lines, three fingerprints-based molecular similarity search methods such as Morgan, MACCS, and AtomPairs fingerprints were implemented in the DeepCancerMap platform. Tanimoto coefficient is used to measure the similarity score of two comparison molecules. Obviously, this module allows users to quickly find compounds similar to the template molecule of interest, which ultimately fish their potential targets or anticancer activity against specific tumor cells. Detailed similarity search results can be summarized in table online and downloaded in CSV format (Fig. 6F).

3.6. Independent external evaluation

To further verify the effectiveness of DeepCancerMap platform, both experimental-based case study and external test set evaluation were conducted as independent additional test. In the experimental-based case study, compounds **4** (Figs. 7B) and **5** (Fig. 7C) identified as novel anticancer agents in our group were used as embodiment cases. Compounds **4** and **5** showed antiproliferative activity against CNE2 tumor cell line with IC_{50} values of 2.99 and 2.03 μM, but they exhibited weak inhibitory activity against A549 cells (IC_{50} values are greater than 50 μM). Two compounds were then submitted on DeepCancerMap platform to predict their inhibition activities against CNE2 and A549 tumor cell lines. Compounds **4** and **5** were predicted as actives for CNE2 cells with the predicted scores of 0.691 and 0.707, while the two compounds were predicted as inactives for A549 cells due to the lower predicted scores of 0.231 and 0.116, respectively. Clearly, the prediction of inhibitory activity at the cancer cell level for the two compounds was consistent with the experimental results (Fig. 7). Additionally, based on Morgan fingerprint, the maximum structural similarity scores between the two compounds and the existing compounds from CEN2 cell modelling dataset were 0.194 and 0.213, indicating that DeepCancerMap could contribute to the discovery of anticancer agents with novel structures.

Conducting the identical approach of collecting and standardizing cytotoxicity experimental data point from ChEMBL database, the bioactivity of small molecules against cancer cells published between September 2021 to March 2023 were integrated as a time-split external test set. Compounds included in training set, validation set, or test set were eliminated to ensure the independence of the external test set. In total, 89,685 cytotoxicity data points on 237 cancer cells under investigation were included in the external test sets. As shown in Table S30, the average AUC value for these time-split external test sets is 0.70. Furthermore, molecules with unexposed scaffolds account for over 49% in average of the external test sets (Table S30). All these results confirmed that these FP-GNN models in DeepCancerMap have promising predictive capability and application value.

4. Conclusion

Designing and discovering new anticancer drugs has been a formidable challenge. In this study, we introduced a versatile, user-friendly webserver, termed DeepCancerMap, which, to our best knowledge, is the first comprehensive platform based on high-quality deep learning models for anticancer drug discovery. Currently, DeepCancerMap consists of 832 robust and accurate predictive models trained on the carefully curated modelling datasets from target- and cell-based pharmacological screening data using the advanced FP-GNN DL method. DeepCancerMap and its local Python software can be used to perform various anticancer drug-related tasks such as virtual screening, inhibitory activity profiling prediction against targets or tumor cell lines, target fishing, and drug repositioning. Collectively, we believe that DeepCancerMap, as a one-stop intelligent platform, can accelerate the design and discovery of new anticancer drugs for experts and non-experts in the field. In addition, we will update DeepCancerMap annually to cover more targets and tumor cells.

Data and code availability

The full datasets, python software package and DeepCancerMap web server are freely available on https://deepcancermap.idruglab.cn/.

Author contributions

Ling Wang conceived and designed the project. Jingxing Wu, Yi Xiao and Mujie Lin contributed to the literature search, data collection, and algorithm architecture realization. Duancheng Zhao was response for analyzing the modelling results and implementation models to webserver. Yirui Li and Hailin Luo were in charge of web-based software construction on front-end and back-end respectively. Chuanqi Tang offered the artistic design. Ling Wang provided support and critically revised the manuscript. All authors have read and agreed to the published version of the manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (81973241) and the Natural Science Foundation of Guangdong Province (2023B1515020042 and 2020A1515010548). We acknowledge the allocation time from the SCUTGrid at South China University of Technology.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ejmech.2023.115401.

