PLOS ONE

RESEARCH ARTICLE

A novel drug repurposing approach for non-small cell lung cancer using deep learning

Bingrui Li☯, Chan Dai☯, Lijun Wang, Hailong Deng, Yingying Li*, Zheng Guan*, Haihong Ni*

Beijing Deep Intelligent Pharma Technologies Co., Ltd, Beijing, China

☯ These authors contributed equally to this work.
* yyli285@iccas.ac.cn (YL); guanzheng828@hotmail.com (ZG); nihaihong@dip-ai.com (HN)

OPEN ACCESS

Citation: Li B, Dai C, Wang L, Deng H, Li Y, Guan Z, et al. (2020) A novel drug repurposing approach for non-small cell lung cancer using deep learning. PLoS ONE 15(6): e0233112. https://doi.org/10.1371/journal.pone.0233112

Editor: Pinyi Lu, Biotechnology HPC Software Applications Institute (BHSAI), UNITED STATES

Received: November 19, 2019

Accepted: April 28, 2020

Published: June 11, 2020

Data Availability Statement: All relevant data are within the manuscript and its supporting informaiton files.

Funding: Beijing Deep Intelligent Pharma Technologies Co., Ltd. provided support for this study in the form of salaries for authors: BRL, CD, LJW, HLD, ZG, YYL, and HNN. The specific roles of these authors are articulated in the 'author contributions' section. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. No

Abstract

Drug repurposing is an attractive and pragmatic way offering reduced risks and development time in the complicated process of drug discovery. In the past, drug repurposing has been largely accidental and serendipitous. The most successful examples so far have not involved a systematic approach. Nowadays, remarkable advances in drugs, diseases and bioinformatic knowledge are offering great opportunities for designing novel drug repurposing approach through comprehensive understanding of drug information. In this study, we introduced a novel drug repurposing approach based on transcriptomic data and chemical structures using deep learning. One strong candidate for repurposing has been identified. Pimozide is an anti-dyskinesia agent that is used for the suppression of motor and phonic tics in patients with Tourette's Disorder. However, our pipeline proposed it as a strong candidate for treating non-small cell lung cancer. The cytotoxicity of pimozide against A549 cell lines has been validated.

Introduction

Although the knowledge and technology of human diseases have developed substantially, the translation of these benefits into therapeutic innovations has been far slower than expected [1, 2]. The main challenges facing global pharmaceutical industries are substantial costs and long consuming time in the process of drug discovery and development [2]. To solve this problem, drug repurposing (also known as drug repositioning, re-tasking or reprofiling) has emerged as an attractive and pragmatic way offering reduced risks and development time [3]. In the past, drug repurposing has been largely accidental and serendipitous. The most successful examples so far have not involved a systematic approach. Nowadays, the boom of drugs, diseases and bioinformatic knowledge is offering great opportunities for designing novel drug repurposing approach [4–8].

On one hand, a number of approaches try to identify an analogue of an existing drug molecule sharing mechanisms of action with the original drug [2, 9, 10]. It means that drugs belong to different therapeutic use classes with similar chemical structures, are most likely to have same indications. This principle can find alternative targets of existing drugs and uncover

additional external funding was received for this study.

Competing interests: The authors of this paper have read the journal's policy and have the following All authors are paid employees of Beijing Deep Intelligent Pharma Technologies Co., Ltd. (https://www.dip-ai.com/). There are no patents, products in development or marketed products associated with this research to declare. This does not alter our adherence to PLOS ONE policies on sharing data and materials.

Abbreviations: BGL, Boost Graph Library; **BIRCH**, Balanced Iterative Reducing Clustering using Hierarchies; **Di-PASS**, DIP®-Pathway Activation Scoring System; **DNNs**, Deep neural networks; **EGFR**, Epidermal growth factor receptor; **FCFP**, Functional-Class Fingerprint; **FDA**, U.S. Food & Drug Administration; **GSEA**, Gene Set Enrichment Analysis; **HIF**, Hypoxia-inducible factor; **IC_{50}**, Half maximal inhibitory concentration; **iPANDA**, *in silico* Pathway Activation Network Decomposition Analysis; **KEGG**, Kyoto Encyclopedia of Genes and Genomes; **LEMONS**, Library for the Enumeration of MOdular Natural Structures; **LINCS**, Library of Integrated Network-Based Cellular Signatures; **MeSH**, Medical Subject Headings; **NIH**, National Institutes of Health; **NSCLC**, Non-small cell lung cancer; **NTG**, Nitroglycerin; **OD**, Optical density; **OPTICS**, Ordering Points To Identify the Clustering Structure; **ReLU**, Rectified linear unit; **RF**, Random forest; **SRB**, Sulforhodmide-B.

potential off-target effects that can be investigated for clinical applications [11]. On the other hand, regardless of the similarity in drug structures, a similar transcriptomic signature between two drugs may share the same indications [2]. These approaches make use of gene expression data before and after drug perturbations to construct a network of transcriptional response and functional properties of drugs [12–14]. Ideally a drug may have the potential to cure a disease if the differential expression profile under drug perturbations is anti-correlated significantly [15].

In this study, we leverage both chemical structures and transcriptomic data for repurposing drugs. The potential candidate drugs identified with both chemical structural and transcriptomic signatures, are more likely to successfully progress through the drug discovery lifecycle. The approach includes the classifying process and the repurposing process. The former relies on a transcriptomic database, based on deep neural networks (DNNs) that is trained on large sets of perturbation samples of X drugs from the Library of Integrated Network-Based Cellular Signatures (LINCS) and links those to 6 therapeutic use classification derived from Medical Subject Headings (MeSH) therapeutic use categories. Based on transcriptional profiles as drug classification, we predict misclassified antineoplastic drugs which belong to another therapeutic use categories in MeSH. Then by chemical structure similarity comparison with non-small cell lung cancer (NSCLC) drugs approved by U.S. Food & Drug Administration (FDA), we obtain a list of potential drugs ranked by the estimation of both pathway activation score and structure similarity score. The proposed candidates for NSCLC are validated experimentally in the lab and discussed in the results. The approach established in this article can be extended to other diseases, holding great promise for shortening the drug discovery process.

Materials and methods

Drug data

MeSH classification (https://www.nlm.nih.gov/mesh/) was utilized to symbolize drugs profiled from LINCS. Only the drugs with the linkage to one disease in "therapeutic use" section were chosen in this study. We used the perturbation samples of 75 drugs and linked them to 6 therapeutic use categories derived from MeSH. The 6 therapeutic use categories include vasodilator agents, anti-dyskinesia agents, anticonvulsants, hypolipidemic agent, anti-asthmatic agent, and antineoplastic agents.

Gene data

In this work, the gene expression data was obtained from the LINCS Project participants (http://www.lincsproject.org/) which was downloaded from NCBI Gene Expression Omnibus (GEO,GSE70138 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE70138)). LINCS is a National Institutes of Health (NIH) program which provides gene-expression profiles across multiple cells and perturbation types, as well as read-outs, at a massive scale. The level 3 (Q2NORM) gene expression of three cell lines: A375, HELA and HT29 were used. (GSE70138_Broad_LINCS_Level3_INF_mlr12k_n345976x12328_2017-03-06.gctx.gz.) The level 3 (Q2NORM) are gene expression profiles of both directly measured landmark transcripts and inferred gene. Landmark genes were normalized by invariant set scaling followed by quantile normalization. The final gene expression dataset has 12,797 genes.

