ARTIFICIAL INTELLIGENCE FOR TRANSLATIONAL PHARMACOLOGY

EDITED BY: Zhi-Liang Ji, Lixia Yao, Kartick Chandra Pramanik and Zhaohui John Cai

PUBLISHED IN: Frontiers in Pharmacology

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88963-888-8
DOI 10.3389/978-2-88963-888-8

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: researchtopics@frontiersin.org

ARTIFICIAL INTELLIGENCE FOR TRANSLATIONAL PHARMACOLOGY

Topic Editors:
Zhi-Liang Ji, Xiamen University, China
Lixia Yao, Mayo Clinic, United States
Kartick Chandra Pramanik, University of Pikeville, United States
Zhaohui John Cai, Celgene (United States), United States

Citation: Ji, Z.-L., Yao, L., Pramanik, K. C., Cai, Z. J., eds. (2020). Artificial Intelligence for Translational Pharmacology. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88963-888-8

Table of Contents

frontiers in Pharmacology

REVIEW
published: 22 May 2018
doi: 10.3389/fphar.2018.00535

Searching Synergistic Dose Combinations for Anticancer Drugs

*Zuojing Yin, Zeliang Deng, Wenyan Zhao and Zhiwei Cao**

Shanghai Tenth People's Hospital, School of Life Sciences and Technology, Tongji University, Shanghai, China

Recent development has enabled synergistic drugs in treating a wide range of cancers. Being highly context-dependent, however, identification of successful ones often requires screening of combinational dose on different testing platforms in order to gain the best anticancer effects. To facilitate the development of effective computational models, we reviewed the latest strategy in searching optimal dose combination from three perspectives: (1) mainly experimental-based approach; (2) Computational-guided experimental approach; and (3) mainly computational-based approach. In addition to the introduction of each strategy, critical discussion of their advantages and disadvantages were also included, with a strong focus on the current applications and future improvements.

Keywords: synergistic combination, optimized dose combination, computational model, feedback system control scheme, regression model

OPEN ACCESS

Edited by:
Zhi-Liang Ji,
Xiamen University, China

Reviewed by:
Qi Liu,
Vanderbilt University Medical Center,
United States
Feng Zhu,
Zhejiang University, China

****Correspondence:***
Zhiwei Cao
zwcao@tongji.edu.cn

Specialty section:
This article was submitted to
Translational Pharmacology,
a section of the journal
Frontiers in Pharmacology

Received: *06 January 2018*
Accepted: *03 May 2018*
Published: *22 May 2018*

Citation:
Yin Z, Deng Z, Zhao W and Cao Z (2018) Searching Synergistic Dose Combinations for Anticancer Drugs. Front. Pharmacol. 9:535. doi: 10.3389/fphar.2018.00535

INTRODUCTION

In current days, combinational drugs have been increasingly used clinically in treating various cancers. Comparing to the traditional single drug approach, combinational strategy is often found with enhancing therapeutic effects or delayed drug resistance, among which synergistic drugs are mostly desired (Chou, 2006). The past few years has witnessed the computational progress in analyzing and predicting synergistic components qualitatively (Han et al., 2017; Sarah, 2017; Sheng et al., 2017). However, the optimal dose of each component needs to be identified before the formula is clinically applied, as different dose combination may lead to different effects even for the same formula (Tallarida and Raffa, 2010). To avoid potential adverse or antagonistic effects, large-scale experiments have to be screened in a huge combinational space of drug concentration which are highly time consuming and laborious. Thus, developing smart methods either experimentally or theoretically are both in urgent need to facilitate the synergistic drug design.

Until the present time, the general experimental criteria to evaluate drug synergy mainly include Loewe isobologram (Chevereau and Bollenbach, 2015), CI index from Median Effect Principle (Chou, 2010), Bliss independence (BI) model (Bansal et al., 2014), Loewe Additivity (LA) model (Lee et al., 2007), and so on. Under defined criteria, substantial data has been accumulated which initiated the computational efforts to predict dose effects of drug combination. Despite of a few algorithm and statistical methods (Calzolari et al., 2008; Deharo and Ginsburg, 2011; Caglar and Pal, 2014; Weiss et al., 2015a), constructing quantitative model to predict synergistic dose remains highly challenging for combinational therapy. To promote future improvements in this area, we reviewed the latest progress in this area covering (1) mainly experimental-based approach; (2) Computational-guided experimental approach; and (3) mainly computational-based approach.

Mainly Experimental-Based Approach

Normally the drug efficacy can be roughly tested via cell viability assay, such as MTT assay and various animal models. But experimental exploration of drug combinations under all dose ratio seems to be unrealistic. Any high throughput technology or heuristic design will significantly save the time and experimental costs by purposely choosing the potential candidate dose.

High Throughput Experimental Screening

In order to identify effective combinations of therapeutic compounds, Borisy et al. (2003) developed a high-throughput screening method to systematically screen of ∼120,000 pairwise combinations for antifungal effects in 2003. The systematic testing began by defining the activity of each compound as a single agent in the assay system. And then, each active compound against all other compounds was tested in dose matrices comprising six concentrations based on EC50. Finally, the possible synergistic dose ratio between the drug pairs would be detected. In this way, this paper proposed a practical application to systematic screening of compounds in disease-relevant phenotypic assays (Borisy et al., 2003). Furthermore, this method also proposed to detect the synergistic effects between constituents within the natural products (Isgut et al., 2017).

Then in 2007, a series of concentration ratios for each drug pair were tested on 10∼20 tumor cell lines via high-throughput screening technology (Mayer and Janoff, 2007). After analyzing the cytotoxicity curves for each, they found that certain dose ratios of combinational drugs can be synergistic, while other ratios of the same agents may be antagonistic (Mayer and Janoff, 2007). Interestingly, high-throughput screening has been applied to tumor organoids system in recent years (Ivanov and Grabowska, 2017; Ivanov et al., 2017; Shahi Thakuri and Tavana, 2017). For instance, colon cancer spheroids were applied for drug synergy between 25 compounds under multiple IC50s, instead of the traditional cell lines (Shahi Thakuri and Tavana, 2017). And animal model of zebrafish was also established for this purpose with the assistance of auto-image analysis technology (Todd et al., 2017).

Fixed Dose Method

To avoid random high-throughput screening, fix dose/ratio method may serve as a starting point to explore when prior information is totally unknown. The dose may be set according to their maximum tolerated doses (MTD) and partial MTDs (Cao and Rustum, 2000; Azrak et al., 2004, 2007; Cao et al., 2005). As early as in 2000, the synergistic effect of Irinotecan and 5-Fluorouracil was studied in the rat model of colon cancer, at the dose of MTDs, 12.5% MTDS, 50% MTDS, and 75% MTDS, respectively (Cao and Rustum, 2000). Another searched the synergistic effect of 200 pairs of antifungal drugs within a dose range between 0 to minimal inhibitory concentration (MIC) in the brewer's yeast (Cokol et al., 2011). It worth to note that, besides dose combination, the time interval and sequential treatment, even the pharmaceutical packaging may influence the effects of drug combination (Azrak et al., 2007; Mohan et al., 2014).

Instead of fixed dose, some studies fixed dose ratios based IC50 when prior information is unknown (Hatakeyama et al., 2014; Zhang et al., 2014). Occasionally, dose ratio may also start from 1:1 to explore the synergistic spectrum for different drugs in different cancer types (Liu et al., 2011).

Computational-Guided Experimental Approach

To avoid exhaustive searching in dose combinational space, computer algorithm was often adopted as a feedback control to suggest next round of experiments design based on preliminary experimental results. Current algorithms for this purpose mainly refers to feedback system control scheme (FSC), which help to converge fast in a huge searching space of multiple drugs with multiple doses. This scheme has been applied to identify the best dose combinations of multiple drugs in various cancer (Liu et al., 2015), and viral infection (Wong et al., 2008).

The procedure of FSC (Tsutsui et al., 2011; Liu et al., 2015; Weiss et al., 2015b) usually includes: (1) Input a number of drugs (usually 5 to 10) with several doses (e.g., 0, IC25, IC50, IC75) for a specific disease; (2) Combine all drugs and their doses to form a large searching space; (3) Random select partial combinations from above space and test experimentally; (4) Update the drug doses by differential evolution algorithm (DE); (5) Repeat (3) and compare latest experimental results to the previous ones; and (6) Choose better experimental results for the next iteration.

Here the detailed heuristic DE algorithm (Tsutsui et al., 2011) is illustrated in **Figure 1**: (1) Choose a drug-dose combination

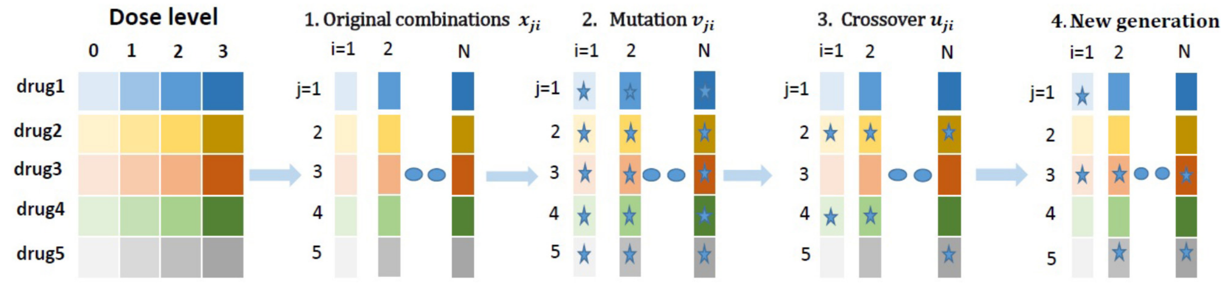


FIGURE 1 | A schematic illustration of differential evolution (DE) algorithm.

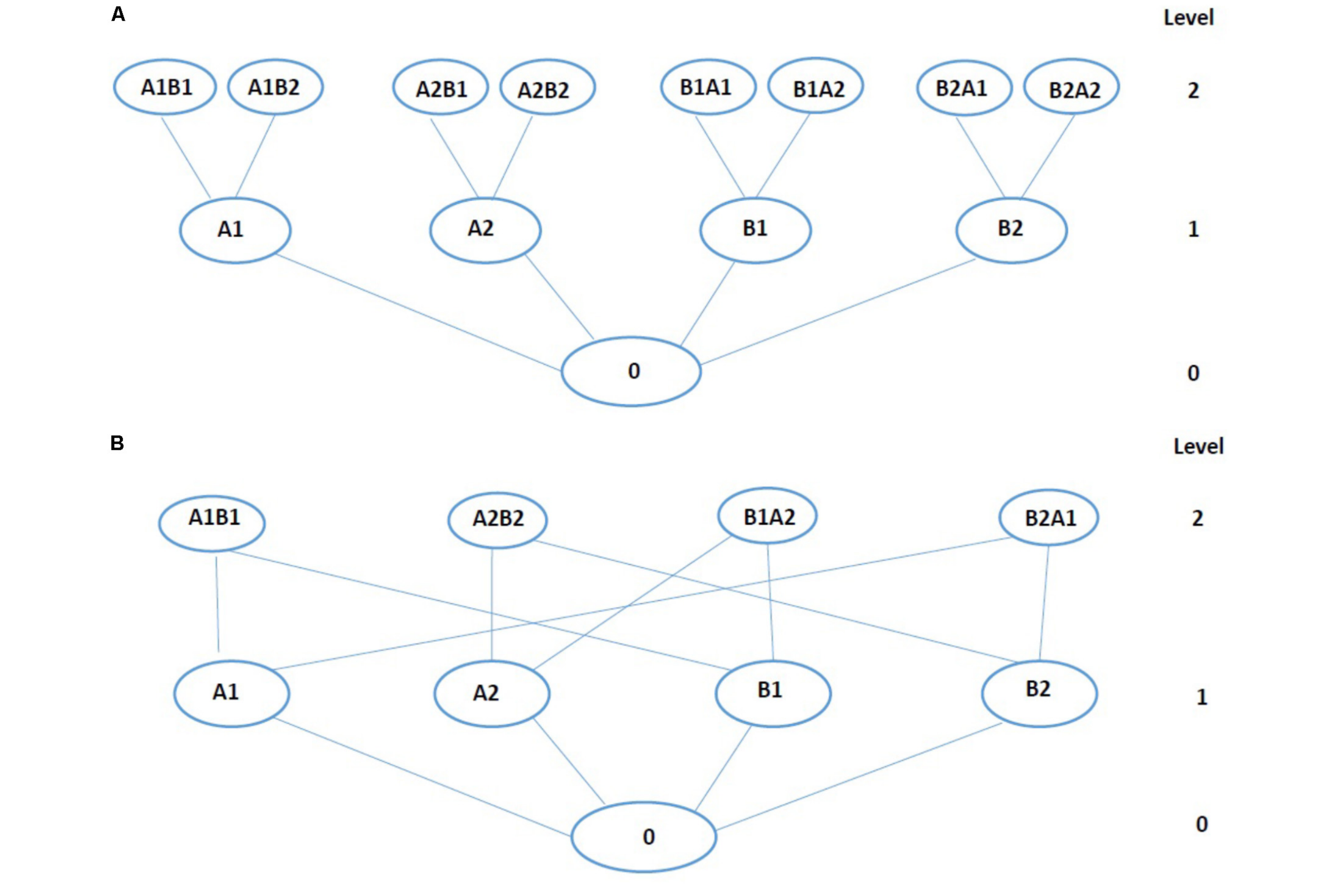


FIGURE 2 | Tree representation of the data. **(A)** The tree with sequential structure. **(B)** The tree with trellis-like structure (alternative version of the **A**). Each circle stands for single drug-dose or combination. Letter and number indicates drug and dose, respectively. For a tree, level 0 is the control (no drug), level 1 is composed of individual drug treatment with two doses, and level 2 is composed of drug-dose combinations. The level depends on the size of the combination.

according to a random algorithm: x_{ji}; (2) Examine the effect through experiments: E(x_{ji}); (3) Mutate the current selected drug dose to v_{ji}; (4) Crossover the current and mutated drug-dose combination ($x_{ji} \times v_{ji}$) to obtain a new drug-dose combination u_{ji}; (5) Examine the effect of the new drug-dose combination: E(u_{ji}); (6) Compare E(u_{ji}) with E(x_{ji}). The new drug-dose combination is u_{ji}, if E(u_{ji}) $>$ E(x_{ji}) and will go into the next iteration cycle.

It can be seen that the features of FSC as several advantages (Nowak-Sliwinska et al., 2016). Firstly, it is phenotypically driven, simpler than genotype-driven methods, and does not require any mechanism information. Secondly it can achieve a fast convergence by using DE algorithm. Despite of that, the experimental testing is still substantial because all input drugs are considered equally in the combination. Thus the improved version of FSC incorporates a regression model to identify those potential synergistic drugs out of the input list before searching optimized dose (Wang et al., 2015; Weiss et al., 2015a).

Recently, FSC was used to screen Nano-diamond modified drugs out of 57 dose combinations and therapeutic dose window was proposed which could optimally inhibit cancer cell lines and protect the normal cell lines (Wang et al., 2015). More application of FSC could be found in prostate cancer and hepatocellular carcinomas (Mohd Abdul Rashid et al., 2015; Jia et al., 2017).

Mainly Computational-Based Approach

Apart from the above approaches, a few mathematical models have been constructed which have been collected as below.

Stochastic searching model

To minimize searching space for optimal dose combination, a few stochastic search algorithms with heuristic ideas have been reported recently (Calzolari et al., 2008; Caglar and Pal, 2014). **Figure 2A** shows an example of stochastic search algorithms, with ideas similar to that of the stack sequential algorithm (Jeline, 1969). An alternative version of the **Figure 2A** tree (**Figure 2B**), eliminates nodes representing redundant drug-dose combinations. Stochastic search algorithms works as this: under search tree structure, the biological score was evaluated at the first level of tree and best single drug Cbest was extracted (Calzolari et al., 2008). Then, the biological scores of Cbest combined with all other drugs were measured and compared with Cbest's to decide the movements of upward or downward. The current best combination was chosen for

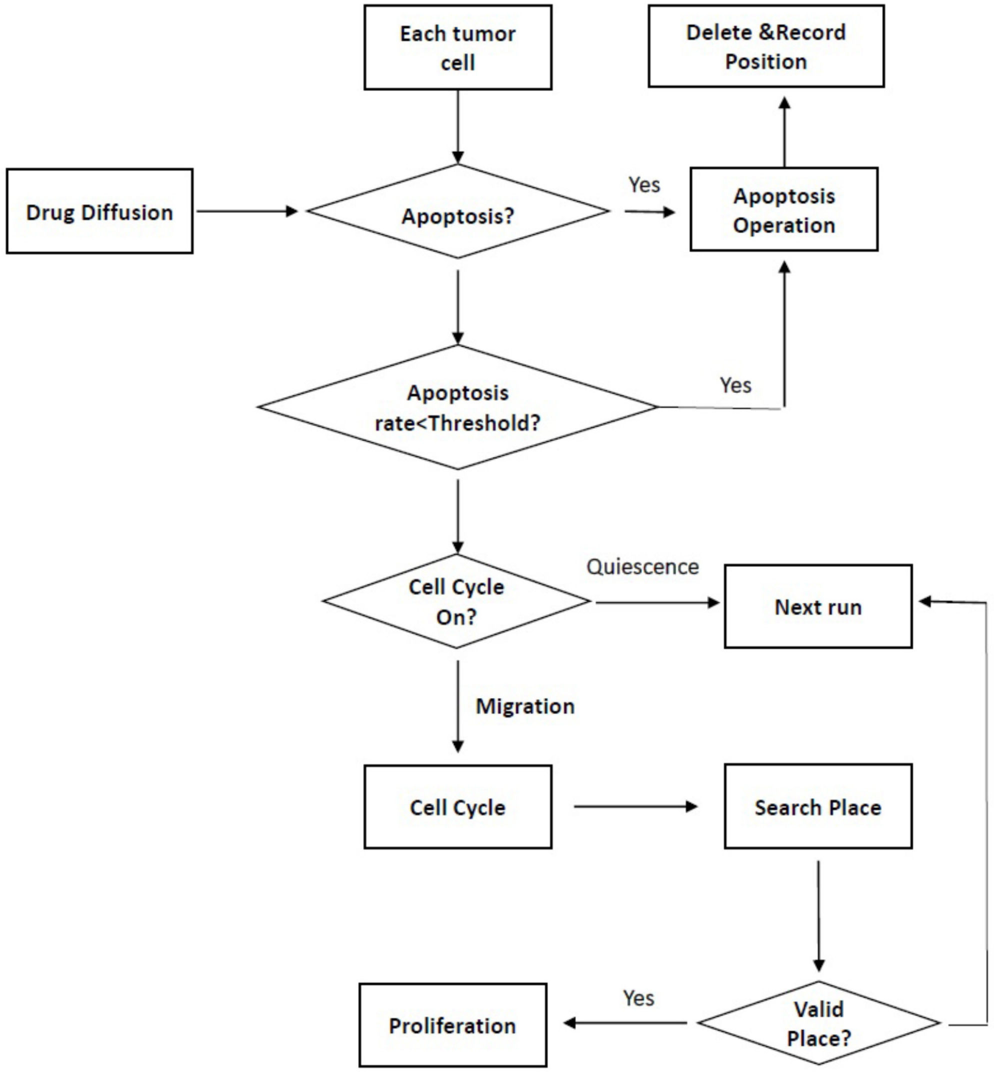


FIGURE 3 | The cellular phenotype switching at the intracellular scale. Apoptosis: Under the drug diffusion, the cell will initiate apoptosis if the simulated apoptosis probability is less than the set threshold. Proliferation: The proliferation will initiate if the cell cycle is on and empty location exist to divide in the mitotic M phase. Migration: A proliferating cell will migrate in the first three phases of cell cycle (G0/G1, S, and G2) under appropriate location. Quiescence: There are two possibilities of quiescence: the cell cannot go through the cell cycle, or the cell cannot find a valid place to divide.

further searching of sub-nodes to get the global optimal combinations. In this way, only one-third of the tests were actually scanned in the Drosophila model of 4 drugs (Calzolari et al., 2008).

Meanwhile, a diversified stochastic search algorithm has been recently proposed to find optimum drug concentrations efficiently without prior normalization of the searching space (Caglar and Pal, 2014). This stochastic algorithm was composed of the initial parallel part and the iteration part. The former was used to generate a rudimentary knowledge of the searching space, while the later was mainly used to search the space repeatedly to update knowledge of new hills that the previous iterations could not locate. After relatively smaller number of iterative steps, the optimized dose combination could be detected for anti-bacteria and anti-cancer effects (Caglar and Pal, 2014).

Statistical model

In addition to stochastic searching, statistical models were also applied to screen the optimal drug-dose combination based on cellular responses (Deharo and Ginsburg, 2011; Weiss et al., 2015a). The logistic regression model showed in equation (1) (Deharo and Ginsburg, 2011) was proposed to predict the EC50s of the drug alone and in combination. And the synergistic effects of six different ergosterol together with the pyrethroid in five selected dose ratios were detected (Deharo and Ginsburg, 2011).

$$f(x, (b, c, d, e)) = c + \frac{d - c}{1 + (x/e)^{\mathrm{e}}} \tag{1}$$

$f(\mathrm{x})$: drug effects; x: dose of drug; b: a measure of the steepness of the curve for the dose equal to the ED50 value; c, d: denote the lower and upper asymptotes of the s-shaped curve; *e*: corresponds to ED50 value.

Different from the logistic regression model, the second-order linear regression model screened out the optimal drug-dose combination by firstly refine drugs which might produce synergistic effect (Chen et al., 2010; Xu et al., 2014; Weiss et al., 2015a; Silva et al., 2016). This model mainly contained the following steps: (1) Establish a stepwise linear regression model describing the relationship between drug doses and effects; (2) Select the drugs most likely to produce synergistic effects according to the model coefficients; (3) Continue to do regression analysis of the drugs selected in (2); (4) Detect final optimal drug combination and dose ratio. Through several cycles, an optimal drug combination toward viability inhibition of renal carcinoma cells from initial 10-drug pool with 4 doses each was detected (Weiss et al., 2015a).

The second-order linear regression model (Weiss et al., 2015a) is showed in equation (2)

$$y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i=1}^{k} \beta_{ii} x_i^2 + \sum_{i=1}^{k} \sum_{j=i+1}^{k} \beta_{ij} x_i x_j + \varepsilon \qquad (2)$$

y: the response variable (i.e., cell viability as percent of control); $\beta_i, \beta_{ii}, \beta_{ij}$: represent the intercept and the coefficients of linear, quadratic, and bilinear terms, respectively; x_i, x_j : independent variables (i.e., drug combination at designed doses); ε: an error term.

Multi-Scale Agent-Based Model

In recently years, the multi-scale agent-based model has been established to evaluate synergistic dose ratios by controlling the fate of cells under different drug combinations (Wang et al., 2013; Qiao et al., 2015). The model simulated the growth process of tumor cells including apoptosis, proliferation, migration, etc. based on some specialized biological regulations to screen the optimal dose combinations with maximal lethality in different dose combinations. Furthermore, the model could not only describe multicellular interaction system and microenvironment in cancer, but also detect synergistic dose with limited experimental data. Usually, the model was established according to discrete dose combination effects to simulate continuous effects under wide range of dose combinations. And the fate of cells was usually described from the intracellular, intercellular, and tissue scales to illustrate the 'phenotypic' switches showed in **Figure 3** (Qiao et al., 2015), cell–cell and cell–microenvironment interaction, respectively (Wang et al., 2013; Qiao et al., 2015).

In 2015, this model was firstly used to choose optimal combinations restoring the balance between osteoclast cells and osteoblast cells as well as killed cancer cells in multiple myeloma Cancer (Qiao et al., 2015). According to the pathogenesis, the behaviors of myeloma cells and two normal cells under the action of multiple cytokines and drug combinations were simulated. Ultimately, the optimal dose ratio of the combination was screened out according to the simulation result.

Besides, artificial intelligence (AI) has had an impact in drug synergy area. Recently, Preuer et al. (2017) developed a novel deep learning method, termed DeepSynergy, to model drug synergy qualitatively using chemical and genomic information, which is based on Neural Networks. This mechanism-free and data-driven method outperformed those previously methods of deep learning within the space of 38 drugs on 39 cell lines. But DeepSynergy didn't make comparison with the other models previously reported, such as RACS (Sun et al., 2015) and other methods in DREAM Challenge (Bansal et al., 2014). RACS, which is semi-supervised, mechanism-guided, and context-dependent combining both genomic and network characteristics, showed a probability concordance of 0.78 compared with 0.61 obtained with the best algorithm reported in DREAM Challenge within the space of 14 compounds on the cell line OCI-Ly3. Furthermore, more computational approaches in qualitatively identifying synergistic drug combinations are summarized by Sheng et al. (2017). Yet AI methods have not been seen in quantitatively screening synergistic dose combinations, which worth further exploration.

PERSPECTIVE

We have summarized the latest development in the area of synergistic dose combinations for Anticancer Drugs. Above accumulated work has paved the way to comprehensive predictive model of optimal dose combination. It should be aware of that, the current searching methods are still limited to local optimization, while more experimental results are needed to validate the computational models. Although challenging, considering below factors may contribute to more effective algorithms. For instance, cancer heterogeneity should be seriously considered in order to achieve better results. Meanwhile, considering the drug response of multiple cells/tissues may minimize the potential side effects of combined drugs to normal tissues. This is particularly important when the drugs are administrated with different time and different order. Coupled with the future development of AI and hardware development, more concrete models are expected to potentially assist the clinical decision of combinational drug dosage to cancer patients.

AUTHOR CONTRIBUTIONS

ZY collected the main papers and wrote the manuscript. ZD and WZ collected the related studies. ZC supervised the whole project and modified the manuscript. All authors read the approved the final manuscript.

FUNDING

This work has been supported by the Fundamental Research Funds for the Central Universities and National Natural Science Foundation of China (No. 31671379).

REFERENCES

Azrak, R. G., Cao, S., Pendyala, L., Durrani, F. A., Fakih, M., Combs, G. F. Jr., et al. (2007). Efficacy of increasing the therapeutic index of irinotecan, plasma and tissue selenium concentrations is methylselenocysteine dose dependent. *Biochem. Pharmacol.* 73, 1280–1287. doi: 10.1016/j.bcp.2006.12.020

Azrak, R. G., Cao, S., Slocum, H. K., Tóth, K., Durrani, F. A., Yin, M. B., et al. (2004). Therapeutic synergy between irinotecan and 5-fluorouracil against human tumor xenografts. *Clin. Cancer Res.* 10, 1121–1129. doi: 10.1158/1078-0432.CCR-0913-3

Bansal, M., Yang, J., Karan, C., Menden, M. P., Costello, J. C., Tang, H., et al. (2014). A community computational challenge to predict the activity of pairs of compounds. *Nat. Biotechnol.* 32, 1213–1222. doi: 10.1038/nbt.3052

Borisy, A. A., Elliott, P. J., Hurst, N. W., Lee, M. S., Lehar, J., Price, E. R., et al. (2003). Systematic discovery of multicomponent therapeutics. *Proc. Natl. Acad. Sci. U.S.A.* 100, 7977–7982. doi: 10.1073/pnas.1337088100

Caglar, M. U., and Pal, R. (2014). A diverse stochastic search algorithm for combination therapeutics. *Biomed. Res. Int.* 2014:873436. doi: 10.1155/2014/873436

Calzolari, D., Bruschi, S., Coquin, L., Schofield, J., Feala, J. D., McCulloch, A. D., et al. (2008). Search algorithms as a framework for the optimization of drug combinations. *PLoS Comput. Biol.* 4:e1000249. doi: 10.1371/journal.pcbi.1000249

Cao, S., Durrani, F. A., and Rustum, Y. M. (2005). Synergistic antitumor activity of capecitabine in combination with irinotecan. *Clin. Colorectal Cancer* 4, 336–343. doi: 10.3816/CCC.2005.n.007

Cao, S., and Rustum, Y. M. (2000). Synergistic antitumor activity of irinotecan in combination with 5-fluorouracil in rats bearing advanced colorectal cancer: role of drug sequence and dose. *Cancer Res.* 60, 3717–3721.

Chen, H. C., Gau, V., Zhang, D. D., Liao, J. C., Wang, F. Y., and Wong, P. K. (2010). Statistical metamodeling for revealing synergistic antimicrobial interactions. *PLoS One* 5:e15472. doi: 10.1371/journal.pone.0015472

Chevereau, G., and Bollenbach, T. (2015). Systematic discovery of drug interaction mechanisms. *Mol. Syst. Biol.* 11:807. doi: 10.15252/msb.20156098

Chou, T. C. (2006). Theoretical basis, experimental design, and computerized simulation of synergism and antagonism in drug combination studies. *Pharmacol. Rev.* 58, 621–681. doi: 10.1124/pr.58.3.10

Chou, T. C. (2010). Drug combination studies and their synergy quantification using the chou-talalay method. *Cancer Res.* 70, 440–446. doi: 10.1158/0008-5472.CAN-09-1947

Cokol, M., Chua, H. N., Tasan, M., Mutlu, B., Weinstein, Z. B., Suzuki, Y., et al. (2011). Systematic exploration of synergistic drug pairs. *Mol. Syst. Biol.* 7:544. doi: 10.1038/msb.2011.71

Deharo, E., and Ginsburg, H. (2011). Analysis of additivity and synergism in the anti-plasmodial effect of purified compounds from plant extracts. *Malar. J.* 10(Suppl. 1):S5. doi: 10.1186/1475-2875-10-S1-S5

Han, K., Jeng, E. E., Hess, G. T., Morgens, D. W., Li, A., and Bassik, M. C. (2017). Synergistic drug combinations for cancer identified in a CRISPR screen for pairwise genetic interactions. *Nat. Biotechnol.* 35, 463–474. doi: 10.1038/nbt.3834

Hatakeyama, Y., Kobayashi, K., Nagano, T., Tamura, D., Yamamoto, M., Tachihara, M., et al. (2014). Synergistic effects of pemetrexed and amrubicin in non-small cell lung cancer cell lines: potential for combination therapy. *Cancer Lett.* 343, 74–79. doi: 10.1016/j.canlet.2013.09.019

Isgut, M., Rao, M., Yang, C., Subrahmanyam, V., Rida, P. C. G., Aneja, R., et al. (2017). Application of combination high-throughput phenotypic screening and target identification methods for the discovery of natural product-based combination drugs. *Med. Res. Rev.* 38, 504–524. doi: 10.1002/med.21444

Ivanov, D. P., and Grabowska, A. M. (2017). Spheroid arrays for high-throughput single-cell analysis of spatial patterns and biomarker expression in 3D. *Sci. Rep.* 7:41160. doi: 10.1038/srep41160

Ivanov, D. P., Grabowska, A. M., and Garnett, M. C. (2017). High-Throughput spheroid screens using volume, resazurin reduction, and acid phosphatase activity. *Methods Mol. Biol.* 1601, 43–59. doi: 10.1007/978-1-4939-6960-9_4

Jeline, F. (1969). Fast sequential decoding algorithm using a stack. *IBM J. Res. Dev.* 13, 675–685. doi: 10.1147/rd.136.0675

Jia, X., Li, Y., Sharma, A., Li, Y., Xie, G., Wang, G., et al. (2017). Application of sequential factorial design and orthogonal array composite design (OACD) to study combination of 5 prostate cancer drugs. *Comput. Biol. Chem.* 67, 234–243. doi: 10.1016/j.compbiolchem.2017.01.010

Lee, J. J., Kong, M., Ayers, G. D., and Lotan, R. (2007). Interaction index and different methods for determining drug interaction in combination therapy. *J. Biopharm. Stat.* 17, 461–480. doi: 10.1080/10543400701199593

Liu, L., Shi, H., Liu, Y., Anderson, A., Peterson, J., Greger, J., et al. (2011). Synergistic effects of foretinib with HER-targeted agents in MET and HER1- or HER2-coactivated tumor cells. *Mol. Cancer Ther.* 10, 518–530. doi: 10.1158/1535-7163.MCT-10-0698

Liu, Q., Zhang, C., Ding, X., Deng, H., Zhang, D., Cui, W., et al. (2015). Preclinical optimization of a broad-spectrum anti-bladder cancer tri-drug regimen via the Feedback System Control (FSC) platform. *Sci. Rep.* 5:11464. doi: 10.1038/srep11464

Mayer, L. D., and Janoff, A. S. (2007). Optimizing combination chemotherapy by controlling drug ratios. *Mol. Interv.* 7, 216–223. doi: 10.1124/mi.7.4.8

Mohan, A., Narayanan, S., Sethuraman, S., and Krishnan, U. M. (2014). Novel resveratrol and 5-fluorouracil coencapsulated in PEGylated nanoliposomes improve chemotherapeutic efficacy of combination against head and neck squamous cell carcinoma. *Biomed. Res. Int.* 2014:424239. doi: 10.1155/2014/424239

Mohd Abdul Rashid, M. B., Toh, T. B., Silva, A., Nurrul Abdullah, L., Ho, C. M., Ho, D., et al. (2015). Identification and optimization of combinatorial glucose metabolism inhibitors in hepatocellular carcinomas. *J. Lab. Autom.* 20, 423–437. doi: 10.1177/2211068215579612

Nowak-Sliwinska, P., Weiss, A., Ding, X., Dyson, P. J., van den Bergh, H., Griffioen, A. W., et al. (2016). Optimization of drug combinations using Feedback System Control. *Nat. Protoc.* 11, 302–315. doi: 10.1038/nprot.2016.017

Preuer, K., Lewis, R. P., Hochreiter, S., Bender, A., Bulusu, K. C., Klambauer, G., et al. (2017). DeepSynergy: predicting anti-cancer drug synergy with Deep Learning. *Bioinformatics* 34, 1538–1546. doi: 10.1093/bioinformatics/btx806

Qiao, M., Wu, D., Carey, M., Zhou, X., and Zhang, L. (2015). Multi-Scale agent-based multiple myeloma cancer modeling and the related study of the balance between osteoclasts and osteoblasts. *PLoS One* 10:e0143206. doi: 10.1371/journal.pone.0143206

Sarah, C. (2017). Cancer: identifying synergistic drug combinations. *Nat. Rev. Drug Discov.* 16:314. doi: 10.1038/nrd.2017.76

Shahi Thakuri, P., and Tavana, H. (2017). Single and combination drug screening with aqueous biphasic tumor spheroids. *SLAS Discov.* 22, 507–515. doi: 10.1177/2472555217698817

Sheng, Z., Sun, Y., Yin, Z., Tang, K., Cao, Z., et al. (2017). Advances in computational approaches in identifying synergistic drug combinations. *Brief. Bioinform.* doi: 10.1093/bib/bbx047. [Epub ahead of print].

Silva, A., Lee, B. Y., Clemens, D. L., Kee, T., Ding, X., Ho, C. M., et al. (2016). Output-driven feedback system control platform optimizes combinatorial therapy of tuberculosis using a macrophage cell culture model. *Proc. Natl. Acad. Sci. U.S.A.* 113, E2172–E2179. doi: 10.1073/pnas.1600812113

Sun, Y., Sheng, Z., Ma, C., Tang, K., Zhu, R., Wu, Z., et al. (2015). Combining genomic and network characteristics for extended capability in predicting synergistic drugs for cancer. *Nat. Commun.* 6:8481. doi: 10.1038/ncomms9481

Tallarida, R. J., and Raffa, R. B. (2010). The application of drug dose equivalence in the quantitative analysis of receptor occupation and drug combinations. *Pharmacol. Ther.* 127, 165–174. doi: 10.1016/j.pharmthera.2010.04.011

Todd, D. W., Philip, R. C., Niihori, M., Ringle, R. A., Coyle, K. R., Zehri, S. F., et al. (2017). A fully automated high-throughput zebrafish behavioral ototoxicity assay. *Zebrafish* 14, 331–342. doi: 10.1089/zeb.2016.1412

Tsutsui, H., Valamehr, B., Hindoyan, A., Qiao, R., Ding, X., Guo, S., et al. (2011). An optimized small molecule inhibitor cocktail supports long-term maintenance of human embryonic stem cells. *Nat. Commun.* 2:167. doi: 10.1038/ncomms1165

Wang, H., Lee, D. K., Chen, K. Y., Chen, J. Y., Zhang, K., Silva, A., et al. (2015). Mechanism-independent optimization of combinatorial nanodiamond and unmodified drug delivery using a phenotypically driven platform technology. *ACS Nano* 9, 3332–3344. doi: 10.1021/acsnano.5b00638

Wang, J., Zhang, L., Jing, C., Ye, G., Wu, H., Miao, H., et al. (2013). Multi-scale agent-based modeling on melanoma and its related angiogenesis analysis. *Theor. Biol. Med. Model.* 10:41. doi: 10.1186/1742-4682-10-41

Weiss, A., Berndsen, R. H., Ding, X., Ho, C. M., Dyson, P. J., van den Bergh, H., et al. (2015a). A streamlined search technology for identification of synergistic drug combinations. *Sci. Rep.* 5:14508. doi: 10.1038/srep14508

Weiss, A., Ding, X., van Beijnum, J. R., Wong, I., Wong, T. J., Berndsen, R. H., et al. (2015b). Rapid optimization of drug combinations for the optimal angiostatic treatment of cancer. *Angiogenesis* 18, 233–244. doi: 10.1007/s10456-015-9462-9

Wong, P. K., Yu, F., Shahangian, A., Cheng, G., Sun, R., and Ho, C. M. (2008). Closed-loop control of cellular functions using combinatory drugs guided by a stochastic search algorithm. *Proc. Natl. Acad. Sci. U.S.A.* 105, 5105–5110. doi: 10.1073/pnas.0800823105

Xu, H. Q., Jaynes, J., and Ding, X. T. (2014). Combining two-level and three-level orthogonal arrays for factor screening and response surface exploration. *Stat. Sin.* 24, 269–289.

Zhang, X. Z., Wang, L., Liu, D. W., Tang, G. Y., and Zhang, H. Y. (2014). Synergistic inhibitory effect of berberine and d-limonene on human gastric carcinoma cell line MGC803. *J. Med. Food* 17, 955–962. doi: 10.1089/jmf.2013.2967

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Pharmacology

ORIGINAL RESEARCH
published: 05 June 2018
doi: 10.3389/fphar.2018.00583

DeCoST: A New Approach in Drug Repurposing From Control System Theory

Thanh M. Nguyen[1]*, *Syed A. Muhammad*[2]*, *Sara Ibrahim*[3], *Lin Ma*[4], *Jinlei Guo*[4], *Baogang Bai*[4] and *Bixin Zeng*[5]*

[1] *Department of Computer and Information Science, Indiana University-Purdue University Indianapolis, Indianapolis, IN, United States,* [2] *Institute of Molecular Biology and Biotechnology, Bahauddin Zakariya University, Multan, Pakistan,* [3] *Department of Biology, School of Science, Indiana University-Purdue University Indianapolis, Indianapolis, IN, United States,* [4] *The 1st School of Medicine and School of Information and Engineering, Wenzhou Medical University, Zhejiang, China,* [5] *Institute of Lasers and Biomedical Photonics, Wenzhou Medical University, Wenzhou, China*

Edited by:
Lixia Yao,
Mayo Clinic, United States

Reviewed by:
Yanshan Wang,
Mayo Clinic, United States
Fuhai Li,
The Ohio State University, United States
Zhichao Liu,
National Center for Toxicological Research (FDA), United States

***Correspondence:**
Thanh M. Nguyen
thamnguy@iupui.edu
Syed A. Muhammad
aunmuhammad78@yahoo.com
Bixin Zeng
z_bixin@163.com

Specialty section:
This article was submitted to Translational Pharmacology, a section of the journal Frontiers in Pharmacology

Received: 15 March 2018
Accepted: 15 May 2018
Published: 05 June 2018

Citation:
Nguyen TM, Muhammad SA, Ibrahim S, Ma L, Guo J, Bai B and Zeng B (2018) DeCoST: A New Approach in Drug Repurposing From Control System Theory. Front. Pharmacol. 9:583. doi: 10.3389/fphar.2018.00583

In this paper, we propose DeCoST (Drug Repurposing from Control System Theory) framework to apply control system paradigm for drug repurposing purpose. Drug repurposing has become one of the most active areas in pharmacology since the last decade. Compared to traditional drug development, drug repurposing may provide more systematic and significantly less expensive approaches in discovering new treatments for complex diseases. Although drug repurposing techniques rapidly evolve from "one: disease-gene-drug" to "multi: gene, dru" and from "lazy guilt-by-association" to "systematic model-based pattern matching," mathematical system and control paradigm has not been widely applied to model the system biology connectivity among drugs, genes, and diseases. In this paradigm, our DeCoST framework, which is among the earliest approaches in drug repurposing with control theory paradigm, applies biological and pharmaceutical knowledge to quantify rich connective data sources among drugs, genes, and diseases to construct disease-specific mathematical model. We use linear–quadratic regulator control technique to assess the therapeutic effect of a drug in disease-specific treatment. DeCoST framework could classify between FDA-approved drugs and rejected/withdrawn drug, which is the foundation to apply DeCoST in recommending potentially new treatment. Applying DeCoST in Breast Cancer and Bladder Cancer, we reprofiled 8 promising candidate drugs for Breast Cancer ER+ (Erbitux, Flutamide, etc.), 2 drugs for Breast Cancer ER- (Daunorubicin and Donepezil) and 10 drugs for Bladder Cancer repurposing (Zafirlukast, Tenofovir, etc.).