References

[1] T. Vos, S.S. Lim, C. Abbafati, et al., Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019, Lancet 396 (2020) 1204–1222.
[2] H.C. Kolberg, C. Villena-Heinsen, M.M. Deml, S. Kraemer, K. Diedrich, M. Friedrich, Relationship between chemotherapy with paclitaxel, cisplatin, vinorelbine and titanocene dichloride and expression of proliferation markers and tumour suppressor gene p53 in human ovarian cancer xenografts in nude mice, Eur. J. Gynaecol. Oncol. 26 (2005) 398–402.
[3] A. Saxena, Cancer chemotherapy and its side effect management, Nurs. J. India 97 (2006) 109–110.
[4] Y. Kato, J. Sato, R. Kato, R. Takata, W. Obara, [Side effect and supportive care to combination of gemcitabine and cisplatin chemotherapy for the advanced urothelial cancer], Nihon Rinsho Jpn. J. Clin. Med. 73 (Suppl 2) (2015) 609–613.
[5] R. Perez-Tomas, Multidrug resistance: retrospect and prospects in anti-cancer drug treatment, Curr. Med. Chem. 13 (2006) 1859–1876.
[6] A.N. Cheng, Y.-K. Lo, Y.-S. Lin, T.-K. Tang, C.-H. Hsu, J.T.-A. Hsu, A.Y.-L. Lee, Identification of novel Cdc7 kinase inhibitors as anti-cancer agents that target the interaction with dbf4 by the fragment complementation and drug repositioning approach, EBioMedicine 36 (2018) 241–251.
[7] D.C. Hooper, G.A. Jacoby, Mechanisms of drug resistance: quinolone resistance: mechanisms of quinolone resistance, Ann. N. Y. Acad. Sci. 1354 (2015) 12–31.
[8] CHEMBL Database Release 29, EMBL-EBI, 2021.
[9] F. Samsdodd, Target-based drug discovery: is something wrong? Drug Discov. Today 10 (2005) 139–147.
[10] G.E. Croston, The utility of target-based discovery, Expet Opin. Drug Discov. 12 (2017) 427–429.
[11] W. Zhang, L. Wang, L. Zhang, W. Chen, X. Chen, M. Xie, G. Yan, X. Hu, J. Xu, J. Zhang, Synthesis and biological evaluation of steroidal derivatives as selective inhibitors of AKR1B10, Steroids 86 (2014) 39–44.
[12] L. Wang, X. Pang, Y. Li, Z. Zhang, W. Tan, RADER: a RApid DEcoy Retriever to facilitate decoy based assessment of virtual screening, Bioinformatics (2016) btw783.
[13] J.G. Moffat, F. Vincent, J.A. Lee, J. Eder, M. Prunotto, Opportunities and challenges in phenotypic drug discovery: an industry perspective, Nat. Rev. Drug Discov. 16 (2017) 531–543.
[14] J. Shang, X. Dai, Y. Li, M. Pistolozzi, L. Wang, HybridSim-VS: a web server for large-scale ligand-based virtual screening using hybrid similarity recognition techniques, Bioinformatics 33 (2017) 3480–3481.
[15] Y. Luo, L. Wang, Discovery and development of ATP-competitive mTOR inhibitors using computational approaches, Curr. Pharmaceut. Des. 23 (2017).
[16] L. Chen, L. Wang, Q. Gu, J. Xu, An in silico protocol for identifying mTOR inhibitors from natural products, Mol. Divers. 18 (2014) 841–852.
[17] D.E. Szymkowski, Too many targets, not enough target validation, Drug Discov. Today 6 (2001) 397.
[18] D.C. Swinney, J. Anthony, How were new medicines discovered? Nat. Rev. Drug Discov. 10 (2011) 507–519.