Signaling pathways and topological weight

We obtained signaling pathways from Kyoto Encyclopedia of Gene Genomics (KEGG) (https://www.kegg.jp/kegg/pathway.html). A signaling pathway in KEGG is graphed with

nodes and directed edges representing genes or proteins, and signal relationships, respectively. The edges are weighted according to activation and inhibition. In this study, we obtained a set of 164 signaling pathways covering 4414 unique genes. The contributions of genes were multiplied by an integer which equals to +1/-1 and +2/-2 in the case of pathway activation/suppression and phosphorylation/dephosphorylation, respectively. KEGG graph (http://www.bioconductor.org/packages/release/bioc/html/KEGGgraph.html) and Boost Graph Library (BGL) R packages [16, 17] were downloaded from Bioconductor (https://www.bioconductor.org/packages/release/bioc/html/RBGL.html). Signaling pathway xml files from KEGG were used as the input for the calculation of topological weight.

Grouping genes into modules

Human database of coexpressed genes COXPRESdb [18] was used in order to obtain the gene modules. We used Euclidean distance matrix to cluster correlation data from COXPRESdb. We used the following equation to calculate distances:

$$R_{ij} = 1 - corr_{ij}$$

Where $corr_{ij}$ is the correlation between expression levels of genes i and j. Ordering Points To Identify the Clustering Structure (OPTICS) [19] was firstly used to identify clusters. Clusters with the average internal pairwise correlation below 0.5 were excluded in our study. Then we merged all the clusters from OPTICS into a group and applied Balanced Iterative Reducing Clustering using Hierarchies (BIRCH) [20] to cluster the group again. Finally, a data set of 177 gene modules including 1271 unique genes was built.

Pathway score matrix

The pathway score matrixes were calculated based on *in silico* pathway activation network decomposition analysis (iPANDA) algorithm [21] with coexpressed genes. The input files were KEGG signaling pathway topological weights and gene list. The output file was an activation score matrix with 3985 samples in 164 signaling pathways.

Classifying process

We selected all samples corresponding to different perturbation concentration, time and cell lines. "Landmark genes" and signaling pathways were new features for training our models. So-called "landmark genes" were derived from the LINCS project. For classification methods, we chose two robust and widely used methods including random forest (RF) and deep neural network (DNN) methods. The feature of RF is that it can build a large number of regression trees and average their predictions, allowing for high dimensional flexible modeling of interactions. Nevertheless, these parameters are not evident for a given data set [22]. Grid search was utilized for hyperparameter optimization. The RF has been trained via nested 3-fold cross validation, the first 66.7% columns of the full data matrix were selected as a training set, while the remaining 33.3% of the columns were reserved as the test set. Using 400 trees and default values for other parameters, the trained model was subsequently used to predict the class of the drugs for test datasets. In this study, DNN has fully connected multilayer perception with 978 input nodes for gene expression and 163 input nodes for signaling pathways. The DNN model was trained via 10-fold cross validation, the first 90% data were served as training set, while the remaining 10% data were reserved as test data. Grid search for hyperparameter optimization of optimal number of layers, nodes of each layer and dropout rejection rate were also utilized in order to compare the performance of RF. The number of layers were trained from 3 to 5,

while the nodes in each layer were started from 100 to 400 in steps of 50. Each layer was started by Glorot uniform approach [23]. At each layer, the dropout rejection ratio was 10% and 50%, respectively. Rectified linear unit (ReLU) [24], as nonlinear function of hidden neurons was utilized to speed up the process of the calculation, especially for the large input. The output nodes number was 6, which was the same with our number of classes. Finally, we found the number of 3 hidden layers with 150 in each with rectified linear activation function was the optimal parameter. The source code can be found on github (https://github.com/ai-diper/DNN-DR).

Repurposing process

Library for the Enumeration of MOdular Natural Structures (LEMONS) is a software package originally designed to enumerate hypothetical modular natural product structures [25]. However, in this study, we utilized LEMONS to compare the structure similarity between two compounds based on Functional-Class Fingerprint (FCFP). Tanimoto coefficient, as one of the best metrics for similarity calculations, was used to compare structure similarities of small molecules. After computing, the drugs from our repurposing drug pool with tanimoto coefficient above than 0.8 were selected for further analysis. The tanimoto coefficient of 0.8 reflected a high probability of two compounds sharing the same activity [26, 27]. Finally, all these potential drugs were ranked according to combinational score of Di-PASS score and chemical structure similarity score.

In vitro experiments

Two NSCLC cell lines (A549 and H157), were used for *in vitro* cytotoxic activity testing using the SRB assay [28]. The rapidly growing cells were harvested, counted, and incubated at the concentration of 1×10^4 cells/well in 96-well plates. After incubation for 48h, the compounds (including pimozide and gemcitabine) were dissolved in culture medium, applied to the culture wells in triplicate, and incubated for 48 h at 37˚C in a humidified incubator. The cultured cells were fixed with 50% (m/V) (50 mg/100 ml) trichloroacetic acid at 4 ˚C for 1 h. Then 0.4% SRB was dissolved in 1% acetic acid for each well for 30 min. 10 mM unbuffered Trisma base solution was used for solubilizing the bound stain with a gyratory shaker. The optical density (OD) of 515 nm was measured spectrophotometrically in a microplate reader. Cytotoxic activity was evaluated by identifying the concentration of compound that was required to inhibit cell growth by 50% (IC_{50}) [29]. Each experiment was operated for three times. We used the following equation to calculate IC_{50}:

$$IC_{50} = [(OD_{control} - OD_{compound})/OD_{control}] \times 100\%.$$

Ethics approval and consent to participate

Cells used in this study were obtained from National Infrastructure of Cell Line Resource and used according to the approved protocols.

Results and discussion

This paper aimed at providing a novel drug repurposing approach to discover alternative NSCLC drugs from existing drugs. In the repurposing process, we not only considered the chemical structural information of molecules, but also leveraged transcriptomic data for discovering new indications (Fig 1). All data including transcriptional data of the reference and perturbation samples were obtained from LINCS. Pathway activation scores were computed