Keywords: drug repurposing, system control, breast cancer, bladder cancer, pathway, expression profile

INTRODUCTION

Drug repurposing (also called drug repositioning) has become one of the most active areas in pharmacology since last decade (Oprea et al., 2011) because this approach could significantly reduce the cost and time to invent a new treatment. Before drug repurposing research became active, it was expected to take about 15 years and $0.8–$1 billion to bring a new drug into the market (Dimasi, 2001) due to many tests and clinical trials in order to be commercially approved by Food and Drug Administration (FDA) (USFDA, 2016). It is expected that the failure probability during clinical trials is about 91.4% (Thomas et al., 2016). One of the key reasons for low productivity in traditional drug development is the lack of systematic evaluation of additional indications (Dudley et al., 2011), which may lead to unexpected side effects and low efficacy. Briefly, drug repurposing finds new indications for known drugs and compounds (Gupta et al., 2013) to reduce the risk of failure and shorten time of discovery. Drug repurposing applies modern computational techniques to digitalize genomic (Power et al., 2014), bioinformatics, chemical informatics (Bisson, 2012) and patients' individual health records (Xu et al., 2014) to offer more systematic evaluation of the chemical compound before entering the laboratory testing and clinical trial steps. In addition, drug repurposing could explore the large set of chemical compounds, which is estimated to be more than 90 million by PubChem statistics (Wang et al., 2014), to reduce the cost of synthesizing new compounds. Prominent successful examples for drug repurposing include Viagra, Avastin, and Rituxan (Dudley et al., 2011).

System biology (Pujol et al., 2010) plays an important role to in the evolvement of drug repurposing evolved from "one: disease-gene-drug" (Durrant et al., 2010) to "multi: gene, drug" (Chou, 2010; Medina-Franco et al., 2013) and from "lazy guilt-by-association" (Campillos et al., 2008; Keiser et al., 2009; Iorio et al., 2010; Gottlieb et al., 2011) to "systematic model-based pattern matching," such as the Broad Institute's Connectivity Maps (CMAP), C2MAP, etc. (Lamb et al., 2006; Hu and Agarwal, 2009; Huang et al., 2012; Jensen et al., 2012; Li and Lu, 2013; Subramanian et al., 2017). System biology reveals connectivity among drug, gene, and diseases (**Figure 1**). In this Figure, the green connectivity shows the types of connectivity for which drug repurposing could utilize to answer the key question: could drug A be re-indicated to treat disease B. The literature and public data sources for these types of connectivity have been thoroughly developed in the recent two decades, such as DrugBank (Law et al., 2013) and SFINX (Andersson et al., 2015) for drug-drug interaction; DrugBank (Law et al., 2013) and STITCH (Kuhn et al., 2012) for drug-gene/protein interaction; BioGRID (Chatr-Aryamontri et al., 2013), STRING (Szklarczyk et al., 2015), HAPPI (Chen et al., 2017), KEGG (Kanehisa et al., 2017) and Reactome (Croft et al., 2011) for protein-protein interaction and human pathway; OMIM (Baxevanis, 2012) and GEO (Barrett et al., 2013) for disease-specific gene curation and analysis; the human disease network (Goh et al., 2007) for disease-disease connectivity; and SIDER for diseases' drug-side-effect (Kuhn et al., 2016). The integration of rich data sources enable mathematical system modeling and analysis in system biology to deepen our understanding and predictive capability for biological processes, disease ontology (Hannon and Ruth, 2014; Goel and Richter-Dyn, 2016; Woodhead et al., 2016) and personalized medicine (Weston and Hood, 2004).

From the mathematical system-model-control-based point of view, there exist a mechanism regulating the gene expression profile. In the healthy condition, the gene expression stays in the stable equilibrium region such that $\mathbf{x}(t) = f\ (\mathbf{x}(t{-}1)) \approx \mathbf{x}(t{-}1)$, where f indicates the expression-regulating mechanism computed from data integration, $\mathbf{x}$ stands for expression and t stands for time. In the disease state, the critical gene expression strays outside the stable region. In this case, without a control (treatment), the expression will be unbounded. The system control algorithms aim to find the sequence of control-treatment that optimally stabilize the expression back to the original equilibrium point, such as linear control (Willems, 1971; Chen et al., 2016), nonlinear control (Bardi and Capuzzo-Dolcetta, 2008; Falcone and Ferretti, 2013), adaptive neural network (Rovithakis and Christodoulou, 1994; Tong et al., 2014). By comparing the real drug treatments with the optimal control-treatment (also called hypo-treatment), we can evaluate the potential efficacy of the drug before being repurposed.

However, applying mathematical system modeling and control in drug repurposing is still in very early steps. There are three key challenges in applying system control approach. First, it is difficult to quantify the gene expression and real drug treatment, as there is very little literature discussing the "normal range" of each gene's expression. Second, constructing a comprehensive and accurate mathematical model to simulate the gene expression change is complicated due to the diversity of gene-gene interaction mechanisms, mutation, and under-discovered data. Third, the biological systems are known for

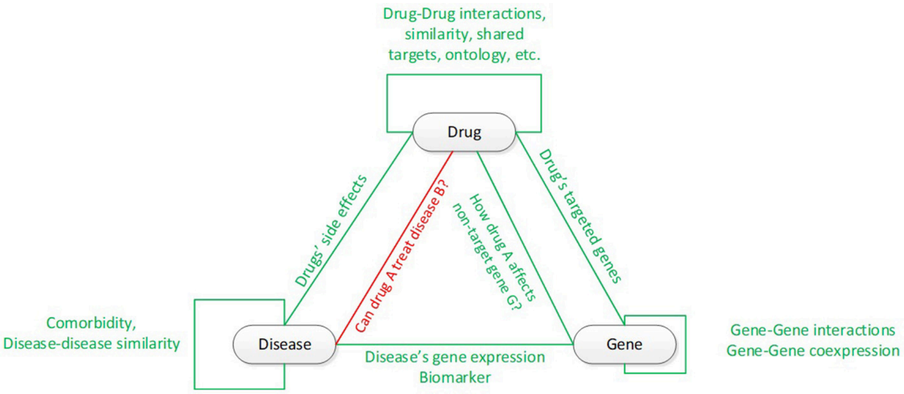

Drug-Drug interactions, similarity, shared targets, ontology, etc.
Drug
Drugs' side effects
Can drug A treat disease B?
How drug A affects non-target gene G?
Drug's targeted genes
Comorbidity, Disease-disease similarity
Disease
Disease's gene expression Biomarker
Gene
Gene-Gene interactions Gene-Gene coexpression

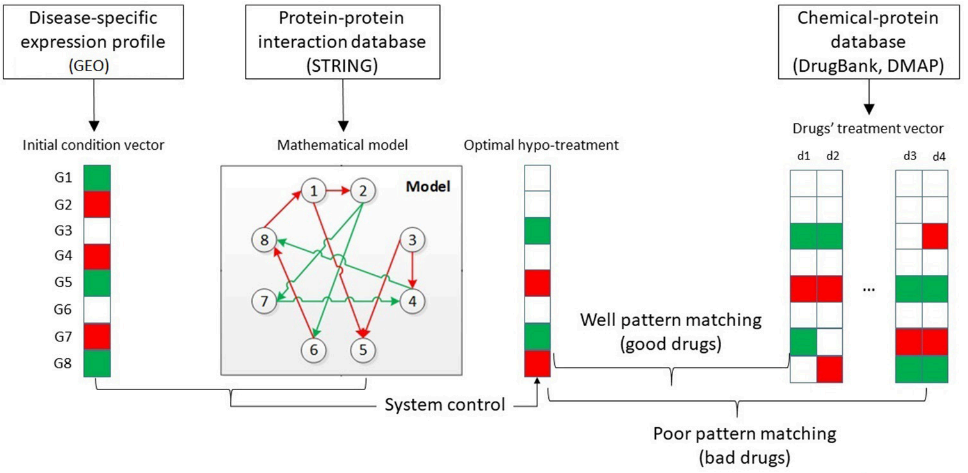
FIGURE 1 | Connectivity among drugs, genes, and diseases. The red line and text show the key connectivity in drug repurposing.

large scale for system control: there may be from hundreds to thousands of genes of interest in a specific disease or biological process.

In this paper, we propose DeCoST (Drug Repurposing from Control System Theory) to apply control system paradigm for drug repurposing purpose, with source code available at https://github.com/thamnguy/DeCoST. The DeCoST framework tackles these challenges above as follow. First, although we could not completely solve the "normal range" challenge, we discretized the gene expression and the connectivity data so that the control-system algorithm could be executed logically without the "normal range" impact. Second, to overcome the comprehensiveness challenge, we utilized the biological and pharmaceutical knowledge and public data sources to quantify the drug-protein interaction and disease-specific gene expression profile. We used the comprehensive public protein-protein databases to setup the mathematical model for the repurposing problem. Third, to reduce the complexity and high-dimensionality of the repurposing problem, we applied the linear-quadratic-regulator method, which is practical in large-scale system control, to compute the hypo-treatment and evaluate the drug therapy. We apply DeCoST in Breast Cancer and Bladder Cancer case studies. Among cancer diseases, Breast Cancer causes the most number of mortality women (Centers for Disease Control Prevention, 2013). Breast Cancer is also the most comprehensively studied disease among cancers, with nearly 20 approved drugs by Food and Drug Administration (FDA). In addition, Breast Cancer has many subtypes, which is ideal for personalized drug repurposing. In contrast, FDA only approves 4 drugs for Bladder Cancer treatment although Bladder Cancer is the fourth most commonly diagnosed cancer in the United States (American Cancer Society, 2017). Therefore, drug development in Bladder Cancer is still an opened and attractive research area. From good performance when classifying between approved drugs and withdrawn drugs, we find 7 compounds that may be promising in Breast Cancer ER-positive subtype, 3 compounds in Breast Cancer ER-negative subtype and 10 compounds in Bladder Cancer for further drug repurposing *in-vivo* study.

METHODS

We developed our drug repurposing framework from the system modeling and control points (**Figure 2**). The framework integrates three types of data. First, from the Disease-specific expression profile, we quantified the expression as the system initial condition vector, where each vector elements specified whether the corresponding gene was overexpressed (red), underexpressed (green) or normally expressed (white). Second, from the protein-protein interaction database, we built the mathematical system model in order to apply the system-control algorithm. The red arrows implies activative; and the green arrow implies inhibitive interactions. Third, from the chemical-protein interaction data, we quantified the treatment vector for each


Disease-specific expression profile (GEO)
Protein-protein interaction database (STRING)
Chemical-protein database (DrugBank, DMAP)
Initial condition vector
Mathematical model
Optimal hypo-treatment
Drugs' treatment vector
d1
d2
d3
d4
G1
G2
G3
G4
G5
G6
G7
G8
Model
1
2
3
4
5
6
7
8
...
Well pattern matching (good drugs)
System control
Poor pattern matching (bad drugs)

$$\mathbf{x}_0 = \begin{bmatrix} -1 \\ 1 \\ 0 \\ 1 \\ -1 \\ 0 \\ 1 \\ -1 \end{bmatrix} \quad \mathbf{A}^* = \begin{bmatrix} 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & -1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad \mathbf{u}_0 = \begin{bmatrix} 0 \\ 0 \\ -1 \\ 0 \\ 1 \\ 0 \\ -1 \\ 1 \end{bmatrix}$$

$$\mathbf{u}_d \quad \begin{array}{cccc} \text{d1} & \text{d2} & \text{d3} & \text{d4} \\ \begin{bmatrix} 0 \\ 0 \\ -1 \\ 0 \\ 1 \\ 0 \\ -1 \\ 0 \end{bmatrix} & \begin{bmatrix} 0 \\ 0 \\ -1 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \end{bmatrix} & \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ -1 \\ 0 \\ 1 \\ -1 \end{bmatrix} & \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ -1 \\ 0 \\ 1 \\ -1 \end{bmatrix} \end{array}$$

FIGURE 2 | Overview of our drug repurposing framework and mathematical representation of drug, protein and interactome data. Red squares: overexpressed genes/drug's activation. Green squares: under expressed genes/drug's inhibition. Red arrow: activated protein-protein interaction. Green arrows: inhibited protein-protein interaction.

drug for later ranking. Using the initial condition vector and the mathematical model, we computed the optimal hypo-treatment. By mapping the pattern of the optimal hypo-treatment and the drugs' treatment vectors, we could rank the drugs and suggest repurposed drugs.

Retrieve the Expression Profile as the Initial Condition Vector

We used GEO2R service (https://www.ncbi.nlm.nih.gov/geo/geo2r/) to analyze GEO dataset for the initial condition vector. The GEO2R service runs on R 3.2.3 platform and utilizes the well-known bioinformatics packages Biobase 2.30.0 (Huber et al., 2015), GEOquery 2.40.0 (Davis and Meltzer, 2007), and limma 3.26.8 (Ritchie et al., 2015). In GEO2R's result, we filtered out genes whose adjusted p-values exceed 0.05. The filtered-out genes were marked with 0 in the initial condition vector. For genes, whose adjusted p-values are less than 0.05, we used the sign of base-10 logarithm fold-change (logFC) in the initial condition vector. In the other words, genes with logFC > 0, which implied that the genes were overexpressed in the disease condition, were marked by 1. Genes with logFC < 0, which implied that the gene were under expressed in the disease condition, were marked by -1.

We chose GSE10886 dataset for expression profile in Breast Cancer case study. GSE10886 is among the largest and most comprehensive Breast Cancer microarray in GEO at the tissue level. After the latest update in January 2013, GSE10886 has 226 samples and including 97 ER-positive-subtype samples, 69 ER-negative-subtype samples, and 32 control samples. We chose GSE31189 dataset for Bladder Cancer expression profile. This dataset contains 52 cancer samples and 40 control samples.

Build Disease-Specific Mathematical System Model From Interactome Data

Due to the availability of public data sources for disease-specific pathway models, we built the disease-specific system model for Breast and Bladder Cancer differently. To avoid potential false-positive, which is a well-known issue in predictive data source, we preferred using the pathway data to construct the mathematical model. For Breast Cancer, we conducted literature search on public curated pathway databases Reactome (Croft et al., 2011) and Wikipathway (Pico et al., 2008) for human disease pathways. In these databases, we only select pathways where the disease name appears in the pathways' titles or description. As the result, we found the Integrated Breast Cancer Pathway (Ibrahim et al., 2015) on Wikipathway. This pathway is among the most comprehensive Breast Cancer human pathway in the literature, which covers 239 genes and 467 interactions. The pathway also integrates 24 Breast Cancer-related pathways, including several signaling network. The entire detail about this pathway could be found in Supplemental Table 2. However, we could not find any pathways having more than 50 genes for Bladder Cancer, which implied low coverage. Therefore, for the Bladder Cancer model, we queried Bladder-Cancer-associated genes from PubMed Gene (https://www.ncbi.nlm.nih.gov/gene), one of the most comprehensive literature collection in biomedical and life sciences. To filter the possible noise during the retrieval process, we used specific query in format: <Disease Name> AND "Homo sapiens"[porgn: __txid9606]. After retrieving the Bladder-Cancer-associated genes, we converted the gene identification to UniProt Knowledge Base Reviewed identification (UniProt, 2013) to filter possible alias. We queried the STRING database v10 (Szklarczyk et al., 2015), one of the most comprehensive interactome databases to retrieve the interactions information among the candidate disease-specific proteins, especially the directionality and mechanism of interactions. To filter out possible noisy information, we limited the search results only on interaction with minimum of 500 confidence score. STRING database covers 7 types of mechanism: activation, expression, inhibition, catalysis, ptmod, binding, reaction.

After retrieving the disease-associated genes and interactions from these models above, we quantified the interactome to finalize the mathematical systems for these diseases. Among the interactions, activation and inhibitions are the mechanisms with the clearest and the most unambiguous impact/directionality. Thus, we quantified the activation mechanisms by $+1$ and the inhibition mechanisms by -1. For the other mechanisms, we quantified them by the default value of 0. The results of this step could be represented by adjacency matrices, as showed in Supplemental Figure 1.

Retrieve Chemical-Protein Interaction for Treatment Vector

For each disease, we curated literature for two set of drugs. The positive set, denoted by D1, includes all drugs which are approved for treatment by Food and Drug Administration (FDA). The negative set, denoted by D2, includes drugs which are withdrawn from disease treatment, or withdrawn/terminated from disease-specific clinical trials due to toxic or inefficient issues. We query https://clinicaltrials.gov/ for clinical trials information. To avoid the complexity of multi-drug and multi-disease treatment, we ignored literature mentioning more than one drug/disease during curation. We also ignored the biotech drugs since this type of drug does not target the molecular level, therefore it is difficult to setup the treatment vector for biotech drugs. **Table 1** summarizes the list of D1 and D2 drugs we curated for Breast Cancer and Bladder Cancer. For Breast Cancer, we found 16 D1 drugs and 7 D2 drugs. In addition, to examine the possible newly therapeutic drugs for Breast Cancer, we referred to 24 drug proposed by Huang et al. (2011) as D3, in which these drugs have been approved for some other diseases by never in trial for Breast Cancer. For Bladder Cancer, we found 3 D1 drugs and 2 D2 drugs. Since we could not find any repurposed drug list for Bladder Cancer in the literature, we selected all of the 421 FDA-approved drugs for non-Bladder-Cancer diseases, which have at least one drug-gene interaction with genes in Bladder Cancer model, as D3 for Bladder Cancer. The entire D3 drug lists for both Breast Cancer and Bladder Cancer could be found in Supplemental Table 1.

We queried the DrugBank (Law et al., 2013) and DMAP (Huang et al., 2015) database for the list of drug-protein

TABLE 1 | Drug lists (D1 and D2) curated for Breast and Bladder cancer.

Disease	Drugs	Drug sets	Disease	Drugs	Drug sets
Breast cancer	Anastrozole	D1	Breast cancer	Trastuzumab	D1
Breast cancer	Cycloheximide	D1	Breast cancer	Vinblastine	D1
Breast cancer	Exemestane	D1	Breast cancer	Diethylstilbestrol	D2
Breast cancer	Fluorouracil	D1	Breast cancer	Dromostanolone	D2
Breast cancer	Fluoxymesterone	D1	Breast cancer	Formestane	D2
Breast cancer	Fulvestrant	D1	Breast cancer	Ixabepilone	D2
Breast cancer	Lapatinib	D1	Breast cancer	Avastin	D2
Breast cancer	Letrozole	D1	Breast cancer	Ethyl Carbamate	D2
Breast cancer	Miltefosine	D1	Breast cancer	Imetelstat	D2
Breast cancer	Paclitaxel	D1	Breast cancer	Tivozanib	D2
Breast cancer	Pamidronate	D1	Bladder cancer	Cisplatin	D1
Breast cancer	Raloxifene	D1	Bladder cancer	Doxorubicin HCl	D1
Breast cancer	Tamoxifen	D1	Bladder cancer	Thiotepa	D1
Breast cancer	Thiotepa	D1	Bladder cancer	Mitomycin C	D2
Bladder cancer	Gemcitabine	D2			

D1, FDA-approved drugs (positive/good drug set); D2, FDA-rejected/withdrawn drugs (negative/bad drug set).

interaction mechanism. DMAP and DrugBank covers 38 mechanisms of drug action. In DMAP, we filtered out interactions with confidence score less than 800 (over 1,000) to avoid noisy information. From biological knowledge, we quantified these mechanisms as showed in **Table 2**. Similar to quantification of protein-protein mechanism of action, an inhibited or similar action is map to −1; and an activated or similar action is map to +1.

TABLE 2 | Quantification of drug-protein mechanism of action in drug-protein interaction databases.

Mechanism of action	Quantification	Mechanism of action	Quantification
Activator	1	Ligand	0
Adduct	0.5	Metabolizer	0
Agonist	1	Modulator	0
Allosteric modulator	0	Multitarget	0
Antagonist	−1	Negative modulator	−1
Antibody	0	Neutralizer	0
Binder	0	Other	0
Chaperone	1	Other/unknown	0
Chelator	0	Partial agonist	1
Cleavage	−1	Partial antagonist	−1
Cofactor	1	Positive allosteric modulator	1
Component of	0	Potentiator	1
Cross-linking/alkylation	0	Product of	0
Incorporation into and destabilization	−1	Reducer	−1
Inducer	1	Stimulator	1
Inhibitor	−1	Suppressor	−1
Inhibitor, competitive	−1	Unknown	0
Inhibitory allosteric modulator	−1	Other terms	0
Intercalation	0	–	–

The Mechanism of Action terminologies are retrieved from drug-target annotation in DrugBank database. Quantification stands for the numerical representation of the Mechanism of Action in the modeling and computing steps.

Construct Disease-Specific Drugs' Therapeutic Scoring for Drug Repurposing Purpose

The key principle in applying system control to evaluate drugs' therapy relies in the following assumption: in disease condition, the gene expressions are derived away from the balanced level of 0. Therefore, a good treatment should reverse the gene expressions in disease condition and stabilize the expressions to the balance level. In **Figure 2**, we illustrate this principle and explain several mathematical notation in a toy example. Based on system biology literature (Alberghina, 2007), we assume that there exists a model governing the gene expressions, which allows us to model the expression using time-series perspective

$$\mathbf{x}(t) = f(\mathbf{x}(t-1), \mathbf{u}(t-1)) \tag{1}$$

where $\mathbf{x} \epsilon \Re^N$ stands for the quantified gene expression of N genes, $\mathbf{u} \epsilon \Re^N$ stands for the quantified treatment and t is the iteration and f is the arbitrary function controlling the expression change. The initial $\mathbf{x}(0)$ is the quantified gene expression in disease condition. In this paper, we choose a linear model for f.

$$\mathbf{x}(t) = \mathbf{A}\mathbf{x}(t-1) + \mathbf{u}(t-1) \tag{2}$$

We chose the linear model because not only it is simple but also it has equilibrium point at the origin: if $\mathbf{x}(t{-}1) = \mathbf{u}(t{-}1) = 0$ then $\mathbf{x}(t) = 0$. This fact implies that when the gene expressions are already at the balance level, treatment is no longer needed. In addition, it is easier to setup a linear system with stability

(Chui and Chen, 2012)

$$If\,||\mathbf{x}(0)|| < \varepsilon \text{ and } \mathbf{u} = 0 \text{ then } ||\,\mathbf{x}(t)\,|| < \varepsilon\forall t \tag{3}$$

where $||\mathbf{x}||$ stands for the second norm of $\mathbf{x}$ and ε is an arbitrary small number. This fact implies the self-adjustment of the gene expression at the control level. We setup matrix $\mathbf{A}$ from quantification of protein-protein mechanism of interactions (section Methods). With temporal matrix $\mathbf{A}^*$ as the result of section Methods

$$\mathbf{A}^*(i,j) \;=\; \begin{cases} -1 \text{ if protein } i \text{ inhibits protein } j \\ 1 \text{ if protein } i \text{ activates protein } j \\ 0 \text{ otherwise} \end{cases} \tag{4}$$

Let λ be the eigenvalue of $\mathbf{A}^*$ with the largest magnitude. By setting up $\mathbf{A}$ as

$$\mathbf{A} = (1/\lambda)\,\mathbf{A}^* \tag{5}$$

We can guarantee the stability of system (2) (Chui and Chen, 2012).

The objective of the linear control is to find a sequence of $\mathbf{u}$(t) such that

$$\mathbf{x}(t) \to 0 \text{ as } t \to \infty \tag{6}$$

Optimal control considers not only how to stabilize $\mathbf{x}$ quickly but also consider the cost-effective of the treatment $\mathbf{u}$. Regarding this point, the optimal linear control aims to minimize

$$J(\mathbf{x}\,(0)) = \sum\nolimits_{t\,=\,0}^{\infty} \left(\mathbf{x}\,(t)^{\boldsymbol{T}}\mathbf{x}\,(t) + \mathbf{u}(t)^{\boldsymbol{T}}\mathbf{u}\,(t)\right) \tag{7}$$

To solve the optimization problem (2–7) we solved the corresponding Riccati equation (Arnold and Laub, 1984)

$$\mathbf{A^T PA - P - A}^T\mathbf{P}\,(\mathbf{P+I})^{-1}\,\mathbf{PA + I} = 0 \tag{8}$$

using DARE algorithm (Arnold and Laub, 1984) in Matlab (https://www.mathworks.com/help/control/ref/dare.html). In (8), P is just an intermediate result containing no biological representation. We compute the treatment vector $\mathbf{u}(t)$ as follow

$$\mathbf{u}\,(t) = -\,(\mathbf{I+P})^{-1}\,\mathbf{PAx}(t) \tag{9}$$

In system control practice, since $\mathbf{u}(t)$ often converges to 0 quickly (Bemporad et al., 2002), the first treatment vector $\mathbf{u}(0) = -\,(\mathbf{I+P})^{-1}\,\mathbf{PAx}(0)$ often plays the most important role in optimally stabilizing the system (2). Therefore, we can consider $\mathbf{u}(0)$ as the optimal hypo-treatment. We compare the similarity between the real drug treatment ($\mathbf{u}_d$) and the hypo-treatment as the therapeutic score $T(d)$ for each drug d as follow

$$T_d = |\mathbf{u}_d^T\text{sign}(\mathbf{u}\,(0))|/|\text{abs}(\mathbf{u}_d)^T\text{abs}(\text{sign}(\mathbf{u}\,(0)))| \tag{10}$$

where abs stand for the absolute value function. Here, T_d ranges between −1 and 1. The numerator $\left|\mathbf{u}_d^T sign(\mathbf{u}\,(0))\right|$ is the matching function between drug d and the optimal hypo-treatment, which is incremented when $\mathbf{u}_d(i)$ and $\mathbf{u}(0)(i)$ are non-zero analog, and decremented when $\mathbf{u}_d(i)$ and $\mathbf{u}(0)(i)$ are opposite. We measured the impact of T_d score by the receiver operating characteristic when we use T_d to classify D1 drugs vs. D2 drugs.

RESULTS

Therapeutic Scores for Breast Cancer Drugs

From the Integrated Breast Cancer Pathway (Ibrahim et al., 2015) on Wikipathway (section Methods) and the Breast Cancer drug list in Supplemental Table 3, we queried 222 drug-protein interactions for the drugs' treatment vectors (Supplemental Table 4). Supplemental Table 5 contains the initial condition vector from GEO2R expression analysis.

Figure 3 shows that the T$_d$ score is able to give appropriate ranking for most of the well-known therapeutic drugs and suggest candidate drugs for repurposing in Breast Cancer ER-positive case. T$_d$ score reflexes the difference between the D1 and D2 drugs with receiver operator characteristic (Hanley and McNeil, 1982) area under the curve (AUC) of 0.76. This result is comparable to the overall result queried from Broad Institute CMAP (Subramanian et al., 2017) on MCF-7, the Breast Cancer ER+ cell line, using the Touchstone tool (https://clue.io/touchstone). Especially on the drugs covered in CMAP, DeCoST achieves AUC of 0.91, which is much higher than the AUC achieved by CMAP (0.79), as showed in the Supplemental Text 1. We did not setup training set and test set for classification because the model construction and T_d calculation does not need the drug categories. The T$_d$ scores for D1 drugs in Breast Cancer ER-negative case are relatively lower than the scores for ER-positive case (**Figure 4**). Comparison detail has been shown in Supplemental Table 5. Using T$_d$ for classifying D1 and D2 drugs yields AUC of 0.68. In fact, clinical trials and literature have showed several drugs which are effective in ER-positive treatment but show little or no impact in ER-negative treatment. For example, Tamoxifen (T_d ER-positive: 0.294, T_d ER-negative: 0.176), which is a selective estrogen receptor modulator, does not prevent ER-negative Breast Cancer, when the estrogen receptor genes do not express (Fabian, 2007; Uray and Brown, 2011).

Therapeutic Scores for Bladder Cancer Drugs

Since we could not find any human pathway with sufficient coverage for Bladder Cancer, our Bladder Cancer system model retrieved the Bladder-Cancer-specific genes from PubMed Gene server. The model contains 738 proteins and 1,241 protein-protein interactions. From 6 drugs in the Bladder Cancer case-study, we retrieved 48 drug-protein interactions for drugs' treatment vector. From GSE31189 gene expression dataset, we found 221 genes whose expression differs from the balance level. Details about the Bladder Cancer system could be found in Supplemental Tables 6–8.

We observed AUC of 1.0 (**Figure 5**) when we used T_d score to classify between D1 and D2 drugs in Bladder Cancer. Here, all of the D1 drugs receive non-negative T_d scores: Cisplatin receives the score of 0.2, Doxorubicin Hydrochloride receives the score of 0.0 and Thiotepa receives the score of 1.0. All of the D2 drugs receive negative T_d scores: Mitomycin C receives the score of −0.2 and Gemcitabine receives the score of −0.09.

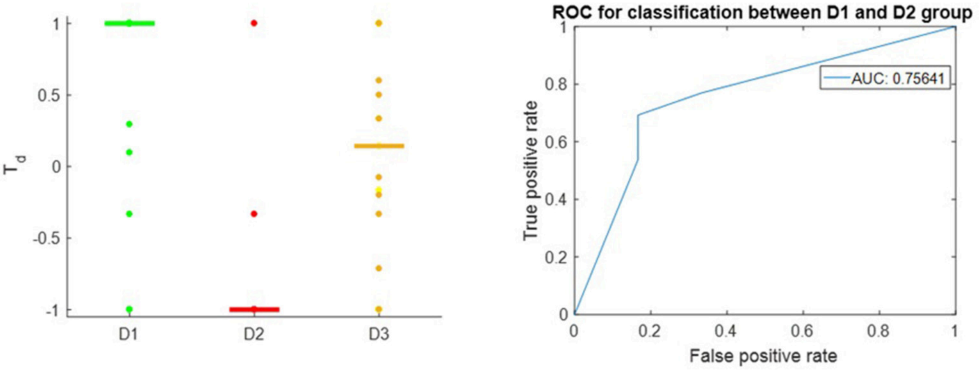


FIGURE 3 | Left: T_d score in Breast Cancer, ER-positive subtype; the horizontal bars in each group stand for median value of T_d. **Right**: ROC of T_d in classifying between D1 drugs and D2 drugs.

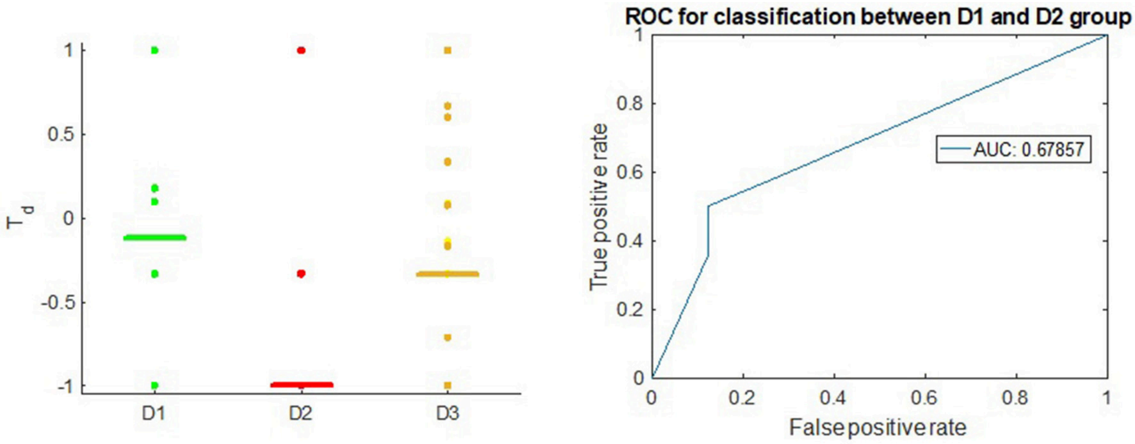


FIGURE 4 | Left: T_d score in Breast Cancer, ER-negative subtype; the horizontal bars in each group stand for median value of T_d. **Right**: ROC of T_d in classifying between D1 drugs and D2 drugs.

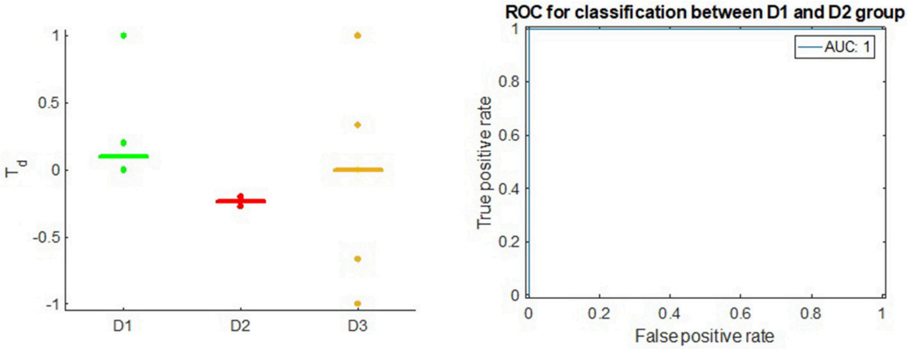


FIGURE 5 | Left: T_d score in Bladder Cancer; the horizontal bars in each group stand for median value of T_d. **Right**: ROC of T_d in classifying between D1 drugs and D2 drugs.

Potential Drugs for Breast Cancer Studies and Biological Insights

From the T_d scores for D3 drugs, our framework suggests 8 drugs (Erbitux, Flutamide, Medrysone, Methylprednisolone, Norethindrone, Prednisolone, Prednisonea, and Vandetanib) with high potential efficacy in Breast Cancer ER+ drug repurposing. Significantly, these drugs do not directly target Estrogen receptor, which is the most well-known approach in Breast Cancer ER+ drug design. Tamoxifen is a typical example of Breast Cancer drugs which slows cancer process by blocking estrogen hormone receptors, preventing hormones from binding to them. About 80% of all breast cancers are ER+: the cancer cells grow in response to the hormone estrogen (Bulut and Altundag, 2015). About 65% of the ER+ cases grow in response to another hormone, progesterone (Hefti et al., 2013). Tumors in ER/PR-positive cases are much more likely to respond to hormone therapy than tumors that are ER/PR-negative. ER+ breast cancer entirely depends on the estrogen for growth and propagation involving genomic and non-genomic pathways. Epidermal growth factor receptor (EGFR) is a receptor found on both normal and tumor cells that is important for cell growth (Herbst, 2004; Khoo et al., 2015). ER-positive (ER+) drugs recommended for repurposing in this framework block the activities and growth of EGFR (**Figure 6A**). These drugs show different mechanism of action with the common objective of the inhibition of the growth of cancerous cells. By adjusting and modifying the known biases of the interactomic networks, our procedure would help to reveal the therapeutic effect of drugs along with effective treatments.

For Breast Cancer ER- case, our framework suggests Daunorubicin and Donepezil as the repurposing candidates. These drugs are independent of estrogen and usually inhibit the cell growth by either interacting with DNA or inhibiting Cholinesterases. Daunorubicin interacts with DNA by intercalation and inhibition of macromolecular biosynthesis (Momparler et al., 1976). This inhibits the progression of the enzyme topoisomerase II, and thereby stopping the process of replication. Donepezil is in a class of cholinesterase inhibitor that improves mental function and fatigue in cancer. The current research focused on recent large-scale efforts to systematically find repositioning candidates and elucidate individual disease mechanisms in cancer (Bruera et al., 2007). Personalized medicine and repositioning both aim to improve the productivity of current drug discovery pipelines. Standard drug discovery strategies can also lead to repositioning opportunities. D1, D2, and D3 drugs (**Table 1**) found to potently modulate the desired activity are repositioning candidates.

Potential Drugs for Bladder Cancer Studies and Biological Insights

From the list of 143 FDA-approved drug with high T_d score, we found 10 candidates drugs (with $T_\mathrm{d} = 1$) whose mechanisms are promising for Bladder Cancer repurposing. The T_d scores for all Bladder Cancer drugs could be found in Supplemental Table 9. The prevalence of drug-repositioning studies has resulted in a variety of innovative computational methods for the identification of new opportunities for the use of old drugs. We sorted the potential list of drugs against bladder cancer. The reprofiling of these drugs followed the same biological mechanisms. For example, Zafirlukast antagonizes ATP-binding cassette and may improve the efficacy of anticancer effects (Sun et al., 2012). Similarly, Tenofovir may reduce the risk of bladder or others cancers while dopamine receptor antagonist Thioridazine inhibits tumor growth (Yin et al., 2015). Losartan

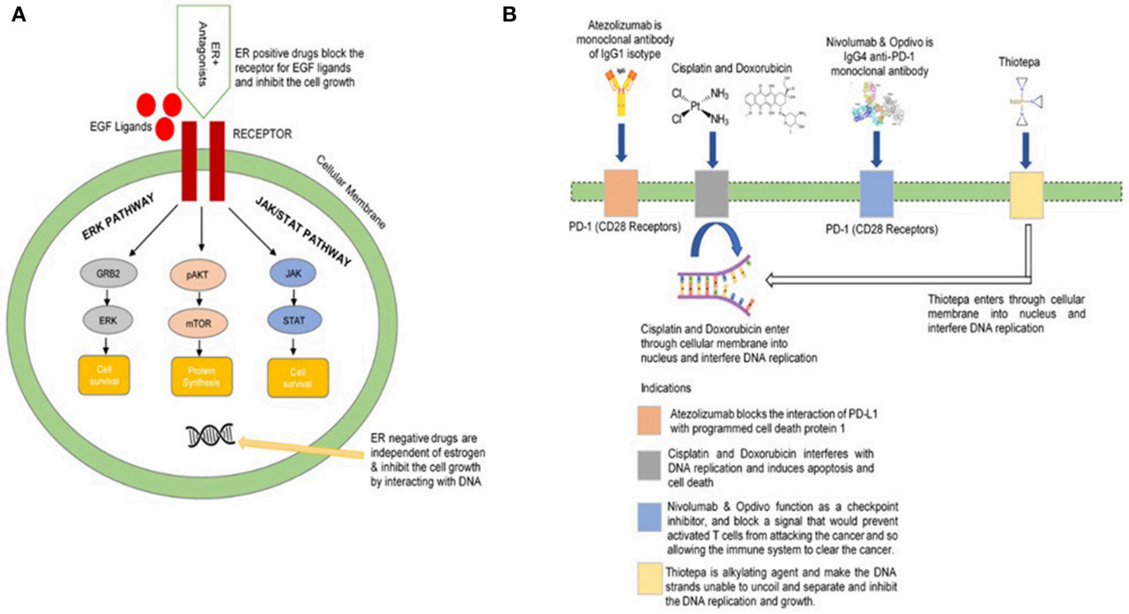

A
ER+ Antagonists
ER positive drugs block the receptor for EGF ligands and inhibit the cell growth
EGF Ligands
RECEPTOR
Cellular Membrane
ERK PATHWAY
JAK/STAT PATHWAY
GRB2
ERK
Cell survival
pAKT
mTOR
Protein Synthesis
JAK
STAT
Cell survival
ER negative drugs are independent of estrogen & inhibit the cell growth by interacting with DNA
B
Atezolizumab is monoclonal antibody of IgG1 isotype
Cisplatin and Doxorubicin
Cl
Cl
Pt
NH3
NH3
Nivolumab & Opdivo is IgG4 anti-PD-1 monoclonal antibody
Thiotepa
PD-1 (CD28 Receptors)
PD-1 (CD28 Receptors)
Thiotepa enters through cellular membrane into nucleus and interfere DNA replication
Cisplatin and Doxorubicin enter through cellular membrane into nucleus and interfere DNA replication
Indications
Atezolizumab blocks the interaction of PD-L1 with programmed cell death protein 1
Cisplatin and Doxorubicin interferes with DNA replication and induces apoptosis and cell death
Nivolumab & Opdivo function as a checkpoint inhibitor, and block a signal that would prevent activated T cells from attacking the cancer and so allowing the immune system to clear the cancer.
Thiotepa is alkylating agent and make the DNA strands unable to uncoil and separate and inhibit the DNA replication and growth.

FIGURE 6 | Illustration of biological mechanism of few FDA approved drugs **(A)** for breast cancer **(B)** for bladder cancer.

is an angiotensin II receptor (AT-II-R) blocker that is widely used by human for blood pressure regulation but it also shows antitumor property (Barreras and Gurk-Turner, 2003). Ciclopirox was first marketed in 1982 as an antifungal agent found in several topical drug products. However, further research demonstrated that it was able to kill bladder cancer cells (Weir et al., 2011). The Atezolizumab, Cisplatin, Doxorubicin, Nivolumab, Opdivo, Thiotepa, and others (**Figure 6B**) are FDA approved drugs which are recommended for bladder cancer.

DISCUSSION

The applications of drug-repositioning studies have brought a variety of new *in silico* approaches in drug designing and development. In most of the studies, the anticancer effect of newly designed drugs usually has been presented *in vitro* as clinical trials are very expensive and time consuming, but remain the only way to validate drug efficiency *in vivo*. Therefore, to establish accurate and effective drug-repositioning framework needs development of new computational techniques. In this work, we discuss and demonstrate the application of control system theory as a computational method to evaluate drug efficacy and repurposing from integrated system biology data. The capability in classification between approved and withdrawn drugs is the fundamental foundation for our framework in drug repurposing. It is important to note that although our AUC of 0.76 and 0.68 in Breast Cancer is inferior compared to the state-of-the-art methods (Cheng et al., 2012; Zheng et al., 2015), our validation is conducted from the pharmaceutical knowledge of drug's efficacy on treatment at the system-pathway level; meanwhile, the other methods often validate at the targeted molecular level. In addition, we set strict criteria in choosing the negative set by only choosing drugs that are rejected or withdrawn from disease-specific clinical trials and treatments. The state-of-the-art methods tend to be more relaxed on the negative set by choosing drug not being used in disease-specific drugs, which may have limitation on repurposing options. In addition, the appropriate assessment of tamoxifen efficacy between Breast Cancer ER+ and Breast Cancer ER- highlights the potential advantages of our framework in personalized drug repurposing. Compare to the approved drugs, the candidate drugs suggested in this work show different promising drug mechanisms which may be useful in future drug design.