[19] E.L. Berg, The future of phenotypic drug discovery, Cell Chem. Biol. 28 (2021) 424–430.
[20] H. Yuan, S. Myers, J. Wang, D. Zhou, J.A. Woo, B. Kallakury, A. Ju, M. Bazylewicz, Y.M. Carter, C. Albanese, N. Grant, A. Shad, A. Dritschilo, X. Liu, R. Schlegel, Use of reprogrammed cells to identify therapy for respiratory papillomatosis, N. Engl. J. Med. 367 (2012) 1220–1227.
[21] S. Neumann, W. Huang, S. Titus, G. Krause, G. Kleinau, A.T. Alberobello, W. Zheng, N.T. Southall, J. Inglese, C.P. Austin, F.S. Celi, O. Gavrilova, C.J. Thomas, B. M. Raaka, M.C. Gershengorn, Small-molecule agonists for the thyrotropin receptor stimulate thyroid function in human thyrocytes and mice, Proc. Natl. Acad. Sci. USA 106 (2009) 12471–12476.
[22] D.C. Swinney, J.A. Lee, Recent advances in phenotypic drug discovery, F1000Research. 9 (2020) 944.
[23] S. Ekins, J. Mestres, B. Testa, *In silico* pharmacology for drug discovery: applications to targets and beyond: *In silico* pharmacology for drug discovery, Br. J. Pharmacol. 152 (2007) 21–37.
[24] J. Burbaum, G.M. Tobal, Proteomics in drug discovery, Curr. Opin. Chem. Biol. 6 (2002) 427–433.
[25] X. Chen, W. Xie, Y. Yang, Y. Hua, G. Xing, L. Liang, C. Deng, Y. Wang, Y. Fan, H. Liu, T. Lu, Y. Chen, Y. Zhang, Discovery of dual FGFR4 and EGFR inhibitors by machine learning and biological evaluation, J. Chem. Inf. Model. 60 (2020) 4640–4652.
[26] M. Yang, B. Tao, C. Chen, W. Jia, S. Sun, T. Zhang, X. Wang, Machine learning models based on molecular fingerprints and an extreme gradient boosting method lead to the discovery of JAK2 inhibitors, J. Chem. Inf. Model. 59 (2019) 5002–5012.
[27] F. Miljković, R. Rodríguez-Pérez, J. Bajorath, Machine learning models for accurate prediction of kinase inhibitors with different binding modes, J. Med. Chem. 63 (2020) 8738–8748.
[28] L. Wang, Y. Li, M. Xu, X. Pang, Z. Liu, W. Tan, J. Xu, Chemical fragment-based CDK4/6 inhibitors prediction and web server, RSC Adv. 6 (2016) 16972–16981.
[29] X. Li, Z. Li, X. Wu, Z. Xiong, T. Yang, Z. Fu, X. Liu, X. Tan, F. Zhong, X. Wan, D. Wang, X. Ding, R. Yang, H. Hou, C. Li, H. Liu, K. Chen, H. Jiang, M. Zheng, Deep learning enhancing kinome-wide polypharmacology profiling: model construction and experiment validation, J. Med. Chem. 63 (2020) 8723–8737.
[30] B. Merget, S. Turk, S. Eid, F. Rippmann, S. Fulle, Profiling prediction of kinase inhibitors: toward the virtual assay, J. Med. Chem. 60 (2017) 474–485.
[31] Y. Luo, R. Zeng, Q. Guo, J. Xu, X. Sun, L. Wang, Identifying a novel anticancer agent with microtubule-stabilizing effects through computational cell-based bioactivity prediction models and bioassays, Org. Biomol. Chem. 17 (2019) 1519–1530.
[32] Q. Guo, Y. Luo, S. Zhai, Z. Jiang, C. Zhao, J. Xu, L. Wang, Discovery, biological evaluation, structure–activity relationships and mechanism of action of pyrazolo [3,4- *b*]pyridin-6-one derivatives as a new class of anticancer agents, Org. Biomol. Chem. 17 (2019) 6201–6214.