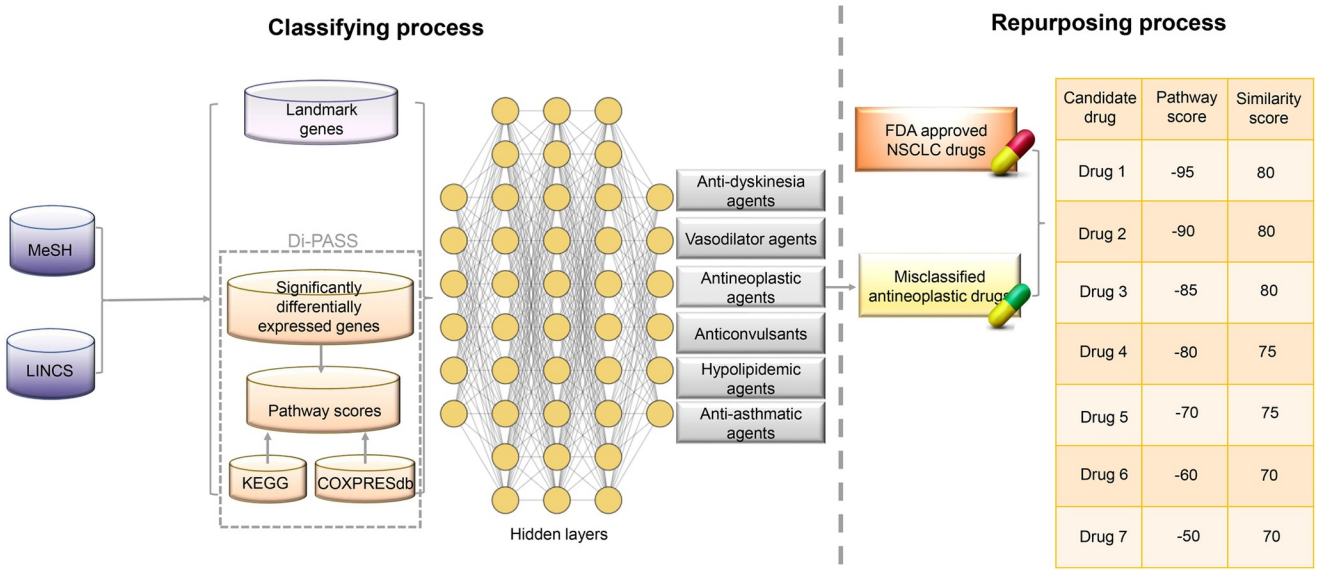

Fig 1. The schematic illustration of study design.

https://doi.org/10.1371/journal.pone.0233112.g001

by DIP® Pathway Activation Scoring System (Di-PASS) based on differentially expressed gene data, Kyoto Encyclopedia of Genes and Genomes (KEGG), and COXPRESdb. In order to reduce biologically relevant dimensions, landmark genes and signaling pathways were chosen as new features for predicting new therapeutic use categories. Based on the classification using transcriptional data only, several misclassified antineoplastic drugs were obtained, which had different therapeutic roles in MeSH. Through combinational ranking of both chemical structures and corresponding pathway activation scores, new potential drugs were discovered. Misclassified antineoplastic drugs are compared with FDA approved NSCLC drugs.

Di-PASS as a tool for pathway activation estimation

When analyzing and organizing data in high dimensional space, the amount of data required to provide a reliable analysis grows exponentially [30]. This phenomenon is called "curse of dimensionality", which is a big challenge in the fields of genomics. Common approaches [31,32] only recognize expression signature patterns of the process, thus failing to capture vital differences among samples that come from complex gene network interactions. To circumvent this obstacle, various pathway analysis techniques have been proposed which can integrate gene expression data into molecular signaling networks, thus reducing biologically relevant dimensions [31]. Some widely used pathway-based algorithms, for instance, Gene Set Enrichment Analysis (GSEA), solely focuses on gene enrichment statistics, without structured sets of genes as pathways [32]. There is still an urgent need to invent a novel analytical method which can build an accurate network of biological signaling from complex transcriptomic data. In this study, we introduced Di-PASS algorithm that integrated different analytical concepts such as gene expression data, gene importance factor [33], and topological weights into a single network, simultaneously exploiting Di-PASS scores for pathway activation estimation (Fig 2). The gene expression data was obtained from the LINCS Project participants. The level 3 (Q2NORM) gene expression of three cell lines: A375, HELA and HT29 was used. After drug perturbation, genes with insignificant change expression were excluded from further analysis. It is well known that several genes exhibit closely connections in their expression level, which can be considered as gene coexpression [18, 34, 35]. COXPRESdb (http://coxpresdb.jp) is a

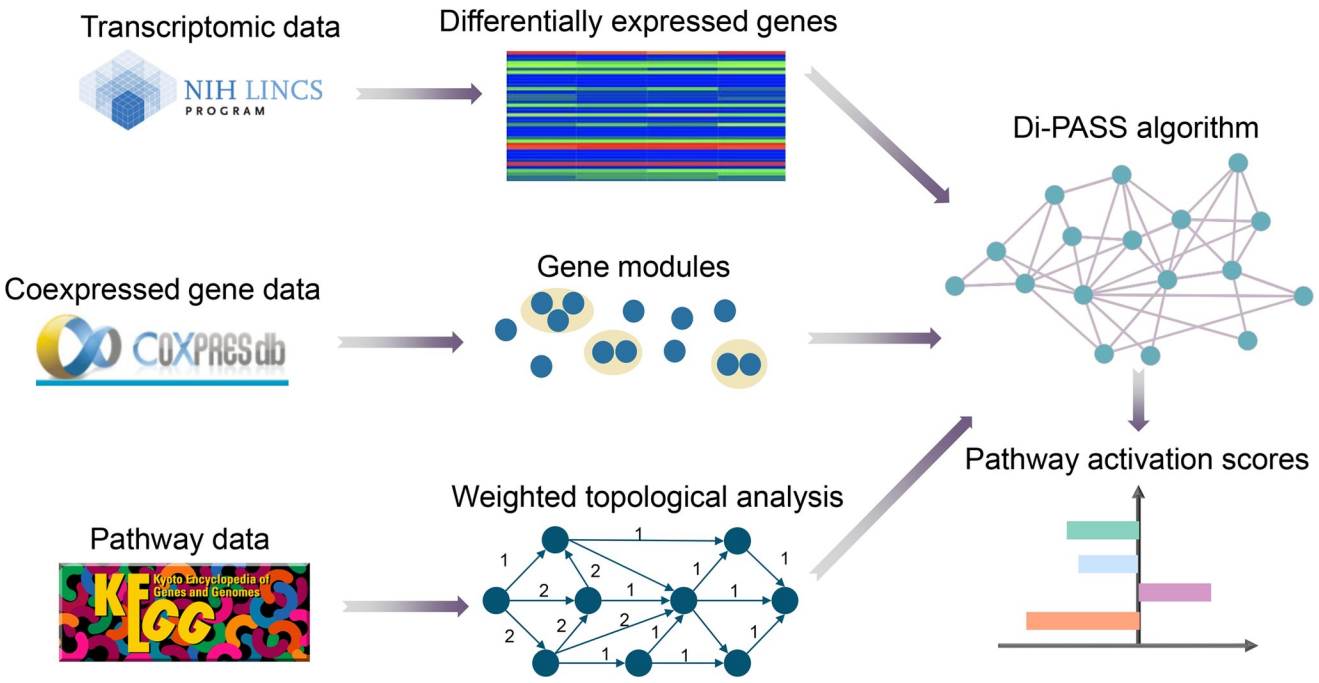

Fig 2. The general scheme of Di-PASS algorithm.

https://doi.org/10.1371/journal.pone.0233112.g002

database which first release coexpression information for human and mouse. In this work, we used COXPRESdb [18, 36] for moduling, because of its reduced false positive relationships in individual gene coexpression data. In addition, topological weight was introduced in order to give more weight to the genes that occupied the central position on the pathway map. In this study, the topological weight of each gene was set proportional to the number of independent paths through the pathway. Then the computation of topological weight was estimated for each module as a whole rather than for individual genes [21]. The pathway data was obtained from KEGG, which constructed manually curated pathway maps that represent current knowledge on biological networks in graph models [37]. The contribution of gene units (either individual genes or gene modules) to pathway activation was calculated as an output of their fold changes and topological weights. In the end, the output results, referring to pathway activation scores (Di-PASS scores), were produced as signed scores showing the direction and intensity of pathway activation.