In our work, although the number of target may be among the key difference between the D1 drugs and the D2 drugs, our analysis shows that the number of drugs' targeted genes and the targeted genes are not the only factors affecting the clinical outcome and predictive results in drug repurposing. As showed in Suppemental Table 3, D1 drugs, on the average, has more targets than D2 drugs. However, D1 drugs for Breast Cancer (average number of targets: 4.8) include both single-target (such as Anastrozole, Exemestane, and Fluorouracil) and multi-target (such as Tamoxifen, Paclitaxel, and Cycloheximide) ones. D2 (average number of targets: 3.3) drugs also contains the single-target (such as Ixabepilone and Avastin) and the multi-target (such as Imetelstat and Diethylstilbestrol). In the result section, DeCoST's evaluation for these drugs showed above is appropriate for their clinical outcome. In addition, drugs targeting the same marker genes do not necessary have the same outcome. For example, both Tamoxifen and Diethylstilbestrol target the estrogen receptors ESR1 and ESR2, which are the marker in Breast Cancer ER+ (Yip and Rhodes, 2014). However, their clinical outcomes and DeCoST's evaluation are opposite, primarily because they have opposite mechanisms on the same targets of estrogen receptors: Tamoxifen is the estrogen inhibitor while Diethylstilbestrol is the estrogen activator. Since Breast Cancer ER+ is strongly associated with the overexpression of estrogen receptors (Yip and Rhodes, 2014), Tamoxifen could have therapeutic outcome because it reverses the disease signature. Meanwhile, Diethylstilbestrol should have poor outcome because it shows the analog to the disease signature.

In this work, we have showed the results between DeCoST and the Broad Institute CMAP, which is among the most well-known and comprehensive platforms for drug repurposing. In addition, our strategy of repurposing is similar to CMAP. Although Supplemental Text 1 shows that our DeCoST has higher AUC than CMAP does, it is inappropriate to conclude that DeCoST is better than the CMAP. There are fundamental differences in conducting experiment making comparison not totally solid. First, the expression profiles acquired by CMAP are at the cell line level; meanwhile, in this work DeCoST acquires the expression profile at the tissue level, which is closer to *in-vivo* studies. Second, due to several factors in experimental design, CMAP does not contains cell line for Breast Cancer ER- and Bladder Cancer. CMAP also covered less number of drugs, compared to the drug list evaluated in this work. Therefore, the key point in comparative evaluation should be on the repurposing hypotheses suggested by these platforms in future *in-vivo* studies and the biological insights of these hypotheses. In our results, we have offered several biological explanations why drugs recommended by DeCoST could be repurposed. Unfortunately, we could not compare between CMAP and DeCoST at this point. DeCoST focuses primarily on recommending drugs that have never been in disease-specific clinical trials; meanwhile, CMAP (https://clue.io/repurposing-app) primarily reports on drugs that has been under early phases of clinical trials. Therefore, we believe that DeCoST could provide complimentary advantages, in addition to CMAP.

The advantages of our framework are established not only by advanced computational method but also by two layers of personalized system (Li and Jones, 2012). In the first layer, the disease-specific gene expression could differ among different patients and subtypes, which results in different initial state condition. In the second layer, different types of disturbance among molecular-molecular interactions could be discovered and represented differently in the system modeling step. In our results, we show that Tamoxifen, which is approved to treat Breast Cancer, may not be effective in treating Breast Cancer ER-. The strong support from literature to this evaluation is a good example of the personalized medicine characteristics. In addition, our framework could easily integrate the results from many other state-of-the-art repurposing approaches such as molecular docking and gene-set enrichment analysis to refine the efficacy prediction. The main idea in this framework, which is based on control system theory, could be applied in many other bioinformatics problem, such as target prioritization and discovering new combination of treatments. In addition, our framework could easily be extended to evaluate combination of treatment, with careful preprocessing the drug-drug interaction data (Ayvaz et al., 2015; Wang et al., 2017).

In addition, our framework shows repurposing capacity at both target level and pathway level. At the target level, we show typical examples for EGFR-targeted and ACHE-targeted drugs. Patients being considered for anti-epidermal EGFR therapy are often screened for mutations in the oncogene KRAS (Hoorens et al., 2010) because a constitutively active KRAS gene downstream of EGFR would not be affected by EGFR inhibition. Many diseases have approved combination regimens, such as metastatic colorectal and bladder cancer and its four-drug FOLFIRI (folinic acid, 5-fluorouracil, irinotecan) with cetuximab regimen (Raoul et al., 2009). Losartan is an angiotensin II receptor (AT-II-R) blocker and this angiotensin-converting enzyme inhibitors (ACE) may have a protective role in bladder and other cancers (Yazdannejat et al., 2016). In the other hand, a typical example at the pathway level is Thioridazine. Thioridazine-induced effects are associated with inhibition of the canonical NFκB pathway.

The limitations in this work are the method to quantify the categorical data from public genomic/proteomic databases and the simplicity of linear system control. First, all of the data are discretized into only three values: −1, 0, and 1, which could lower the resolution of the final drug therapeutic score. Second, the linear system control approach needs to assume that the gene expression transition could be approximate closely by a linear equation, which is still unverified due to the scarcity of time-series gene expression data. Therefore, when applying into another repurposing problem, biologists and pharmacologists should apply deeper domain knowledge to increase the resolution of discrete quantification. Furthermore, mathematical nonlinear system identification and reinforcement learning, which are popular approach in unknown system control, could be used to increase the accuracy of system modeling and make the system more personalized. Integration of other resources, such as drugs, genes, and systems associated with side-effects (Kuhn et al., 2016; Maier et al., 2018) and high-throughput screening (Deftereos et al., 2011; Macarron et al., 2011) would also be valuable expansions of this work in the future. Also, the computational complexity of DeCoST is generally high (expected O(n^8), where n is the number of genes in the model). This complexity is manageable with most of the existing biological pathway model (expect about 400 genes). However, this could be a bottleneck if the number of genes raises to several thousands.

In addition, the advantages of our framework in personalized medicine may associate with the reproducibility issues (Draghici et al., 2006; Frye et al., 2015). As mentioned, the disease-specific gene expression could differ among different patients and subtypes. Therefore, we could not completely guarantee that applying our framework on different gene expression data and on different interactome data sources (Chatr-Aryamontri et al., 2013; Szklarczyk et al., 2015) would return the same result. Therefore, by reproducibility, we can only guarantee that given a specific gene expression profile and an interactome data source, we can always produce the same result. In this work, we have tried to tackle the reproducibility issue by using tight criteria to select the positive/negative drug set, by maintaining the relevance and coverage of the disease-specific model, and by choosing the expression data set with high number of samples.

CONCLUSION

In this work, we have developed DeCoST, one of the first techniques from system control paradigm, to tackle the drug repurposing challenges. We showed that DeCoST could appropriately retrieve the clinical outcomes of drugs treating personalized Breast Cancer and Bladder Cancer. From the good retrieval result, DeCoST suggests repurposing 8-candidate drugs for Breast and 10 drugs for Bladder Cancer with biological insights. This framework would be promising to discover new therapeutic strategies to treat other cancer diseases.

AUTHOR CONTRIBUTIONS

TN designed the study (including the mathematical details), curated the Bladder Cancer drug dataset and analyzed the computational results. SM validated and provided biological insights for the results. SI constructed the Breast Cancer pathway model and collected the drug clinical outcomes for Breast Cancer. LM processed the expression and protein-protein interaction data for Bladder Cancer. LM, JG, BB, and BZ implemented the system control algorithm used in the paper. All authors contributed to the manuscript writing and edition.

ACKNOWLEDGMENTS

The authors thank Dr. Jake Chen from Informatics Institute—School of Medicine—the Alabama University of Birmingham for helpful comments in the study design. The authors also thank Wenzhou Medical University, Zhejiang, China for covering the publication cost for this paper.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphar.2018.00583/full#supplementary-material

REFERENCES

Alberghina, L. (2007). *Systems Biology: Definitions and Perspectives*. Berlin: Springer Science and Business Media.

American Cancer Society (2017). *Key Statistics for Bladder Cancer*. American Cancer Society.

Andersson, M. L., Bottiger, Y., Bastholm-Rahmner, P., Ovesjo, M. L., Veg, A., and Eiermann, B. (2015). Evaluation of usage patterns and user perception of the drug-drug interaction database SFINX. *Int. J. Med. Inform.* 84, 327–333. doi: 10.1016/j.ijmedinf.2015.01.013

Arnold, W. F. III, and Laub, A. J. (1984). Generalized eigenproblem algorithms and software for algebraic Riccati equations. *Proc. IEEE* 72, 1746–1754. doi: 10.1109/PROC.1984.13083

Ayvaz, S., Horn, J., Hassanzadeh, O., Zhu, Q., Stan, J., Tatonetti, N. P., et al. (2015). Toward a complete dataset of drug-drug interaction information from publicly available sources. *J. Biomed. Inform.* 55, 206–217. doi: 10.1016/j.jbi.2015.04.006

Bardi, M., and Capuzzo-Dolcetta, I. (2008). *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations*. Boston, MA: Springer Science and Business Media.

Barreras, A., and Gurk-Turner, C. (2003). Angiotensin II receptor blockers. *Proc. Bayl. Univ. Med. Cent.* 16, 123–126. doi: 10.1080/08998280.2003.11927893

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., et al. (2013). NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.* 41, D991–D995. doi: 10.1093/nar/gks1193

Baxevanis, A. D. (2012). Searching online mendelian inheritance in man (OMIM) for information on genetic loci involved in human disease. *Curr. Protoc. Hum. Genet.* 13, 11–10. doi: 10.1002/0471250953.bi0102s27

Bemporad, A., Morari, M., Dua, V., and Pistikopoulos, E. N. (2002). The explicit linear quadratic regulator for constrained systems. *Automatica* 38, 3–20. doi: 10.1016/S0005-1098(01)00174-1

Bisson, W. H. (2012). Drug repurposing in chemical genomics: can we learn from the past to improve the future? *Curr. Top Med. Chem.* 12, 1883–1888. doi: 10.2174/1568026128045473 44

Bruera, E., El Osta, B., Valero, V., Driver, L. C., Pei, B. L., Shen, L., et al. (2007). Donepezil for cancer fatigue: a double-blind, randomized, placebo-controlled trial. *J. Clin. Oncol.* 25, 3475–3481. doi: 10.1200/JCO.2007.10.9231

Bulut, N., and Altundag, K. (2015). Does estrogen receptor determination affect prognosis in early stage breast cancers? *Int. J. Clin. Exp. Med.* 8, 21454–21459.

Campillos, M., Kuhn, M., Gavin, A. C., Jensen, L., J., and Bork, P. (2008). Drug target identification using side-effect similarity. *Science* 321, 263–266. doi: 10.1126/science.1158140

Centers for Disease Control and Prevention (2013). *Breast Cancer Statistics, October 23, 2013*. Available online at: http://www.cdc.gov/cancer/breast/statistics/

Chatr-Aryamontri, A., Breitkreutz, B. J., Heinicke, S., Boucher, L., Winter, A., and Tyers, M. (2013). The BioGRID interaction database: 2013 update. *Nucleic Acids Res.* 41, D816–D823. doi: 10.1093/nar/gks1158

Chen, J., Pandey, R., and Nguyen, T. M. (2017). HAPPI-2: a comprehensive and high-quality map of human annotated and predicted protein interactions. *BMC Genomics* 18:182 doi: 10.1186/s12864-017-3512-1

Chen, M. Z., Zhang, L., Su, H., and Chen, G. (2016). Stabilizing solution and parameter dependence of modified algebraic Riccati equation with application to discrete-time network synchronization. *IEEE Trans. Automat. Contr.* 61, 228–233. doi: 10.1109/TAC.2015.2434011

Cheng, F., Liu, C., Jiang, J., Lu, W., Li, W., Liu, G., et al. (2012). Prediction of drug-target interactions and drug repositioning via network-based inference. *PLoS Comput. Biol.* 8:e1002503. doi: 10.1371/journal.pcbi.1002503

Chou, T.-C. (2010). Drug combination studies and their synergy quantification using the Chou-Talalay method. *Cancer Res.* 70, 440–446. doi: 10.1158/0008-5472.CAN-09-1947

Chui, C. K., and Chen, G. (2012). *Linear Systems and Optimal Control*. Berlin: Springer Science and Business Media.

Croft, D., O'Kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., et al. (2011). Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.* 39, D691–D697. doi: 10.1093/nar/gkq1018

Davis, S., and Meltzer, P. S. (2007). GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics* 23, 1846–1847. doi: 10.1093/bioinformatics/btm254

Deftereos, S. N., Andronis, C., Friedla, E. J., Persidis, A., and Persidis, A. (2011). Drug repurposing and adverse event prediction using high-throughput literature analysis. *Wiley Interdiscip. Rev. Syst. Biol. Med.* 3, 323–334. doi: 10.1002/wsbm.147

Dimasi, J. A. (2001). New drug development in the United States from 1963 to 1999. *Clin. Pharmacol. Ther.* 69, 286–296. doi: 10.1067/mcp.2001.115132

Draghici, S., Khatri, P., Eklund, A. C., and Szallasi, Z. (2006). Reliability and reproducibility issues in DNA microarray measurements. *Trends Genet.* 22, 101–109. doi: 10.1016/j.tig.2005.12.005

Dudley, J. T., Deshpande, T., and Butte, A. J. (2011). Exploiting drug-disease relationships for computational drug repositioning. *Brief. Bioinformatics* 12, 303–311. doi: 10.1093/bib/bbr013

Durrant, J. D., Amaro, R. E., Xie, L., Urbaniak, M. D., Ferguson, M. A., Haapalainen, A., et al. (2010). A multidimensional strategy to detect polypharmacological targets in the absence of structural and sequence homology. *PLoS Comput. Biol.* 6:e1000648. doi: 10.1371/journal.pcbi.1000648

Fabian, C. J. (2007). The what, why and how of aromatase inhibitors: hormonal agents for treatment and prevention of breast cancer. *Int. J. Clin. Pract.* 61, 2051–2063. doi: 10.1111/j.1742-1241.2007.01587.x

Falcone, M., and Ferretti, R. (2013). *Semi-Lagrangian Approximation Schemes for Linear and Hamilton—Jacobi Equations*. Philadelphia, PA: SIAM.

Frye, S. V., Arkin, M. R., Arrowsmith, C. H., Conn, P. J., Glicksman, M. A., Hull-Ryde, E. A., et al. (2015). Tackling reproducibility in academic preclinical drug discovery. *Nat. Rev. Drug Discov.* 14, 733–734. doi: 10.1038/nrd4737

Goel, N. S., and Richter-Dyn, N. (2016). *Stochastic Models in Biology*. London, UK: Elsevier.

Goh, K. I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., and Barabasi, A. L. (2007). The human disease network. *Proc. Natl. Acad. Sci. U.S.A.* 104, 8685–8690. doi: 10.1073/pnas.0701361104

Gottlieb, A., Stein, G. Y., Ruppin, E., and Sharan, R. (2011). PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Mol. Syst. Biol.* 7:496. doi: 10.1038/msb.2011.26

Gupta, S. C., Sung, B., Prasad, S., Webb, L. J., and Aggarwal, B. B. (2013). Cancer drug discovery by repurposing: teaching new tricks to old dogs. *Trends Pharmacol. Sci.* 34, 508–517. doi: 10.1016/j.tips.2013.06.005

Hanley, J. A., and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143, 29–36. doi: 10.1148/radiology.143.1.7063747

Hannon, B., and Ruth, M. (2014). *Modeling Dynamic Biological Systems*. London, UK: Springer.

Hefti, M. M., Hu, R., Knoblauch, N. W., Collins, L. C., Haibe-Kains, B., Tamimi, R. M., et al. (2013). Estrogen receptor negative/progesterone receptor positive breast cancer is not a reproducible subtype. *Breast Cancer Res.* 15:R68. doi: 10.1186/bcr3462

Herbst, R. S. (2004). Review of epidermal growth factor receptor biology. *Int. J. Radiat. Oncol. Biol. Phys.* 59(Suppl.), 21–26. doi: 10.1016/j.ijrobp.2003.11.041

Hoorens, A., Jouret-Mourin, A., Sempoux, C., Demetter, P., De Hertogh, G., and Teugels E. (2010). Accurate KRAS mutation testing for EGFR-targeted therapy in colorectal cancer: emphasis on the key role and responsibility of pathologists. *Acta Gastroenterol. Belg.* 73, 497–503.

Hu, G., and Agarwal, P. (2009). Human disease-drug network based on genomic expression profiles. *PLoS ONE* 4:e6536. doi: 10.1371/journal.pone.0006536

Huang, H., Nguyen, T., Ibrahim, S., Shantharam, S., Yue, Z., and Chen, J. Y. (2015). DMAP: a connectivity map database to enable identification of novel drug repositioning candidates. *BMC Bioinformatics* 16(Suppl. 13):S4. doi: 10.1186/1471-2105-16-S13-S4

Huang, H., Wu, X., Pandey, R., Li, J., Zhao, G., Ibrahim, S., et al. (2012). C(2)Maps: a network pharmacology database with comprehensive disease-gene-drug connectivity relationships. *BMC Genomics* 13(Suppl. 6):S17. doi: 10.1186/1471-2164-13-S6-S17

Huang, H., Xiaogang, W., Ibrahim, S., Kenzie, M. M., and Chen, J. Y. (2011). "Predicting drug efficacy based on the integrated breast cancer pathway model," in *2011 IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS),* (San Antonio, TX), 42–45.

Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., et al. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods* 12, 115–121. doi: 10.1038/nmeth.3252

Ibrahim, S., Hanspers, K., Willighagen, E., Wagle, P., Chen, J., Digles, D., et al. (2015). *Integrated Breast Cancer Pathway (Homo sapiens)*. Wikipathway.org, Wikipathway.org.

Iorio, F., Bosotti, R., Scacheri, E., Belcastro, V., Mithbaokar, P., Ferriero, R., et al. (2010). Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proc. Natl. Acad. Sci. U.S.A.* 107, 14621–14626. doi: 10.1073/pnas.1000138107

Jensen, P. B., Jensen, L. J., and Brunak, S. (2012). Mining electronic health records: towards better research applications and clinical care. *Nat. Rev. Genet.* 13, 395–405. doi: 10.1038/nrg3208

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 45, D353–D361. doi: 10.1093/nar/gkw1092

Keiser, M. J., Setola, V., Irwin, J. J., Laggner, C., Abbas, A. I., and Roth, B. L. (2009). Predicting new molecular targets for known drugs. *Nature* 462, 175–181. doi: 10.1038/nature08506

Khoo, C., Rogers, T. M., Fellowes, A., Bell, A., and Fox, S. (2015). Molecular methods for somatic mutation testing in lung adenocarcinoma: EGFR and beyond. *Transl Lung Cancer Res.* 4, 126–141. doi: 10.3978/j.issn.2218-6751.2015.01.10

Kuhn, M., Letunic, I., Jensen, L. J., and Bork, P. (2016). The SIDER database of drugs and side effects. *Nucleic Acids Res.* 44, D1075–D1079. doi: 10.1093/nar/gkv1075

Kuhn, M., Szklarczyk, D., Franceschini, A., von Mering, C., Jensen, L. J., and Bork, P. (2012). STITCH 3: zooming in on protein-chemical interactions. *Nucleic Acids Res.* 40, D876–D880. doi: 10.1093/nar/gkr1011

Lamb, J., Crawford, E. D., Peck, D., Modell, J. W., Blat, I. C., and Golub, T. R. (2006). The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* 313, 1929–1935. doi: 10.1126/science.1132939

Law, V., Knox, C., Djoumbou, Y., Jewison, T., Guo, A. C., and Wishart, D. S. (2013). DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.* 42, D1091–D1097. doi: 10.1093/nar/gkt1068

Li, J., and Lu, Z. (2013). Pathway-based drug repositioning using causal inference. *BMC Bioinformatics* 14(Suppl. 16):S3. doi: 10.1186/1471-2105-14-S16-S3

Li, Y. Y., and Jones, S. J. (2012). Drug repositioning for personalized medicine. *Genome Med.* 4:27. doi: 10.1186/gm326

Macarron, R., Banks, M. N., Bojanic, D., Burns, D. J., Cirovic, D. A., and Sittampalam, G. S. (2011). Impact of high-throughput screening in biomedical research. *Nat. Rev. Drug Discov.* 10, 188–195. doi: 10.1038/nrd3368

Maier, L., Pruteanu, M., Kuhn, M., Zeller, G., Telzerow, A., Anderson, E., et al. (2018). Extensive impact of non-antibiotic drugs on human gut bacteria. *Nature* 555, 623–628. doi: 10.1038/nature25979

Medina-Franco, J. L., Giulianotti, M. A., Welmaker, G. S., and Houghten, R. A. (2013). Shifting from the single to the multitarget paradigm in drug discovery. *Drug Discov. Today* 18, 495–501. doi: 10.1016/j.drudis.2013.01.008

Momparler, R. L., Karon, M., Siegel, S. E., and Avila, F. (1976). Effect of adriamycin on DNA, RNA, and protein synthesis in cell-free systems and intact cells. *Cancer Res.* 36, 2891–2895.

Oprea, T. I., Bauman, J. E., Bologa, C. G., Buranda, T., Chigaev, A., and Sklar, L. A. (2011). Drug repurposing from an academic perspective. *Drug Discov. Today Ther. Strateg.* 8, 61–69. doi: 10.1016/j.ddstr.2011.10.002

Pico, A. R., Kelder, T., van Iersel, M. P., Hanspers, K., Conklin, B. R., and Evelo, C. (2008). WikiPathways: pathway editing for the people. *PLoS Biol.* 6:e184. doi: 10.1371/journal.pbio.0060184

Power, A., Berger, A. C., and Ginsburg, G. S. (2014). Genomics-enabled drug repositioning and repurposing: insights from an IOM Roundtable activity. *JAMA* 311, 2063–2064. doi: 10.1001/jama.2014.3002

Pujol, A., Mosca, R., Farres, J., and Aloy, P. (2010). Unveiling the role of network and systems biology in drug discovery. *Trends Pharmacol. Sci.* 31, 115–123. doi: 10.1016/j.tips.2009.11.006

Raoul, J. L., Van Laethem, J. L., Peeters, M., Brezault, C., Husseini, F., Cals, L., et al. (2009). Cetuximab in combination with irinotecan/5-fluorouracil/folinic acid (FOLFIRI) in the initial treatment of metastatic colorectal cancer: a multicentre two-part phase I/II study. *BMC Cancer* 9:112.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., and Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43:e47. doi: 10.1093/nar/gkv007

Rovithakis, G. A., and Christodoulou, M. A. (1994). Adaptive control of unknown plants using dynamical neural networks. *IEEE Trans. Syst. Man Cybernet.* 24, 400–412. doi: 10.1109/21.278990

Subramanian, A., Narayan, R., Corsello, S. M., Peck, D. D., Natoli, T. E., Golub, T. R. et al. (2017). A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* 171, 1437.e17–1452.e17. doi: 10.1016/j.cell.2017.10.049

Sun, Y. L., Kathawala, R. J., Singh, S., Zheng, K., Talele, T. T., Chen, Z. S., et al. (2012). Zafirlukast antagonizes ATP-binding cassette subfamily G member 2-mediated multidrug resistance. *Anticancer Drugs* 23, 865–873. doi: 10.1097/CAD.0b013e328354a196

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., von Mering, C., et al. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* 43, D447–D452. doi: 10.1093/nar/gku1003

Thomas, D., Burns, J., Audette, J., Carrol, A., Dow-Hygelund, C., and Hay, M. (2016). *Clinical development success rates 2006–2015.* San Diego, CA: Biomedtracker; Washington, DC: BIO/Bend: Amplion.

Tong, S., Wang, T., Li, Y., and Zhang, H. (2014). Adaptive neural network output feedback control for stochastic nonlinear systems with unknown dead-zone and unmodeled dynamics. *IEEE Trans. Cybern.* 44, 910–921. doi: 10.1109/TCYB.2013.2276043

UniProt, C. (2013). Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Res.* 41, D43–D47. doi: 10.1093/nar/gks1068

Uray, I. P., and Brown, P. H. (2011). Chemoprevention of hormone receptor-negative breast cancer: new approaches needed. *Recent Results Cancer Res.* 188, 147–162. doi: 10.1007/978-3-642-10858-7_13

USFDA. (2016). *Step 3: Clinical Research*. Available online at: http://www.fda.gov/ForPatients/Approvals/Drugs/ucm405622.htm

Wang, Y., Liu, S., Rastegar-Mojarad, M., Wang, L., Shen, F., Liu, F., et al. (2017). "Dependency and AMR embeddings for drug-drug interaction extraction from biomedical literature," in *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics* (Boston, MA: ACM).

Wang, Y., Suzek, T., Zhang, J., Wang, J., He, S., Cheng, T., et al. (2014). PubChem BioAssay: 2014 update. *Nucleic Acids Res.* 42, D1075–D1082. doi: 10.1093/nar/gkt978

Weir, S. J., Patton, L., Castle, K., Rajewski, L., Kasper, J., and Schimmer, A. D. (2011). The repositioning of the anti-fungal agent ciclopirox olamine as a novel therapeutic agent for the treatment of haematologic malignancy. *J. Clin. Pharm. Ther.* 36, 128–134. doi: 10.1111/j.1365-2710.2010.01172.x

Weston, A. D., and Hood, L. (2004). Systems biology, proteomics, and the future of health care: toward predictive, preventative, and personalized medicine. *J. Proteome Res.* 3, 179–196. doi: 10.1021/pr0499693

Willems, J. (1971). Least squares stationary optimal control and the algebraic Riccati equation. *IEEE Trans. Automat. Contr.* 16, 621–634. doi: 10.1109/TAC.1971.1099831

Woodhead, J. L., Watkins, P. B., Howell, B. A., Siler, S. Q., and Shoda, L. K. (2016). The role of quantitative systems pharmacology modeling in the prediction and explanation of idiosyncratic drug-induced liver injury. *Drug Metab. Pharmacokinet.* 32, 40–45. doi: 10.1016/j.dmpk.2016.11.008

Xu, H., Aldrich, M. C., Chen, Q., Liu, H., Peterson, N. B., Ruan, X., et al. (2014). Validating drug repurposing signals using electronic health records: a case study of metformin associated with reduced cancer mortality. *J. Am. Med. Inform. Association* 2014:002649. doi: 10.1136/amiajnl-2014-002649

Yazdannejat, H., Hosseinimehr, S. J., Ghasemi, A., Pourfallah, T. A., and Rafiei, A. (2016). Losartan sensitizes selectively prostate cancer cell to ionizing radiation. *Cell Mol. Biol.* 62, 30–33.

Yin, T., He, S., Shen, G., Ye, T., Guo, F., and Wang, Y. (2015). Dopamine receptor antagonist thioridazine inhibits tumor growth in a murine breast cancer model. *Mol. Med. Rep.* 12, 4103–4108. doi: 10.3892/mmr.2015.3967

Yip, C. H., and Rhodes, A. (2014). Estrogen and progesterone receptors in breast cancer. *Future Oncol.* 10, 2293–2301. doi: 10.2217/fon.14.110

Zheng, C., Guo, Z., Huang, C., Wu, Z., Li, Y., Chen, X., et al. (2015). Large-scale direct targeting for drug repositioning and discovery. *Sci. Rep.* 5:11970. doi: 10.1038/srep11970

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer YW and handling Editor declared their shared affiliation.

frontiers in Pharmacology

ORIGINAL RESEARCH
published: 26 June 2018
doi: 10.3389/fphar.2018.00681

Discovery of the Consistently Well-Performed Analysis Chain for SWATH-MS Based Pharmacoproteomic Quantification

*Jianbo Fu[1], Jing Tang[1,2], Yunxia Wang[1], Xuejiao Cui[1,2], Qingxia Yang[1,2], Jiajun Hong[1], Xiaoxu Li[1,2], Shuang Li[1,2], Yuzong Chen[3], Weiwei Xue[2] and Feng Zhu[1,2]**

[1] College of Pharmaceutical Sciences, Zhejiang University, Hangzhou, China, [2] School of Pharmaceutical Sciences and Collaborative Innovation Center for Brain Science, Chongqing University, Chongqing, China, [3] Bioinformatics and Drug Design Group, Department of Pharmacy, Center for Computational Science and Engineering, National University of Singapore, Singapore, Singapore

Sequential windowed acquisition of all theoretical fragment ion mass spectra (SWATH-MS) has emerged as one of the most popular techniques for label-free proteome quantification in current pharmacoproteomic research. It provides more comprehensive detection and more accurate quantitation of proteins comparing with the traditional techniques. The performance of SWATH-MS is highly susceptible to the selection of processing method. Till now, ≥27 methods (transformation, normalization, and missing-value imputation) are sequentially applied to construct numerous analysis chains for SWATH-MS, but it is still not clear which analysis chain gives the optimal quantification performance. Herein, the performances of 560 analysis chains for quantifying pharmacoproteomic data were comprehensively assessed. Firstly, the most complete set of the publicly available SWATH-MS based pharmacoproteomic data were collected by comprehensive literature review. Secondly, substantial variations among the performances of various analysis chains were observed, and the consistently well-performed analysis chains (CWPACs) across various datasets were for the first time generalized. Finally, the log and power transformations sequentially followed by the total ion current normalization were discovered as one of the best performed analysis chains for the quantification of SWATH-MS based pharmacoproteomic data. In sum, the CWPACs identified here provided important guidance to the quantification of proteomic data and could therefore facilitate the cutting-edge research in any pharmacoproteomic studies requiring SWATH-MS technique.

Keywords: pharmacoproteomics, SWATH-MS, processing method, transformation, normalization

OPEN ACCESS

Edited by:
Zhi-Liang Ji,
Xiamen University, China

Reviewed by:
Dariusz Mrozek,
Silesian University of Technology, Poland
Qing-Chuan Zheng,
Jilin University, China

****Correspondence:***
Feng Zhu
zhufeng@zju.edu.cn;
zhufeng.ns@gmail.com

Specialty section:
This article was submitted to Translational Pharmacology, a section of the journal Frontiers in Pharmacology

Received: *27 April 2018*
Accepted: *05 June 2018*
Published: *26 June 2018*

Citation:
Fu J, Tang J, Wang Y, Cui X, Yang Q, Hong J, Li X, Li S, Chen Y, Xue W and Zhu F (2018) Discovery of the Consistently Well-Performed Analysis Chain for SWATH-MS Based Pharmacoproteomic Quantification. Front. Pharmacol. 9:681. doi: 10.3389/fphar.2018.00681

INTRODUCTION

The pharmacoproteomics has been widely applied to various aspects of current pharmaceutical researches by discovering disease-related genes (Mrozek et al., 2013; Quiros et al., 2017; Zeng et al., 2017) or new drug targets (Li et al., 2018; Saei et al., 2018), constructing pharmacology screening model (Hauser et al., 2005), and revealing the drug mechanism of action (Yue et al., 2016; Zhu et al., 2018), resistance (Paul et al., 2016), and toxicity (Tan et al., 2017; Wang et al., 2017b). Recent findings uncover its potentials to fulfill the promise that the pharmacogenomics has not accomplished yet (D'Alessandro and Zolla, 2010; Chambliss and Chan, 2016; Yang et al., 2016).

As a newly emerging technique (Anjo et al., 2017), the *sequential windowed acquisition of all theoretical fragment ion mass spectra* (SWATH-MS) has been reported to provide much more comprehensive detection and accurate quantitation of proteins compared to the traditional techniques used in pharmacoproteomic analyses (Zhu et al., 2008b; Tao et al., 2015; Aebersold and Mann, 2016; Li et al., 2016a; Anjo et al., 2017), and it thus becomes one of the most popular techniques for target discovery (Li et al., 2016b; Xu et al., 2016; Anjo et al., 2017), drug/lead quantification (Roemmelt et al., 2015) and identification (Scheidweiler et al., 2015; Wang et al., 2015; Aratyn-Schaus and Ramanathan, 2016; Li B. et al., 2017), construction of assay library for targeted proteomic analysis (Schubert et al., 2015), and quantitative protein profiling (Krasny et al., 2018) for recognizing drug-induced alterations (Roemmelt et al., 2015; Xue et al., 2016).

However, due to the interdependent nature among multiple acquisition parameters (dwell time, duty cycle, precursor isolation window width, and mass range), the protein quantification based on SWATH-MS is reported to be limited in dynamic range (Anjo et al., 2017) and in turn low in accuracy (Gillet et al., 2012; Huang et al., 2015; Shi et al., 2016; Yang et al., 2017; Xue et al., 2018b). The problems above can be even worse considering the innate complexity of clinical samples (Jamwal et al., 2017), small amount of proteins (Sajic et al., 2015), and low abundance of drug-metabolizing enzymes (Jamwal et al., 2017). To cope with these problems, a variety of popular quantification tools, including *DIA-Umpire* (Sajic et al., 2015), *OpenSWATH* (Rost et al., 2014), *Skyline* (MacLean et al., 2010), *Spectronaut* (Bruderer et al., 2015), and *SWATH2.0* (Li S. et al., 2017), and dozens of subsequent processing methods (transformation, normalization, and missing-value imputation) are developed to enhance the accuracy of SWATH-MS (Navarro et al., 2016). Recent reports further reveal that SWATH-MS' accuracies depend heavily on the specific quantification tool/processing method used in a particular study (Navarro et al., 2016), and the protein quantification can significantly benefit from comparative benchmarking of the performance of these tools and methods (Gatto et al., 2016; Zheng et al., 2016). Therefore, it is urgently needed to assess the performances of tools/methods for discovering the optimal one(s) for SWATH-MS based pharmacoproteomic studies.

The performance of various quantification tools has already been systematically evaluated by benchmark SWATH-MS data (Navarro et al., 2016). Among those tools, only 2 (*OpenSWATH* and *Skyline*) are non-commercial ones, and the *OpenSWATH* (Rost et al., 2014) is of the most popular one used to quantify SWATH-MS based pharmacoproteomic data (Rost et al., 2014; Parker et al., 2015; Weisser and Choudhary, 2017). So far, $\geq$4 transformation, $\geq$15 normalization, and $\geq$6 missing-value imputation algorithms (Guo et al., 2015; Li et al., 2016c; Ori et al., 2016; Wu et al., 2016; Tan et al., 2017; Wang et al., 2017a) have been sequentially applied to process pharmacoproteomic data. Among these algorithms, four for normalizing label-free proteomic data have been assessed to identify the best performed one (Callister et al., 2006) and six for missing-value imputation have been evaluated to discover the one enhancing proteomic quantifications in the differential expression analysis (Valikangas et al., 2017). Appropriate integrations of the processing methods into a sequential analysis chain are reported to improve the quantification accuracies (Karpievitch et al., 2012; Chawade et al., 2015; Valikangas et al., 2017) with some chains identified as highly accurate in particular pharmacoproteomic studies (Guo et al., 2015; Ori et al., 2016; Tan et al., 2017; Zheng et al., 2017). For example, log transformation followed by median normalization performs well in identifying the therapeutic target/pathway for *Down syndrome* (Sullivan et al., 2017), endogenous toxins inducing the haploinsufficiency of tumor suppressor (Tan et al., 2017) and biological mechanism underlying the role of proteins played in *Alzheimer's disease* (Khoonsari et al., 2016). Since the processing methods are sequentially used to form the integrated analysis chain (Guo et al., 2015; Ori et al., 2016; Tan et al., 2017), any performance assessment aiming solely at transformation, normalization, or imputation may not be able to reflect the overall performance of the whole analysis chain. Considering the huge amount of possible analysis chains [560 in total, taking non-transformation, non-normalization, and non-imputation into account adopted by previous studies (Guo et al., 2015; Liu et al., 2015; Wu et al., 2016)] by randomly integrating those processing methods, it is therefore essential to comprehensively evaluate the performance of all analysis chains to identify the optimal one for specific pharmacoproteomic dataset. However, no such analysis has been conducted yet.

In this study, the performances of all possible analysis chains integrating 4 transformation, 15 normalization, and 6 imputation algorithms were comprehensively assessed by their precisions based on the proteomes among replicates (Kuharev et al., 2015; Navarro et al., 2016; Chignell et al., 2018; Muller et al., 2018). Systematic literature review on the popular quantification tool *OpenSWATH* firstly yielded seven SWATH-MS based benchmark pharmacoproteomic datasets of varied sample sizes (from 6 to 116). To the best of our knowledge, these seven provided the most complete set of the publicly available pharmacoproteomic data based on the SWATH-MS technique. Secondly, the performance of analysis chains was assessed by each dataset. Thirdly, the analysis chains consistently performed well across all datasets were identified for the first time and compared with those popular chains frequently applied in current pharmacoproteomic studies. Finally, the consistently well-performed analysis chains were further discussed based on their performances. The analysis chains identified in and the corresponding findings of this study provided important guidance to current pharmacoproteomic studies.

MATERIALS AND METHODS

Collection of SWATH-MS Based Benchmark Pharmacoproteomic Datasets

A systematic literature review on the popular quantification tool *OpenSWATH* and the analysis on the datasets provided in the PRIDE database (Navarro et al., 2016) were collectively conducted to find SWATH-MS based benchmark pharmacoproteomic datasets. Firstly, PRIDE database was searched against by keyword "SWATH-MS." Together with the literature review on the resulting projects, 85 projects were identified as based on SWATH-MS, among which 76 and 9 projects were acquired by TripleTOF instruments 5600 and 6600, respectively. Secondly, several criteria were used to guarantee the availability and processability of the raw proteomic data, which included (1) complete set of raw data files, (2) well-defined parameters (isolation scheme, range of retention time, and transition settings), (3) availability of spectral library and protein database to search against, and (4) clear description on sample groups. The application of these criteria on the resulting PRIDE projects yielded seven SWATH-MS based benchmark pharmacoproteomic datasets of varied sample sizes (**Table 1**), which covered both TripleTOF instruments (5600 and 6600) of all 85 projects. Therefore, these datasets can be recognized as representatives of SWATH-MS based pharmacoproteomic data. To the best of our knowledge, these datasets provided the most complete set of SWATH-MS based pharmacoproteomic data.

Processing Methods for Data Transformation, Normalization, and Imputation

So far, ≥4 transformation, ≥15 normalization, and ≥6 missing-value imputation algorithms (Guo et al., 2015; Li et al., 2016c; Ori et al., 2016; Wu et al., 2016; Tan et al., 2017; Wang et al., 2017a) have been reported to be sequentially and frequently used to process pharmacoproteomic data. Based on our comprehensive literature review, their corresponding applications to current proteomic research were discussed in Supplementary Method S1. These 25 methods include 4 ***transformation***: *Box-cox* (Sakia, 1992), *Cube Root* (Wen et al., 2017), *Log* (De Livera et al., 2012), and *Power* (Zhang, 2014), 15 ***normalization***: *Auto Scaling* (Kohl et al., 2012), *Cyclic Loess* (Zhu et al., 2012b), *EigenMS* (Zhu et al., 2009), *Locally Weighted Scatterplot Smoothing* (Wilson et al., 2003), *Mean* (Andjelkovic and Thompson, 2006), *Median* (Bolstad et al., 2003), *Median Absolute Deviation* (Matzke et al., 2011), *Pareto* (Zhu et al., 2010), *Probabilistic Quotient* (Dieterle et al., 2006), *Quantile* (Callister et al., 2006), *Robust Linear Regression* (Hong et al., 2016), *Total Ion Current* (Gaspari et al., 2016), *Trimmed Mean of M Values* (Lin et al., 2016), *VSN* (Huber et al., 2002), and *Z-score* (Cheadle et al., 2003), and 6 ***imputation***: *Background* (Chai et al., 2014), *Bayesian Principal* (Chai et al., 2014), *Censored* (Valikangas et al., 2017), *K-nearest Neighbor* (Zhu et al., 2008a), *Singular Value Decomposition* (Alter et al., 2000), and *Zero Imputation* (Gan et al., 2006). As shown in the Supplementary Method S1, due to their popularity in current pharmacoproteomic studies, these 25 methods were included, sequentially applied, and analyzed in this study. Each method was abbreviated by a three-letter code which was demonstrated in Supplementary Table S1.

Assessing Analysis Chain Using the Precision Based on Proteomes Among Replicates

Diverse methods for proteomic data processing (transformation, normalization, and imputation) profoundly affected the precision of protein quantification which was frequently assessed using the value of pooled intragroup median absolute deviation (PMAD) of reported protein intensity among replicates (Chawade et al., 2014; Kuharev et al., 2015; Valikangas et al., 2018; Yu et al., 2018). Particularly, the PMAD was designed to demonstrate the capacity of each analysis chain to reduce the variation among replicates, and therefore to enhance the technical reproducibility (Chawade et al., 2014). The lower value of PMAD denoted the more thorough removal of the experimentally induced noise and indicated better precision of the corresponding analysis chain (Valikangas et al., 2018). So far, PMAD value within the range of ≤0.3, >0.3 & ≤0.7, and >0.7 was generally accepted as with

TABLE 1 | Seven SWATH-MS based benchmark pharmacoproteomic datasets collected for the analysis of this study.

Datasets	PRIDE ID	Sample size and Dataset description	Analysis Chain	Instrument
Nat. Biotechnol. 34:1130-6, 2016	PXD002952	3 samples of 65% human, 30% yeast, and 5% *E. coli* proteins 3 samples of 65% human, 15% yeast, and 20% *E. coli* proteins	LOG-MED-???	TripleTOF 6600
Cell Rep. 20:1229-41, 2017	PXD003278	6 siRNA-treated Cal51 cell samples 6 PRPF8-depleted Cal51 cell samples	LOG-QUA-NON	TripleTOF 5600
Cell. 169:1105-18, 2017	PXD006106	10 formaldehyde treated HeLa Kyoto cell samples 10 formaldehyde untreated HeLa Kyoto cell samples	LOG-MED-NON	TripleTOF 5600
Nat Med. 21:407-13, 2015	PXD000672	18 tumorous kidney tissue biopsies 18 non-tumorous kidney tissue biopsies	LOG-QUA-NON	TripleTOF 5600
Sci Rep. 7:14818, 2017	PXD004880	18 plasma samples from individuals with *Down syndrome* 18 plasma samples from healthy controls	LOG-MED-NON	TripleTOF 5600
Cell Rep. 18:3219-26, 2017	PXD003972	20 wild type mouse samples 20 knock-in mouse samples expressing endogenous GRB2	LOG-???-???	TripleTOF 5600
Mol Syst. Biol. 11:786, 2015	PXD001064	72 blood samples of monozygotic twins 44 blood samples of dizygotic twins	???-RLR-BAK	TripleTOF 5600

All datasets were from PRIDE database (Navarro et al., 2016). Each method in the analysis chain was abbreviated by a three-letter code as demonstrated in Supplementary Table S1, and ??? indicated that the corresponding method was not specified in the corresponding study of the dataset.

superior, good, and poor precision, respectively (Chawade et al., 2014; Valikangas et al., 2018), which had gradually become a popular metric for assessing the precision of processing methods in OMICs (Chawade et al., 2014; Valikangas et al., 2018).