[33] S. He, D. Zhao, Y. Ling, H. Cai, Y. Cai, J. Zhang, L. Wang, Machine learning enables accurate and rapid prediction of active molecules against breast cancer cells, Front. Pharmacol. 12 (2021).
[34] A. Zhavoronkov, Y.A. Ivanenkov, A. Aliper, M.S. Veselov, V.A. Aladinskiy, A. V. Aladinskaya, V.A. Terentiev, D.A. Polykovskiy, M.D. Kuznetsov, A. Asadulaev, Y. Volkov, A. Zholus, R.R. Shayakhmetov, A. Zhebrak, L.I. Minaeva, B. A. Zagribelnyy, L.H. Lee, R. Soll, D. Madge, L. Xing, T. Guo, A. Aspuru-Guzik, Deep learning enables rapid identification of potent DDR1 kinase inhibitors, Nat. Biotechnol. 37 (2019) 1038–1040.
[35] A. Yoshimori, Y. Asawa, E. Kawasaki, T. Tasaka, S. Matsuda, T. Sekikawa, S. Tanabe, M. Neya, H. Natsugari, C. Kanai, Design and synthesis of DDR1 inhibitors with a desired pharmacophore using deep generative models, ChemMedChem 16 (2021) 955–958.
[36] X. Tan, C. Li, R. Yang, S. Zhao, F. Li, X. Li, L. Chen, X. Wan, X. Liu, T. Yang, X. Tong, T. Xu, R. Cui, H. Jiang, S. Zhang, H. Liu, M. Zheng, Discovery of pyrazolo[3,4-d] pyridazinone derivatives as selective DDR1 inhibitors via deep learning based design, synthesis, and biological evaluation, J. Med. Chem. 65 (2022) 103–119.
[37] Z. Wang, T. Ran, F. Xu, C. Wen, S. Song, Y. Zhou, H. Chen, X. Lu, Deep learning-driven scaffold hopping in the discovery of Akt kinase inhibitors, Chem. Commun. Camb. Engl. 57 (2021) 10588–10591.
[38] M. Assmann, M. Bal, M. Craig, J. D'Oyley, L. Phillips, H. Triendl, P.A. Bates, U. Bashir, P. Ruprah, N. Shaker, V. Stojevic, A novel machine learning approach uncovers new and distinctive inhibitors for cyclin-dependent kinase 9, BioRxiv (2020), 2020.03.18.996538.
[39] Y. Zhou, Y. Zhang, X. Lian, F. Li, C. Wang, F. Zhu, Y. Qiu, Y. Chen, Therapeutic target database update 2022: facilitating drug discovery with enriched comparative data of targeted agents, Nucleic Acids Res. 50 (2022) D1398–D1407.
[40] M. Zheng, Z. Liu, X. Yan, Q. Ding, Q. Gu, J. Xu, LBVS: an online platform for ligand-based virtual screening using publicly accessible databases, Mol. Divers. 18 (2014) 829–840.
[41] X. Zeng, P. Zhang, Y. Wang, C. Qin, S. Chen, W. He, L. Tao, Y. Tan, D. Gao, B. Wang, Z. Chen, W. Chen, Y.Y. Jiang, Y.Z. Chen, CMAUP: a database of collective molecular activities of useful plants, Nucleic Acids Res. 47 (2019) D1118–D1127.
[42] F.R. Fields, S.D. Freed, K.E. Carothers, M.N. Hamid, D.E. Hammers, J.N. Ross, V. R. Kalwajtys, A.J. Gonzalez, A.D. Hildreth, I. Friedberg, S.W. Lee, Novel antimicrobial peptide discovery using machine learning and biophysical selection of minimal bacteriocin domains, Drug Dev. Res. 81 (2020) 43–51.
[43] Q. Ye, X. Chai, D. Jiang, L. Yang, C. Shen, X. Zhang, D. Li, D. Cao, T. Hou, Identification of active molecules against *Mycobacterium tuberculosis* through machine learning, Briefings Bioinf. 22 (2021) bbab068.
[44] G.-H. Li, J.-F. Huang, CDRUG: a web server for predicting anticancer activity of chemical compounds, Bioinforma. Oxf. Engl. 28 (2012) 3334–3335.
[45] H. Cai, H. Zhang, D. Zhao, J. Wu, L. Wang, FP-GNN: a versatile deep learning architecture for enhanced molecular property prediction, Briefings Bioinf. (2022) bbac408.
[46] Z. Wu, B. Ramsundar, E.N. Feinberg, J. Gomes, C. Geniesse, A.S. Pappu, K. Leswing, V. Pande, MoleculeNet: a benchmark for molecular machine learning, Chem. Sci. 9 (2018) 513–530.
[47] L. Wang, X. Le, L. Li, Y. Ju, Z. Lin, Q. Gu, J. Xu, Discovering new agents active against methicillin-resistant Staphylococcus aureus with ligand-based approaches, J. Chem. Inf. Model. 54 (2014) 3186–3197.
[48] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, in: Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., ACM, San Francisco California USA, 2016, pp. 785–794.
[49] D. Chicco, G. Jurman, The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation, BMC Genom. 21 (2020) 6.
[50] A. Tropsha, 4.07 - predictive quantitative structure–activity relationship modeling, in: J.B. Taylor, D.J. Triggle (Eds.), Compr. Med. Chem. II, Elsevier, Oxford, 2007, pp. 149–165.
[51] B. Bienfait, P. Ertl, JSME: a free molecule editor in JavaScript, J. Cheminf. 5 (2013) 24.
[52] D. Ai, J. Wu, H. Cai, D. Zhao, Y. Chen, J. Wei, J. Xu, J. Zhang, L. Wang, A multi-task FP-GNN framework enables accurate prediction of selective PARP inhibitors, Front. Pharmacol. 13 (2022), 971369.
[53] B.V. Lichitsky, A.N. Komogortsev, A.A. Dudinov, M.M. Krayushkin, E.N. Khodot, A. V. Samet, E.A. Silyanova, L.D. Konyushkin, A.S. Karpov, D. Gorses, T. Radimerski, M.N. Semenova, A.S. Kiselyov, V.V. Semenov, Benzimidazolyl-pyrazolo[3,4- *b*] pyridinones, selective inhibitors of MOLT-4 leukemia cell growth and sea urchin embryo spiculogenesis: target quest, ACS Comb. Sci. 21 (2019) 805–816.
[54] D.V. Tsyganov, A.V. Samet, E.A. Silyanova, V.I. Ushkarov, A.E. Varakutin, N. B. Chernysheva, R.N. Chuprov-Netochin, A.A. Khomutov, A.S. Volkova, S. V. Leonov, M.N. Semenova, V.V. Semenov, Synthesis and antiproliferative activity of triphenylphosphonium derivatives of natural allylpolyalkoxybenzenes, ACS Omega 7 (2022) 3369–3383.