The input data for Di-PASS algorithm includes three parts. The first part started from differential expression analysis. After drug perturbation, only significantly differentially expressed genes were selected for further analysis. Coexpressed gene data was obtained from COXPRESdb. Coexpressed genes were grouped into modules, and were given topological weights according to their importance. The pathway data was obtained from KEGG. The contribution of gene units (both gene modules and individual genes) to pathway activation was calculated as a product of their fold changes and topological weights. The contributions were multiplied by a discrete coefficient which equals to +1 or -1 in the case of pathway activation or suppression, and +2 or -2 in the case of phosphorylation or dephosphorylation, respectively. In the end, the output results, referring to pathway activation scores, were produced as signed scores showing the direction and intensity of pathway activation. The source code of Di-PASS is available (https://github.com/ai-diper/Di-PASS).

Classifying process

When dealing with transcriptional data, we used 3985 drug perturbation samples for 3 cell lines: A375, HELA and HT29 from the Broad LINCS database (S1 Table). All the samples were classified to 6 therapeutic use categories based on MeSH classification of 75 specified drugs (S2 Table). Those drugs belonging to only one category were chosen in our study. In order to solve the problem of "curse of dimensionality [30]", we utilized landmark genes and signaling pathways as new features to predict drug classifications.

For landmark genes analysis, we obtained a data set including normalized gene expression data for 978 landmark genes which captured approximately 80% of the information and possessed great inferential value, according to the authors of LINCS projects [38]. Based on this feature data, we built a classifier using deep learning methods to evaluate the performance of classification on complicated drug action patterns recognition across different therapeutic indications. DNN was the first classifier model we chose. As a high-level representation model of deep learning, DNN outperformed traditional machine learning approaches in the field of automatic task-optimal features learning from deluged data sets [12]. We compared the results from DNN with random forest (RF) via nested 3-fold cross validation for several hyperparameters. Finally, we chose 6 abundant categories: antineoplastic, vasodilator, anti-dyskinesia, anticonvulsants, hypolipidemic, and anti-asthmatic for the class classification. On 6-class classification, DNN and RF performed with mean F1 scores of 0.41 and 0.35, respectively (Fig 3).

For signaling pathways analysis, we used Di-PASS algorithm that performed quantitative estimation of pathway activation intensity. The sign of the Di-PASS scores indicated whether the pathway was activated or suppressed. All the same perturbation samples with those used in landmark genes analysis were estimated with this tool and each sample was computed for 164 signaling pathways (see Methods section for more details). DNN trained on signaling pathway data performed with 10-fold cross-validation mean F1 score of 0.43, while RF performed 0.34 (Fig 3). The results indicated that DNN outperformed RF in classification analysis.

Considering the low mean F1 score based on the single feature, we combined landmark genes and signaling pathways as new features for classification. DNN performed with mean F1 score of 0.59, while RF performed 0.46 (Fig 3). The results indicated that the combinational features with landmark genes and signaling pathways outperformed the single feature-based methods. DNN was more suitable to classify drugs into therapeutic use categories rather than

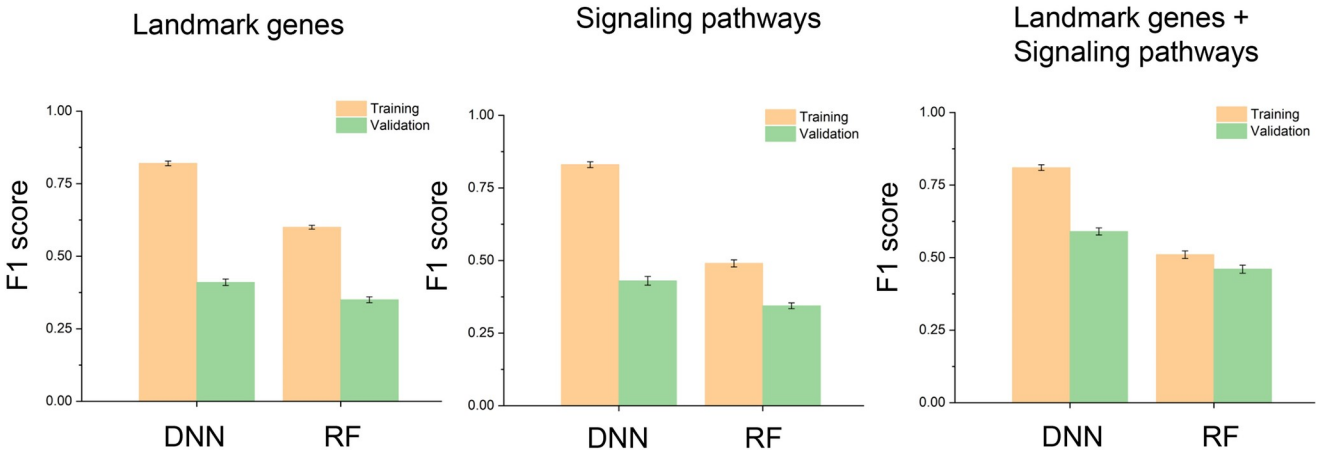

Fig 3. Classification results. Classification performance of DNN and RF trained on landmark gene, signaling pathways, and combinatorial features for 6 drug classes, respectively. Training and validation set were shown in orange and green colors.