Performance Assessment Among Various Analysis Chains by Hierarchical Clustering

Pooled intragroup median absolute deviation values of 560 possible analysis chains across the seven benchmark datasets were firstly calculated. Fifty-one out of these 560 analysis chains reported error for processing at least one of the benchmark datasets. Therefore, the hierarchical clustering of the remaining 509 analysis chains with calculatable results of all seven PMADs was conducted to identify the relationship among the performances of various analysis chains. Particularly, PMAD values of a specific analysis chain among 7 datasets were used to form a 7-dimensional vector. Then, hierarchical clustering was applied to investigate the relationship among those 509 vectors, and therefore among the corresponding analysis chains. To measure the distance between any 2 vectors, the *Euclidean distance* was adopted, which could be demonstrated as below:

$$\text{Euclidean distance}\,(a,b)=\sqrt{\sum\nolimits_{\mathrm{i}=1}^{\mathrm{n}}\left(a_{\mathrm{i}}-b_{\mathrm{i}}\right)^{2}}$$

where *i* denoted each dimension of the analysis chain *a* and *b*. The clustering algorithm applied here was Ward's minimum variance algorithm (Barer and Harwood, 1999), which was designed to minimize the total within-cluster variance. Ward's minimum variance module in R package (Tippmann, 2015) was used. To visualize the hierarchical tree graph among those 509 analysis chains, the tree generator *iTOL* was used to generate and display the hierarchical tree structure (Letunic and Bork, 2016).

RESULTS AND DISCUSSION

Ranking the Analysis Chains Based on Their Performances on Each Benchmark

The performances of each analysis chain on the seven SWATH-MS based benchmark datasets (**Table 1**) were assessed by measuring the corresponding PMAD values. As shown in **Figure 1**, the performances of 509 analysis chains ($\log_{10}$ PMAD, *Y*-axis) with calculatable PMAD values were measured and ranked (*X*-axis). Because some analysis chains may not be able to result in a PMAD value, there were slight variations among the number of analysis chains for different benchmark datasets (from 530 to 560). Taking the dataset shown in the center of **Figure 1** as an example (Nat Med. 21:407-13, 2015), a total of 558 analysis chains were assessed and ranked, and the performance of different analysis chains varied significantly (PMAD from 1.8×10^{-15} to 2.0×10^{5}). With reference to the frequently adopted cutoff (PMAD = 0.7) for differentiating the analysis chains of good and poor precision (Chawade et al., 2014; Valikangas et al., 2018), 203 (36.4%) out of these 558 analysis chains were ranked as well-performed. Similar to this dataset (Nat Med. 21:407-13, 2015), the performance of different analysis chains for the other datasets also differentiated substantially (PMAD from 1.7×10^{-16} to 3.4×10^{5}) with 38.8%∼49.7% of the analysis chains ranked as well-performed.

The specific analysis chains for each benchmark dataset adopted in the corresponding original studies were identified by literature review (**Table 1**). Particularly, 4 out of these datasets were with the clearly defined analysis chain (LOG-QUA-NON, LOG-MED-NON, LOG-QUA-NON, and LOG-MED-NON for PXD003278, PXD006106, PXD000672, and PXD004880, respectively), while the remaining 3 datasets were with incomplete information of the adopted analysis chain (LOG-MED-???, LOG-???-???, and ???-RLR-BAK for the datasets of PXD002952, PXD003972, and PXD001064, respectively). Taking the same dataset in the middle of **Figure 1** as an example (Nat Med. 21:407-13, 2015), the red dot indicated the PMAD of the analysis chain adopted by this study and its corresponding ranking among all 558 analysis chains. As shown, the adopted chain (LOG-QUA-NON) in this study was ranked to be the 156th well-performed one (PMAD = 0.598) showing its capacity to reduce variations among replicates and thus enhance technical reproducibility (Chawade et al., 2014). However, there were 155 chains performed better than the adopted one (PMAD from 1.8×10^{-15} to 0.595) with POW-TMM-ZER chain performed the best. Similar to this example dataset, the analysis chains adopted by the corresponding studies of PXD003278, PXD006106, and PXD004880 were ranked 162nd, 154th, and 164th well-performed ones, which demonstrated appropriate selection of analysis chain in previous studies. However, there were still more than a hundred chains performed better than the adopted ones, which may further enhance the accuracy of SWATH-MS based protein quantification. For the studies with incomplete information of the adopted chain (PXD002952, PXD003972, and PXD001064), the possible integrations based on the known information were highlighted by multiple red dots. 1 (20%) out of 5, 28 (25%) out of 112, and 7 (100%) out of 7 integrations were within the ranges of well-performance for PXD002952, PXD003972, and PXD001064, respectively.

Analysis Chains Consistently Well-Preformed Across All Benchmark Datasets

The performances of 20 representative analysis chains across different datasets were illustrated in **Figure 2**. PMAD within the ranges of ≤0.3, >0.3 & ≤0.7, and >0.7 was generally accepted as with superior, good, and poor performance, respectively (Chawade et al., 2014; Valikangas et al., 2018), which was illustrated by a circle of various diameters (the smaller diameter denoted the lower PMAD value). As shown, the performances of specific chain among various datasets varied significantly. Particularly, the LOG-PQN-BPC performed superior, good, and poor in 3, 3, and 1 datasets, respectively, and POW-ZSC-ZER performed superior, good, and poor in 1, 5, and 1 datasets, respectively. These results demonstrated a certain level of variations among the seven datasets for each analysis chain. However, as shown in **Figure 2**, there were some chains

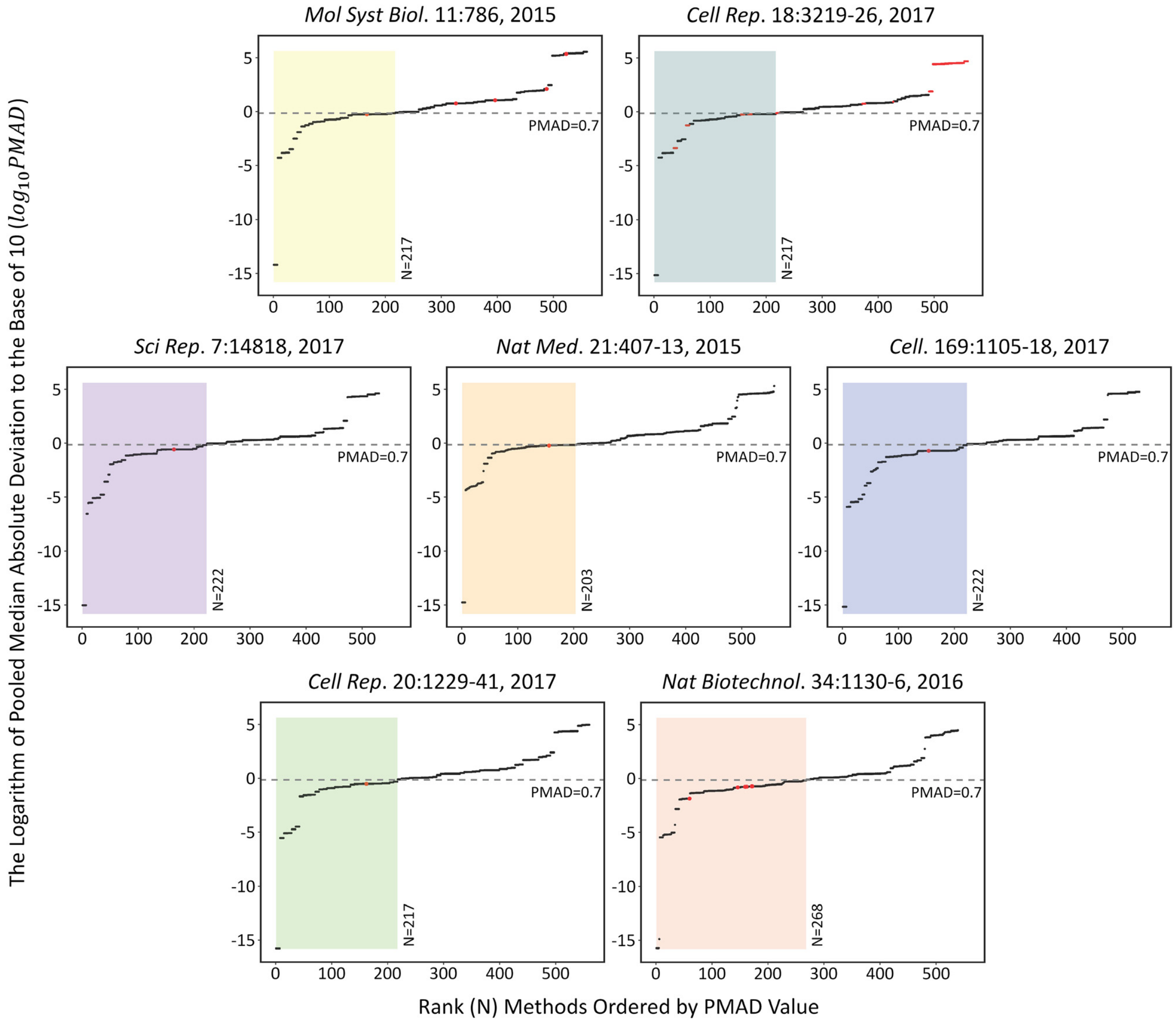


FIGURE 1 | The performances of each analysis chain on those seven SWATH-MS based benchmark datasets assessed by measuring the corresponding PMAD values [>500 analysis chains ($\log_{10}$ PMAD, *Y*-axis) were measured and ranked (*X*-axis)]. Since some analysis chains may not be able to result in a specific PMAD value, there were slight variations among the number of analysis chains for different benchmark datasets (from 530 to 560). Detail information on these seven datasets were provided in **Table 1**.

performed consistently across different benchmark datasets. For instance, CUB-TIC-BAK and CUB-VSN-CEN performed superior in all datasets, while 2 other chains (NON-CYC-ZER and NON-MEA-SVD) performed poor in all seven benchmarks. It was of great interests to explore dataset-independent properties underlying the consistency across datasets, which thus inspired us to further investigate the similarity among performances of different analysis chains.

Since the type of instrument (TripleTOF 5600 and 6600) covered by seven benchmark datasets were the same as that of 85 SWATH-MS based projects, those datasets could be recognized as representative datasets of SWATH-MS based pharmacoproteomic data. Thus, the discovery of analysis chain performed consistently well across the various datasets might give great insights into the selection of the most appropriate analysis chain in SWATH-MS based proteomic study. To identify such chains performed consistently well across datasets, the hierarchical clustering with the ward algorithm (Barer and Harwood, 1999; Zhu et al., 2011; Fu et al., 2018; Xue et al., 2018a) was used to identify the "consistently well-performed" analysis chains (CWPACs) based on their PMAD values across different datasets. Theoretically, there were 560 possible analysis chains by randomly integrating 5 transformation, 16 normalization, and 7 imputation algorithms (including non-transformation, non-normalization, and non-imputation). 51 (9.1%) out of these 560 were with at least one PMAD value of the seven datasets unavailable due to the calculation error. Then, the PMAD values of the remaining 509 analysis chains were applied for clustering analysis. As illustrated in **Figure 3**, six partitions of the analysis chains (A_1, A_2, A_3, B, C, and D) were identified. The PMADs meeting the "well-performed" criterion ($\leq$0.7) were displayed

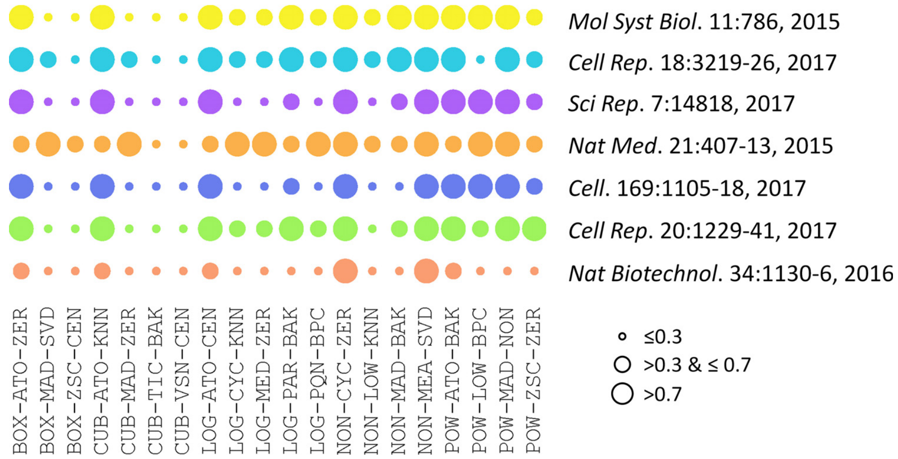


FIGURE 2 | Performances of 20 representative analysis chains across different datasets measured by PMAD values. The PMAD values within the ranges of ≤0.3, >0.3 & ≤0.7, and >0.7 was generally accepted as with superior, good, and poor performance, respectively (Chawade et al., 2014; Valikangas et al., 2018), which was illustrated by the circles of different diameters (the smaller circle diameter indicated the lower PMAD value).

by blue color, with the $\log_{10}$ PMAD ≤ -5 set as exact blue and the larger $\log_{10}$ PMAD gradually fading toward white (PMAD = 0.7). Meanwhile, those "poor-performed" PMADs (>0.7) were colored by orange, with $\log_{10}$ PMAD ≥ 5 set as exact orange and the smaller PMAD gradually fading toward white (PMAD = 0.7).

The analysis chains in the partition A_1, A_2, and A_3 were "consistently well-performed" across all datasets (**Figure 3**). For partition A_1, 320 (99.4%) out of 322 PMAD values were ≤0.1, and the remaining PMADs were ≤0.7 (Supplementary Figure S1). For partition A_2, 288 (52.7%), 209 (38.3%), and 40 (7.3%) out of those 546 PMAD values were ≤0.1, ≤0.3, and ≤0.7, respectively (Supplementary Figure S2). In partition A_3, 187 (46.1%) and 183 (45.1%) out of 406 PMADs were ≤0.3 and ≤0.7, respectively (Supplementary Figure S3). In summary, 608 (47.7%), 396 (31.1%), and 225 (17.7%) out of all 1,274 PMADs in the partition combined by A_1, $A_{2,}$ and A_3 were ≤0.1, ≤0.3, and ≤0.7, respectively, indicating an extremely high percentage (96.5%) of the PMAD values meeting the widely adopted cutoff (PMAD = 0.7) for differentiating the chain of good and poor performances (Chawade et al., 2014; Valikangas et al., 2018). Comprehensive literature review on the 85 SWATH-MS based proteomic projects further identified the analysis chains adopted by their corresponding studies (Supplementary Table S2). In total, there were 55 analysis chains previously applied in proteomic studies, which were mapped to and labeled on **Figure 3** (pink triangles). As illustrated, 7 (12.7%), 9 (16.4%), and 21 (38.2%) out of the 55 analysis chains previously adopted were within the partition A_1, $A_{2,}$ and A_3, respectively, which indicated that the majority (67.3%) of these analysis chains were the CWPACs.

As shown in Supplementary Figure S4, the percentage of each processing method adopted by the previous proteomic studies were analyzed. *Log Transformation* was the only transformation method used in SWATH-MS based proteomic studies, and was widely recognized as powerful in quantifying thousands of proteins (Rao et al., 2011; De Livera et al., 2012; Wisniewski et al., 2012; Zhu et al., 2012a; Feng et al., 2014). For normalizations, *Median Normalization*, *Total Ion Current*, and *Quantile Normalization* were the top-3 ranked methods in their popularity. The *Median* and *Quantile Normalization* were frequently adopted in MS-based label-free proteomic analyses (Callister et al., 2006), while the *Total Ion Current* was reported to be preferably used in the proteomic profiling based on MALDI- and SELDI-TOF mass spectra (Borgaonkar et al., 2010). For imputation, *K-nearest Neighbor* and *Background Imputation* accounted for >80% of the SWATH-MS based proteomic studies adopting imputation methods. Among those methods used in proteomic studies (4 transformation, 15 normalization, and 6 missing-value imputation), Supplementary Figure S4 showed that some methods were adopted seldomly in SWATH-MS based proteomic studies (such as *Box-Cox Transformation*, *Pareto Scaling*, and *Singular Value Decomposition*). Therefore, it is of great interests to discover whether there are other methods suitable or demonstrating enhanced performance in SWATH-MS based proteomic analysis.

Fifty-three analysis chains consistently performed poor among datasets were also discovered by **Figure 3** (partition D), all of which did not adopt any transformation method in their analysis. In total, 101 out of the 509 analysis chains (**Figure 3**) adopted non-transformation, and 53 (52.5%), 10 (9.9%), 11 (10.9%), 14 (13.9%), 6 (5.9%), and 7 (6.9%) out of these 101 chains were within the partition D, C, B, A_3, A_2, and A_1, respectively. These results demonstrated the important roles played by transformation methods in the quantification performance of analysis chains.

Contribution of Each Processing Method to the Performance of Analysis Chain

With the discovery of a variety of CWPACs based on those independent benchmark datasets, it was interesting to go back

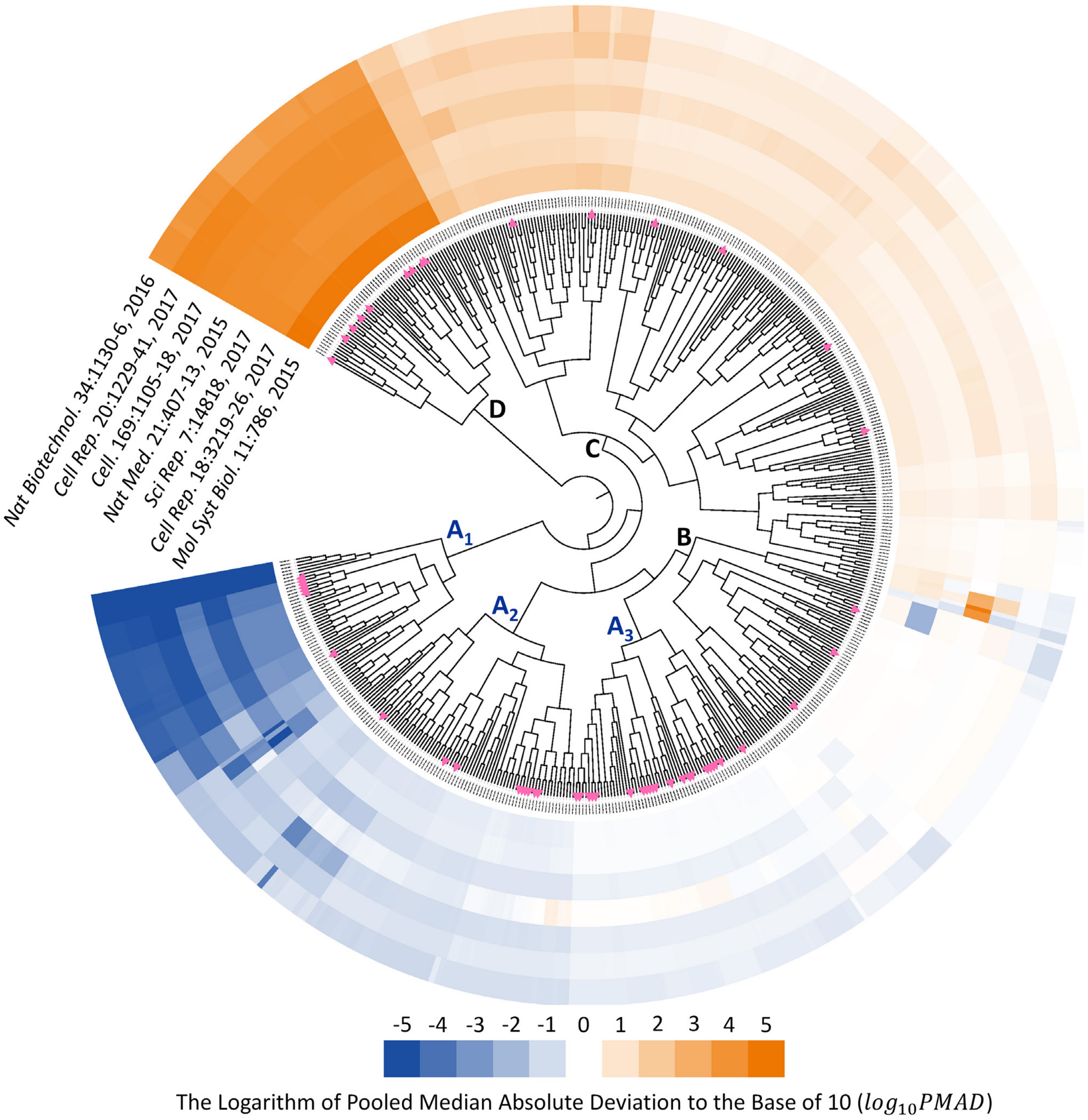


FIGURE 3 | Six partitions of analysis chains (A_1, A_2, A_3, B, C, and D) were identified based on their PMAD values. PMAD values meeting the "well-performed" criterion ($\leq$0.7) were displayed in blue color, with the $\log_{10}$ PMAD ≤ -5 set as exact blue and the larger PMADs gradually fading toward white (PMAD = 0.7). Meanwhile, the "poor-performed" PMAD values (>0.7) were all colored in orange, with $\log_{10}$ PMAD ≥ 5 set as exact orange and the smaller PMAD gradually fading toward white. The pink triangles indicated the analysis chains adopted by previous published SWATH-MS based proteomic studies.

to each processing method used to integrate these CWPACs, which might be able to discover processing methods with significant contributions to the performance of CWPACs. Therefore, all CWPACs listed in Supplementary Figures S1–S3 were investigated by analyzing their corresponding processing methods. As shown in **Figure 4**, the percentage of each method appeared in 3 different partitions (A_1 & A_2 & A_3, A_1 & A_2, and A_1) were analyzed. For transformation, the percentage of *Power Transformation* significantly increased from 7% to 10% to 29% with the gradual narrow down of partitions (from A_1 & A_2 & A_3 to A_1 & A_2 to A_1), which showed significantly enhanced role played by this transformation to achieve good performance in protein quantifications. However, *Log Transformation* decreased greatly from 41% to 25% to

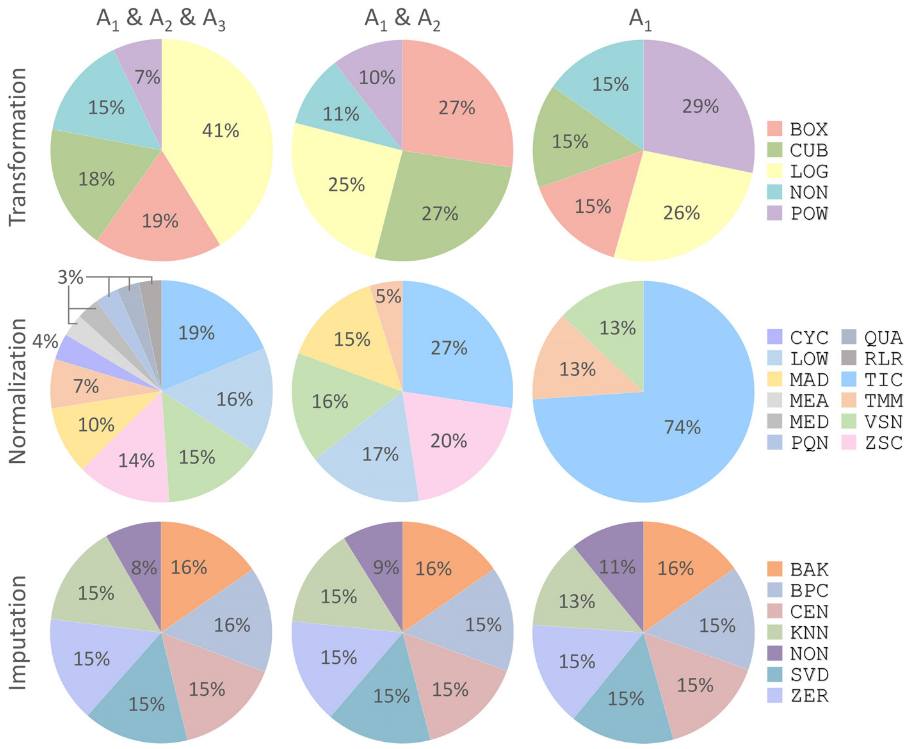


FIGURE 4 | Percentages of each processing method (transformation, normalization, and imputation) appeared in three different partitions (A_1 & A_2 & A_3, A_1 & A_2, and A_1) shown in **Figure 3**. Each processing method was abbreviated by a three-letter code as demonstrated in Supplementary Table S1.

26%. This indicated that *Log Transformation* contributed most to the CWPACs compared to other transformations. But when it came to the superior performance (partition A_1 with PMAD $\leq$ 0.1), its contribution decreased and ranked as the second. For normalization, the *Total Ion Current* method stood out among all methods as the one with the highest contribution to CWPAC. With gradual narrow down of partitions (from A_1 & A_2 & A_3 to A_1 & A_2 to A_1), the importance of *Total Ion Current* method was enhanced significantly from 19% to 27% to 74%. For imputation, methods were almost evenly distributed with no clear change among different partitions. This indicated that each imputation method contributed equally to CWPACs, and the selection of any of those methods could not make statistical difference in protein quantification. Due to the equal contribution of imputation methods, it was essential to focus on selecting the appropriate combinations of transformation and normalization methods to achieve the optimal performance of analysis chains, which included POW-TMM, LOG-TIC, BOX-TIC, CUB-TIC, NON-TIC, POW-TIC, and LOG-VSN (Supplementary Figure S1).

CONCLUSION

Based on the most complete set of the publicly available pharmacoproteomic data generated by SWATH-MS technique, this study revealed a substantial variation among the performances of various analysis chains applied for pharmacoproteomic quantification, and the analysis chains performed consistently well across a diverse set of publicly available pharmacoproteomic data were discovered. As a result, log and power transformations sequentially followed by total ion current normalization were discovered as one of the best performed analysis chains applied for the SWATH-MS based pharmacoproteomic quantification. In summary, the identified analysis chains provided important guidance to current proteomic research and could thus facilitate the cutting-edge research in any proteomic studies requiring SWATH-MS technique.

AUTHOR CONTRIBUTIONS

FZ conceived the idea and supervised the work. JF, JT, and YW performed the research. JF, XC, QY, JH, XL, SL, YC, and WX prepared and analyzed the data. FZ and JF wrote the manuscript. All authors have read and approved this manuscript.

FUNDING

Funded by the research support of the Precision Medicine Project of the National Key Research and Development Plan of China

(2016YFC0902200); Innovation Project on Industrial Generic Key Technologies of Chongqing (cstc2015zdcy-ztzx120003); and Fundamental Research Funds for the Central Universities (10611CDJXZ238826, CDJZR14468801, and CDJKXB14011).

REFERENCES

Aebersold, R., and Mann, M. (2016). Mass-spectrometric exploration of proteome structure and function. *Nature* 537, 347–355. doi: 10.1038/nature19949

Alter, O., Brown, P. O., and Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. U.S.A.* 97, 10101–10106. doi: 10.1073/pnas.97.18.10101

Andjelkovic, V., and Thompson, R. (2006). Changes in gene expression in maize kernel in response to water and salt stress. *Plant Cell Rep.* 25, 71–79. doi: 10.1007/s00299-005-0037-x

Anjo, S. I., Santa, C., and Manadas, B. (2017). SWATH-MS as a tool for biomarker discovery: from basic research to clinical applications. *Proteomics* 17:1600278. doi: 10.1002/pmic.201600278

Aratyn-Schaus, Y., and Ramanathan, R. (2016). Advances in high-resolution MS and hepatocyte models solve a long-standing metabolism challenge: the loratadine story. *Bioanalysis* 8, 1645–1662. doi: 10.4155/bio-2016-0094

Barer, M. R., and Harwood, C. R. (1999). Bacterial viability and culturability. *Adv. Microb. Physiol.* 41, 93–137. doi: 10.1016/S0065-2911(08)60166-6

Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19, 185–193. doi: 10.1093/bioinformatics/19.2.185

Borgaonkar, S. P., Hocker, H., Shin, H., and Markey, M. K. (2010). Comparison of normalization methods for the identification of biomarkers using MALDI-TOF and SELDI-TOF mass spectra. *OMICS* 14, 115–126. doi: 10.1089/omi.2009.0082

Bruderer, R., Bernhardt, O. M., Gandhi, T., Miladinovic, S. M., Cheng, L. Y., Messner, S., et al. (2015). Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues. *Mol. Cell. Proteomics* 14, 1400–1410. doi: 10.1074/mcp.M114.044305

Callister, S. J., Barry, R. C., Adkins, J. N., Johnson, E. T., Qian, W. J., Webb-Robertson, B. J., et al. (2006). Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics. *J. Proteome Res.* 5, 277–286. doi: 10.1021/pr050300l

Chai, L. E., Law, C. K., Mohamad, M. S., Chong, C. K., Choon, Y. W., Deris, S., et al. (2014). Investigating the effects of imputation methods for modelling gene networks using a dynamic bayesian network from gene expression data. *Malays. J. Med. Sci.* 21, 20–27.

Chambliss, A. B., and Chan, D. W. (2016). Precision medicine: from pharmacogenomics to pharmacoproteomics. *Clin. Proteomics* 13:25. doi: 10.1186/s12014-016-9127-8

Chawade, A., Alexandersson, E., and Levander, F. (2014). Normalyzer: a tool for rapid evaluation of normalization methods for omics data sets. *J. Proteome Res.* 13, 3114–3120. doi: 10.1021/pr401264n

Chawade, A., Sandin, M., Teleman, J., Malmstrom, J., and Levander, F. (2015). Data processing has major impact on the outcome of quantitative label-free LC-MS analysis. *J. Proteome Res.* 14, 676–687. doi: 10.1021/pr500665j

Cheadle, C., Vawter, M. P., Freed, W. J., and Becker, K. G. (2003). Analysis of microarray data using Z score transformation. *J. Mol. Diagn.* 5, 73–81. doi: 10.1016/S1525-1578(10)60455-2

Chignell, J. F., Park, S., Lacerda, C. M. R., De Long, S. K., and Reardon, K. F. (2018). Label-free proteomics of a defined, binary co-culture reveals diversity of competitive responses between members of a model soil microbial system. *Microb. Ecol.* 75, 701–719. doi: 10.1007/s00248-017-1072-1

D'Alessandro, A., and Zolla, L. (2010). Pharmacoproteomics: a chess game on a protein field. *Drug Discov. Today* 15, 1015–1023. doi: 10.1016/j.drudis.2010.10.002

De Livera, A. M., Dias, D. A., De Souza, D., Rupasinghe, T., Pyke, J., Tull, D., et al. (2012). Normalizing and integrating metabolomics data. *Anal. Chem.* 84, 10768–10776. doi: 10.1021/ac302748b

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphar.2018.00681/full#supplementary-material

Dieterle, F., Ross, A., Schlotterbeck, G., and Senn, H. (2006). Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics. *Anal. Chem.* 78, 4281–4290. doi: 10.1021/ac051632c

Feng, C., Wang, H., Lu, N., Chen, T., He, H., Lu, Y., et al. (2014). Log-transformation and its implications for data analysis. *Shanghai Arch. Psychiatry* 26, 105–109. doi: 10.3969/j.issn.1002-0829.2014.02.009

Fu, T., Zheng, G., Tu, G., Yang, F., Chen, Y., Yao, X., et al. (2018). Exploring the binding mechanism of metabotropic glutamate receptor 5 negative allosteric modulators in clinical trials by molecular dynamics simulations. *ACS Chem. Neurosci.* doi: 10.1021/acschemneuro.8b00059 [Epub ahead of print].

Gan, X., Liew, A. W., and Yan, H. (2006). Microarray missing data imputation based on a set theoretic framework and biological knowledge. *Nucleic Acids Res.* 34, 1608–1619. doi: 10.1093/nar/gkl047

Gaspari, M., Chiesa, L., Nicastri, A., Gabriele, C., Harper, V., Britti, D., et al. (2016). Proteome speciation by mass spectrometry: characterization of composite protein mixtures in milk replacers. *Anal. Chem.* 88, 11568–11574. doi: 10.1021/acs.analchem.6b02848

Gatto, L., Hansen, K. D., Hoopmann, M. R., Hermjakob, H., Kohlbacher, O., and Beyer, A. (2016). Testing and validation of computational methods for mass spectrometry. *J. Proteome Res.* 15, 809–814. doi: 10.1021/acs.jproteome.5b00852

Gillet, L. C., Navarro, P., Tate, S., Rost, H., Selevsek, N., Reiter, L., et al. (2012). Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol. Cell. Proteomics* 11:O111.016717. doi: 10.1074/mcp.O111.016717

Guo, T., Kouvonen, P., Koh, C. C., Gillet, L. C., Wolski, W. E., Rost, H. L., et al. (2015). Rapid mass spectrometric conversion of tissue biopsy samples into permanent quantitative digital proteome maps. *Nat. Med.* 21, 407–413. doi: 10.1038/nm.3807

Hauser, D. S., Stade, M., Schmidt, A., and Hanauer, G. (2005). Cardiovascular parameters in anaesthetized guinea pigs: a safety pharmacology screening model. *J. Pharmacol. Toxicol. Methods* 52, 106–114. doi: 10.1016/j.vascn.2005.03.003

Hong, M. G., Lee, W., Nilsson, P., Pawitan, Y., and Schwenk, J. M. (2016). Multidimensional normalization to minimize plate effects of suspension bead array data. *J. Proteome Res.* 15, 3473–3480. doi: 10.1021/acs.jproteome.5b01131

Huang, Q., Yang, L., Luo, J., Guo, L., Wang, Z., Yang, X., et al. (2015). SWATH enables precise label-free quantification on proteome scale. *Proteomics* 15, 1215–1223. doi: 10.1002/pmic.201400270

Huber, W., von Heydebreck, A., Sultmann, H., Poustka, A., and Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* 18, S96–S104. doi: 10.1093/bioinformatics/18.suppl_1.S96

Jamwal, R., Barlock, B. J., Adusumalli, S., Ogasawara, K., Simons, B. L., and Akhlaghi, F. (2017). Multiplex and label-free relative quantification approach for studying protein abundance of drug metabolizing enzymes in human liver microsomes using SWATH-MS. *J. Proteome Res.* 16, 4134–4143. doi: 10.1021/acs.jproteome.7b00505

Karpievitch, Y. V., Dabney, A. R., and Smith, R. D. (2012). Normalization and missing value imputation for label-free LC-MS analysis. *BMC Bioinformatics* 13:S5. doi: 10.1186/1471-2105-13-S16-S5

Khoonsari, P. E., Haggmark, A., Lonnberg, M., Mikus, M., Kilander, L., Lannfelt, L., et al. (2016). Analysis of the cerebrospinal fluid proteome in Alzheimer's Disease. *PLoS One* 11:e0150672. doi: 10.1371/journal.pone.0150672

Kohl, S. M., Klein, M. S., Hochrein, J., Oefner, P. J., Spang, R., and Gronwald, W. (2012). State-of-the art data normalization methods improve NMR-based metabolomic analysis. *Metabolomics* 8, 146–160. doi: 10.1007/s11306-011-0350-z

Krasny, L., Bland, P., Kogata, N., Wai, P., Howard, B. A., Natrajan, R. C., et al. (2018). SWATH mass spectrometry as a tool for quantitative profiling of the matrisome. *J. Proteomics* doi: 10.1016/j.jprot.2018.02.026 [Epub ahead of print].

Kuharev, J., Navarro, P., Distler, U., Jahn, O., and Tenzer, S. (2015). In-depth evaluation of software tools for data-independent acquisition based label-free quantification. *Proteomics* 15, 3140–3151. doi: 10.1002/pmic.201400396

Letunic, I., and Bork, P. (2016). Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* 44, W242–W245. doi: 10.1093/nar/gkw290

Li, B., Tang, J., Yang, Q., Cui, X., Li, S., Chen, S., et al. (2016a). Performance evaluation and online realization of data-driven normalization methods used in LC/MS based untargeted metabolomics analysis. *Sci. Rep.* 6:38881. doi: 10.1038/srep38881

Li, B., Tang, J., Yang, Q., Li, S., Cui, X., Li, Y., et al. (2017). NOREVA: normalization and evaluation of MS-based metabolomics data. *Nucleic Acids Res.* 45, W162–W170. doi: 10.1093/nar/gkx449

Li, S., Cao, Q., Xiao, W., Guo, Y., Yang, Y., Duan, X., et al. (2017). Optimization of acquisition and data-processing parameters for improved proteomic quantification by sequential window acquisition of all theoretical fragment ion mass spectrometry. *J. Proteome Res.* 16, 738–747. doi: 10.1021/acs.jproteome.6b00767

Li, Y. H., Wang, P. P., Li, X. X., Yu, C. Y., Yang, H., Zhou, J., et al. (2016b). The human kinome targeted by fda approved multi-target drugs and combination products: a comparative study from the drug-target interaction network perspective. *PLoS One* 11:e0165737. doi: 10.1371/journal.pone.0165737

Li, Y. H., Xu, J. Y., Tao, L., Li, X. F., Li, S., Zeng, X., et al. (2016c). SVM-Prot 2016: a web-server for machine learning prediction of protein functional families from sequence irrespective of similarity. *PLoS One* 11:e0155290. doi: 10.1371/journal.pone.0155290

Li, Y. H., Yu, C. Y., Li, X. X., Zhang, P., Tang, J., Yang, Q., et al. (2018). Therapeutic target database update 2018: enriched resource for facilitating bench-to-clinic research of targeted therapeutics. *Nucleic Acids Res.* 46, D1121–D1127. doi: 10.1093/nar/gkx1076

Lin, Y., Golovnina, K., Chen, Z. X., Lee, H. N., Negron, Y. L., Sultana, H., et al. (2016). Comparison of normalization and differential expression analyses using RNA-Seq data from 726 individual *Drosophila melanogaster*. *BMC Genomics* 17:28. doi: 10.1186/s12864-015-2353-z

Liu, Y., Buil, A., Collins, B. C., Gillet, L. C., Blum, L. C., Cheng, L. Y., et al. (2015). Quantitative variability of 342 plasma proteins in a human twin population. *Mol. Syst. Biol.* 11:786. doi: 10.15252/msb.20145728

MacLean, B., Tomazela, D. M., Shulman, N., Chambers, M., Finney, G. L., Frewen, B., et al. (2010). Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* 26, 966–968. doi: 10.1093/bioinformatics/btq054

Matzke, M. M., Waters, K. M., Metz, T. O., Jacobs, J. M., Sims, A. C., Baric, R. S., et al. (2011). Improved quality control processing of peptide-centric LC-MS proteomics data. *Bioinformatics* 27, 2866–2872. doi: 10.1093/bioinformatics/btr479

Mrozek, D., Malysiak-Mrozek, B., and Siaznik, A. (2013). search GenBank: interactive orchestration and ad-hoc choreography of Web services in the exploration of the biomedical resources of the National Center For Biotechnology Information. *BMC Bioinformatics* 14:73. doi: 10.1186/1471-2105-14-73

Muller, F., Fischer, L., Chen, Z. A., Auchynnikava, T., and Rappsilber, J. (2018). On the reproducibility of label-free quantitative cross-linking/mass spectrometry. *J. Am. Soc. Mass Spectrom.* 29, 405–412. doi: 10.1007/s13361-017-1837-2

Navarro, P., Kuharev, J., Gillet, L. C., Bernhardt, O. M., MacLean, B., Rost, H. L., et al. (2016). A multicenter study benchmarks software tools for label-free proteome quantification. *Nat. Biotechnol.* 34, 1130–1136. doi: 10.1038/nbt.3685

Ori, A., Iskar, M., Buczak, K., Kastritis, P., Parca, L., Andres-Pons, A., et al. (2016). Spatiotemporal variation of mammalian protein complex stoichiometries. *Genome Biol.* 17:47. doi: 10.1186/s13059-016-0912-5

Parker, S. J., Rost, H., Rosenberger, G., Collins, B. C., Malmstrom, L., Amodei, D., et al. (2015). Identification of a set of conserved eukaryotic internal retention time standards for data-independent acquisition mass spectrometry. *Mol. Cell. Proteomics* 14, 2800–2813. doi: 10.1074/mcp.O114.042267

Paul, D., Chanukuppa, V., Reddy, P. J., Taunk, K., Adhav, R., Srivastava, S., et al. (2016). Global proteomic profiling identifies etoposide chemoresistance markers in non-small cell lung carcinoma. *J. Proteomics* 138, 95–105. doi: 10.1016/j.jprot.2016.02.008

Quiros, P. M., Prado, M. A., Zamboni, N., D'Amico, D., Williams, R. W., Finley, D., et al. (2017). Multi-omics analysis identifies ATF4 as a key regulator of the mitochondrial stress response in mammals. *J. Cell Biol.* 216, 2027–2045. doi: 10.1083/jcb.201702058

Rao, H. B., Zhu, F., Yang, G. B., Li, Z. R., and Chen, Y. Z. (2011). Update of PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Res.* 39, W385–W390. doi: 10.1093/nar/gkr284

Roemmelt, A. T., Steuer, A. E., and Kraemer, T. (2015). Liquid chromatography, in combination with a quadrupole time-of-flight instrument, with sequential window acquisition of all theoretical fragment-ion spectra acquisition: validated quantification of 39 antidepressants in whole blood as part of a simultaneous screening and quantification procedure. *Anal. Chem.* 87, 9294–9301. doi: 10.1021/acs.analchem.5b02031

Rost, H. L., Rosenberger, G., Navarro, P., Gillet, L., Miladinovic, S. M., Schubert, O. T., et al. (2014). OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nat. Biotechnol.* 32, 219–223. doi: 10.1038/nbt.2841

Saei, A. A., Sabatier, P., Guler Tokat, U., Chernobrovkin, A., Pirmoradian, M., and Zubarev, R. A. (2018). Comparative proteomics of dying and surviving cancer cells improves the identification of drug targets and sheds light on cell life/death decisions. *Mol. Cell. Proteomics* 17, 1144–1155. doi: 10.1074/mcp.RA118.000610

Sajic, T., Liu, Y., and Aebersold, R. (2015). Using data-independent, high-resolution mass spectrometry in protein biomarker research: perspectives and clinical applications. *Proteomics Clin. Appl.* 9, 307–321. doi: 10.1002/prca.201400117

Sakia, R. M. (1992). The box-cox transformation technique - a review. *Statistician* 41, 169–178. doi: 10.2307/2348250

Scheidweiler, K. B., Jarvis, M. J., and Huestis, M. A. (2015). Nontargeted SWATH acquisition for identifying 47 synthetic cannabinoid metabolites in human urine by liquid chromatography-high-resolution tandem mass spectrometry. *Anal. Bioanal. Chem.* 407, 883–897. doi: 10.1007/s00216-014-8118-8

Schubert, O. T., Gillet, L. C., Collins, B. C., Navarro, P., Rosenberger, G., Wolski, W. E., et al. (2015). Building high-quality assay libraries for targeted analysis of SWATH MS data. *Nat. Protoc.* 10, 426–441. doi: 10.1038/nprot.2015.015

Shi, T., Song, E., Nie, S., Rodland, K. D., Liu, T., Qian, W. J., et al. (2016). Advances in targeted proteomics and applications to biomedical research. *Proteomics* 16, 2160–2182. doi: 10.1002/pmic.201500449

Sullivan, K. D., Evans, D., Pandey, A., Hraha, T. H., Smith, K. P., Markham, N., et al. (2017). Trisomy 21 causes changes in the circulating proteome indicative of chronic autoinflammation. *Sci. Rep.* 7:14818. doi: 10.1038/s41598-017-13858-3

Tan, S. L. W., Chadha, S., Liu, Y., Gabasova, E., Perera, D., Ahmed, K., et al. (2017). A class of environmental and endogenous toxins induces BRCA2 haploinsufficiency and genome instability. *Cell* 169, 1105–1118. doi: 10.1016/j.cell.2017.05.010

Tao, L., Zhu, F., Xu, F., Chen, Z., Jiang, Y. Y., and Chen, Y. Z. (2015). Co-targeting cancer drug escape pathways confers clinical advantage for multi-target anticancer drugs. *Pharmacol. Res.* 102, 123–131. doi: 10.1016/j.phrs.2015.09.019

Tippmann, S. (2015). Programming tools: adventures with R. *Nature* 517, 109–110. doi: 10.1038/517109a

Valikangas, T., Suomi, T., and Elo, L. L. (2017). A comprehensive evaluation of popular proteomics software workflows for label-free proteome quantification and imputation. *Brief. Bioinform.* doi: 10.1093/bib/bbx054 [Epub ahead of print].