https://doi.org/10.1371/journal.pone.0233112.g003

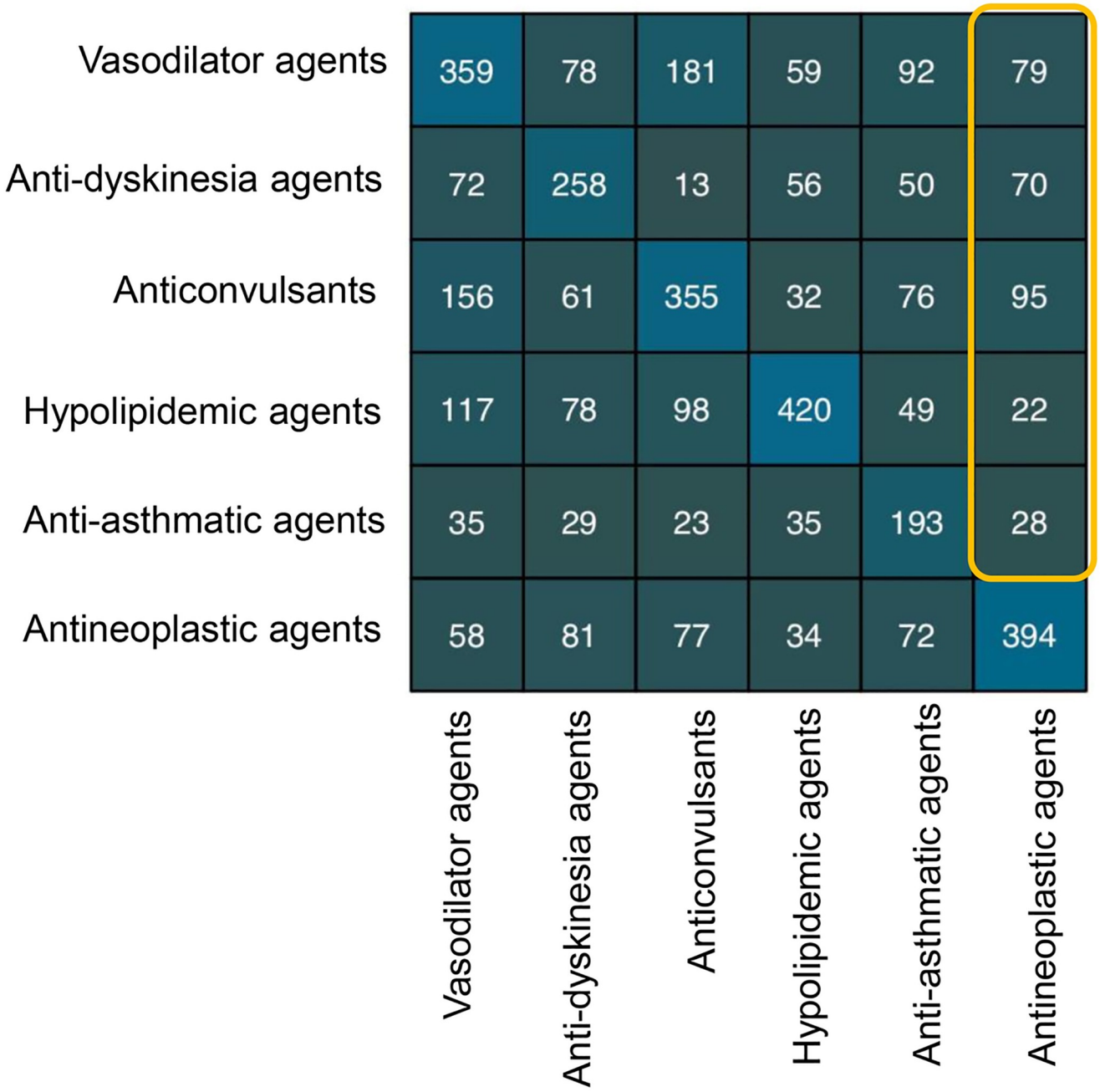

Fig 4. Validation confusion matrix. Validation confusion matrix illustrated DNN classification performance over a set of drugs profiled for A375, HELA and HT29 cell lines, belonging to 6 therapeutic use categories in MeSH. C(i,j) element was the number of samples of how many times i was the truth and j was predicted. The box circled the misclassified false positive drugs.

https://doi.org/10.1371/journal.pone.0233112.g004

RF. Thus, DNN with combinational features was selected as the optimal condition for therapeutic use prediction. Finally, we illustrated separability of therapeutic use categories by confusion matrix (Fig 4) utilizing DNN model with landmark genes and signaling pathways as combinational features. Here we observed that vasodilator agents were relatively often misclassified as anticonvulsants and antineoplastic agents. However, the imperfect accuracy here may not be a bad thing. These misclassified false positive drugs are likely to represent a possibility

for drug repurposing. For example, nitroglycerin (NTG), as vasodilator agents according to the MeSH therapeutic use section, was discovered to be an antineoplastic agent for its capability of reducing hypoxia-inducible factor (HIF)-1α levels in hypoxic tumor tissues [39]. Another misclassified example is vasodilator agent propranolol, which displays remarkable anti-cancer effects on cellular proliferation and invasion [40]. Here we chose misclassified antineoplastic agents as potential drugs for new indication of treating NSCLC (S3 Table).

Repurposing process

The commonly used drug repurposing approach is the analogue based approach that aims to identify an analogue of an existing drug molecule sharing mechanisms of action with the original drug [2, 9]. The similarity of chemical structure may suggest shared biological or physicochemical activities. Thanks to iterative improvements, the process of using marketed drug structures as basis for repurposing, leads to increased safety and efficacy of therapeutic agents [41]. Our repurposing process started with finding similar chemical structured drugs with FDA approved drugs. First, we identify an original compound from FDA approved drugs with specific indications. Concerning the disease, we chose NSCLC as it occupied a large part of lung cancer [42]. 15 FDA approved drugs for NSCLC were selected from the human disease database (https://www.malacards.org/) as original compounds and compared with the misclassified antineoplastic drugs in our drug pool (S4 Table). To compute the similarity between two compounds, we utilized Library for the Enumeration of Modular Natural Structure (LEMONS) algorithm designed by Skinnider and Magarvey [25]. Tanimoto coefficient is identified as one of the best similarity matrixes, when estimating the similarities [43]. In this study, the drugs with tanimoto index above 0.8 were picked for further analysis (S5 Table). From 164 signaling pathways information of drugs embedded by Di-PASS, three pathways closely related to NSCLC therapies, including NSCLC pathway, epidermal growth factor receptor (EGFR) tyrosine kinase inhibitor resistance, and pathway in cancer were collected [44]. We listed all the drugs with negative Di-PASS scores in these three pathways from misclassified antineoplastic agents in S6 Table. Finally, the candidate drugs owning similar structures with FDA approved NSCLC drugs and negative Di-PASS score in NSCLC related pathways were obtained (Fig 5).

Take the first pair of drugs for example, pimozide was the repurposed potential drug produced by our pipeline. Gemcitabine is one of the most promising anti-cancer chemotherapy drugs, which has shown most activity in the treatment of NSCLC [45]. As an analogue, it is very possible for pimozide to share similar biological and physiochemical activities with gemcitabine. Not surprisingly, the pathway activation scores of pimozide showed -69.89, -13.83, and -20.76 in pathways of NSCLC, EGFR tyrosine kinase inhibitor resistance, and pathway in cancer, respectively. As mentioned previously, Di-PASS scores represented the intensity of pathway activating effects with positive/negative signs corresponding to activation/suppression. The high negative value of Di-PASS score indicates strong suppression effect of pimozide to NSCLC related pathways. Here an exciting potential drug candidate with strong evidences of both chemical structure and pathway signatures in new indication was discovered.