Valikangas, T., Suomi, T., and Elo, L. L. (2018). A systematic evaluation of normalization methods in quantitative label-free proteomics. *Brief. Bioinform.* 19, 1–11. doi: 10.1093/bib/bbw095

Wang, P., Fu, T., Zhang, X., Yang, F., Zheng, G., Xue, W., et al. (2017a). Differentiating physicochemical properties between NDRIs and sNRIs clinically important for the treatment of ADHD. *Biochim. Biophys. Acta* 1861, 2766–2777. doi: 10.1016/j.bbagen.2017.07.022

Wang, P., Yang, F., Yang, H., Xu, X., Liu, D., Xue, W., et al. (2015). Identification of dual active agents targeting 5-HT1A and SERT by combinatorial virtual screening methods. *Biomed. Mater. Eng.* 26, S2233–S2239. doi: 10.3233/BME-151529

Wang, P., Zhang, X., Fu, T., Li, S., Li, B., Xue, W., et al. (2017b). Differentiating physicochemical properties between addictive and nonaddictive ADHD drugs revealed by molecular dynamics simulation studies. *ACS Chem. Neurosci.* 8, 1416–1428. doi: 10.1021/acschemneuro.7b00173

Weisser, H., and Choudhary, J. S. (2017). Targeted feature detection for data-dependent shotgun proteomics. *J. Proteome Res.* 16, 2964–2974. doi: 10.1021/acs.jproteome.7b00248

Wen, B., Mei, Z., Zeng, C., and Liu, S. (2017). metaX: a flexible and comprehensive software for processing metabolomics data. *BMC Bioinformatics* 18:183. doi: 10.1186/s12859-017-1579-y

Wilson, D. L., Buckley, M. J., Helliwell, C. A., and Wilson, I. W. (2003). New normalization methods for cDNA microarray data. *Bioinformatics* 19, 1325–1332. doi: 10.1093/bioinformatics/btg146

Wisniewski, J. R., Ostasiewicz, P., Dus, K., Zielinska, D. F., Gnad, F., and Mann, M. (2012). Extensive quantitative remodeling of the proteome between normal colon tissue and adenocarcinoma. *Mol. Syst. Biol.* 8:611. doi: 10.1038/msb.2012.44

Wu, J. X., Song, X., Pascovici, D., Zaw, T., Care, N., Krisp, C., et al. (2016). SWATH mass spectrometry performance using extended peptide MS/MS assay libraries. *Mol. Cell. Proteomics* 15, 2501–2514. doi: 10.1074/mcp.M115.055558

Xu, J., Wang, P., Yang, H., Zhou, J., Li, Y., Li, X., et al. (2016). Comparison of FDA approved kinase targets to clinical trial ones: insights from their system profiles and drug-target interaction networks. *Biomed Res. Int.* 2016:2509385. doi: 10.1155/2016/2509385

Xue, W., Wang, P., Li, B., Li, Y., Xu, X., Yang, F., et al. (2016). Identification of the inhibitory mechanism of FDA approved selective serotonin reuptake inhibitors: an insight from molecular dynamics simulation study. *Phys. Chem. Chem. Phys.* 18, 3260–3271. doi: 10.1039/c5cp05771j

Xue, W., Wang, P., Tu, G., Yang, F., Zheng, G., Li, X., et al. (2018a). Computational identification of the binding mechanism of a triple reuptake inhibitor amitifadine for the treatment of major depressive disorder. *Phys. Chem. Chem. Phys.* 20, 6606–6616. doi: 10.1039/c7cp07869b

Xue, W., Yang, F., Wang, P., Zheng, G., Chen, Y., Yao, X., et al. (2018b). What contributes to serotonin-norepinephrine reuptake inhibitors' dual-targeting mechanism? the key role of transmembrane domain 6 in human serotonin and norepinephrine transporters revealed by molecular dynamics simulation. *ACS Chem. Neurosci.* 9, 1128–1140. doi: 10.1021/acschemneuro.7b00490

Yang, F. Y., Fu, T. T., Zhang, X. Y., Hu, J., Xue, W. W., Zheng, G. X., et al. (2017). Comparison of computational model and X-ray crystal structure of human serotonin transporter: potential application for the pharmacology of human monoamine transporters. *Mol. Simul.* 43, 1089–1098. doi: 10.1080/08927022.2017.1309653

Yang, H., Qin, C., Li, Y. H., Tao, L., Zhou, J., Yu, C. Y., et al. (2016). Therapeutic target database update 2016: enriched resource for bench to clinical drug target and targeted pathway information. *Nucleic Acids Res.* 44, D1069–D1074. doi: 10.1093/nar/gkv1230

Yu, C. Y., Li, X. X., Yang, H., Li, Y. H., Xue, W. W., Chen, Y. Z., et al. (2018). Assessing the performances of protein function prediction algorithms from the perspectives of identification accuracy and false discovery rate. *Int. J. Mol. Sci.* 19:E183 doi: 10.3390/ijms19010183

Yue, Q., Feng, L., Cao, B., Liu, M., Zhang, D., Wu, W., et al. (2016). Proteomic analysis revealed the important role of vimentin in human cervical carcinoma hela cells treated with gambogic acid. *Mol. Cell. Proteomics* 15, 26–44. doi: 10.1074/mcp.M115.053272

Zeng, X., Ding, N., Rodriguez-Paton, A., and Zou, Q. (2017). Probability-based collaborative filtering model for predicting gene-disease associations. *BMC Med. Genomics* 10:76. doi: 10.1186/s12920-017-0313-y

Zhang, Z. (2014). Recombinant human activated protein C for the treatment of severe sepsis and septic shock: a study protocol for incorporating observational evidence using a Bayesian approach. *BMJ Open* 4:e005622. doi: 10.1136/bmjopen-2014-005622

Zheng, G., Xue, W., Wang, P., Yang, F., Li, B., Li, X., et al. (2016). Exploring the inhibitory mechanism of approved selective norepinephrine reuptake inhibitors and reboxetine enantiomers by molecular dynamics study. *Sci. Rep.* 6:26883. doi: 10.1038/srep26883

Zheng, G., Xue, W., Yang, F., Zhang, Y., Chen, Y., Yao, X., et al. (2017). Revealing vilazodone's binding mechanism underlying its partial agonism to the 5-HT1A receptor in the treatment of major depressive disorder. *Phys. Chem. Chem. Phys.* 19, 28885–28896. doi: 10.1039/c7cp05688e

Zhu, F., Han, B., Kumar, P., Liu, X., Ma, X., Wei, X., et al. (2010). Update of TTD: therapeutic target database. *Nucleic Acids Res.* 38, D787–D791. doi: 10.1093/nar/gkp1014

Zhu, F., Han, L., Zheng, C., Xie, B., Tammi, M. T., Yang, S., et al. (2009). What are next generation innovative therapeutic targets? Clues from genetic, structural, physicochemical, and systems profiles of successful targets. *J. Pharmacol. Exp. Ther.* 330, 304–315. doi: 10.1124/jpet.108.149955

Zhu, F., Han, L. Y., Chen, X., Lin, H. H., Ong, S., Xie, B., et al. (2008a). Homology-free prediction of functional class of proteins and peptides by support vector machines. *Curr. Protein Pept. Sci.* 9, 70–95. doi: 10.2174/138920308783565697

Zhu, F., Li, X. X., Yang, S. Y., and Chen, Y. Z. (2018). Clinical success of drug targets prospectively predicted by in silico study. *Trends Pharmacol. Sci.* 39, 229–231. doi: 10.1016/j.tips.2017.12.002

Zhu, F., Ma, X. H., Qin, C., Tao, L., Liu, X., Shi, Z., et al. (2012a). Drug discovery prospect from untapped species: indications from approved natural product drugs. *PLoS One* 7:e39782. doi: 10.1371/journal.pone.0039782

Zhu, F., Qin, C., Tao, L., Liu, X., Shi, Z., Ma, X., et al. (2011). Clustered patterns of species origins of nature-derived drugs and clues for future bioprospecting. *Proc. Natl. Acad. Sci. U.S.A.* 108, 12943–12948. doi: 10.1073/pnas.1107336108

Zhu, F., Shi, Z., Qin, C., Tao, L., Liu, X., Xu, F., et al. (2012b). Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery. *Nucleic Acids Res.* 40, D1128–D1136. doi: 10.1093/nar/gkr797

Zhu, F., Zheng, C. J., Han, L. Y., Xie, B., Jia, J., Liu, X., et al. (2008b). Trends in the exploration of anticancer targets and strategies in enhancing the efficacy of drug targeting. *Curr. Mol. Pharmacol.* 1, 213–232. doi: 10.2174/1874467210801030213

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Pharmacology

ORIGINAL RESEARCH
published: 24 July 2018
doi: 10.3389/fphar.2018.00778

Therapeutic Effect of Repurposed Temsirolimus in Lung Adenocarcinoma Model

*Hsuen-Wen Chang[1], Min-Ju Wu[1], Zih-Miao Lin[2], Chueh-Yi Wang[1], Shu-Yun Cheng[1], Yen-Kuang Lin[3], Yen-Hung Chow[4], Hui-Ju Ch'ang[5] and Vincent H. S. Chang[2]**

[1] Laboratory Animal Center, Office of Research and Development, Taipei Medical University, Taipei, Taiwan, [2] The PhD Program for Translational Medicine, College of Medical Science and Technology, Taipei Medical University, Taipei, Taiwan, [3] Biostatistics Research Center, Taipei Medical University, Taipei, Taiwan, [4] National Institutes of Infectious Diseases and Vaccinology, National Health Research Institutes, Zhunan, Taiwan, [5] National Institute of Cancer Research, National Health Research Institutes, Zhunan, Taiwan

OPEN ACCESS

Edited by:
*Zhaohui John Cai,
Celgene (United States),
United States*

Reviewed by:
*Michele Samaja,
Università degli Studi di Milano, Italy
Michela De Bellis,
Università degli Studi di Bari Aldo
Moro, Italy*

***Correspondence:**
*Vincent H. S. Chang
vinhschang@tmu.edu.tw*

Specialty section:
*This article was submitted to
Translational Pharmacology,
a section of the journal
Frontiers in Pharmacology*

Received: *16 March 2018*
Accepted: *26 June 2018*
Published: *24 July 2018*

Citation:
Chang H-W, Wu M-J, Lin Z-M, Wang C-Y, Cheng S-Y, Lin Y-K, Chow Y-H, Ch'ang H-J and Chang VHS (2018) Therapeutic Effect of Repurposed Temsirolimus in Lung Adenocarcinoma Model. Front. Pharmacol. 9:778. doi: 10.3389/fphar.2018.00778

Lung cancer is one of the major cause of cancer-related deaths worldwide. The poor prognosis and resistance to both radiation and chemotherapy urged the development of potential targets for lung cancer treatment. In this study, using a network-based cellular signature bioinformatics approach, we repurposed a clinically approved mTOR inhibitor for renal cell carcinomans, temsirolimus, as the potential therapeutic candidate for lung adenocarcinoma. The PI3K-AKT-mTOR pathway is known as one of the most frequently dysregulated pathway in cancers, including non-small-cell lung cancer. By using a well-documented lung adenocarcinoma mouse model of human pathophysiology, we examined the effect of temsirolimus on the growth of lung adenocarcinoma *in vitro* and *in vivo*. In addition, temsirolimus combined with reduced doses of cisplatin and gemcitabine significantly inhibited the lung tumor growth in the lung adenocarcinoma mouse model compared with the temsirolimus alone or the conventional cisplatin–gemcitabine combination. Functional imaging techniques and microscopic analyses were used to reveal the response mechanisms. Extensive immunohistochemical analyses were used to demonstrate the apparent effects of combined treatments on tumor architecture, vasculature, apoptosis, and the mTOR-pathway. The present findings urge the further exploration of temsirolimus in combination with chemotherapy for treating lung adenocarcinoma.

Keywords: mTOR inhibitor, drug repositioning, temsirolimus, lung adenocarcinoma, chemotherapy

INTRODUCTION

Lung cancer is one of the most common forms of cancer and remains the number one cause of cancer-related deaths worldwide among men and women. Based on histological differentiation, there are two major types of lung cancers: small-cell lung cancer (SCLC) and non-small-cell lung cancer (NSCLC). NSCLCs are further divided into squamous cell carcinomas (SCCs), pulmonary adenocarcinomas (ADC), and large-cell carcinomas. Among them, lung ADC is the most prevalent form of NSCLC (Teng, 2005; Chang et al., 2017). Lung cancer has a dismal prognosis of 15%, mainly attributed to ineffective early detection and lack of therapeutic options for metastatic disease (Molina et al., 2008). This has spurred efforts for the development of molecularly targeted therapies.

The definition of drug repositioning is to identify new indications from existing drugs or compounds to treat a different disease. In addition to being time- and cost-efficient, drug repositioning offers a more favorable risk-versus-reward trade-off of the available drug development strategies. Because the existing drugs have already been tested in terms of safety, dosage, and toxicity, they can often enter clinical trials much more rapidly than newly developed drugs (Ashburn and Thor, 2004). Computational drug repositioning is deemed as an alternative and effective way to identifying novel connections between diseases and existing drugs (Hurle et al., 2013). The increase in drug-target information and advances in systems pharmacology approaches have led to an increase in the success of *in silico* drug repositioning. In particular, large-scale genomics databases, such as the Connectivity Map, provide abundant information on the modes of action of drugs, which are reflected in the transcriptomic responses to chemical perturbation (Vempati et al., 2014). Recently, a similar but highly expanded version of a chemical genomics dataset was publicly released by the National Institutes of Health Library of Integrated Network-Based Cellular Signatures (NIH LINCS) program. This dataset includes gene expression signatures and protein binding, cellular phenotypic, and phosphoproteomics profiles resulted from chemical or genetic perturbation. Specifically, it presents the gene expression profiles of approximately 1000 landmark genes (L1000) in response to more than 20,000 chemical perturbations across many cell lines. Additionally, transcriptome-level expression profiles of approximately 20,000 genes have been computationally inferred using 1000 landmark genes (Vempati et al., 2014).

In this study, we compared the transcriptome profiles obtained from a well-documented mouse lung cancer model (Chang et al., 2017) and used the LINCS L1000 cellular signature bioinformatics approach to identify clinically approved candidate drugs to treat ADC. By using this strategy, we identified temsirolimus, a mTOR inhibitor approved for renal cell carcinoma, as a potential therapeutic agent for the treatment of lung tumor. In a study using mouse model xenografted with human NSCLC cells (A549, H1299, and H358), it was found that temsirolimus could inhibit the growth of subcutaneous tumors, as well as to prolong the survival of mice having pleural dissemination of cancer cells due to its anti-proliferative effect (Ohara et al., 2011). Temsirolimus also has been used on a case report (Vichai and Kirtikara, 2006) with lung adenocarcinoma harboring specific gene mutation; it was also noted to restore radio-sensitivity in lung adenocarcinoma cell lines (Ushijima et al., 2015). Two updated phase two clinical trials of temsirolimus (Study 1: Neratinib with and without temsirolimus for patients with HER2 activating mutations in non-small cell lung cancer. Study 2: Temsirolimus and pemetrexed for recurrent or refractory non-small cell lung cancer.) were found from webpage searching[1], either as monotherapy or combined therapy with another drug. Although there are more than 40 inhibitors of the PI3K-AKT-mTOR signaling pathway have reached different stages of clinical development, only a few have been approved for clinical use (Skehan et al., 1990). However, an *in vivo* systemic evaluation of the lung tumor inhibitory effect of temsirolimus was lack. Here we assessed the combination of the mTOR inhibitor temsirolimus with the first-line chemotherapy for advanced NSCLC, cisplatin, and gemcitabine, to reduce cytotoxicity and enhance the therapeutic response.

MATERIALS AND METHODS

Microarray Analysis

Total RNA was extracted from tissue samples or cells by using TRIzol® Reagent (Sigma, St. Louis, MO, United States) by following the manufacturer's instructions. Total RNA (0.2 μg) was amplified as previously mentioned (Chang et al., 2017) for microarray analysis by using a microarray scanner (Agilent Technologies, Santa Clara, CA, United States). A total of 155 differentially expressed genes were identified in the Tg-3m mice, of which 126 genes were upregulated (a log2 fold change of $\geq$0.6) and 29 genes were downregulated (a log2 fold change of ≤ -0.6). A total of 123 differentially expressed genes were identified in the Tg-6m mice, of which 105 genes were upregulated (a log2 fold change of $\geq$0.6) and 18 genes were downregulated (a log2 fold change of ≤ -0.6).

LINCS Perturbagen Signature Comparisons

The LINCS L1000 is one of the complete drug treatment expression profile databases and currently contains more than a million gene expression profiles of chemically perturbed human cell lines (Blois et al., 2011; Li et al., 2014). First, for comparing the gene expression signatures in transgenic mice with LINCS data sets, the gene expression data were ranked according to the log2 fold changes. We retrieved the top 100 and bottom 100 most differentially expressed genes as gene expression signatures in both Tg-3m and Tg-6m mice. Then, we transferred the mouse gene symbols to homologous human gene symbols by using the HomoloGene database (Gabriel et al., 2016). Next, we queried the homologous human genes against the LINCS database by using sig_query and sig_summly in the LINCS C3 server. Finally, we annotated the returned results by combining the DrugBank (Vichai and Kirtikara, 2006) and PubChem (Ushijima et al., 2015) results to provide detailed perturbagen information.

Functional Annotation of Differentially Expressed Genes

To discuss the gene ontology and Kyoto Encyclopedia of Genes and Genomes pathways involved in transgenic mice, we analyzed the differentially expressed genes by using the Database for Annotation, Visualization and Integrated Discovery (DAVID, version 6.7[2]) (Huang da et al., 2009) application programming interfaces (APIs). A p-value of 0.05 was set as the threshold, which was calculated using Fisher's exact test.

[1] https://clinicaltrials.gov/beta/

[2] david.ncifcrf.gov

Animals and Ethics Statement

Murine lung adenocarcinoma models were maintained as previously mentioned (Chang et al., 2017) in a specific pathogen-free environment at the animal facility of Taipei Medical University. Experimental uses of mice were approved by the Institutional Animal Care and Use Committee of Taipei Medical University (Approved Proposal No. LAC-2014-0217). All experiments were conducted in accordance with relevant guidelines and regulations. The mice were monitored daily for physiological conditions. Tumor growths were monitored using micro-CT on a weekly basis. Mice were anesthetized by administering 5% isoflurane followed by 2% isoflurane through the inhalation route for maintenance during the imaging process. Total lung volumes were measured and analyzed using CTAn software (v.1.15), and mice were euthanized when the total lung volumes were less than 120 mm^3. At the endpoint of the experiment (the 16th week), the tested mice were euthanized by administering 100% CO_2 through inhalation to minimize their suffering.

Cell Cycle and Apoptosis Assays

The effects of temsirolimus and chemotherapy on the cell cycle and apoptosis were evaluated by seeding tumor cells into 6-well plates at a density of 5×10^4 per well. The cells were treated accordingly and incubated for 24 h followed by a phosphate-buffered saline (PBS) wash. The cell cycle phases were determined using a Muse cell analyzer (Merck Millipore, Darmstadt, Germany) and a Muse Cell Cycle Assay Kit (Merck Millipore, Darmstadt, Germany) according to the manufacturer's instructions. Cell apoptosis was analyzed using Annexin V Dead cell reagent (Merck Millipore, Darmstadt, Germany) according to the manufacturer's instructions. An average of at least 10,000 cells was analyzed for each condition. Triplicate independent experiments were conducted.

Protein Preparation and Western Blotting

Protein extraction and Western blotting analysis were performed as previously mentioned (Chang et al., 2017). The blots were immunostained with 1:1000 of anti-p-mTOR (Ser2448) antibody (2971, Cell Signaling, Danvers, MA, United States). After incubation with horseradish peroxidase-conjugated secondary antibody (1:4000 of goat antirabbit IgG, GTX213110-01, GeneTex, Irvine, CA, United States), protein bands were visualized with an enhanced chemiluminescent reagent.

Micro-CT

Mice were anesthetized with an induction flow dose of 3% isoflurane and oxygen mixture, following a maintaining flow dose of 1%. The chest area was scanned at one time through *in vivo* micro-CT (Bruker SkyScan 1176, Kontich, Belgium). Image scanning was performed in resolution of 35 μm. The instrument setting was at a voltage of 50 kVp, a current of 500 μA, and an exposure time of 50 ms with a 0.5-mm aluminum filter. To prevent artifacts caused by cardiac and respiratory motion, images were captured using the synchronization mode. Sections were reconstructed using a graphics processing unit-based NRecon software. The tumor volume inside the lung area was separated and analyzed using CTAn software (Bruker SkyScan, Kontich, Belgium). The cross-sectional images were obtained using DataViewer software (Bruker Skyscan, Kontich, Belgium).

Histology and Immunohistochemistry

Mouse lung tumors were removed and prepared for paraffin-embedded sectioning immunohistochemistry (IHC) staining was performed as previously mentioned (Chang et al., 2017). After antigen retrieval, primary antibody dilutions were prepared in the blocking buffer (10% bovine serum albumin with 0.1% Triton-100 in PBS) as follows: 1:200 of anti-Ki67 antibody (ab15580, Abcam, Cambridge, MA, United States), 1:250 of anti-CD34 antibody (ab81289, Abcam, Cambridge, MA, United States), 1:100 of p-mTOR (ab109268, Abcam, Cambridge, MA, United States), and 1:400 of p-S6RP antibody (2211, Cell Signaling, Danvers, MA, United States). Immunochemical signals were detected using a MultiLink Detection Kit (BioGenex, Fremont, CA, United States). The peroxidase reaction was developed with diaminobenzidine, and sections were counterstained using Mayer's hematoxylin. The intensity of positive signal areas was measured using ImageJ software (IJ 1.46r).

Statistical Analyses

SAS version 9.3 for Windows (SAS Institute, Cary, NC, United States) was used for data manipulation and visualization. The means are used to describe the central tendency of continuous variables while standard deviations are used to depict the variation. One-way ANOVA and the Bonferroni *post hoc* multiple comparison tests were of inhibitory effects among different treatments. All statistical analyses were two sided, and $p < 0.05$ was considered as statistically significant. p-Values were depicted using asterisks, with $^*p < 0.05$, $^{**}p < 0.01$.

RESULTS

Data Processing and Drug Repositioning

To compare the gene expression signatures from different stages of lung tumors, microarray results of Tg-3m and Tg-6m tumors (Chang et al., 2017) were subjected to the LINCS L1000 data sets, and the gene expression data were ranked according to the log2 fold changes. The top 100 and bottom 100 most differentially expressed genes were retrieved as gene expression signatures in both Tg-3m and Tg-6m mice. The mouse gene symbols were then converted to homologous human gene symbols by using the HomoloGene database (Coordinators, 2016). Next, the homologous human genes were queried against the LINCS database by using sig_query and sig_summly in the LINCS C3 server. The returned results were annotated by combining the DrugBank (Wishart et al., 2006) and PubChem (Kim et al., 2016) results to obtain detailed drug information (**Figure 1**). The drugs that negatively (K score = −1) correlated with the gene expression from both Tg-3m and Tg-6m lung tumor cell lines were selected for further screening. The data regarding the drugs were then manually curated from DrugBank and PubMed

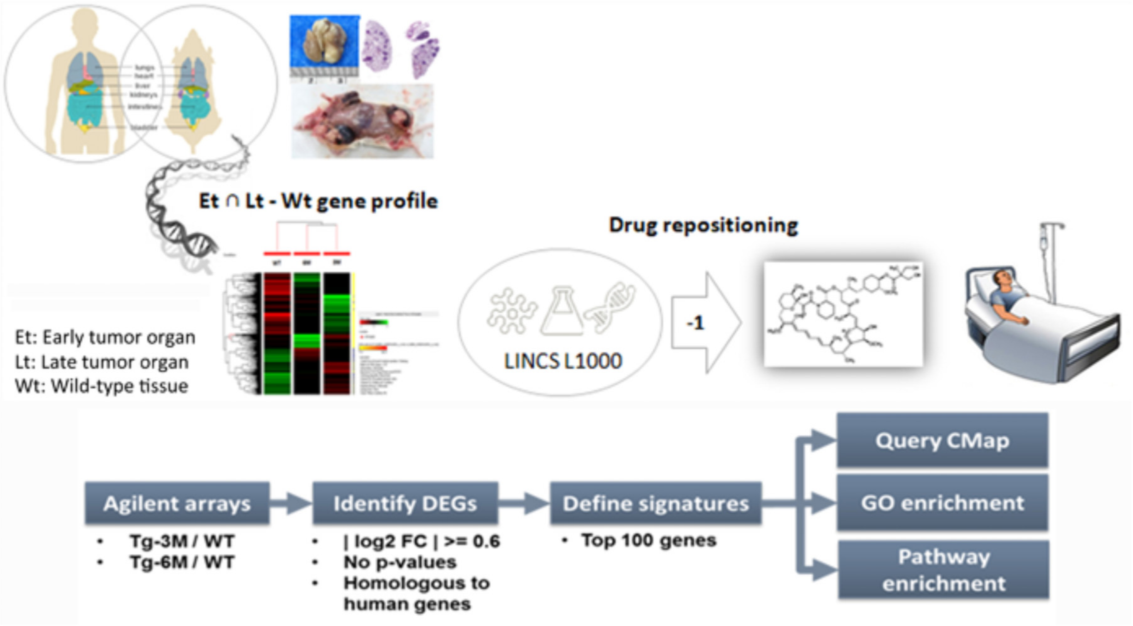


FIGURE 1 | Bioinformatics-based drug-repositioning approach to identify candidate drugs. Schematic representation of the bioinformatics workflow by using the LINCS L1000 data set for the repositioning approach to identifying potential candidate drugs for the treatment of NSCLC. The microarray results of Tg-3m and Tg-6m tumors were subjected to the LINCS L1000 data sets to obtain the most differentially expressed genes. The mouse gene symbols were then converted to human homologous genes and annotated by combining the DrugBank and PubChem results to obtain detailed drug information. The drugs that negatively (K score = −1) correlated with the gene expression from both Tg-3m and Tg-6m lung tumor cell lines were selected for further screening.

by searching for keywords and abstracts that explicitly described their association with cancers. The repositioned drug candidates are listed in **Table 1**. This list contained a wide range of drugs, including some antineoplastic agents used for cancers other than lung cancer, suggesting that the use of these agents in clinics may affect the gene expression signature of lung cancer (Kerr et al., 2007; Lin et al., 2007; Gallotta et al., 2010; Wynne and Djakiew, 2010; Endo et al., 2014; Li et al., 2016). We focused on the top-scoring candidates and clinically approved antineoplastic drugs. This analysis led to the identification of temsirolimus, a U.S. Food and Drug Administration (FDA)-approved mTOR inhibitor for renal cell carcinoma, which was repositioned from both stages of lung tumor cells and was tested in combination with thoracic radiation in NSCLC (Waqar et al., 2014).

Temsirolimus Treatment Leads to G_0/G_1 Cell Cycle Arrest

To understand whether temsirolimus treatment is lethal to lung tumor cells at both early and late stages, we performed flow cytometry to analyze the cell cycle distribution in Tg-3m (**Figure 2A**) and Tg-6m (**Figure 2B**) cell lines treated with temsirolimus at different concentrations (2.5, 5.0, and 10 μM). Temsirolimus treatment increased the cell population in the G_0/G_1 phase in both Tg-3m and Tg-6m cell lines but did not cause significant cell death (**Figure 3**). Taken together, these results suggest that temsirolimus suppressed the proliferation of Tg-3m and Tg-6m cells through its cytostatic effect and not through cytotoxicity.

Efficacy of Temsirolimus, Cisplatin, and Gemcitabine in mTOR Pathway and Cytotoxicity

The efficacy of temsirolimus, cisplatin, and gemcitabine (each at 10 μM) alone and in combination was evaluated in Tg-3m (**Figure 3A**) and Tg-6m (**Figure 3B**) cells. To evaluate the effect of temsirolimus on activation regulation in the mTOR pathway, we examined the phosphorylation of mTOR (s2448) by using Western blot analysis. Gemcitabine or cisplatin treatment did not alter the phosphorylation of mTOR. However, treatment with temsirolimus alone markedly suppressed the activation of mTOR in Tg-6m than in Tg-3m cells. When cells were treated with temsirolimus combined with cisplatin and gemcitabine, the effect of mTOR suppression was evident. The apoptotic cell death in H1299 human NSCLC cell line was presented in **Supplementary Figure S3**.

To evaluate the cytotoxic effect of temsirolimus, we examined the total cell apoptotic rate by using annexin V staining. In human NSCLC cell line H1299 treated with temsirolimus alone caused about 25% cell death, when combined with cisplatin and gemcitabine showed enhanced cytotoxicity by approximately 10% in G + C and 15% in G + C + T (p = 0.02 and 0.003, respectively) (**Supplementary Figure S4**). Treatment with gemcitabine alone induces higher cytotoxicity in Tg-6m than in Tg-3m cells; however, treatment with cisplatin alone did not reveal any substantial difference. Treatment with gemcitabine plus cisplatin revealed similar apoptotic results in both cell lines. Although treatment with temsirolimus alone did not cause

TABLE 1 | List of drug repositioning candidates.

Name	K score	Original indication	Cancer indication/clinical trials+
		Drug repositioning from Tg-3m tumor cells	
Mesoridazine	−1	Antipsychotic	N/A
Dexamethasone	−1	Anti-inflammatory; steroids	Myeloma
Nilotinib	−1	Antineoplastic	Leukemia
Testosterone	−1	Anabolic	Prostate cancer
Temsirolimus	−0.99	Antineoplastic	Renal cell carcinoma
Prazosin	−0.98	Adrenergic	Prostate cancer
Pipamperone	−0.95	Antipsychotic	N/A
Rifabutin	−0.94	Antibiotic	Lung cancer
Omeprazole	−0.94	Anti-ulcer	Head and neck cancer
Cytarabine	−0.94	Antineoplastic	Leukemia
Timolol	−0.94	Adrenergic	N/A
Rofecoxib	−0.94	Analgesics	Colorectal cancer
Ibuprofen	−0.94	Analgesics	Lung cancer; prostate cancer
Ranitidine	−0.94	Anti-ulcer	Myeloma, renal cell carcinoma
		Drug repositioning from Tg-6m tumor cells	
Triamcinolone	−1	Anti-inflammatory; steroids	N/A
Flurbiprofen	−0.98	Analgesics	Prostate cancer
Rimonabant	−0.98	Antiobesity	Leukemia
Tamoxifen	−0.98	Anti-estrogen; antineoplastic	Breast cancer
Temsirolimus	−0.98	Antineoplastic	Renal cell carcinoma
Nicorandil	−0.98	Vasodialator	N/A

+*Information obtained from DrugBank (https://www.drugbank.ca/).*

cytotoxicity, it enhanced the cisplatin and gemcitabine-induced apoptosis in both cell lines significantly ($p < 0.05$; **Figures 3A,B**).

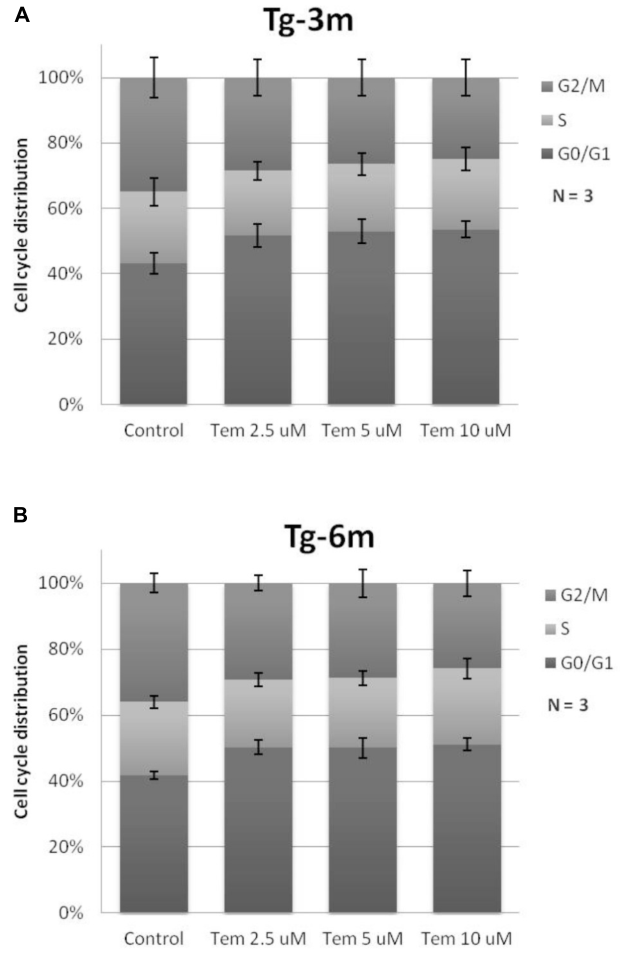


FIGURE 2 | Cytostatic effect caused by temsirolimus at different concentrations in both Tg-3m **(A)** and Tg-6m **(B)** lung tumor cell lines. Temsirolimus treatment resulted in the cell arrest at the G1 phase in a concentration-dependent manner. The representative data showed the results from three independent experiment.

Treatment Effects of Temsirolimus, Cisplatin, and Gemcitabine on Tumor Growth

To investigate the effect of temsirolimus, cisplatin, and gemcitabine on tumor growth, we used a therapeutic approach with a previously documented NSCLC mouse model (Chang et al., 2017). The mice were divided into three groups ($n = 5$): the control group (no treatment), the group that received a low dosage of cisplatin and gemcitabine (low-dose C + G), and the group that received temsirolimus combined with a low dosage of cisplatin and gemcitabine (mix T + C + G). The mice were treated at the age of 9 weeks for 8 weeks. Both treatments were administered weekly through the tail vein, and micro-CT imaging was performed to follow up tumor growths (**Figure 4A**). The imaging on week 15 was postponed because of regular maintenance of the scanner. On week 16, the mice were sacrificed and their lungs were removed for histopathological analysis.

The tumor growth rate was calculated by normalizing each tumor volume to the baseline tumor volume of each mouse at the beginning of week 9. The tumor growth was slightly reduced in the low C + G group, whereas it was markedly inhibited in the mix (T + C + G) group. In addition, the tumor growth significantly declined after 4 weeks' treatment in the mix (T + C + G) group with $p \leq 0.05$ (weeks 13–16). Smaller and reduced lung tumors were also noted in hematoxylin and eosin (H&E)-stained lung sections (**Figures 4B,C**). Collectively, the weekly administration of temsirolimus combined with low doses of cisplatin and gemcitabine effectively reduced the growth of lung tumors.

Treatment Effects on General Tumor Characteristics and the mTOR-Pathway

At the end of the experiment (week 16), all lungs were dissected and immunohistochemically analyzed to assess and quantify the microscopic effects of combined therapies with or without temsirolimus on general tumor characteristics (H&E

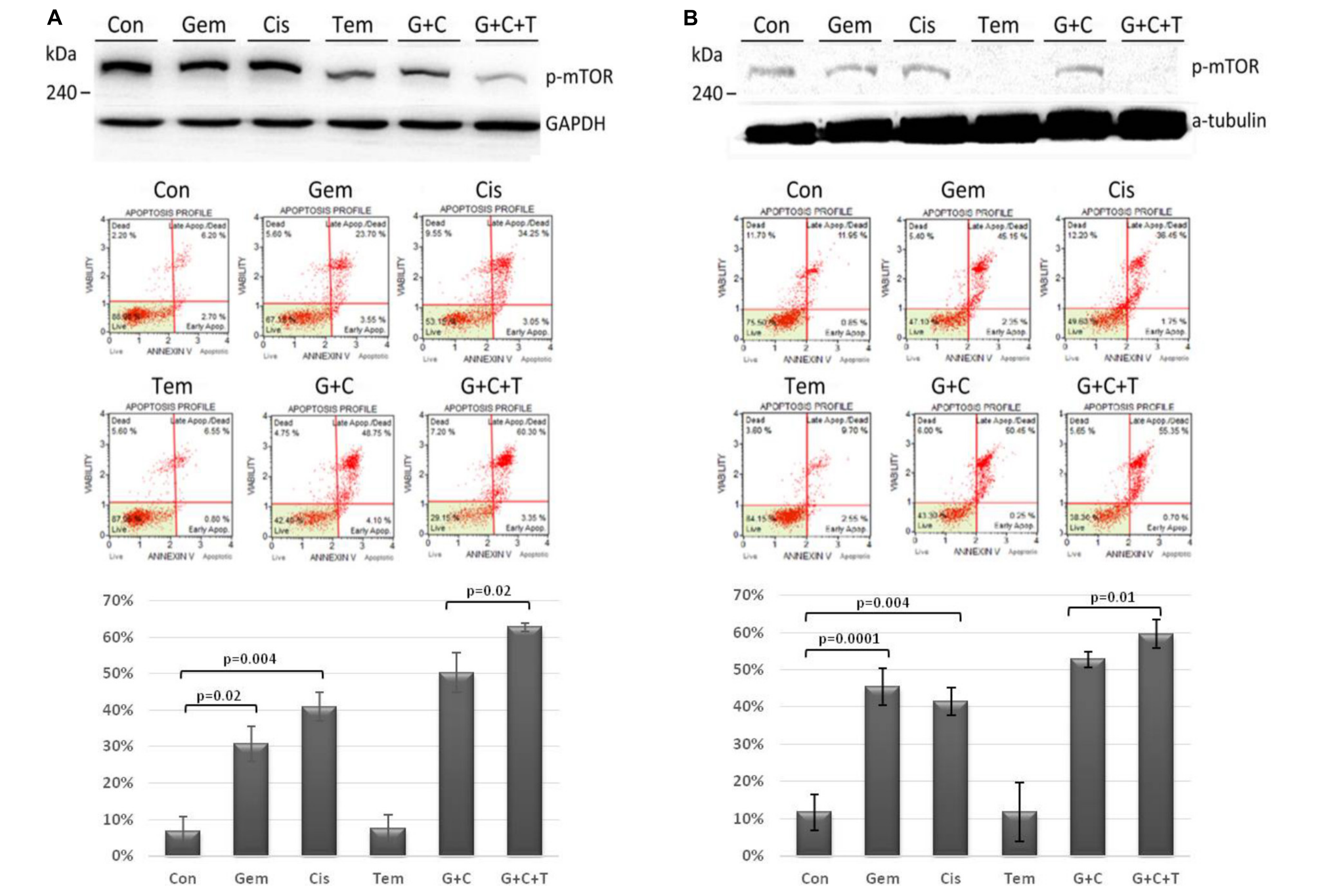


FIGURE 3 | Temsirolimus combined with cisplatin and gemcitabine induced significant apoptotic cell death through the inhibition of p-mTOR in both Tg-3m **(A)** and Tg-6m **(B)** lung tumor cell lines. Although temsirolimus treatment alone did not cause cell apoptosis, when combined with cisplatin and gemcitabine, it significantly enhanced the cytotoxicity by approximately 10% in Tg-3m and Tg-6m cells (*p* = 0.02 and 0.01, respectively), which was higher than that caused by the doublet of cisplatin and gemcitabine. Either gemcitabine or cisplatin alone also showed statistical significance from control. Con: control, Gem: gemcitabine, Cis: cisplatin, Tem: temsirolimus, G + C: gemcitabine + cisplatin, G + C + T: gemcitabine + cisplatin + temsirolimus.

stain; Ki-67 and CD34) and to identify possible mechanisms for the observed differences in growth inhibition. H&E staining revealed viable tumor mass within the lung parenchyma in untreated tumors, with immune cell infiltration. Residual tumor mass within the lung parenchyma with congestion, hyaline deposition, and immune cell infiltration were observed in low-dose C + G treated tumors. Scattered viable tumor cells with nuclear pleomorphism within the lung parenchyma revealed foamy macrophages and giant cells when treated with combined T + C + G after chemotherapy (magnification: 100×; **Figure 5**). Ki-67 staining revealed condensed signals of proliferating tumor cells in untreated control tumors. Treatment with low-dose C + G resulted in a lower fraction of proliferating cells, whereas that with combined T + C + G demonstrated diffused proliferating signals (magnification: 300×). CD34 staining demonstrated disruptive angiogenetic architectures in the low-dose and mix groups compared with the untreated control groups (magnification: 400×). In addition to general tumor characteristics, we investigated specific treatment effects on the mTOR pathway by evaluating p-mTOR and pS6RP in all lung tumors (magnification: 400×; **Figure 6**). The quantitative bar charts represent the positively stained areas of the whole image above, revealed that both treatments inhibited the tumor proliferation marker of Ki67. The combination treatment with temsirolimus markedly inhibited angiogenesis compared with low-dose chemotherapy. Quantitative stained areas demonstrated reduced p-mTOR signaling in both the treated groups, whereas the p-S6BP signal was higher. The statistical analysis of the intensity of positive signals from three selected views of each IHC-stained section demonstrated similar results (**Supplementary Figure S5**). Whether the p-S6BP signaling resulted from heterogeneous tumor cells remains to be investigated.

DISCUSSION

Chemotherapy is one of the most important treatment methods for advanced NSCLC, and cisplatin-based combinations are usually used as standard regimens. The combination of one

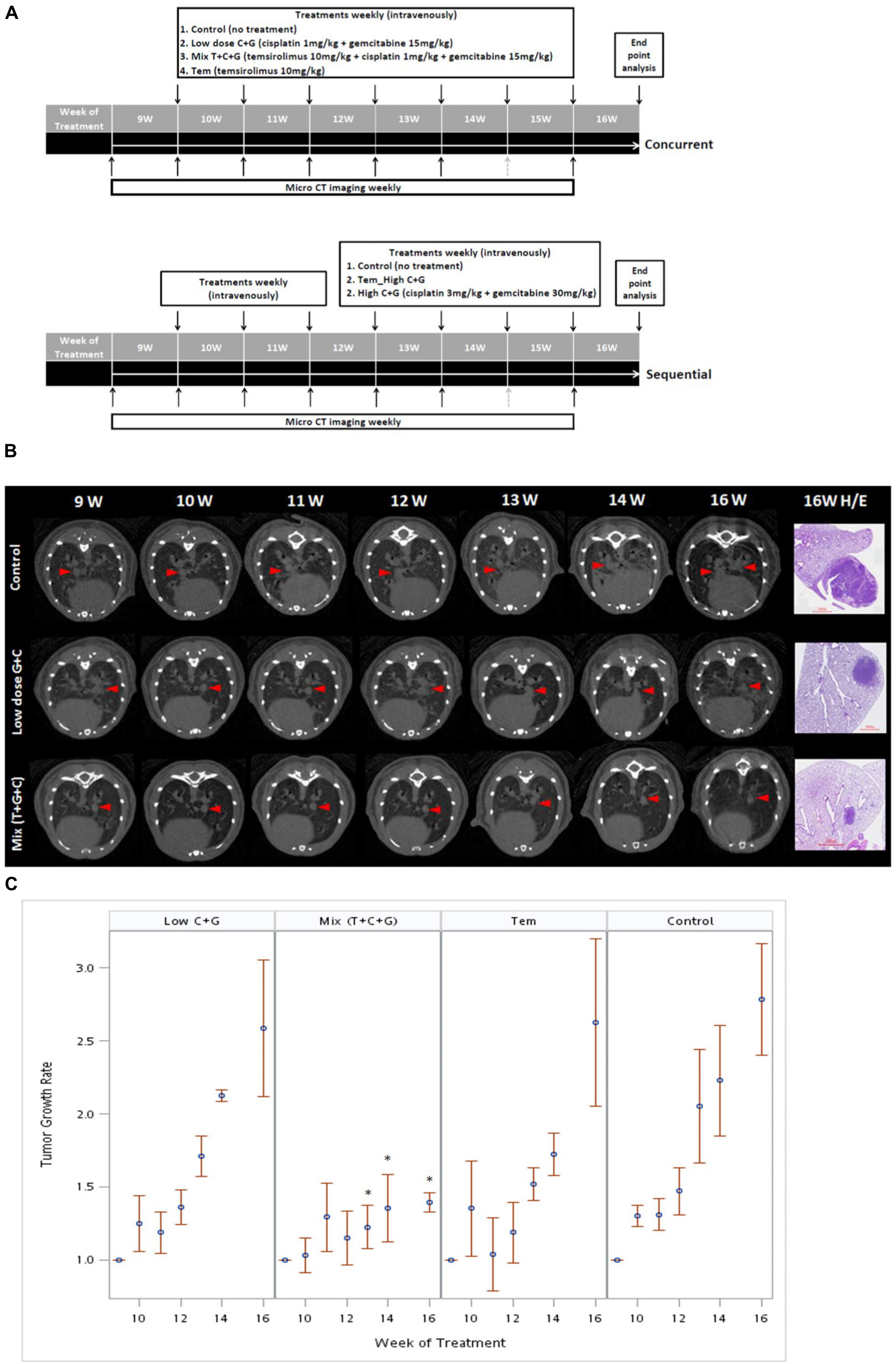


FIGURE 4 | Treatment sequences of temsirolimus alone (Tem), low-dose chemotherapy (Low C + G) and temsirolimus combined with low-dose chemotherapy (Mix T + C + G) in the lung tumor mouse model. Mice were treated at the age of 9 weeks for 7 weeks. The lung tumor growths were monitored using micro-CT every week, except for the 15th week because of regular maintenance of the scanner (gray-dotted arrow). Concurrent and sequential administration of treatments were depicted **(A)**. Red arrow heads indicate the monitored tumors compared with the corresponding H&E-stained histopathologic sections at the endpoint **(B)**. The endpoint H&E-stained sectioned sections were also displayed as inset in **Figure 5**. The tumor growth rate was calculated by normalizing each tumor volume to the baseline tumor volume of each mouse at the beginning of week 9. The effect of different treatments: Tem, Low C + G and Mix T + C + G in lung tumor growth inhibition in time periods were displayed **(C)**. The significance of tumor growth inhibition among each treatment was analyzed and it was found significant after week 13 in the Mix (T + C + G) group compared to the control group (*$p \leq 0.05$).

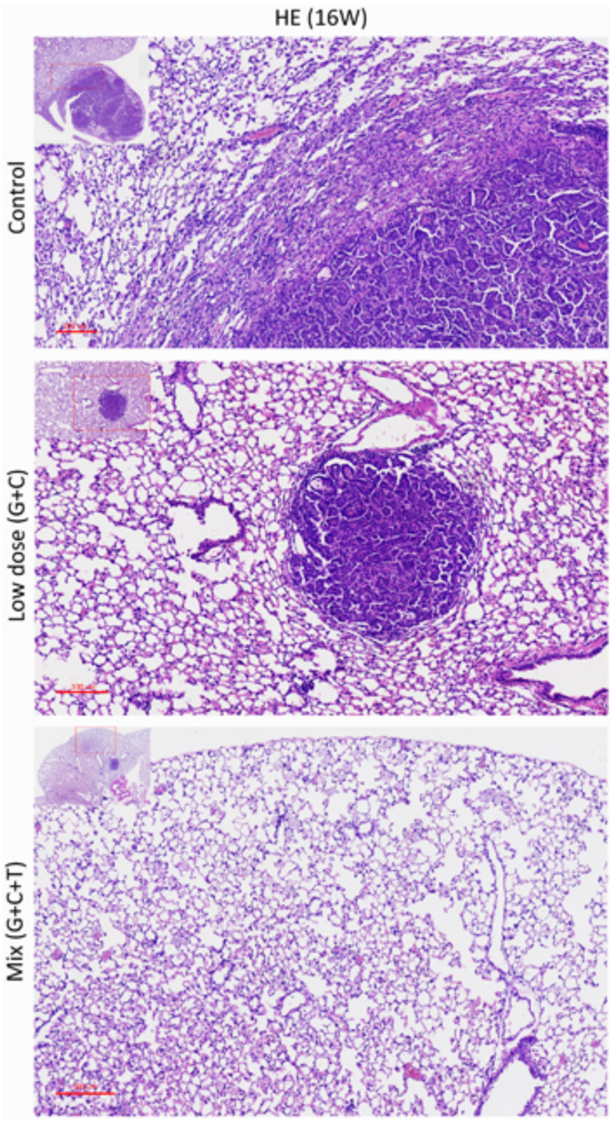


FIGURE 5 | Histopathologic views of treatment effects. Representative sections of the lung tissue of mice after various treatments. Viable tumor mass within the lung parenchyma with immune cell infiltration (control). Residual tumor mass within the lung parenchyma with congestion, hyaline deposition, and immune cell infiltration (low-dose C + G). Scattered tumor cells with nuclear pleomorphism within the lung parenchyma with congestion, foamy macrophages, and giant cells (mix T + C + G). H&E-stained slides of sections from mice with lung tumors were assessed by pathologists blinded to the treatment and outcome. Magnification: 100× and 40× (inset).

or more agents with a platinum compound resulted in high response rates and prolonged survival (Schiller et al., 2002; Ruiz-Ceja and Chirino, 2017). Gemcitabine was approved by FDA in 1996 with DNA synthesis inhibition. Gemcitabine is indicated in combination with cisplatin as the first-line treatment of patients with advanced NSCLC (Ruiz-Ceja and Chirino, 2017). Common cisplatin plus gemcitabine treatment-related adverse events are hematologic toxicity and gastrointestinal reaction. Hematologic toxicity mainly included decreased white blood cells and platelets. Gastrointestinal reactions mainly included nausea and vomiting (Ai et al., 2016). However, the high toxicity induced by cisplatin-based doublets urges research on alternative treatments. In this study, we used the LINCS L1000 database and a well-characterized lung adenocarcinoma mouse model to repurpose existing drugs for lung adenocarcinoma. By using this approach, we identified the mTOR inhibitor, temsirolimus, which has been approved by the FDA for renal cell carcinoma, as a potential therapeutic agent. In our results, both temsirolimus-treated early (Tg-3m) and late-stage (Tg-6m) lung tumor cell lines demonstrated cell cycle arrest at the G0/G1 phase. The treatment with temsirolimus alone markedly suppressed mTOR activation in Tg-6m than in Tg-3m cells. When temsirolimus was combined with cisplatin and gemcitabine, the effect of mTOR suppression was evident. Additionally, temsirolimus combined with gemcitabine and cisplatin not only suppressed the phosphorylation of mTOR but also significantly improved cell death in Tg-3m and Tg-6m cell lines compared with gemcitabine plus cisplatin.

As reported by Khuri colleagues (Li et al., 2014), mTOR inhibition triggers rapid and sustained activation of the PI3K/Akt survival pathway in the human lung and other types of cancer cells; therefore, the combination of mTOR-targeted therapy with drugs that block PI3K/Akt activation might also be reasonable. In a reported phase II study, temsirolimus was administered as a single agent in 52 patients with untreated NSCLC on a weekly basis. The clinical benefit rate was 35%, with a confirmed partial response of 8% and stable disease of 27%. Although these results did not satisfy the protocol-defined criteria for success, they evidenced the clinical activity of temsirolimus as a single agent in NSCLC (Reungwetwattana et al., 2012). In a phase I study, temsirolimus was combined with weekly thoracic radiation, which proved the tolerance (Waqar et al., 2014). Because temsirolimus has demonstrated considerable activity in clinical studies, we hypothesize that it works synergistically with the first-line NSCLC chemotherapy cisplatin plus gemcitabine.

In animal studies, optimizing the cytostatic agent temsirolimus with cycle-active chemotherapy is important for maximizing the clinical benefit. Therefore, we designed concurrent and sequential administration of temsirolimus with either low or high doses of chemotherapy in a mouse lung adenocarcinoma model. In the concurrent schedule, administration of low-dose chemotherapy and temsirolimus (T + C + G) demonstrated greater inhibition of tumor growth compared with low-dose chemotherapy alone (C + G) in the mouse model. In the sequential schedule in which temsirolimus alone was administrated weekly for 3 weeks prior to the administration of high-dose chemotherapy (3 mg/kg cisplatin + 30 mg/kg gemcitabine) in the following weeks, the effect of tumor growth inhibition was less significant (**Supplementary Figures S1**, **S2**). Collectively, our study revealed that concurrent administration of low-dose chemotherapy and temsirolimus is more effective in suppressing lung tumor growth, which may be advantageous to reduce the cytotoxicity caused by standard chemotherapy. The histopathologic evaluation of endpoint H&E-stained lung tumor sections revealed that

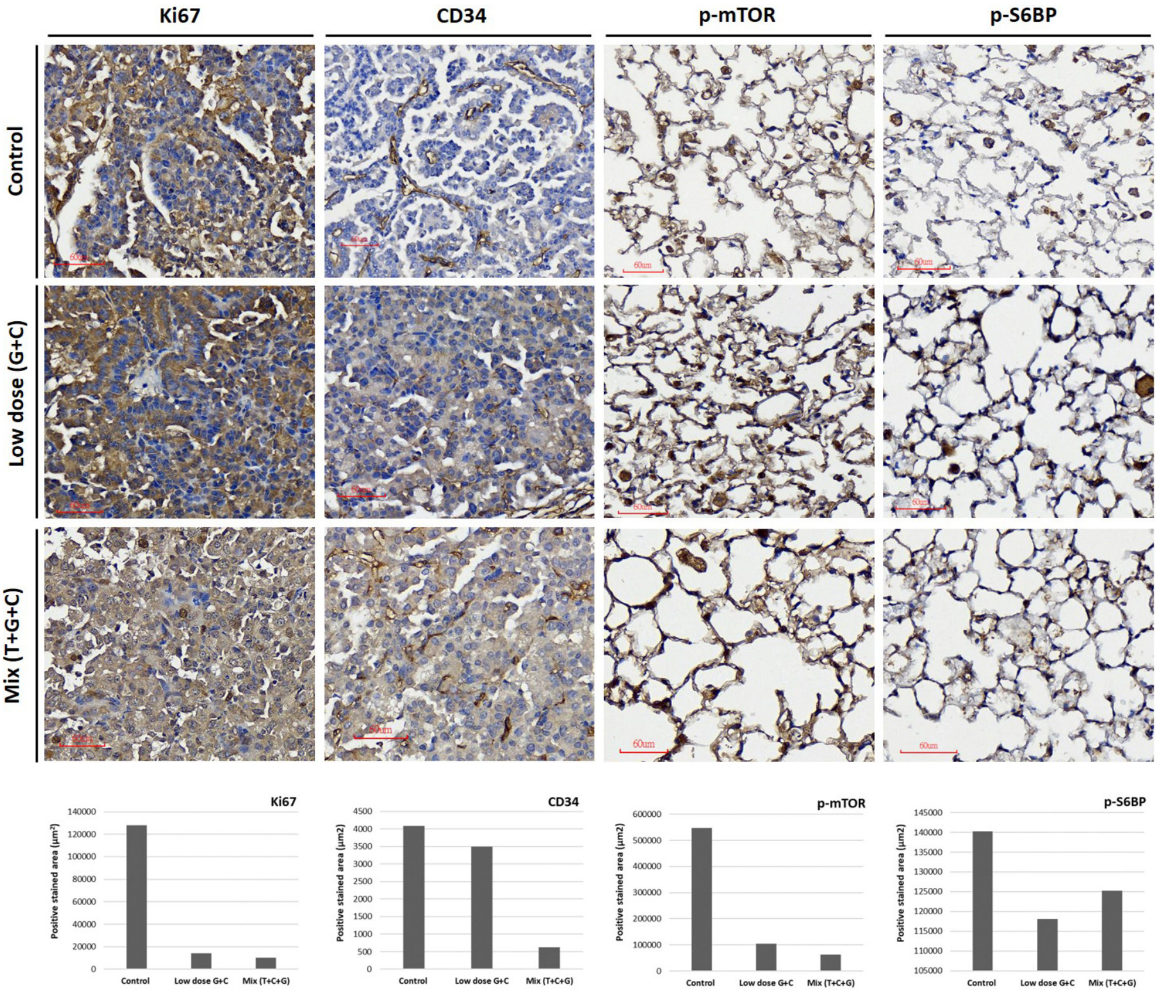


FIGURE 6 | Treatment effects on tumor proliferation, angiogenesis, and the mTOR-pathway. Representative images of Ki67, CD34, p-mTOR, and p-S6RP expression were immunostained using specific antibodies as indicated. Tumor proliferative signal of Ki67 was more condensed in untreated tumors and more diffuse in low-dose and mixed treatment groups. The angiogenetic architecture was more intact in untreated tumors compared with treated groups, as analyzed using CD34 staining. The inhibition of p-mTOR expression was higher in the mixed treatment groups. The phosphorylation of S6RP was also examined as a downstream target of the mTOR-pathway. The p-S6RP expression was reduced after both treatments. Image magnification: 300× in Ki67 and 400× in CD34, p-mTOR, and p-S6RP. Quantitative analysis of Ki67, CD34, p-mTOR, and p-S6RP expression in IHC-stained section were analyzed and present. The whole positive stained areas (µm^2) of each representative image were measured using ImageJ software and visualized as bar chart below.

the tumors were associated with an extensive response to the T + C + G treatment compared with low-dose C + G treatment. Common tumorigenic and angiogenetic markers (Ki67 and CD34) were apparently inhibited after the T + C + G treatment compared with low-dose C + G treatment. These results proved the tumor inhibition efficacy of temsirolimus combined with low-dose chemotherapy. The mTOR phosphorylation inhibition was higher in the mixed treatment. Moreover, the phosphorylation of the ribosomal protein S6 (p-S6RP), one of the targets downstream of the mTOR pathway, was reduced after both treatments. The examination of phosphorylated mTOR and S6RP suggested their sensitivity to temsirolimus.

The clinical benefits of chemotherapy are limited by drug resistance and systemic toxicity. Temsirolimus was reported to restore cisplatin sensitivity in lung cancer cell lines by blocking the translation of proteins that are involved in cisplatin resistance (Blois et al., 2011). The cytostatic effect of temsirolimus was also demonstrated by introducing temsirolimus as a molecular-targeted agent with the potential for inhibiting tumor cell repopulation (Fung et al., 2009). However, the pulmonary toxicity was associated with mTOR inhibitors as many other drugs, including anticancer agents (Blois et al., 2011; Li et al., 2014). Proper chemotherapeutic strategy management and clinical pulmonary symptom diagnosis should be taken account when administration with mTOR inhibitors. Our study demonstrated that a combination of low-dose chemotherapy and temsirolimus treatment was more effective in inhibiting tumor growth than a doublet chemotherapy regimen in the mouse lung tumor model. In addition, the concurrent administration of the combined treatment was more efficacious than the sequential administration of these agents at a higher dose. Our study results suggest that the combination of low-dose chemotherapy and temsirolimus treatment might be beneficial in the treatment of lung adenocarcinoma, which warrants further investigation.

AUTHOR CONTRIBUTIONS

H-WC and VC conceived the experiments. M-JW and Z-ML conducted the experiments. C-YW conducted the micro-CT imaging. S-YC conducted the tissue embedding and histopathology. H-WC, H-JC, and VC analyzed the results. Y-KL assisted on statistical analysis. H-WC wrote up the manuscript. Y-HC and VC provided comments on the manuscript. All authors reviewed the manuscript.

ACKNOWLEDGMENTS

This work was partially supported by the foundation of the Ph.D. program for Translational Medicine, College of Medical Science and Technology from Taipei Medical University.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphar.2018.00778/full#supplementary-material

FIGURE S1 | Effects of high-dose chemotherapy and the sequential administration of temsirolimus followed by high-dose chemotherapy in lung tumor growth inhibition. The tumor growth inhibition was significant in the first 3 weeks of treatment with temsirolimus. However, the tumor growth inhibition efficacies between the two chemotherapy regimens were similar **(A)**. Effect of various treatments on tumor growth rate **(B)**. The tumor growth inhibition was most significant after the mixed treatment (T + C + G) with a lower dose of chemotherapy, which is beneficial in reducing the cytotoxic effect. ($^{*}p \leq 0.05$; $^{**}p \leq 0.01$).

FIGURE S2 | Effect of temsirolimus alone (orange line) compared to different regimes of chemotherapy in lung tumor growth inhibition. The tumor growth inhibition of different treatments was depicted with statistical significance. While the tumor growth was inhibited moderately using temsirolimus alone (orange line), combined temsirolimus with low-dose chemotherapy (blue line, *p*-value = 0.019) inhibited the tumor growth significantly. Combined temsirolimus with low-dose chemotherapy treatment also showed significance over the treatment using temsirolimus alone or high-dose chemotherapy (black line), with *p*-value = 0.037 and 0.032, respectively.

FIGURE S3 | Temsirolimus combined with cisplatin and gemcitabine induced apoptotic cell death in H1299 human NSCLC cell line. Although temsirolimus treatment alone caused about 25% cell death, when combined with cisplatin and gemcitabine, it significantly enhanced the cytotoxicity by approximately 10% in G + C and 15% in G + C + T (*p*-value = 0.02 and 0.003, respectively). Con: control, Gem: gemcitabine, Cis: cisplatin, Tem: temsirolimus, G + C: gemcitabine+cisplatin, G + C + T: gemcitabine + cisplatin + temsirolimus.

FIGURE S4 | SRB cytotoxicity assay performed in Tg-3m **(A)** and Tg-6m **(B)** cell lines. The results showed either temsirolimus alone or combined with cisplatin and gemcitabine displayed significant cytotoxic effects in both Tg-3m and Tg-6m cell lines. The results of SRB assay demonstrated the *in vitro* correlation of cytotoxicity and cell death caused by chemotherapeutic agents tested.

FIGURE S5 | Statistical quantification of IHC-stained sections of Ki67, CD37, p-mTOR, and p-S6RP. The intensity of positive signals from each IHC-stained section were selected from three different views (red-lined squares) and analyzed using Image J software with IHC toolbox as plugins.

REFERENCES

Ai, D., Guan, Y., Liu, X.-J., Zhang, C.-F., Wang, P., Liang, H.-L., et al. (2016). Clinical comparative investigation of efficacy and toxicity of cisplatin plus gemcitabine or plus Abraxane as first-line chemotherapy for stage III/IV non-small-cell lung cancer. *Onco Targets Ther.* 9, 5693–5698. doi: 10.2147/OTT.S109683

Ashburn, T. T., and Thor, K. B. (2004). Drug repositioning: identifying and developing new uses for existing drugs. *Nat. Rev. Drug Discov.* 3, 673–683. doi: 10.1038/nrd1468

Blois, J., Smith, A., and Josephson, L. (2011). The slow cell death response when screening chemotherapeutic agents. *Cancer Chemother. Pharmacol.* 68, 795–803. doi: 10.1007/s00280-010-1549-9

Chang, H. W., Lin, Z. M., Wu, M. J., Wang, L. Y., Chow, Y. H., Jiang, S. S., et al. (2017). Characterization of a transgenic mouse model exhibiting spontaneous lung adenocarcinomas with a metastatic phenotype. *PLoS One* 12:e0175586. doi: 10.1371/journal.pone.0175586

Coordinators, N. R. (2016). Database resources of the national center for biotechnology information. *Nucleic Acids Res.* 44, D7–D19. doi: 10.1093/nar/gkv1290

Endo, H., Yano, M., Okumura, Y., and Kido, H. (2014). Ibuprofen enhances the anticancer activity of cisplatin in lung cancer cells by inhibiting the heat shock protein 70. *Cell Death Dis.* 5:e1027. doi: 10.1038/cddis.2013.550

Fung, A. S., Wu, L., and Tannock, I. F. (2009). Concurrent and sequential administration of chemotherapy and the Mammalian target of rapamycin inhibitor temsirolimus in human cancer cells and xenografts. *Clin. Cancer Res.* 15, 5389–5395. doi: 10.1158/1078-0432.CCR-08-3007

Gabriel, D., Gordon, L. B., and Djabali, K. (2016). Temsirolimus partially rescues the hutchinson-gilford progeria cellular phenotype. *PLoS One* 11:e0168988. doi: 10.1371/journal.pone.0168988

Gallotta, D., Nigro, P., Cotugno, R., Gazzerro, P., Bifulco, M., and Belisario, M. A. (2010). Rimonabant-induced apoptosis in leukemia cell lines: activation of caspase-dependent and -independent pathways. *Biochem. Pharmacol.* 80, 370–380. doi: 10.1016/j.bcp.2010.04.023

Huang da, W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4, 44–57. doi: 10.1038/nprot.2008.211

Hurle, M. R., Yang, L., Xie, Q., Rajpal, D. K., Sanseau, P., and Agarwal, P. (2013). Computational drug repositioning: from data to therapeutics. *Clin. Pharmacol. Ther.* 93, 335–341. doi: 10.1038/clpt.2013.1

Kerr, D. J., Dunn, J. A., Langman, M. J., Smith, J. L., Midgley, R. S. J., Stanley, A., et al. (2007). Rofecoxib and cardiovascular adverse events in adjuvant treatment of colorectal cancer. *N. Engl. J. Med.* 357, 360–369. doi: 10.1056/NEJMoa071841

Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., et al. (2016). PubChem substance and compound databases. *Nucleic Acids Res.* 44, D1202–D1213. doi: 10.1093/nar/gkv951

Li, J., Huang, Y., Gao, Y., Wu, H., Dong, W., and Liu, L. (2016). Antibiotic drug rifabutin is effective against lung cancer cells by targeting the eIF4E-β-catenin axis. *Biochem. Biophys. Res. Commun.* 472, 299–305. doi: 10.1016/j.bbrc.2016.02.120

Li, S., Kong, Y., Si, L., Chi, Z., Cui, C., Sheng, X., et al. (2014). Phosphorylation of mTOR and S6RP predicts the efficacy of everolimus in patients with metastatic renal cell carcinoma. *BMC Cancer* 14:376. doi: 10.1186/1471-2407-14-376

Lin, S.-C., Chueh, S.-C., Hsiao, C.-J., Li, T.-K., Chen, T.-H., Liao, C.-H., et al. (2007). Prazosin displays anticancer activity against human prostate cancers: targeting DNA and cell cycle. *Neoplasia* 9, 830–839. doi: 10.1593/neo.07475

Molina, J. R., Yang, P., Cassivi, S. D., Schild, S. E., and Adjei, A. A. (2008). Non–small cell lung cancer: epidemiology, risk factors, treatment, and survivorship. *Mayo Clin. Proc.* 83, 584–594. doi: 10.4065/83.5.584

Ohara, T., Takaoka, M., Toyooka, S., Tomono, Y., Nishikawa, T., Shirakawa, Y., et al. (2011). Inhibition of mTOR by temsirolimus contributes to prolonged survival of mice with pleural dissemination of non-small-cell lung cancer cells. *Cancer Sci.* 102, 1344–1349. doi: 10.1111/j.1349-7006.2011.01967.x

Reungwetwattana, T., Molina, J. R., Mandrekar, S. J., Allen-Ziegler, K., Rowland, K. M., Reuter, N. F., et al. (2012). Brief report: a phase II "window-of-opportunity" frontline study of the MTOR inhibitor, temsirolimus given as a single agent in patients with advanced NSCLC, an NCCTG study. *J. Thorac. Oncol.* 7, 919–922. doi: 10.1097/JTO.0b013e31824de0d6

Ruiz-Ceja, K. A., and Chirino, Y. I. (2017). Current FDA-approved treatments for non-small cell lung cancer and potential biomarkers for its detection. *Biomed. Pharmacother.* 90, 24–37. doi: 10.1016/j.biopha.2017.03.018

Schiller, J. H., Harrington, D., Belani, C. P., Langer, C., Sandler, A., Krook, J., et al. (2002). Comparison of four chemotherapy regimens for advanced non-small-cell lung cancer. *N. Engl. J. Med.* 346, 92–98. doi: 10.1056/NEJMoa011954

Skehan, P., Storeng, R., Scudiero, D., Monks, A., Mcmahon, J., Vistica, D., et al. (1990). New colorimetric cytotoxicity assay for anticancer-drug screening. *J. Natl. Cancer Inst.* 82, 1107–1112. doi: 10.1093/jnci/82.13.1107

Teng, X.-D. (2005). [World Health Organization classification of tumours, pathology and genetics of tumours of the lung]. *Chin. J. Pathol.* 34, 544–546.

Ushijima, H., Suzuki, Y., Oike, T., Komachi, M., Yoshimoto, Y., Ando, K., et al. (2015). Radio-sensitization effect of an mTOR inhibitor, temsirolimus, on lung adenocarcinoma A549 cells under normoxic and hypoxic conditions. *J. Radiat. Res.* 56, 663–668. doi: 10.1093/jrr/rrv021

Vempati, U. D., Chung, C., Mader, C., Koleti, A., Datar, N., and Vidović, D. (2014). Metadata Standard and Data Exchange Specifications to Describe, Model, and Integrate Complex and Diverse High-Throughput Screening Data from the Library of Integrated Network-based Cellular Signatures (LINCS). *J. Biomol. Screen* 19, 803–816. doi: 10.1177/1087057114522514

Vichai, V., and Kirtikara, K. (2006). Sulforhodamine B colorimetric assay for cytotoxicity screening. *Nat. Protoc.* 1, 1112–1116. doi: 10.1038/nprot.2006.179

Waqar, S. N., Robinson, C., Bradley, J., Goodgame, B., Rooney, M., Williams, K., et al. (2014). A phase I study of temsirolimus and thoracic radiation in non–small-cell lung cancer. *Clin. Lung Cancer* 15, 119–123. doi: 10.1016/j.cllc.2013.11.007

Wishart, D. S., Knox, C., Guo, A. C., Shrivastava, S., Hassanali, M., Stothard, P., et al. (2006). DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* 34, D668–D672. doi: 10.1093/nar/gkj067

Wynne, S., and Djakiew, D. (2010). NSAID inhibition of prostate cancer cell migration is mediated by Nag-1 Induction via the p38 MAPK-p75(NTR) pathway. *Mol. Cancer Res.* 8, 1656–1664. doi: 10.1158/1541-7786.MCR-10-0342

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Pharmacology

ORIGINAL RESEARCH
published: 15 August 2018
doi: 10.3389/fphar.2018.00918

Exploring the Mechanism of Flavonoids Through Systematic Bioinformatics Analysis

Tianyi Qiu[1]†, Dingfeng Wu[2]†, LinLin Yang[3], Hao Ye[4,5], Qiming Wang[2], Zhiwei Cao[2] and Kailin Tang[2]**

[1] Institute of Biomedical Sciences, Fudan University, Shanghai, China, [2] School of Life Sciences and Technology, Tongji University, Shanghai, China, [3] Hebei Key Laboratory of Metabolic Diseases and Clinical Medicine Research Center, Hebei General Hospital, Hebei, China, [4] Sinotech Genomics Ltd., Shanghai, China, [5] East China University of Science and Technology, Shanghai, China

Flavonoids are the largest class of plant polyphenols, with common structure of diphenylpropanes, consisting of two aromatic rings linked through three carbons and are abundant in both daily diets and medicinal plants. Fueled by the recognition of consuming flavonoids to get better health, researchers became interested in deciphering how flavonoids alter the functions of human body. Here, systematic studies were performed on 679 flavonoid compounds and 481 corresponding targets through bioinformatics analysis. Multiple human diseases related pathways including cancers, neuro-disease, diabetes, and infectious diseases were significantly regulated by flavonoids. Specific functions of each flavonoid subclass were further analyzed in both target and pathway level. Flavones and isoflavones were significantly enriched in multi-cancer related pathways, flavan-3-ols were found focusing on cellular processing and lymphocyte regulation, flavones preferred to act on cardiovascular related activities and isoflavones were closely related with cell multisystem disorders. Relationship between chemical constitution fragment and biological effects indicated that different side chain could significantly affect the biological functions of flavonoids subclasses. Results will highlight the common and preference functions of flavonoids and their subclasses, which concerning their pharmacological and biological properties.

Keywords: flavonoids, mechanism of action, pathway analysis, protein–protein interaction network, structure activity relationship

OPEN ACCESS

Edited by:
Zhi-Liang Ji,
Xiamen University, China

Reviewed by:
Shi-Bing Su,
Shanghai University of Traditional Chinese Medicine, China
Haifeng Chen,
Shanghai Jiao Tong University, China

****Correspondence:***
Zhiwei Cao
zwcao@tongji.edu.cn
Kailin Tang
kltang@tongji.edu.cn

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Translational Pharmacology, a section of the journal Frontiers in Pharmacology

Received: *07 February 2018*
Accepted: *26 July 2018*
Published: *15 August 2018*

Citation:
Qiu T, Wu D, Yang L, Ye H, Wang Q, Cao Z and Tang K (2018) Exploring the Mechanism of Flavonoids Through Systematic Bioinformatics Analysis. Front. Pharmacol. 9:918. doi: 10.3389/fphar.2018.00918

INTRODUCTION

Flavonoids are a family of phenolic substances sharing the same backbone structure of 2-pheny1-1,4-benzopyronemay, which are very abundant in nature, being accumulated in regular human diets including flowers (Zhang and Ma, 2018), fruits (Chang et al., 2018), vegetables, tea, wine (Matveeva et al., 2018), and so on (Szmitko and Verma, 2005). With the basic core scaffold, flavonoids have been demonstrated to exhibit relevant biological properties involving strong activity for anti-oxidant (Pietta, 2000), anti-allergy (Kawai et al., 2007; Castell et al., 2014), anti-inflammatory (Nijveldt et al., 2001; Serafini et al., 2010; Matias et al., 2014), anti-microbial (Cushnie and Lamb, 2005), and anti-obesity (Hughes et al., 2008) effects. Also, flavonoids have been reported to have effect on reducing the risk of cardiovascular disease (Hooper et al., 2008; Mulvihill and Huff, 2010; Feliciano et al., 2015) and cancers (Yao et al., 2011; Batra and Sharma, 2013), ameliorating cognition (Spencer et al., 2009; Williams and Spencer, 2012) and neuro-protection in Alzheimer's disease (Bakhtiari et al., 2017; Mohebali et al., 2018). Moreover, it is also found that flavonoids act as agonist or antagonist depending on the estrogen concentrations to regulate estrogenic-like activity (Breinholt et al., 1999; Hwang et al., 2006).

On the basis of common core scaffold, various combinations of substituent chemical groups on different positions may lead to structure diversity of flavonoids. This diversity can be further increased with possible variations of different functional groups, such as hydroxyl, methoxyl, carbonyl, and olefinic groups (Gontijo et al., 2017). According to the structure variations, flavonoids can be generally assigned into six main subclasses: flavones, flavonols, flavanones, flavanols, flavan-3-ols, and isoflavones (Ross and Kasum, 2002), for which the chemical properties depend on their structural classes, degrees of hydroxylation, substitutions, conjugation, and degree of polymerization (Kumar and Pandey, 2013). However, the functional similarities and differences, as well as the structure basis of different functions for flavonoids subclasses are not fully revealed yet.

In this study, a comprehensive bioinformatics analysis was performed based on a large-scale dataset including 679 flavonoids and 481 corresponding targets to decipher the mechanism of action (MOA) of flavonoids with a new perspective. Results illustrated the structure activity relationship of different flavonoids subclasses, which hint the protective roles of flavonoids subclasses in different human diseases. With the accumulation of flavonoids and corresponding targets, it is possible to comprehensively investigate the MOA of flavonoids in a systematic level and interpret the therapeutic mechanism to guide the drug discovery from natural flavonoid products.

MATERIALS AND METHODS

Dataset

Flavonoids and Corresponding Targets

A total number of 5,006 chemical structures of natural plant products were derived from Natural Product Activity and Species Source Database (NPASS) (Zeng et al., 2018). Among them, main types of flavonoids including flavones, flavonols, flavanones, flavanonol, isoflavones, and flavan-3-ols were categorized according to the scaffold structures derived by cheminformatics software-RDKit (Landrum, 2010), which were illustrated in **Figure 1A**. Further, corresponding direct targets of flavonoids were selected from 5,337 targets of natural plant products in NPASS. After that, 679 flavonoids and 481 corresponding targets were selected and listed in **Supplementary Table 1**. Number of targets for different flavonoid subclasses were illustrated in **Figure 1B**.

Enrichment Analysis of Flavonoids' Targets

Diversity Analysis of Natural Flavonoid Products' Targets

Targets of natural flavonoid products were mapped into Kyoto Encyclopedic of Genes and Genomes (KEGGs) (Kanehisa et al., 2012) and Gene Ontology (GO) (Ashburner et al., 2000) through Metascape (Tripathi et al., 2015) to analyze their enrichment pathways. Then, the enrichment pathways were generated for six flavonoid subclasses.

Specific Pathway Enrichment Analysis of Natural Flavonoids Products

To distinguish the specific pathway of flavonoids from other natural plant products, permutation test was implemented 1,000 times to identify the specific pathway of flavonoids' targets by setting the 4,327 other natural plant products as background.

Pharmacology Network Analysis

Protein–protein interaction (PPI) networks of flavonoids' targets were generated and modularized through Metascape (Tripathi et al., 2015). Further, the bio-functional similarity and difference between networks of six subclasses were compared based on the main functional modules. Then, PPI enrichment analysis was carried out with the following databases including BioGrid (Chatr-Aryamontri et al., 2017), InWeb_IM (Li et al., 2017), and OmniPath (Turei et al., 2016). The densely connected network components was identified by Molecular Complex Detection (MCODE) algorithm (Bader and Hogue, 2003) and viewed by Cytoscape (Shannon et al., 2003).

Structure–Activity Relationship Analysis

In order to analyze the structure–activity relationship, basic physicochemical properties including molecular mass (weight), lipid water distribution coefficient (LogP), hydrogen bond receptor (NumHAcceptors), hydrogen bond donor (NumHDonors), rotatable bond (NumRotatableBonds), topological molecular polarity surface area (TPSA) and Lipinski's Rule of five were calculated for different natural flavonoid products through RDKit (Landrum, 2010).

Also, the core scaffold and side chains of each natural flavonoid products were derived according to their chemical structures. Since flavones, flavonols, flavanones, flavanonol, and flavan-3-ols share the same core scaffold, the structure–activity relationships of above five subclasses were analyzed. Then, according to GO (Ashburner et al., 2000), the bio-functional annotation of each structure segment can be obtained. Further, to identify the association between chemical structure of flavonoid subclasses and biological function, structure–activity relationship was further analyzed through Apriori algorithm (Agrawal and Srikant, 1994). Here, the minimum support parameter was set as 0.01 and the minimum confidence was set as 0.5 for calculation.

RESULTS

Pathway Enrichment Analysis of Flavonoids' Targets

The biological function of flavonoids' target was deciphered through pathway enrichment analysis based on the background pathway dataset (**Figure 2** and **Supplementary Table 2**). Results showed that, the targets of flavonoids were enriched in multiple essential pathways including metabolism, genetic information

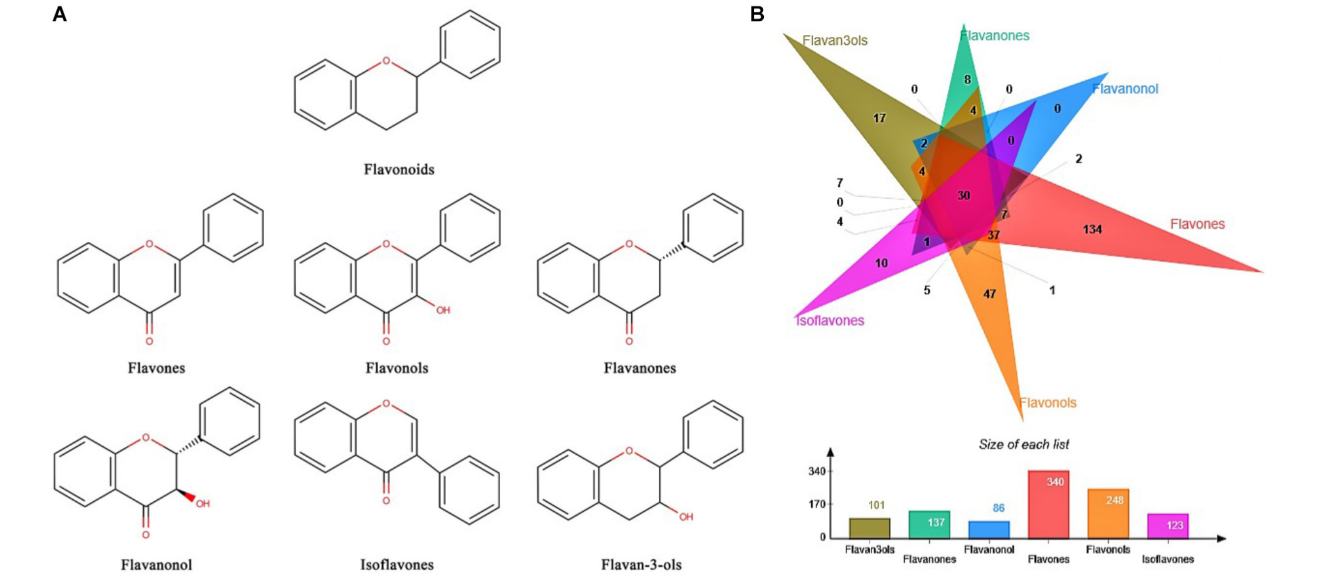


FIGURE 1 | Structures and targets information of flavonoids. **(A)** Core scaffold structures of six flavonoid subclasses. **(B)** Target number of different flavonoid subclasses.