To further validate the proposed candidate for repurposing, we did additional experiments in lab (Fig 6). Pimozide was evaluated using the sulforhodmide-B (SRB) assay [28] in two NSCLC cell lines (A549 and H157), and the antitumor activity was compared with the positive control (gemcitabine). These cell lines were chosen because they were characterized earlier for sensitivity to gemcitabine [46]. The *in vitro* results suggested that pimozide had comparable IC_{50} with gemcitabine and possibly warranted further evaluation in vivo. Both cell lines of A549 and H157 indicated that pimozide possessed activity for inhibiting NSCLC cells. The

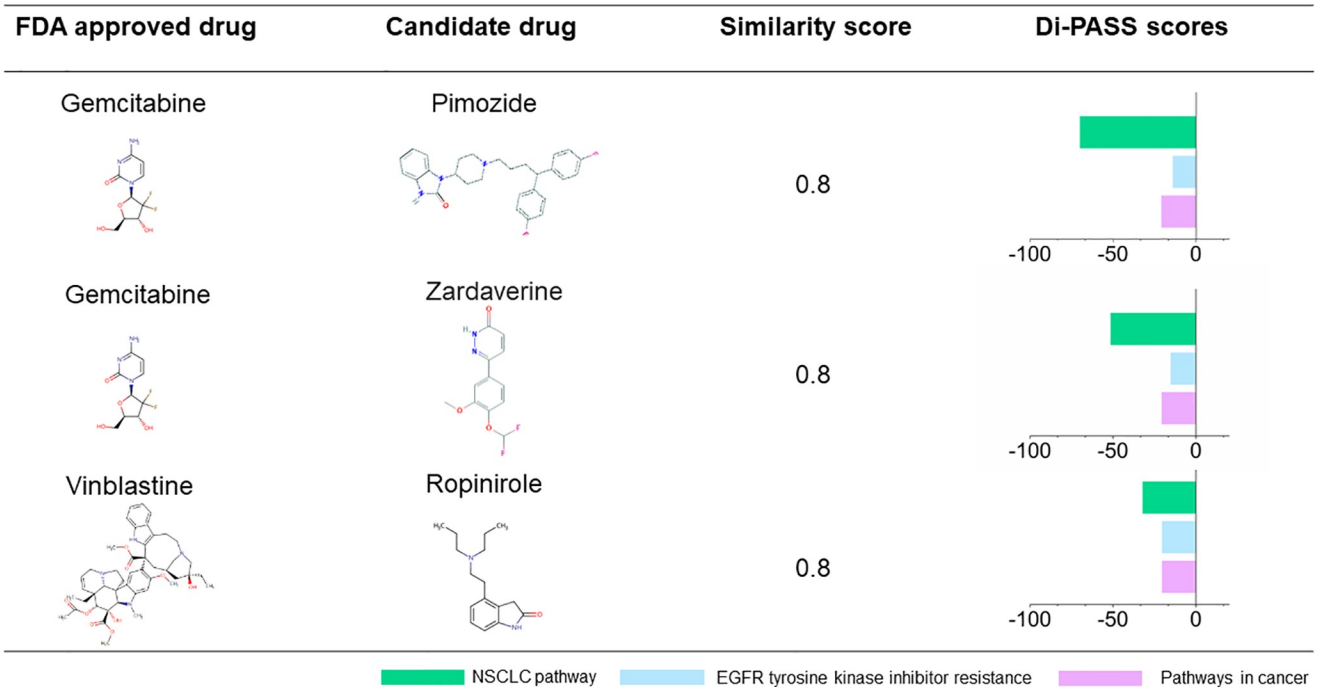

Fig 5. Top candidate drugs ranked by similarity score and Di-PASS scores. Tanimoto coefficient was used for measuring similarity of chemical structures. Di-PASS scores represented the intensity of activation/depression of drugs in NSCLC related pathways.

https://doi.org/10.1371/journal.pone.0233112.g005

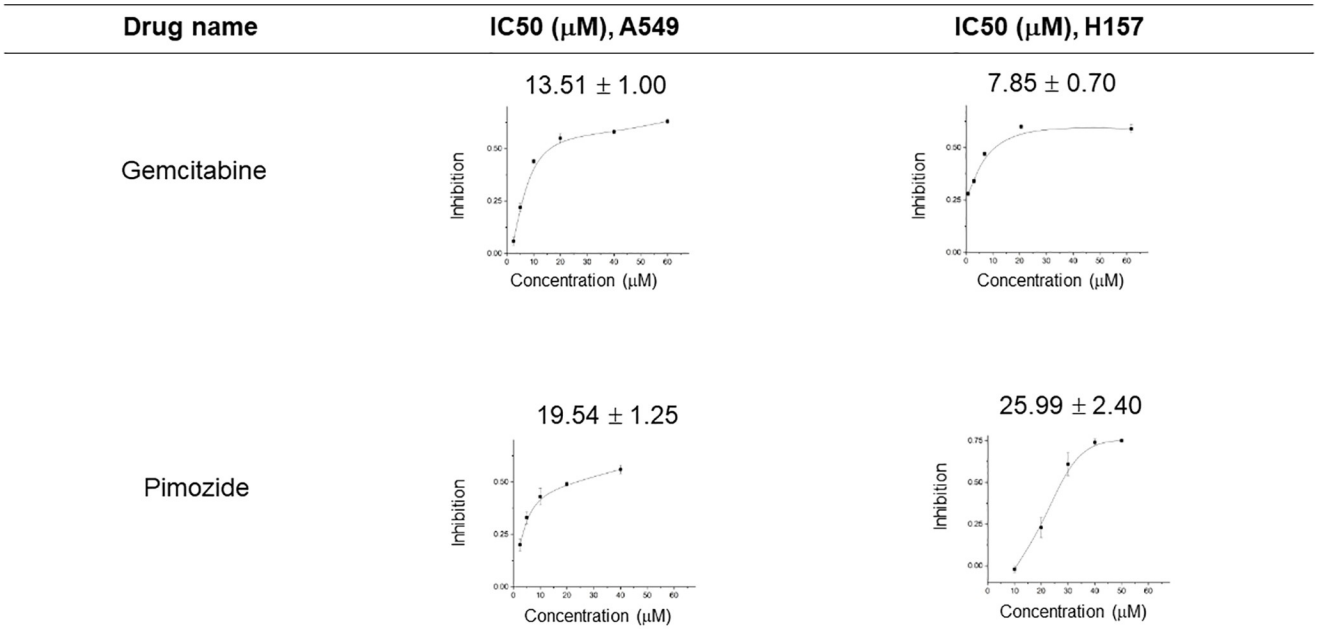

Fig 6. Cytotoxic activity of gemcitabine and pimozide on two NSCLC cell lines (A549 and H157). Test data expressed as the mean of three experiments.

https://doi.org/10.1371/journal.pone.0233112.g006

results were consistent with some earlier research based on pimozide [47, 48]. It is well known that pimozide is an anti-dyskinesia agent, which is used to reduce uncontrolled movements or outbursts of words caused by Tourette syndrome. However, we found that pimozide can be repurposed for treating NSCLC. Our study provided a pragmatic way to redirect traditional drugs to novel therapeutic uses. The other drug candidates were analyzed in the same way (S7 Table).

Conclusions

We have proposed a novel drug repurposing approach based on transcriptomic data and chemical structures using DNN model. We have successfully repurposed pimozide from an anti-dyskinesia agent to an anti-NSCLC drug, which is supported by the *in vitro* experiments. The incorporation of both chemical structures and pathway activation scores when redirecting drugs to new therapeutic roles is proved to be an effective drug repurposing approach.