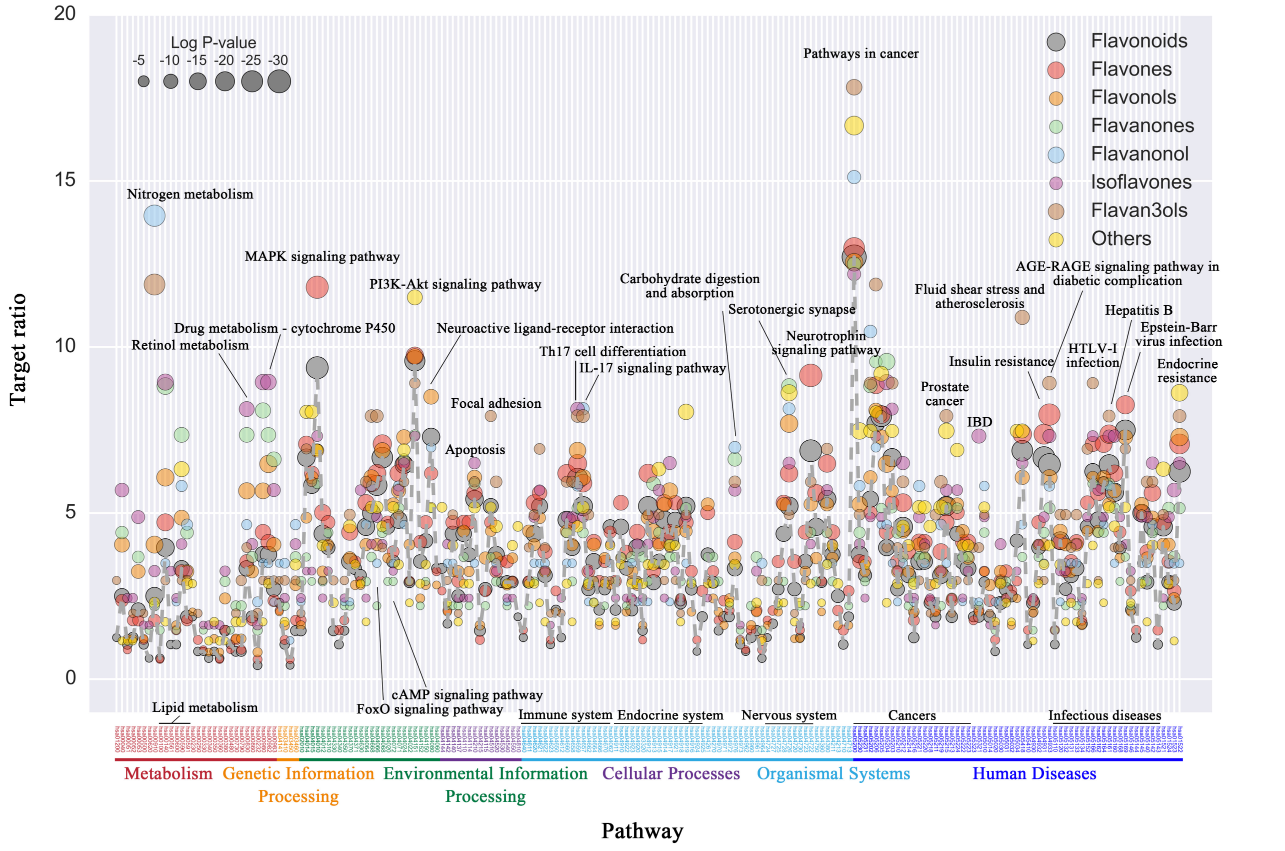


FIGURE 2 | KEGG pathway enrichment analysis of natural flavonoid products' targets. Here, *X*-axis represents the enriched pathways (p-value < 0.05), which were categorized according to KEGG classification. *Y*-axis represents target proportion of flavonoids in each pathway (number of flavonoids' target in pathway/total number of flavonoids' targets), the size of each nodes represents the significance of enrichment level (–LogP). Flavonoids and all six subclasses were marked in different colors.

processing, environmental information processing, cellular process, organismal systems, and multiple pathways which were related to human diseases such as infectious diseases and cancer. For instance, in environmental information processing, flavonoids were enriched in multiple cell signaling pathways including MAPK signaling pathway, PI3K-Akt signaling pathway, FoxO signaling pathway and cAMP signaling pathway. In cellular processes, flavonoids can significantly regulate pathways such as apoptosis, focal adhesion, cell cycle, and autophagy. Further, it can be found that flavonoids' targets were significantly enriched in several organismal systems including immune system, endocrine system and nervous system. Especially for immune system related pathways, flavonoids were enriched in Th17 cell differentiations, IL-17 signaling pathway, Toll-like, and NOD-like signaling pathways. Besides, multiple flavonoids' targets can be found in the endocrine system pathways, such as progesterone-medicated oocyte maturation, GnRH signaling, oxytocin signaling and thyroid hormone signaling pathways. Also, nervous system-related pathways such as serotonergic synapse, and neurotrophin signaling pathways were enriched by corresponding targets. Moreover, flavonoids' targets existed in pathways of essential human diseases such as multi-cancer, insulin resistance and infectious diseases including HTLV-1 infection, Epstein–Barr virus infection and Hepatitis B.

Besides the common enrichment pathways, different flavonoid subclasses illustrated different preference. For instance, targets of flavanonol and flavan-3-ols were more significantly enriched in nitrogen metabolism pathways than other subclasses. Targets of isoflavones, flavanones, and flavonols were enriched in metabolism pathways such as lipid, retinol, and drug metabolism pathway. Flavones' targets were significantly enriched in MAPK signaling pathway and neurotrophin signaling pathway, which means natural flavone products may have therapeutic effects on neurological-related diseases. Pervious researches indicated that flavones such as apigenin and luteolin could activate Nrf2-antioxidant response element (ARE)-mediated gene expression and induce anti-inflammatory activities through the PI3K and MAPK signaling pathways (Paredes-Gonzalez et al., 2015). Also, both compounds could significantly increase the endogenous mRNA and protein level of Nrf2 and Nrf2 targeting genes with important effects on hemo oxygenase-1 (HO-1) expression, thus, led to cytoprotective effects and neurite outgrowth (Lin et al., 2010; Zhao et al., 2013; Zhang et al., 2015). In addition, corresponding targets of flavan-3-ols and flavanonol were enriched in cancer-related pathways. Natural flavan-3-ol products such as (-)-epigallocatechin gallate (EGCG), (-)-epicatechin gallate (ECG), (-)-epigallocatechin (EGC), and (-)-epicatechin (EC) were discovered flavan-3-ols from green tea, which could provide possible prevention of cancers (Henning et al., 2013; Yang C.S. et al., 2014). Although the flavonoids contain similar biological function based on the same core scaffold, above results indicated the different biological functions of flavonoid subclasses with different chemical structures. Thus, the therapeutic selection and clinical application for flavonoid subclasses were different from each other.

Functional Difference Between Flavonoids and Other Natural Plant Products

To further discover the functional difference between flavonoids and other natural plant products, the specific enrichment pathway of flavonoids' targets were analyzed by setting other natural plant products as background. Results showed that, flavonoids were enriched in cancer-related pathways compared with other natural products (**Figure 3** and **Supplementary Table 3**). Among them, isoflavones and flavones were enriched in multi-cancer related pathways, flavan-3-ols can regulate the pathway of microRNA in cancer and isoflavones significantly enriched in the pathway of breast cancer, indicating the potential anti-cancer preferences of flavonoid subclasses. It can be noticed that natural flavan-3-ol products such as EGCG could alter epigenetic processes through DNA methylation, histone modification and miRNA regulation such as miR-92, miR-93, miR-106, miR-7-1, miR-34a, and miR-99a (Chakrabarti et al., 2012), which could provide anti-cancer and cardiovascular protections (Henning et al., 2013; Yang C.S. et al., 2014). Also, soy isoflavones, including genistein, daidzein, and their corresponding glucosides were reported to reduce the risk of breast cancer through meta-analysis (Yamamoto et al., 2003; Dong and Qin, 2011). *In vitro*, these isoflavones could significantly restrain the growth of human breast cancer cells (Peterson and Barnes, 1991). In addition to cancer related pathways, flavonoids were enriched in metabolic, steroid hormone biosynthesis, replication and repair, adherence junction, insulin signaling and several diseases related pathways. Meanwhile, existential discrepancy was found between different flavonoid subclasses. For example, although flavonoids showed biological functions on multiple nervous system diseases, only flavones were significantly enriched in Alzheimer' disease related pathways. Previous researches showed that the derivatives of flavone acting at different target could elicit varied pharmacological properties with various substitution patterns, including anti-oxidant, anti-cancer activity, neuroprotective activity (Singh et al., 2014). Those derivatives also showed good binding affinity to Aβ aggregates and high brain penetration, which illustrate potential therapeutic utilities for Alzheimer's disease (Ono et al., 2005, 2007). Also, isoflavones were significantly enriched in Huntington's diseases related pathways, which may related with isoflavones-mediated autophagy (Pierzynowska et al., 2017, 2018). Generally, both common and discrepancy were found in flavonoid enriched pathways which represent the specific biological function and potential therapeutic utility of different flavonoid subclasses.

Network Pharmacology and Modularization Analysis

To globally view the enrichment pathways of flavonoid, the network of all enriched targets for six flavonoids' subclass were analyzed and decomposed into eight modules. In **Figure 4**, the size of each node represents the ratio (=number of target related compounds/total number of compounds) of targets in here. Targets mapped in the same modules were marked in the same

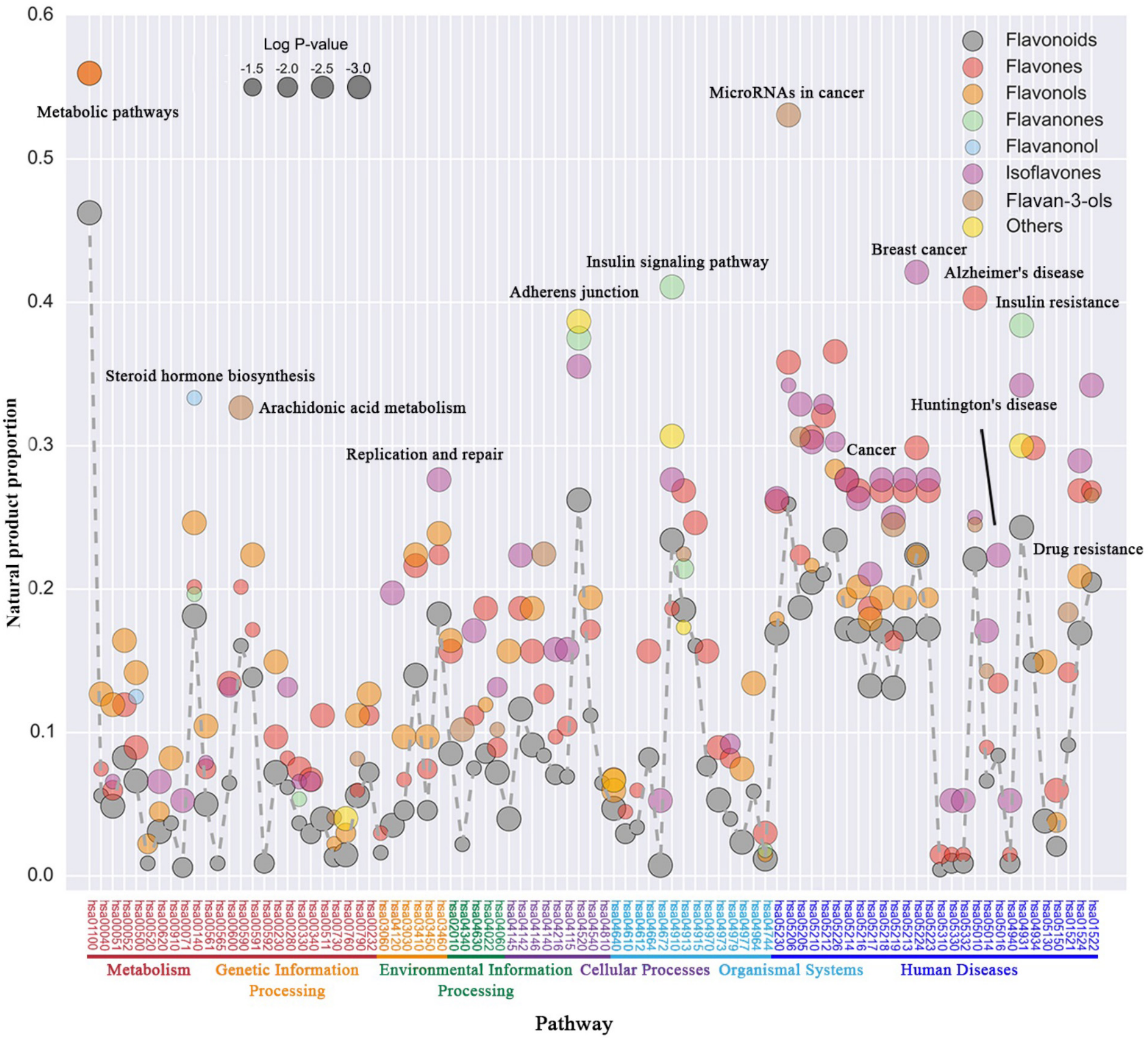


FIGURE 3 | Specific KEGG pathway enrichment analysis of natural flavonoid products' targets. Here, *X*-axis represents the enriched pathways (p-value < 0.05), which were categorized according to KEGG classification. *Y*-axis represents compound proportion of flavonoids in each pathway (number of target related compounds/total number of compounds in each class), the size of each nodes represented the significance of enrichment level (–LogP). Flavonoids and all six subclasses were marked in different colors.

color and network of targets in six major modules were analyzed through KEGG and GO to discover the function of flavonoid (**Table 1**), top 10 enriched pathways and GO terms were list in **Supplementary Table 4**. Entrez Gene ID and symbol in each module were listed in **Supplementary Table 5**.

Targets in module 1 were mainly enriched on epoxygenase P450 pathway, VEGF signaling pathway, fluid shear stress and atherosclerosis pathway, which related with cardiovascular regulations such as vascular dilatation. Meanwhile, the enrichment of pathways for FoxO signaling, mitotic cell cycle regulation, cell cycle arrest, negative regulation of cell cycle showed that cell cycle related functions can also been reflected in module 1. For module 2, targets were significantly enriched on pathways for T-cell receptor signaling, NF-kappa B signaling, inflammatory mediator regulation of TRP channels, immune response-regulating cell surface receptor signaling, immune response activating cell surface receptor signaling and immune response-activating signal transduction, which indicated the function of module 2 was related to immune inflammation. Further, it can be noticed that targets in module 3 were significantly enriched on pathways of insulin resistance, peptidyl-serine phosphorylation, peptidyl-serine modification, cellular response to nitrogen compound, cellular response to organonitrogen compound, positive regulation of kinase activity and cellular response to hormone stimulus, which illustrated the function of module 3 were closely related with functions of cell response to stimulation and hormone regulation. Moreover, the enrichment in neuroactive ligand-receptor interaction, cAMP

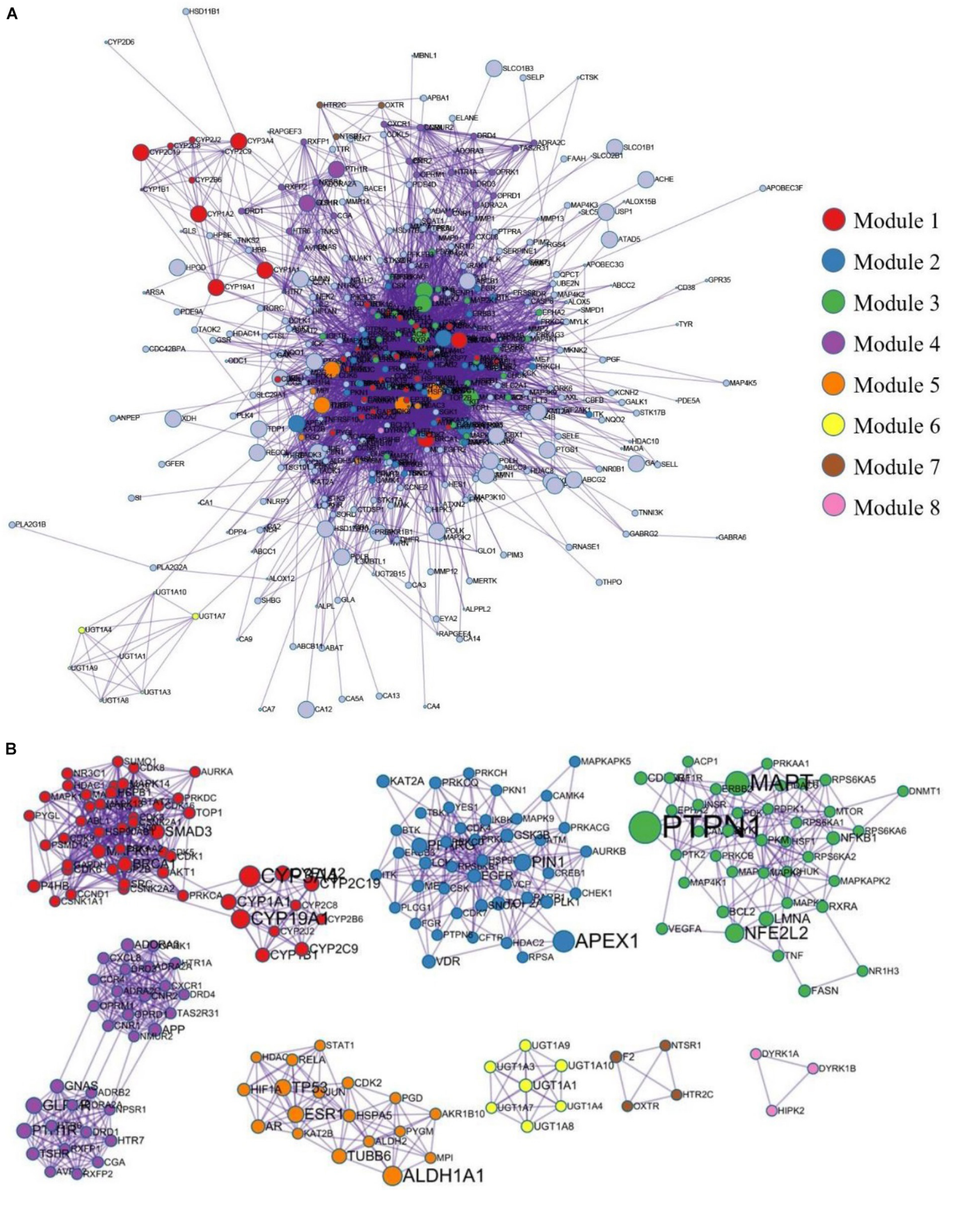


FIGURE 4 | Protein–protein interaction (PPI) network of natural flavonoid products' targets. **(A)** General PPI network of natural flavonoid products' targets. **(B)** Modularized PPI network of natural flavonoid products' targets. Size of each node represents the ratio (=number of target related compounds/total number of compounds) of targets. Different modules were marked in different colors.

TABLE 1 | Main functions of six major modules.

Modules	Main function
1	Cardiovascular regulations; cell cycle regulation
2	Immune and inflammation response
3	Cell response to stimulation and hormone regulation
4	Neuromodulation and signal transduction
5	ancer and viral-related diseases
6	Process of glyoxylic acid metabolism

signaling, calcium signaling, serotonergic synapse, cGMP-PKG signaling and dopaminergic synapse pathways reflected the targets in module 4 were related to neuromodulation and signal transduction. In addition, module 5 was found to relate with human diseases such as cancer and viral-related diseases since most targets were enriched in viral carcinogenesis pathway, cancer and infectious disease related pathways. For module 6, targets were enriched in flavonoids glucuronidation, glucuronate pathway, ascorbate and aldarate metabolism, pentose and glucuronate interconversions, which meant functions of module 6 were mainly embodied in the process of glyoxylic acid metabolism.

Module Mapping of Different Flavonoid Subclasses

In order to understand the function differences among flavonoid subclass, targets of six flavonoid subclasses were mapped into above modules. Major nodes reflected the common targets for each flavonoid subclass.

Start from flavan-3-ols, the targets were mainly distributed in module 1, 2, 3, and 5, and generally enriched in pathways of cancer, fluid shear stress and atherosclerosis, AGE-RAGE signaling in diabetic complications, which related with cardiovascular, cell cycle regulation and cancer (**Supplementary Figure 1**). For example, MAPK 14 in module 3 was found to participate in multiple cellular processes including cell proliferation, differentiation, transcriptional regulation and development (Young, 2013). Also, it can be noted that MAPK 14 may related with atherosclerosis (Cheng et al., 2017). Further, BCL2 in module 2 was a therapeutic target for chronic lymphocytic leukemia since it can regulate lymphocyte in blood by hindering cell apoptosis (Ruefli-Brasse and Reed, 2017; Tahir et al., 2017) and PGD in module 5 was related with human cervical carcinoma (Lee et al., 2014).

Targets of flavanones mainly distributed in module 1, 4, and 5, several scattered in modules 2 and 3, which illustrated the pharmacological activities of flavanones for anti-cancer and anti-oxidant (**Supplementary Figure 2**). For example, nodes such as CYP1A1, CYP1A2, CYP1B1 in module 1 belonged to cytochrome P450 (CYPs) family, which could enrich in epoxygenase P450 pathway and relate with cardiovascular-related functions such as vascular ectasia. APEX1 in module 2 was found affecting cancer RNA metabolism and triple-negative breast cancer (Antoniali et al., 2017; Chen et al., 2017).

Targets of flavanonol were relatively less than the others and separated in different modules, which means the function of flavanonol are quite scattered (**Supplementary Figure 3**). Similar to flavanones, nodes such as CYPs in module 1 and APEX1 in module 2 were also detected in flavanonol, which indicated the potential function of it on cardiovascular and cancer related functions. Meanwhile, MAPT in module 2 was found closely related with neurodegenerative diseases, including Parkinson's disease (Beevers et al., 2017).

Target of flavones (**Supplementary Figure 4**) and flavonols (**Supplementary Figure 5**) were distributed in all six major modules, which indicated the broad function of compounds from those two subclasses. Besides common nodes such as CYPs, APEX1, MAPT, which reflect the same function for cancer, cardiovascular and neurodegenerative as other flavonoid subclasses flavones contains other nodes such as ALDH1A1 in module 5, which reflect potential associations with cancer invasion (Yao et al., 2017; Li et al., 2018).

Specifically, isoflavones are a type of naturally occurring isoflavonoids, which act as phytoestrogens in mammals, their targets were mainly distributed in module 1, 2, and 5 (**Supplementary Figure 6**). Previous researches indicated that BRCA1 in module 1 was associated with risk of estrogen-receptor-negative breast cancer (Milne et al., 2017), NFE2L2 in module 3 was related with cell multisystem disorder (Huppke et al., 2017), and TP53 in module 5 was related with human immunodeficiency virus-related head and neck squamous cell carcinoma (Gleber-Netto et al., 2018).

Structure Activity Relationship Analysis of Flavonoids

In order to explore the cause of the functional similarity and difference among flavonoids' subclasses, the structure activity relationship of different flavonoids were analyzed. By calculating the structural and physic-chemical properties of natural flavonoid products, the structure difference of flavonoids' subclasses can be discovered to conjecture the potential effects of their biological functions (**Supplementary Figure 7**). Results showed that the rotatable bonds (NumRotatableBonds) and molecular weight (Weight) in different subclasses are quite similar, however, difference can be detected in H-bond acceptor (HAcceptor), H-bond donor (HDonor), lipid-water partition coefficient (LogP), and Topological polarity surface area (TPSA) for different flavonoid subclasses. LogP and TPSA could affect the absorption and distribution of drug, which should contain a certain degree of dissolution and appropriate lipid water distribution to be effective. Further, to provide nervous system activity, drug with larger liposolubility may be easier to pass the blood–brain barrier (BBB) (Yang Y. et al., 2014). The non-polar structural fragments such as alkyl group, halogen atom and aliphatic ring in chemical molecules will increase the liposolubility of molecules. Meanwhile, TPSA has a great impact on the cell penetration of drug molecules. In that case, the TPSA should be relatively lower for drugs which needs to across BBB and act on the receptors of central nervous system (Mehdipour and Hamidi, 2009). Natural products of flavones, flavanones, and isoflavones contain larger LogP and lower TPSA than other flavonoids, which indicated the potential activities to across the

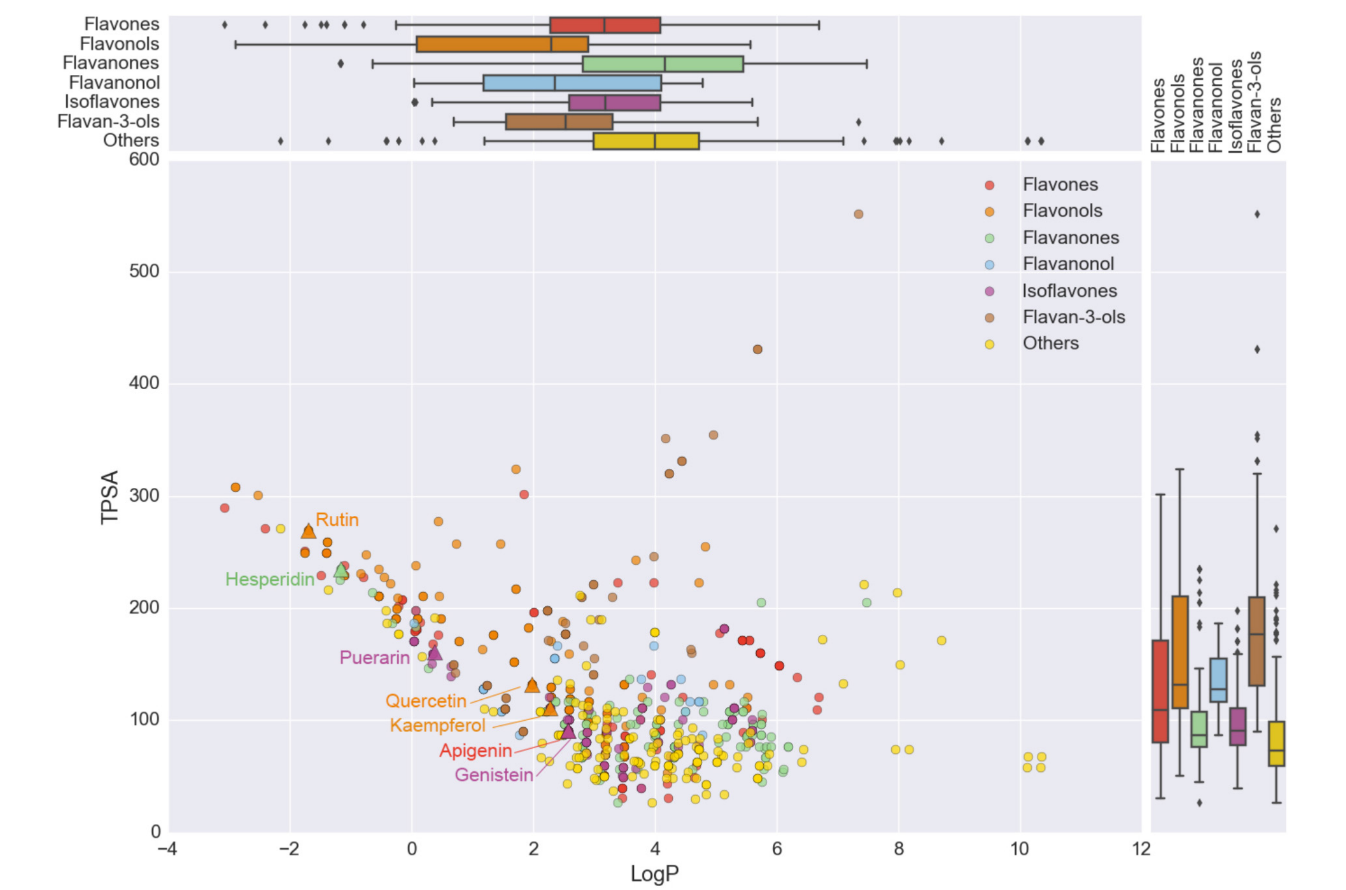


FIGURE 5 | Distribution of LogP and TPSA in different flavonoids. Different color represents different flavonoid subclasses. *X*-axis represents the value of LogP, while *Y*-axis represents the value of TPSA.

BBB (**Figure 5**). For example, apigenin of flavones, quercetin, and genistein of isoflavones, hesperidin of flavanones and rutin, quercetin, and kaempferol of flavonols would have the ability to across the BBB (**Figure 5**). Among them, genistein and apigenin could provide stronger ability to across the BBB since their larger LogP and lower TPSA (Yang Y. et al., 2014), which indicated the potential ability of other flavonoids meets the appropriate value of LogP and TPSA.

Further, in order to evaluate the drug-likeness of flavonoids, Lipinski's Rule of Five (ROF) of different flavonoid subclasses were analyzed (**Supplementary Figure 8**). It can be found that for flavones, flavonols, flavanones, flavanonol, and isoflavones, near half of the compound can pass ROF, while for flavan-3-ols the percentage of ROF-passed compounds is extremely low, which indicated different drug-likeness of flavonoid subclasses.

By excavating the relationship between chemical constitution fragment and biological effects through Apriori (Agrawal and Srikant, 1994), results showed that the core scaffold and side chain in flavonoids can significantly affect the biological functions (**Figure 6**). For example, in rule 01–07, side chain such as hydroxyl in position 1 on the core scaffold structure of flavanones may assist the negative regulation of PERK-mediated unfolded protein response. Also, in rule 09–13, hydroxyl side chain in position 1, 3, 10, 11, and 12 on the core scaffold structure of flavonols closely related with error-prone translesion synthesis. Among them, natural products such as myricetin, robinetin, tricetin could against hydrogen peroxide-induced DNA damage and might reduce the risk of multiple cancers (Huang and Ferraro, 1992; Shelby et al., 1997). Meanwhile, natural products with core scaffold of flavonols and oxygen methyl on different positions as side chain illustrated the bio-function of cellular iron ion homeostasis (rule 14–17) and microtubule-based process (rule 18–20). Besides above rule of generality, individual rules can also be found in **Figure 6**. For example, the core scaffold of flavones combined with hydroxyl side chain in position 1, 3, 11, and 12 will related with base-excision repair and base-free sugar-phosphate removal (rule 21). Previous studies indicated that the number and position of glycoside and hydroxyl groups in flavonoids would affects the ability of permeation (Yang Y. et al., 2014). We also found that hydroxyl side chain in position 1 and 3 combined with hydrocarbyl side chain in position 2 which related with neurotransmitter receptor biosynthetic process (rule 22) will increase the liposolubility and enhance its transmembrane abilities. It can be noted that, the bio-activity of molecules which meet rule 22 is enhanced over 14.68 times than other molecules in flavonoids' families (**Supplementary**

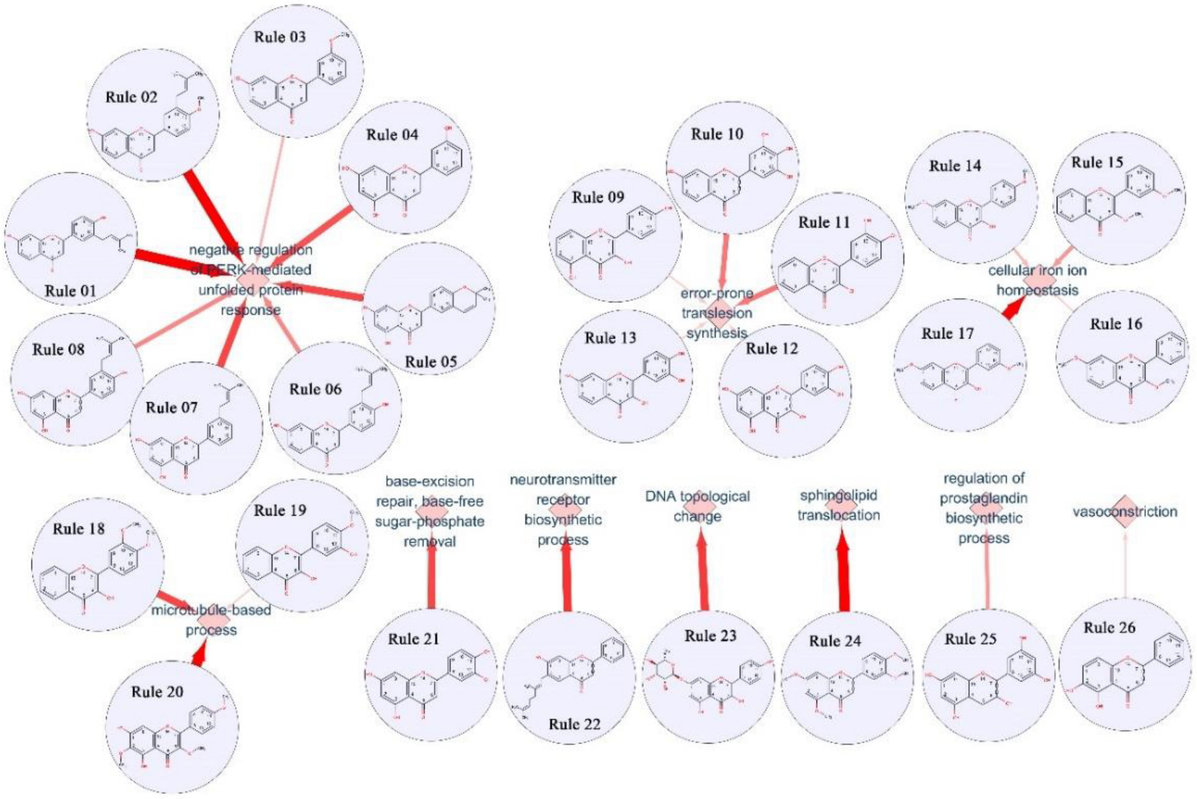

Rule 01
Rule 02
Rule 03
Rule 04
Rule 05
Rule 06
Rule 07
Rule 08
negative regulation of PERK-mediated unfolded protein response
Rule 09
Rule 10
Rule 11
Rule 12
Rule 13
error-prone translesion synthesis
Rule 14
Rule 15
Rule 16
Rule 17
cellular iron ion homeostasis
Rule 18
Rule 19
Rule 20
microtubule-based process
base-excision repair, base-free sugar-phosphate removal
neurotransmitter receptor biosynthetic process
DNA topological change
sphingolipid translocation
regulation of prostaglandin biosynthetic process
vasoconstriction
Rule 21
Rule 22
Rule 23
Rule 24
Rule 25
Rule 26

FIGURE 6 | Illustration of relationship between structure patterns and functions according to structure–activity relationship analysis of flavonoids.

Table 6). Analogously, flavonols meet rule 23, which contains pentose in position 1 and hydroxyl in position 3, 11 will increase the bioactivity for 32.37 times than others for the function of DNA topological change. Natural products such as kaempferol glycoside could targeting the DNA topoisomerase, which closely related with DNA replication and cell cycle (Vega et al., 2007; Baikar and Malpathak, 2010). Rule 24 indicate multiple oxygen methyl in side chain will benefit to the function of sphingolipid translocation. Moreover, rule 25 and rule 26 illustrate the different side chain components may have potential affects for regulation of prostaglandin biosynthetic process and vasoconstriction.

DISCUSSION

In this article, comprehensive analysis was proposed to explore the MOA of natural flavonoid products and results indicated that flavonoids could affect essential pathways in several categories such metabolism, genetic information processing, environmental information processing, cellular processes, organismal systems, and human diseases related pathways. Among them, the enrichment in human diseases-related pathways illustrated the multifaceted therapeutic applications of flavonoids which could affect multiple human diseases such as cancers, neuro-disease, diabetes and infectious diseases. By compared with other natural plant products, flavonoids could significantly enrich in the pathways of breast cancer, Huntington's disease, Alzheimer's disease, insulin resistance and drug resistance. Also, after systemically analysis of targets for different flavonoids subclasses, it can be found that targets such as MAPT, APEX1, and ALDH1A1, which were closely related with nervous system and cancer, were significantly enriched in almost all flavonoid subclasses. In that case, the multifaceted therapeutic ability indicates the utility of flavonoids for cancer and nervous system related drug discoveries.

Besides common biological functions, specific functions of different flavonoids subclasses were also analyzed and detected in both target and pathway level. For example, flavones and isoflavones were significantly enriched in multi-cancer related pathways than others, which indicate the potential therapeutic utility in cancer treatment. Also, flavan-3-ols were found on cellular processing and lymphocyte regulation, flavones specifically acted on cardiovascular related activities and isoflavones were closely related with cell multisystem disorders. Different structural and physic-chemical properties of natural flavonoid products may relate with the functional differences and can be detected in physic-chemical properties including H-bond acceptor, H-bond donor, lipid-water partition coefficient and topological polarity surface area. It can be noted that LogP and TPSA are closely related with absorption and distribution of chemical components in drugs, since appropriate solubility and lipid water distribution coefficient play essential roles in drug efficacy (Avdeef, 2001). For example, drugs which were activated in central nervous system requires larger liposolubility, which could be increased by non-polar structural fragments such as alkyl, halogen atom and aliphatic ring in chemical molecules. Meanwhile, TPSA can affect the cell penetration of drug molecules. Previous research indicated that in order to pass through the BBB and activate on the receptors in central nervous system, the polar surface areas of drug should be less than 90 square angstroms (van de Waterbeemd et al., 1998). Thus, natural products in flavonoids, flavones, flavanones, and isoflavones, which contains larger LogP and lower TPSA, have the ability to pass through the BBBs with potential activities.

Since flavonoids contain the same core scaffold, the functional difference was mainly related with the substituent groups. Relationship between chemical constitution fragment and biological effects indicated that different side chain can significantly affect the activity of flavonoids on the same target. Flavonoids with structures meet the corresponding rules will enhance the bioactivity of molecules for dozens of times. For 26 rules summarized in this article, the bioactivities were increased over three times at least. Among them, seven rules could enhance the bioactivities for over 10 times, and two rules (rule 23 and 26) could increase the activities for 30 times (**Supplementary Table 6**). Considering the substituent groups and positions of side chain, the relationship between structure and bioactivity analyzed in here may help to enhance the understanding of flavonoids and its potential ability for new drug discovery.

AUTHOR CONTRIBUTIONS

TQ and KT wrote the manuscript. DW and LY conceived and designed the experiments. HY and TQ analyzed and interpreted the results. TQ and QW modified the manuscript. ZC and KT supervised the project. All authors have read and approved the final version.

FUNDING

This work was supported in part by National Key R&D Program of China (SQ2017YFC170310), National Natural Science Foundation of China (31671379), the National Postdoctoral Program for Innovative Talents (BX201600033), and the China Postdoctoral Science Foundation Funded Project (2017M611451).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphar.2018.00918/full#supplementary-material

REFERENCES

Agrawal, R., and Srikant, R. (1994). "Fast algorithms for mining association rules in large databases," in *Proceedings of the International Conference on Very Large Data Bases*, Santiago, 487–499.