Although the Di-PASS algorithm can predict drugs with transcriptomic information (S8 Table provided entire drug list we can predict), it has been challenging to build assembled large-scale drug-transcriptomic datasets. In addition, further *in vitro* or *in vivo* experiments are still needed to warrant our predictions. Despite the above limitations, the approach can be extended to other diseases and drugs to identify novel therapeutic relationships.

Supporting information

S1 Table. Number of drugs selected for A375, HELA and HT29 cell lines according to MeSH therapeutic use section. Number of samples is shown in brackets.
(XLSX)

S2 Table. MeSH category stratification binary matrix of 75 significantly perturbed drugs. Every drug belongs only to one category.
(XLSX)

S3 Table. Misclassified drugs from antineoplastic agents.
(TSV)

S4 Table. FDA approved drugs of NSCLC from drug bank.
(TSV)

S5 Table. The drugs with tanimoto index above 0.8.
(XLSX)

S6 Table. The drugs with negative Di-PASS scores in three pathways related to NSCLC.
(XLSX)

S7 Table. Cytotoxic activity of potential drugs on two NSCLC cell lines (A549 and H157). SRB test data expressed as the mean of three experiments.
(XLSX)

S8 Table. Drug list that Di-PASS can predict.
(CSV)

S9 Table. An input file example of Di-PASS.
(CSV)

Author Contributions

Data curation: Bingrui Li, Chan Dai.

Formal analysis: Bingrui Li.

Supervision: Zheng Guan.

Validation: Lijun Wang, Hailong Deng.

Writing – original draft: Yingying Li.

Writing – review & editing: Yingying Li, Haihong Ni.

References

1. Pammolli F, Magazzini L, Riccaboni M. The productivity crisis in pharmaceutical R&D. Nat Rev Drug Discov. 2011; 10:428. https://doi.org/10.1038/nrd3405 PMID: 21629293
2. Pushpakom S, Iorio F, Eyers PA, Escott KJ, Hopper S, Wells A, et al. Drug repurposing: progress, challenges and recommendations. Nat Rev Drug Discov. 2018. https://doi.org/10.1038/nrd.2018.168 PMID: 30310233
3. Ashburn TT, Thor KB. Drug repositioning: identifying and developing new uses for existing drugs. Nat Rev Drug Discov. 2004; 3:673. https://doi.org/10.1038/nrd1468 PMID: 15286734
4. Xue H, Li J, Xie H, Wang Y. Review of drug repositioning approaches and resources. Int J Biol Sci. 2018; 14(10):1232–44. https://doi.org/10.7150/ijbs.24612 PMID: 30123072
5. Oprea TI, Overington JP. Computational and practical aspects of drug repositioning. ASSAY Drug Dev Tech. 2015; 13(6):299–306. https://doi.org/10.1089/adt.2015.29011.tiodrrr PMID: 26241209
6. Moridi M, Ghadirinia M, Sharifi-Zarchi A, Zare-Mirakabad F. The assessment of efficient representation of drug features using deep learning for drug repositioning. BMC Bioinformatics. 2019; 20(1):577. https://doi.org/10.1186/s12859-019-3165-y PMID: 31726977
7. Zeng X, Zhu S, Liu X, Zhou Y, Nussinov R, Cheng F. deepDR: a network-based deep learning approach to in silico drug repositioning. Bioinformatics. 2019; 35(24):5191–8. Epub 2019/05/23. https://doi.org/10.1093/bioinformatics/btz418 PMID: 31116390
8. Zhao K, So HC, Using drug expression profiles and machine learning approach for drug repurposing. computational methods for drug repurposing. Methods Mol Biol. 2019; 1903:219–237. https://doi.org/10.1007/978-1-4939-8955-3_13. PMID: 30547445
9. Wermuth CG. Similarity in drugs: reflections on analogue design. Drug Discov Today. 2006; 11(7–8):348–54. Epub 2006/04/04. https://doi.org/10.1016/j.drudis.2006.02.006 PMID: 16580977
10. Shabana KM, Abdul Nazeer KA, Pradhan M, Palakal M. A computational method for drug repositioning using publicly available gene expression data. BMC Bioinformatics. 2015; 16(17):S5. https://doi.org/10.1186/1471-2105-16-S17-S5 PMID: 26679199
11. Keiser MJ, Setola V, Irwin JJ, Laggner C, Abbas AI, Hufeisen SJ, et al. Predicting new molecular targets for known drugs. Nature. 2009; 462:175. https://doi.org/10.1038/nature08506 PMID: 19881490
12. Aliper A, Plis S, Artemov A, Ulloa A, Mamoshina P, Zhavoronkov A. Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data. Mol Pharmaceut 2016; 13(7):2524–30. https://doi.org/10.1021/acs.molpharmaceut.6b00248 PMID: 27200455
13. Iskar M, Zeller G, Blattmann P, Campillos M, Kuhn M, Kaminska KH, et al. Characterization of drug-induced transcriptional modules: towards drug repositioning and functional understanding. Mol Syst Biol. 2013; 9(1):662. https://doi.org/10.1038/msb.2013.20 PMID: 23632384
14. Kutalik Z, Beckmann JS, Bergmann S. A modular approach for integrative analysis of large-scale gene-expression and drug-response data. Nat Biotechnol. 2008; 26:531. https://doi.org/10.1038/nbt1397 https://www.nature.com/articles/nbt1397#supplementary-information. PMID: 18464786
15. Wu Z, Wang Y, Chen L, editors. A new method to identify repositioned drugs for prostate cancer. 2012 IEEE 6th International Conference on Systems Biology (ISB); 2012 18–20 Aug. 2012.
16. Brandes U. A faster algorithm for betweenness centrality. J Math Sociol. 2001; 25(2):163–77. https://doi.org/10.1080/0022250X.2001.9990249
17. Zhang JD, Wiemann S. KEGGgraph: a graph approach to KEGG PATHWAY in R and bioconductor. Bioinformatics. 2009; 25(11):1470–1. https://doi.org/10.1093/bioinformatics/btp167 PMID: 19307239
18. Okamura Y, Aoki Y, Obayashi T, Tadaka S, Ito S, Narise T, et al. COXPRESdb in 2015: coexpression database for animal species by DNA-microarray and RNAseq-based expression data with multiple

quality assessment systems. Nucleic Acids Res. 2015; 43(Database issue):D82–6. Epub 2014/11/14. https://doi.org/10.1093/nar/gku1163 PMID: 25392420

19. Ankerst M, Breunig MM, Kriegel H, Sander J. OPTICS: Ordering Points To Identify the Clustering Structure. ACM Press. 1999:49–60.
20. Zhang T, Ramakrishnan R, Linvy M. BIRCH: A new data clustering algorithm and its applications. Data Min Knowl Disc. 1997; 1(2):141–82
21. Ozerov IV, Lezhnina KV, Izumchenko E, Artemov AV, Medintsev S, Vanhaelen Q, et al. In silico Pathway Activation Network Decomposition Analysis (iPANDA) as a method for biomarker development. Nat Commun. 2016; 7:13427. https://doi.org/10.1038/ncomms13427 PMID: 27848968
22. Wager S, Athey S. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. J Am Stat Assoc. 2018; 113(523):1228–42.
23. Glorot X. Bengio Y. Understanding the difficulty of training deep feedforward neural networks. Int Conf Artif Intell Stat. 2010:249–56
24. Nair V, Hinton GE. Rectified linear units improve restricted boltzmann machines. ICML. 2010
25. Skinnider MA, Dejong CA, Franczak BC, McNicholas PD, Magarvey NA. Comparative analysis of chemical similarity methods for modular natural products with a hypothetical structure enumeration algorithm. J Cheminform. 2017; 9(1):46. Epub 2017/11/01. https://doi.org/10.1186/s13321-017-0234-y PMID: 29086195
26. Maggiora G, Vogt M, Stumpfe D, Bajorath J. Molecular similarity in medicinal chemistry. J Med Chem. 2014; 57(8):3186–204. Epub 2013/10/25. https://doi.org/10.1021/jm401411z PMID: 24151987
27. Patterson DE, Cramer Rd Fau—Ferguson AM, Ferguson Am Fau—Clark RD, Clark Rd Fau—Weinberger LE, Weinberger LE. Neighborhood behavior: a useful concept for validation of "molecular diversity" descriptors. (0022–2623 (Print))
28. Rubinstein LV, Shoemaker RH, Paull KD, Simon RM, Tosini S, Skehan P, et al. Comparison of in vitro anticancer-drug-screening data generated with a tetrazolium assay versus a protein assay against a diverse panel of human tumor cell lines. JNCI-J Natl Cancer I. 1990; 82(13):1113–7. https://doi.org/10.1093/jnci/82.13.1113 PMID: 2359137
29. Jung M, Lee Y, Moon HI, Jung Y, Jung H, Oh M. Total synthesis and anticancer activity of highly potent novel glycolipid derivatives. Eur J Med Chem. 2009; 44(8):3120–9. https://doi.org/10.1016/j.ejmech.2009.03.007 PMID: 19342127
30. Hira ZM, Gillies DF. A review of feature selection and feature extraction methods applied on microarray data. Adv Bioinformatics. 2015; 2015:198363. https://doi.org/10.1155/2015/198363 PMID: 26170834
31. Hodos RA, Kidd BA, Shameer K, Readhead BP, Dudley JT. In silico methods for drug repurposing and pharmacology. WIRES Syst Biol Med. 2016; 8(3):186–210. https://doi.org/10.1002/wsbm.1337 PMID: 27080087
32. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. P Natl Acad Sci USA. 2005; 102(43):15545
33. Blum AL, Rivest RL. Training a 3-node neural network is NP-complete. Neural Networks. 1992; 5 (1):117–27. https://doi.org/10.1016/S0893-6080(05)80010-3
34. Aoki K, Ogata Y, Shibata D. Approaches for extracting practical information from gene co-expression networks in plant biology. Plant Cell Physiol. 2007; 48(3):381–90. https://doi.org/10.1093/pcp/pcm013 PMID: 17251202
35. Rung J, Brazma A. Reuse of public genome-wide gene expression data. Nat Rev Genet. 2012; 14:89. https://doi.org/10.1038/nrg3394 https://www.nature.com/articles/nrg3394#supplementary-information. PMID: 23269463
36. Obayashi T, Kagaya Y, Aoki Y, Tadaka S, Kinoshita K. COXPRESdb v7: a gene coexpression database for 11 animal species supported by 23 coexpression platforms for technical evaluation and evolutionary inference. Nucleic Acids Res. 2019; 47(D1):D55–D62. https://doi.org/10.1093/nar/gky1155 PMID: 30462320
37. Zhang JD, Wiemann S. KEGGgraph: a graph approach to KEGG PATHWAY in R and bioconductor. Bioinformatics. 2009; 25(11):1470–1. https://doi.org/10.1093/bioinformatics/btp167 PMID: 19307239
38. Zhavoronkov A, Mamoshina P, Vanhaelen Q, Scheibye-Knudsen M, Moskalev A, Aliper A. Artificial intelligence for aging and longevity research: Recent advances and perspectives. Ageing Res Rev. 2019; 49:49–66. Epub 2018/11/26. https://doi.org/10.1016/j.arr.2018.11.003 PMID: 30472217
39. Sukhatme V, Bouche G, Meheus L, Sukhatme VP, Pantziarka P. Repurposing Drugs in Oncology (ReDO)-nitroglycerin as an anti-cancer agent. Ecancermedicalscience. 2015; 9:568. https://doi.org/10.3332/ecancer.2015.568 PMID: 26435741

40. Pantziarka P, Bouche G, Sukhatme V, Meheus L, Rooman I, Sukhatme VP. Repurposing Drugs in Oncology (ReDO)-Propranolol as an anti-cancer agent. Ecancermedicalscience. 2016; 10:680. https://doi.org/10.3332/ecancer.2016.680 PMID: 27899953

41. Proudfoot JR. Drugs, leads, and drug-likeness: an analysis of some recently launched drugs. Bioorg Med Chem Lett 2002; 12(12):1647–50. https://doi.org/10.1016/S0960-894X(02)00244-5. PMID: 12039582

42. Chen W, Zheng R, Zeng H, Zhang S. Epidemiology of lung cancer in China. Thorac Cancer. 2015; 6 (2):209–15. https://doi.org/10.1111/1759-7714.12169 PMID: 26273360

43. Bajusz D, Racz A, Heberger K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? J Cheminform. 2015; 7:20. https://doi.org/10.1186/s13321-015-0069-3 PMID: 26052348

44. Brambilla E, Gazdar A. Pathogenesis of lung cancer signalling pathways: roadmap for therapies. Eur Respir J. 2009; 33(6):1485–97. Epub 2009/06/02. https://doi.org/10.1183/09031936.00014009 PMID: 19483050

45. Toschi L, Finocchiaro G, Bartolini S, Gioia V, Cappuzzo F. Role of gemcitabine in cancer therapy. Future Oncol. 2005; 1(1):7–17. https://doi.org/10.1517/14796694.1.1.7 PMID: 16555971

46. Radi M, Adema AD, Daft JR, Cho JH, Hoebe EK, Alexander L-EMM, et al. In Vitro Optimization of Non-Small Cell Lung Cancer Activity with Troxacitabine, l-1,3-Dioxolane-cytidine, Prodrugs. J. Med Chem. 2007; 50(9):2249–53. https://doi.org/10.1021/jm0612923 PMID: 17419604

47. Goncalves JM, Silva CAB, Rivero ERC, Cordeiro MMR. Inhibition of cancer stem cells promoted by Pimozide. Clin Exp Pharmacol Physiol. 2019; 46(2):116–25. https://doi.org/10.1111/1440-1681.13049 PMID: 30383889.

48. Fortney K, Griesman J, Kotlyar M, Pastrello C, Angeli M, Sound-Tsao M, et al. Prioritizing therapeutics for lung cancer: an integrative meta-analysis of cancer gene signatures and chemogenomic data. PLoS Comput Biol. 2015; 11(3):e1004068. https://doi.org/10.1371/journal.pcbi.1004068 PMID: 25786242