Antoniali, G., Serra, F., Lirussi, L., Tanaka, M., D'Ambrosio, C., Zhang, S. H., et al. (2017). Mammalian APE1 controls miRNA processing and its interactome is linked to cancer RNA metabolism. *Nat. Commun.* 8:797. doi: 10.1038/s41467-017-00842-8

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat. Genet.* 25, 25–29. doi: 10.1038/75556

Avdeef, A. (2001). Physicochemical profiling (solubility, permeability and charge state). *Curr. Top. Med. Chem.* 1, 277–351. doi: 10.2174/1568026013395100

Bader, G. D., and Hogue, C. W. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* 4:2. doi: 10.1186/1471-2105-4-2

Baikar, S., and Malpathak, N. (2010). Secondary metabolites as DNA topoisomerase inhibitors: a new era towards designing of anticancer drugs. *Pharmacogn. Rev.* **4**, 12–26. doi: 10.4103/0973-7847.65320

Bakhtiari, M., Panahi, Y., Ameli, J., and Darvishi, B. (2017). Protective effects of flavonoids against Alzheimer's disease-related neural dysfunctions. *Biomed. Pharmacother.* 93, 218–229. doi: 10.1016/j.biopha.2017.06.010

Batra, P., and Sharma, A. K. (2013). Anti-cancer potential of flavonoids: recent trends and future perspectives. *3 Biotech* 3, 439–459. doi: 10.1007/s13205-013-0117-5

Beevers, J. E., Lai, M. C., Collins, E., Booth, H. D. E., Zambon, F., Parkkinen, L., et al. (2017). MAPT genetic variation and neuronal maturity alter isoform expression affecting axonal transport in iPSC-derived dopamine neurons. *Stem Cell Rep.* 9, 587–599. doi: 10.1016/j.stemcr.2017.06.005

Breinholt, V., Hossaini, A., Brouwer, C., and Larsen, J. (1999). In vitro and in vivo estrogenic activity of dietary flavonoids: importance of bioavailability and metabolism. *J. Med. Food* 2, 227–229. doi: 10.1089/jmf.1999.2.227

Castell, M., Perez-Cano, F. J., Abril-Gil, M., and Franch, A. (2014). Flavonoids on allergy. *Curr. Pharm. Des.* 20, 973–987. doi: 10.2174/13816128113199990041

Chakrabarti, M., Khandkar, M., Banik, N. L., and Ray, S. K. (2012). Alterations in expression of specific microRNAs by combination of 4-HPR and EGCG inhibited growth of human malignant neuroblastoma cells. *Brain Res.* 1454, 1–13. doi: 10.1016/j.brainres.2012.03.017

Chang, S. K., Alasalvar, C., and Shahidi, F. (2018). Superfruits: phytochemicals, antioxidant efficacies, and health effects - A comprehensive review. *Crit. Rev. Food Sci. Nutr.* [Epub ahead of print]. doi: 10.1080/10408398.2017.1422111

Chatr-Aryamontri, A., Oughtred, R., Boucher, L., Rust, J., Chang, C., Kolas, N. K., et al. (2017). The BioGRID interaction database: 2017 update. *Nucleic Acids Res.* 45, D369–D379. doi: 10.1093/nar/gkw1102

Chen, T., Liu, C., Lu, H., Yin, M., Shao, C., Hu, X., et al. (2017). The expression of APE1 in triple-negative breast cancer and its effect on drug sensitivity of olaparib. *Tumour Biol.* 39:1010428317713390. doi: 10.1177/1010428317713390

Cheng, F., Twardowski, L., Fehr, S., Aner, C., Schaeffeler, E., Joos, T., et al. (2017). Selective p38alpha MAP kinase/MAPK14 inhibition in enzymatically modified LDL-stimulated human monocytes: implications for atherosclerosis. *FASEB J.* 31, 674–686. doi: 10.1096/fj.201600669R

Cushnie, T. P. T., and Lamb, A. J. (2005). Antimicrobial activity of flavonoids. *Int. J. Antimicrob. Aging* 26, 343–356. doi: 10.1016/j.ijantimicag.2005.09.002

Dong, J. Y., and Qin, L. Q. (2011). Soy isoflavones consumption and risk of breast cancer incidence or recurrence: a meta-analysis of prospective studies. *Breast Cancer Res. Treat.* 125, 315–323. doi: 10.1007/s10549-010-1270-8

Feliciano, R. P., Pritzel, S., Heiss, C., and Rodriguez-Mateos, A. (2015). Flavonoid intake and cardiovascular disease risk. *Curr. Opin. Food Sci.* 2, 92–99. doi: 10.1016/j.cofs.2015.02.006

Gleber-Netto, F. O., Zhao, M., Trivedi, S., Wang, J., Jasser, S., McDowell, C., et al. (2018). Distinct pattern of TP53 mutations in human immunodeficiency virus-related head and neck squamous cell carcinoma. *Cancer* 124, 84–94. doi: 10.1002/cncr.31063

Gontijo, V. S., Dos Santos, M. H., and Viegas, C., Jr. (2017). Biological and chemical aspects of natural biflavonoids from plants: a brief review. *Mini Rev. Med. Chem.* 17, 834–862. doi: 10.2174/1389557517666161104130026

Henning, S. M., Wang, P., Carpenter, C. L., and Heber, D. (2013). Epigenetic effects of green tea polyphenols in cancer. *Epigenomics* 5, 729–741. doi: 10.2217/epi.13.57

Hooper, L., Kroon, P. A., Rimm, E. B., Cohn, J. S., Harvey, I., Le Cornu, K. A., et al. (2008). Flavonoids, flavonoid-rich foods, and cardiovascular risk: a meta-analysis of randomized controlled trials. *Am. J. Clin. Nutr.* 88, 38–50. doi: 10.1093/ajcn/88.1.38

Huang, M. T., and Ferraro, T. (1992). Phenolic-compounds in food and cancer prevention. *Acs Symp. Ser.* 507, 8–34. doi: 10.1021/bk-1992-0507.ch002

Hughes, L. A. E., Arts, I. C. W., Ambergen, T., Brants, H. A. M., Dagnelie, P. C., et al. (2008). Higher dietary flavone, flavonol, and catechin intakes are associated with less of an increase in BMI over time in women: a longitudinal analysis from the Netherlands Cohort Study. *Am. J. Clin. Nutr.* 88, 1341–1352.

Huppke, P., Weissbach, S., Church, J. A., Schnur, R., Krusen, M., Dreha-Kulaczewski, S., et al. (2017). Activating de novo mutations in NFE2L2 encoding NRF2 cause a multisystem disorder. *Nat. Commun.* 8:818. doi: 10.1038/s41467-017-00932-7

Hwang, C. S., Kwak, H. S., Lim, H. J., Lee, S. H., Kang, Y. S., Choe, T. B., et al. (2006). Isoflavone metabolites and their in vitro dual functions: they can act as an estrogenic agonist or antagonist depending on the estrogen concentration. *J. Steroid Biochem.* 101, 246–253. doi: 10.1016/j.jsbmb.2006.06.020

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* 40, D109–D114. doi: 10.1093/nar/gkr988

Kawai, M., Hirano, T., Higa, S., Arimitsu, J., Maruta, M., Kuwahara, Y., et al. (2007). Flavonoids and related compounds as anti-allergic substances. *Allergol. Int.* 56, 113–123. doi: 10.2332/allergolint.R-06-135

Kumar, S., and Pandey, A. K. (2013). Chemistry and biological activities of flavonoids: an overview. *Sci. World J.* 2013:162750. doi: 10.1155/2013/162750

Landrum, G. (2010). *RDKit: Open-Source Cheminformatics*. Available at: http://www.rdkit.org/

Lee, M., Nam, E. S., Jung, S. H., Kim, S. Y., Lee, S. J., Yoon, J. H., et al. (2014). 1p36.22 region containing PGD gene is frequently gained in human cervical cancer. *J. Obstet. Gynaecol. Res.* 40, 545–553. doi: 10.1111/jog.12193

Li, T. B., Wernersson, R., Hansen, R. B., Horn, H., Mercer, J., Slodkowicz, G., et al. (2017). A scored human protein-protein interaction network to catalyze genomic interpretation. *Nat. Methods* 14, 61–64. doi: 10.1038/nmeth.4083

Li, Y., Chen, T., Zhu, J., Zhang, H., Jiang, H., and Sun, H. (2018). High ALDH activity defines ovarian cancer stem-like cells with enhanced invasiveness and EMT progress which are responsible for tumor invasion. *Biochem. Biophys. Res. Commun.* 495, 1081–1088. doi: 10.1016/j.bbrc.2017.11.117

Lin, C. W., Wu, M. J., Liu, I. Y., Su, J. D., and Yen, J. H. (2010). Neurotrophic and cytoprotective action of luteolin in PC12 cells through ERK-dependent induction of Nrf2-driven HO-1 expression. *J. Agric. Food Chem.* 58, 4477–4486. doi: 10.1021/jf904061x

Matias, A., Nunes, S. L., Poejo, J., Mecha, E., Serra, A. T., Madeira., et al. (2014). Antioxidant and anti-inflammatory activity of a flavonoid-rich concentrate recovered from *Opuntia ficus-indica* juice. *Food Funct.* 5, 3269–3280. doi: 10.1039/C4FO00071D

Matveeva, O., Bogie, J. F. J., Hendriks, J. J. A., Linker, R. A., Haghikia, A., and Kleinewietfeld, M. (2018). Western lifestyle and immunopathology of multiple sclerosis. *Ann. N. Y. Acad. Sci.* 1417, 71–86. doi: 10.1111/nyas.13583

Mehdipour, A. R., and Hamidi, M. (2009). Brain drug targeting: a computational approach for overcoming blood-brain barrier. *Drug Discov. Today* 14, 1030–1036. doi: 10.1016/j.drudis.2009.07.009

Milne, R. L., Kuchenbaecker, K. B., Michailidou, K., Beesley, J., Kar, S., Lindstrom, S., et al. (2017). Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer. *Nat. Genet.* 49, 1767–1778. doi: 10.1038/ng.3785

Mohebali, N., Shahzadeh Fazeli, S. A., Ghafoori, H., Farahmand, Z., MohammadKhani, E., Vakhshiteh, F., et al. (2018). Effect of flavonoids rich extract of *Capparis spinosa* on inflammatory involved genes in amyloid-beta peptide injected rat model of Alzheimer's disease. *Nutr. Neurosci.* 21, 143–150. doi: 10.1080/1028415X.2016.1238026

Mulvihill, E. E., and Huff, M. W. (2010). Antiatherogenic properties of flavonoids: implications for cardiovascular health. *Can. J. Cardiol.* 26, 17a–21a. doi: 10.1016/S0828-282X(10)71056-4

Nijveldt, R. J., van Nood, E., van Hoorn, D. E. C., Boelens, P. G., van Norren, K., and van Leeuwen, P. A. M. (2001). Flavonoids: a review of probable mechanisms of action and potential applications. *Am. J. Clin. Nutr.* 74, 418–425. doi: 10.1093/ajcn/74.4.418

Ono, M., Maya, Y., Haratake, M., Ito, K., Mori, H., and Nakayama, M. (2007). Aurones serve as probes of P-amyloid plaques in Alzheimer's disease. *Biochem. Biophys. Res. Commun.* 361, 116–121. doi: 10.1016/j.bbrc.2007.06.162

Ono, M., Yoshida, N., Ishibashi, K., Haratake, M., Arano, Y., Mori, H., and Nakayama, M. (2005). Radioiodinated flavones for in vivo imaging of beta-amyloid plaques in the brain. *J. Med. Chem.* 48, 7253–7260. doi: 10.1021/jm050635e

Paredes-Gonzalez, X., Fuentes, F., Jeffery, S., Saw, C. L., Shu, L., Su, Z. Y., and Kong, A. N. (2015). Induction of NRF2-mediated gene expression by dietary phytochemical flavones apigenin and luteolin. *Biopharm. Drug Dispos.* 36, 440–451. doi: 10.1002/bdd.1956

Peterson, G., and Barnes, S. (1991). Genistein inhibition of the growth of human breast cancer cells: independence from estrogen receptors and the multi-drug resistance gene. *Biochem. Biophys. Res. Commun.* 179, 661–667. doi: 10.1016/0006-291X(91)91423-A

Pierzynowska, K., Gaffke, L., Hac, A., Mantej, J., Niedzialek, N., Brokowska, J., and Wegrzyn, G. (2018). Correction of Huntington's disease phenotype by genistein-induced autophagy in the cellular model. *Neuromol. Med.* 20, 112–123. doi: 10.1007/s12017-018-8482-1

Pierzynowska, K., Mantej, J., Niedzialek, N., Hac, A., and Wegrzyn, G. (2017). Stimulation of autophagy by genistein reduces levels of mutant huntingtin in cellular models of Huntington disease. *Mol. Genet. Metab.* 120:S107. doi: 10.1016/j.ymgme.2016.11.273

Pietta, P. G. (2000). Flavonoids as antioxidants. *J. Nat. Prod.* 63, 1035–1042. doi: 10.1021/np9904509

Ross, J. A., and Kasum, C. M. (2002). Dietary flavonoids: bioavailability, metabolic effects, and safety. *Annu. Rev. Nutr.* 22, 19–34. doi: 10.1146/annurev.nutr.22.111401.144957

Ruefli-Brasse, A., and Reed, J. C. (2017). Therapeutics targeting Bcl-2 in hematological malignancies. *Biochem. J.* 474, 3643–3657. doi: 10.1042/BCJ20170080

Serafini, M., Peluso, I., and Raguzzini, A. (2010). Session 1: antioxidants and the immune system Flavonoids as anti-inflammatory agents. *Proc. Nutr. Soc.* 69, 273–278. doi: 10.1017/S002966511000162X

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303

Shelby, M. D., Lohman, P. H. M., and Hayatsu, H. (1997). Policy on plant extracts and other complex mixtures for submissions to mutation research - genetic toxicology and environmental mutagenesis. *Mutat. Res.Genet. Toxicol. Environ. Mutagen.* 390:R5.

Singh, M., Kaur, M., and Silakari, O. (2014). Flavones: an important scaffold for medicinal chemistry. *Eur. J. Med. Chem.* 84, 206–239. doi: 10.1016/j.ejmech.2014.07.013

Spencer, J. P., Vauzour, D., and Rendeiro, C. (2009). Flavonoids and cognition: the molecular mechanisms underlying their behavioural effects. *Arch. Biochem. Biophys.* 492, 1–9. doi: 10.1016/j.abb.2009.10.003

Szmitko, P. E., and Verma, S. (2005). Cardiology patient pages. Red wine and your heart. *Circulation* 111, e10–e11. doi: 10.1161/01.CIR.0000151608.29217.62

Tahir, I. M., Iqbal, T., Jamil, A., and Saqib, M. (2017). Association of BCL-2 with oxidative stress and total antioxidant status in pediatric acute lymphoblastic leukemia. *J. Biol. Regul. Homeost. Agents* 31, 1023–1027.

Tripathi, S., Pohl, M. O., Zhou, Y., Rodriguez-Frandsen, A., Wang, G., Stein, D. A., et al. (2015). Meta- and orthogonal integration of influenza "OMICS" data defines a role for UBR4 in virus budding. *Cell Host Microbe* 18, 723–735. doi: 10.1016/j.chom.2015.11.002

Turei, D., Korcsmaros, T., and Saez-Rodriguez, J. (2016). OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nat. Methods* 13, 965–967. doi: 10.1038/nmeth.4077 van de Waterbeemd, H., Camenisch, G., Folkers, G., Chretien, J. R., and Raevsky, O. A. (1998). Estimation of blood-brain barrier crossing of drugs using molecular size and shape, and H-bonding descriptors. *J. Drug Target.* 6, 151–165. doi: 10.3109/10611869808997889

Vega, M. R. G., Esteves-Souza, A., Vieira, I. J. C., Mathias, L., Braz-Filho, R., and Echevarria, A. (2007). Flavonoids from *Annona dioica* leaves and their effects in Ehrlich carcinoma cells, DNA-topoisomerase I and II. *J. Braz. Chem. Soc.* 18, 1554–1559. doi: 10.1590/S0103-50532007000800016

Williams, R. J., and Spencer, J. P. (2012). Flavonoids, cognition, and dementia: actions, mechanisms, and potential therapeutic utility for Alzheimer disease. *Free Radic. Biol. Med.* 52, 35–45. doi: 10.1016/j.freeradbiomed.2011.09.010

Yamamoto, S., Sobue, T., Kobayashi, M., Sasaki, S., Tsugane, S., and Japan Public Health Center-Based Prospective Study on Cancer Cardiovascular Diseases Group (2003). Soy, isoflavones, and breast cancer risk in Japan. *J. Natl. Cancer Inst.* 95, 906–913. doi: 10.1093/jnci/95.12.906

Yang, C. S., Chen, G., and Wu, Q. (2014). Recent scientific studies of a traditional Chinese medicine, tea, on prevention of chronic diseases. *J. Tradit. Complement. Med.* 4, 17–23. doi: 10.4103/2225-4110.124326

Yang, Y., Bai, L., Li, X. R., Xiong, J., Xu, P. X., Guo, C. Y., et al. (2014). Transport of active flavonoids, based on cytotoxicity and lipophilicity: an evaluation using the blood-brain barrier cell and Caco-2 cell models. *Toxicol. In Vitro* 28, 388–396. doi: 10.1016/j.tiv.2013.12.002

Yao, H., Xu, W. Z., Shi, X. L., and Zhang, Z. (2011). Dietary flavonoids as cancer prevention agents. *J. Environ. Sci. Health C Environ. Carcinog. Ecotoxicol. Rev.* 29, 1–31. doi: 10.1080/10590501.2011.551317

Yao, J., Jin, Q., Wang, X. D., Zhu, H. J., and Ni, Q. C. (2017). Aldehyde dehydrogenase 1 expression is correlated with poor prognosis in breast cancer. *Medicine* 96:e7171. doi: 10.1097/MD.0000000000007171

Young, P. R. (2013). Perspective on the discovery and scientific impact of p38 MAP kinase. *J. Biomol. Screen.* 18, 1156–1163. doi: 10.1177/1087057113497401

Zeng, X., Zhang, P., He, W., Qin, C., Chen, S., Tao, L., et al. (2018). NPASS: natural product activity and species source database for natural product research, discovery and tool development. *Nucleic Acids Res.* 46, D1217–D1222. doi: 10.1093/nar/gkx1026

Zhang, H., and Ma, Z. F. (2018). Phytochemical and pharmacological properties of *Capparis spinosa* as a medicinal plant. *Nutrients* 10:E116. doi: 10.3390/nu10020116

Zhang, T., Su, J., Guo, B., Wang, K., Li, X., and Liang, G. (2015). Apigenin protects blood-brain barrier and ameliorates early brain injury by inhibiting TLR4-mediated inflammatory pathway in subarachnoid hemorrhage rats. *Int. Immunopharmacol.* 28, 79–87. doi: 10.1016/j.intimp.2015.05.024

Zhao, L., Wang, J. L., Liu, R., Li, X. X., Li, J. F., and Zhang, L. (2013). Neuroprotective, anti-amyloidogenic and neurotrophic effects of apigenin in an Alzheimer's disease mouse model. *Molecules* 18, 9949–9965. doi: 10.3390/molecules18089949

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Pharmacology

ORIGINAL RESEARCH
published: 12 September 2018
doi: 10.3389/fphar.2018.01017

A Hybrid Interpolation Weighted Collaborative Filtering Method for Anti-cancer Drug Response Prediction

Lin Zhang[1], *Xing Chen*[1]*, *Na-Na Guan*[2], *Hui Liu*[1] *and Jian-Qiang Li*[2]

[1] *School of Information and Control Engineering, China University of Mining and Technology, Xuzhou, China,* [2] *College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China*

OPEN ACCESS

Edited by:
Lixia Yao,
Mayo Clinic, United States

Reviewed by:
Yanshan Wang,
Mayo Clinic, United States
Chen Li,
Xi'an Jiaotong University, China
Chen Wang,
Mayo Clinic, United States

***Correspondence:**
Xing Chen
xingchen@amss.ac.cn

Specialty section:
This article was submitted to Translational Pharmacology, a section of the journal Frontiers in Pharmacology

Received: *09 April 2018*
Accepted: *22 August 2018*
Published: *12 September 2018*

Citation:
Zhang L, Chen X, Guan N-N, Liu H and Li J-Q (2018) A Hybrid Interpolation Weighted Collaborative Filtering Method for Anti-cancer Drug Response Prediction. Front. Pharmacol. 9:1017. doi: 10.3389/fphar.2018.01017

Individualized therapies ask for the most effective regimen for each patient, while the patients' response may differ from each other. However, it is impossible to clinically evaluate each patient's response due to the large population. Human cell lines have harbored most of the same genetic changes found in patients' tumors, thus are widely used to help understand initial responses of drugs. Based on the more credible assumption that similar cell lines and similar drugs exhibit similar responses, we formulated drug response prediction as a recommender system problem, and then adopted a hybrid interpolation weighted collaborative filtering (HIWCF) method to predict anti-cancer drug responses of cell lines by incorporating cell line similarity and drug similarity shown from gene expression profiles, drug chemical structure as well as drug response similarity. Specifically, we estimated the baseline based on the available responses and shrunk the similarity score for each cell line pair as well as each drug pair. The similarity scores were then shrunk and weighted by the correlation coefficients drawn from the know response between each pair. Before used to find the K most similar neighbors for further prediction, they went through the case amplification strategy to emphasize high similarity and neglect low similarity. In the last step for prediction, cell line-oriented and drug-oriented collaborative filtering models were carried out, and the average of predicted values from both models was used as the final predicted sensitivity. Through 10-fold cross validation, this approach was shown to reach accurate and reproducible outcome for those missing drug sensitivities. We also found that the drug response similarity between cell lines or drugs may play important role in the prediction. Finally, we discussed the biological outcomes based on the newly predicted response values in GDSC dataset.

Keywords: anti-cancer drug response, drug response prediction, recommender system, collaborative filtering, interpolation weighted method

INTRODUCTION

One of the top challenges in individualized therapies is the choice of the most effective chemotherapeutic regimen for each patient, while the administration of ineffective chemotherapy may increase mortality and decrease quality of life in cancer patients (Chen et al., 2013). Thus, it is urgent to evaluate each patients' possible response to each chemotherapeutic regimen to make sure the regimens applied are most likely to be effective. To address this problem, extensive patient drug screening projects need to be carried out so as to unveil significant drug response patterns. However, the large populations of cancer patients with numerous drugs has become the bottleneck.

To circumvent this issue in the context of cancer, some large drug screening projects have been carried out using cancer cell lines instead of individual cancer patients. These are NCI-60 panel, Genomics of Drug Sensitivity in Cancer (GDSC) and the Cancer Cell Line Encyclopedia (CCLE) projects (Boyd and Paull, 1995; Barretina et al., 2012; Yang et al., 2013). The NCI-60 study was pioneered by the US National Cancer Institute (NCI) to assemble the NCI60 tumor cell line panel, which has been assayed for its sensitivity to over 130,000 compounds and had been extensively profiled at the biological level (Shoemaker, 2006). It has been useful for the development of computational approaches aiming at linking drug sensitivity with genotype profiles together (Shoemaker et al., 1988; Weinstein et al., 1997; Garnett et al., 2012). The GDSC project is, to date, the largest public resource for information on drug sensitivity in human cancer cell lines and molecular markers of drug response. It pioneered the combination of drug and cell line information, including gene expression, gene copy number variations, and mutation profiles for drug sensitivity prediction (Garnett et al., 2012; Yang et al., 2013). It systematically addressed the issue of predictive biomarker identification by collectively analyzing the clinically-relevant human cell lines and their pharmacological profiles for corresponding cancer drugs. The other widely used database, CCLE (Barretina et al., 2012), collects gene expression, chromosomal copy number and massively parallel sequencing data from 947 human cancer cell lines, coupled with pharmacological profiles for 24 anti-cancer drugs across 479 of the cell lines. It allows identification of genetic, lineage, and gene expression-based predictors of drug sensitivity.

Corresponding to the large-scale datasets screened on cultured human cell line panels, many computational methods have been developed for the elucidation of the response mechanism of anti-cancer drugs, most commonly are multivariate linear regression (LASSO and elastic net regularizations) and nonlinear regression (e.g., neural networks and some kernel based methods; Barretina et al., 2012; Garnett et al., 2012; Heiser et al., 2012; Menden et al., 2013; Yang et al., 2013; Costello et al., 2014). Deamen et al. used least squares-support vector machine and random forest to identify drug response associated molecular features in breast cancer (Daemen et al., 2013). Based on the NCI-60 panel, a weighted voting classification model, an ensemble regression model using Random Forest as well as a simultaneous machine learning modeling of chemical and cell line information have been developed to predict anti-cancer drug sensitivity (Staunton et al., 2001; Riddick et al., 2011; Cortes-Ciriano et al., 2016). Based on the GDSC dataset, Ammad-uddin et al. developed a kernelized Bayesian matrix factorization (KBMF) method to integrate genomic and chemical properties as well as drug target information for drug sensitivity prediction (Ammad-ud-din et al., 2014). Sheng et al predicted unseen drug responses by calculating a weighted average of observed drug responses based on drug specific cell line similarity and drug structure similarity (Breese et al., 1998). Liu et al. proposed a dual-layer cell line drug integrated network (DLN) model, which integrated both cell line and drug similarity network data, to predict the missing drug response (Zhang et al., 2015). Wang et al. proposed HNMDRP method, incorporating gene expression, chemical structure as well as drug target and protein-protein interaction information to predict missing values of drug responses in cell lines (Zhang et al., 2018). Based on the transcriptomic data from both GDSC and CCLE, Kim et al. developed a network-based classifier for predicting sensitivity of cell lines to anti-cancer drugs (Kim et al., 2016). Base on the same whole datasets, Wang et al. proposed a similarity-regularized matrix factorization (SRMF) method for drug response prediction, which incorporates similarities of drugs and of cell lines simultaneously (Wang et al., 2017). Stanfield et al. proposed a heterogeneous network based method to predict the interaction between cell line-drug pairs (Stanfield et al., 2017). They classified the interaction between each cell line-drug pairs into sensitive and resistant, thus, turned the prediction problem into classification. Current methods have taken the similarity of genomic or transcriptomic profiles as well as drug structure into consideration for similarity definition, which were often defined by calculating the Pearson correlation coefficient for genomic profiles, or Jaccard coefficient for drug chemical fingerprint in present studies and are called as *COEF* in the following for short. However, the similarity that exhibited through drug sensitivity, which can be defined by calculating the Pearson correlation coefficient based on drug response sensitivity, has not been considered yet and is called as *RPCC* for short in the following. Not to mention the combination of *COEF* and *RPCC*, which is called as *MRPCC* (Multiplication of *COEF* and *RPCC*) for short throughout the paper. Drug-target interaction and PPI network have also been considered to improve the prediction performance (Chen et al., 2012; Stanfield et al., 2017).

Regarding the relatively more credible assumption that similar cell lines and similar drugs exhibit similar drug responses (Zhang et al., 2015), the prediction of missing drug response can be considered as a typical Recommender System (RS) (Adomavicius and Tuzhilin, 2005). Typically, in a recommender system, there is a set of users and a set of items. Each user rates a set of items by some values. The recommender system attempts to profile user preferences and tries to model the interaction between users and items, which is exactly what we want in the issue of drug response prediction. The cell lines correspond to users while drugs correspond to items. From the RS perspective, the similarity shown through drug sensitivity is also very important for missing value prediction. Thus, we improved an RS technique, Hybrid Interpolation Weighted Collaborative Filtering (HIWCF)

(The acronym list defined in this paper is shown in **Table 1**), for drug response prediction, which incorporates similarities of drugs and of cell lines in additional to the known drug response simultaneously (The key source code and ready to use CCLE and GDSC datasets are provided at https://github.com/laureniezhang/HIWCF). To demonstrate its effectiveness, we compared HIWCF with SRMF and KBMF, which have been proved to show higher performance than typical similarity-based methods. The evaluation metrics used were averaged Pearson correlation coefficient (PCC) and averaged root mean square error (RMSE) over all drugs. The results on GDSC and CCLE drug response datasets by 10-fold cross validation showed that similarity defined based on drug response is more dependable for unknown response prediction, and the incorporation of gene expression profile, drug response, and drug structure similarity help to better improve the prediction performance. Finally, HIWCF was applied to impute the unknown drug response values in GDSC dataset for further evaluation.

MATERIALS AND METHODS

Data and Preprocessing

In this paper, two datasets, both consisting of large scale genomic expression profiles, pharmacologic profiling of drug compounds, as well as the experimentally determined drug response measurements IC50 values (the concentration of a drug compound that reached the absolute inhibition of 50% *in vitro*, given as natural log of μM) or experimental activity areas were used for performance evaluation. Large scale genomic expression profiles were normalized across cell lines to draw the similarity matrix of cell lines. The chemical structures of drug compounds were used to draw the similarity matrix of drugs.

The first dataset is from GDSC project (http://www.cancerrxgene.org/), consisting of 139 drugs and a panel of 790 cancer cell lines (release 5.0). We selected 652 cell lines for which both drug response data and gene expression were available, and 135 drugs whose SDF format (encoding the chemical structure of the drugs) were available. The drug response is given with IC50 values (70,676 data points, matrix 80.3% complete).

TABLE 1 | Acronym list.

Acronym	Detailed description
HIWCF	Hybrid Interpolation Weighted Collaborative Filtering
$COEF_c$	Pearson Correlation Coefficient drawn from cell line gene expression profile
$COEF_d$	Jaccard Correlation Coefficient drawn from drug chemical fingerprint
$RPCC_c$	Pearson Correlation Coefficient between cell lines drawn from drug response matrix
$RPCC_d$	Pearson Correlation Coefficient between drugs drawn from drug response matrix
$RPCC$	Refers to $RPCC_c$ or $RPCC_d$. It depends on the context.
$MRPCC_c$	Multiplication of $COEF_c$ with $RPCC_c$, used as final similarity score between cell lines.
$MRPCC_d$	Multiplication of $COEF_d$ with $RPCC_d$, used as final similarity score between drugs.
$MRPCC$	Refers to $MRPCC_c$ or $MRPCC_d$. It depends on the context.

The second dataset consists of 1,036 human cancer cell lines and 24 drugs, which is from CCLE project (http://www.broadinstitute.org/ccle). We also selected 491 cell lines and 23 drugs following the same rule used in GDSC dataset. The drug response is given with activity areas (10,870 data points, matrix 96.25% complete). Both ready to use datasets are submitted to Github at https://github.com/laureniezhang/HIWCF.

Problem Formulation

We basically treat anti-cancer drug response prediction as a RS problem where each cell line-drug pair is the typical user-item pair. Based on the finding that similar cell lines by gene expression profiles exhibit similar response to the same drug (Zhang et al., 2015), we proposed a weighted interpolation collaborative filtering method to approximate the sensitivity of cell line u to drug i. For convenience, we reserve special indexing letters for distinguishing cell lines from items: for cell lines u, v, and for drugs i, j. We are given cell line drug response about m cell lines and n drugs, arranged as an $m \times n$ matrix $R = \{r_{ui}\}_{1\le u\le m, 1\le i\le n}$, where higher value of activity area or lower value of IC50 means a better sensitivity of a cell line to a given drug.

Baseline Estimate Strategy

Since typical CF data often exhibit large user and item effects, that means systematic tendencies for some users to give higher ratings than others, and for some items to receive higher ratings than others, we first adjusted the rating data by accounting for these effects, which we include in the baseline estimate strategy. Let μ denotes the overall average drug response, we denote the estimated baseline for an unknown rating $\hat{r}_{ui}$ as b_{ui}, which accounts for the above-mentioned user and item effects.

$$b_{ui} = \mu + b_u + b_i \tag{1}$$

The parameters b_u and b_i indicate the observed deviations of cell line u and drug i, respectively, from the average.

In order to get the baseline formulation, for each drug i, we set:

$$b_u = \frac{\sum_{i\in U(u,i)} (r_{ui} - \mu - b_i)}{\lambda_3 + |U(u,i)|} \tag{2}$$

Then, for each cell line u, we set:

$$b_i = \frac{\sum_{u\in U(u,i)} (r_{ui} - \mu)}{\lambda_2 + |U(u,i)|} \tag{3}$$

where $U(u,i)$ is the set of cell lines who responses to drug i, or the set of drugs who have responses in cell line u, and $|U(u,i)|$ means the number of elements in set$U(u,i)$. $\lambda_2$and λ_3 are regularization parameters that help to shrink the averages b_u and b_i toward zero. They are set to 5 and 2, respectively in the following simulation process.

TABLE 2 | The comparison results between HIWCF with different similarity definition (MRPCC/RPCC/COEF), SRMF, and KBMF obtained under 10-fold cross validation on CCLE dataset.

Methods		Drug-averaged PCC_S/R	Drug-averaged RMSE_S/R	Drug-averaged PCC	Drug-averaged RMSE
HIWCF	MRPCC	0.80(±0.07)	0.66(±0.21)	0.74(±0.08)	0.53(±0.15)
	RPCC	0.80(±0.06)	0.67(±0.22)	0.73(±0.08)	0.54(±0.16)
	COEF	0.74(±0.06)	0.76(±0.27)	0.66(±0.06)	0.60(±0.20)
SRMF		0.78(±0.07)	0.74(±0.23)	0.71(±0.09)	0.57(±0.18)
KBMF		0.65(±0.10)	0.81(±0.20)	0.71(±0.10)	0.64(±0.17)

TABLE 3 | The comparison results between HIWCF with different similarity definition (MRPCC/RPCC/COEF), SRMF, and KBMF obtained under 10-fold cross validation on GDSC dataset.

Methods		Drug-averaged PCC_S/R	Drug-averaged RMSE_S/R	Drug-averaged PCC	Drug-averaged RMSE
HIWCF	MRPCC	0.68(±0.14)	1.88(±0.54)	0.58(±0.15)	1.51(±0.39)
	RPCC	0.68(±0.14)	1.87(±0.53)	0.58(±0.15)	1.50(±0.38)
	COEF	0.57(±0.15)	2.12(±0.60)	0.46(±0.14)	1.66(±0.43)
SRMF		0.71(±0.15)	1.73(±0.46)	0.62(±0.16)	1.43(±0.36)
KBMF		0.59(±0.14)	2.00(±0.51)	0.49(±0.14)	1.59(±0.42)

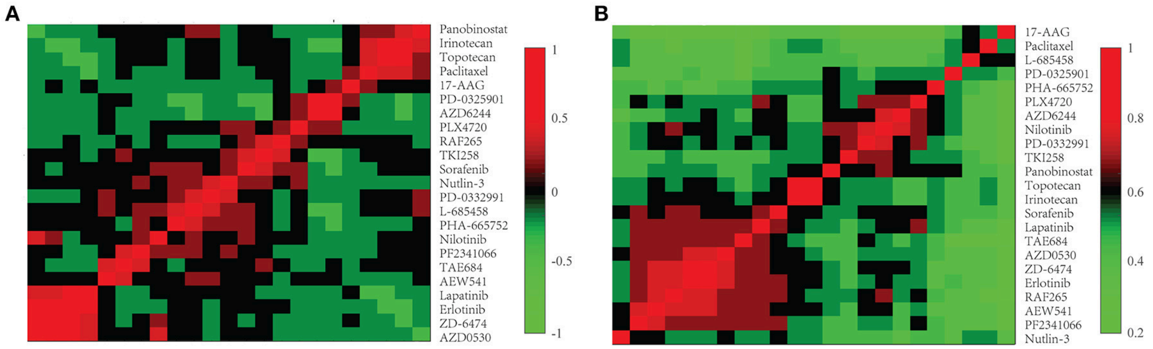


FIGURE 1 | The drug similarity *RPCC* and *COEF* of 23 drugs in CCLE dataset. **(A)** The plot shows *RPCC* similarity for 23 drugs in CCLE dataset. **(B)** The plot shows *COEF* similarity for 23 drugs in CCLE dataset.

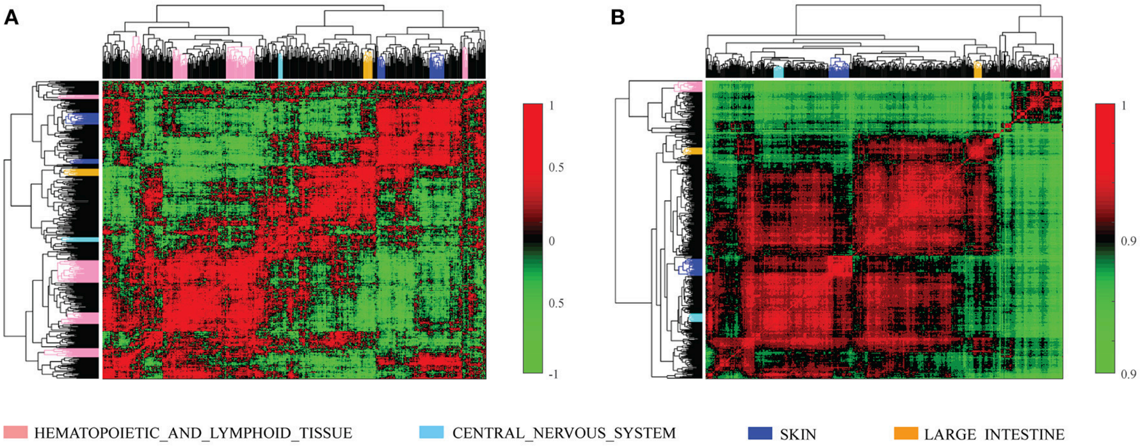


FIGURE 2 | The cell line similarity *RPCC* and *COEF* of 491 cell lines in CCLE dataset. **(A)** The plot shows *RPCC* similarity for 491 cell lines in CCLE dataset. **(B)** The plot shows *COEF* similarity for 491 cell lines in CCLE dataset.

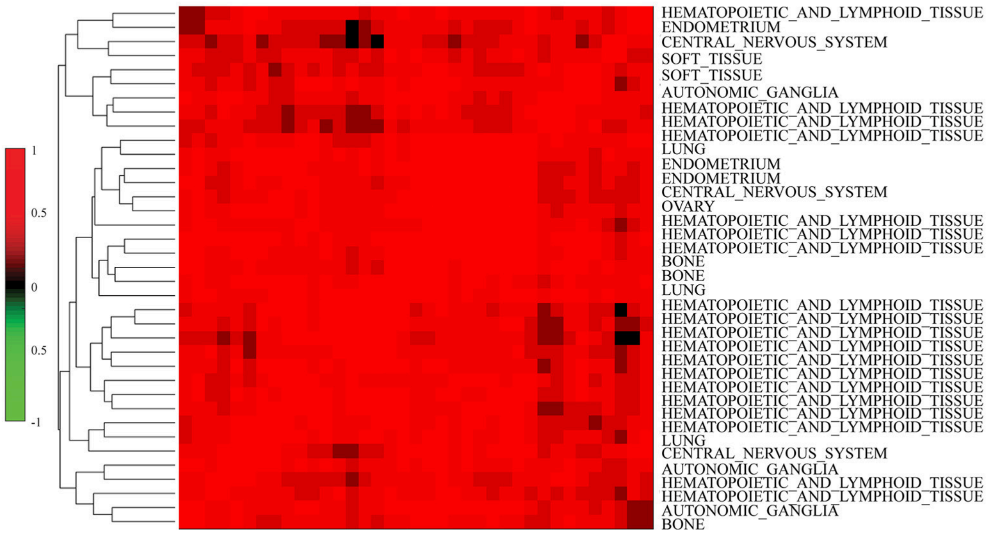


FIGURE 3 | Similar cell lines are more likely to be clustered into the same group (have similar similarity score) based on *MRPCC* similarity score. Most cell lines in the plot were collected from hematopoietic and lymphoid tissues.

Similarity Definition

The similarity matrixes are required for identification of K nearest neighbors. The original similarity of cell lines was drawn based on the Pearson correlation coefficient between the gene expression profiles of cell line u and v, which is indicated as $COEF_{c_{uv}}$. The c in the subscript refers to cell line-oriented. The similarity of drugs was drawn based on the Jaccard coefficient between the drug chemical structures of drug i and j, which is indicated as $COEF_{d_{ij}}$. The d in the subscript refers to drug-oriented.

However, to some extent, the similarity between cell line u and v can also be shown from their drug response. Thus, in this paper, we investigated the performance of different similarity definitions for drug response prediction. To be more specific, the similarity of cell line u and v, indicated as$MRPCC_{c_{uv}}$, was defined as the multiplication of $COEF_{c_{uv}}$and$RPCC_{c_{uv}}$, which helps the cell line pairs with consistent similarity in gene expression and drug response to get higher rank for unknown response prediction.

$$MRPCC_{c_{uv}} \leftarrow COEF_{c_{uv}} \times RPCC_{c_{uv}} \tag{4}$$

where $COEF_{c_{uv}}$ was defined as the their gene expression profile's Pearson correlation, while $RPCC_{c_{uv}}$ was defined as the correlation between the response IC50 value of cell line u and v.

$$RPCC_{c_{uv}} = \frac{\sum (R_{u\bullet} - \bar{R}_{u\bullet})(R_{v\bullet} - \bar{R}_{v\bullet})}{\sqrt{\sum (R_{u\bullet} - \bar{R}_{u\bullet})^2 \sum (R_{v\bullet} - \bar{R}_{v\bullet})^2}} \tag{5}$$

where $R_{u\bullet}$ represents the response value of the u-th cell line, and $\bar{R}_{u\bullet}$ represents the mean of the u-th cell line's response.

In the same way, the similarity between drug i and j, indicated as $MRPCC_{d_{ij}}$, was defined as the multiplication of $COEF_{d_{ij}}$ and $RPCC_{d_{ij}}$.

$$MRPCC_{d_{ij}} = COEF_{d_{ij}} \times RPCC_{d_{ij}} \tag{6}$$

where $COEF_{d_{ij}}$ was defined as their drug chemical fingerprint's Jaccard coefficient, while $RPCC_{d_{ij}}$ was defined as the Pearson correlation coefficient between response IC50 values of drug i and j.

$$RPCC_{d_{ij}} = \frac{\sum (R_{\bullet i} - \bar{R}_{\bullet i})(R_{\bullet j} - \bar{R}_{\bullet j})}{\sqrt{\sum (R_{\bullet i} - \bar{R}_{\bullet i})^2 \sum (R_{\bullet j} - \bar{R}_{\bullet j})^2}} \tag{7}$$

where $R_{\bullet i}$ represents the response value of the i-th drug, and $\bar{R}_{\bullet i}$ represents the mean of the i-th drug's response.

In order to avoid the bias caused by the different level of support (different number of known responses) for each cell line-drug pair, we also went through a shrunk procedure for similarity score, which is denoted by (Koren, 2010):

$$w_{i,j} \leftarrow \frac{|U(i,j)|}{|U(i,j)| + \lambda_4} w_{i,j} \tag{8}$$

where $|U(i,j)|$ is the number of cell lines who have responses to both drug i and j, or the number of drugs who have responses from both cell line i and j. $w_{ij}$is the similarity $MRPCC_c$ defined in (4) and $MRPCC_d$ in (6). $\lambda_4$is a constant, which is set as 50 in the experiments.

In the following, we adopted a case amplification strategy, which refers to a transform applied to the weights used in the following collaborative filtering prediction, to reduce the noise in the data. The transform emphasizes high weights and punishes low weights by (Breese et al., 1998):

$$w_{i,j} \leftarrow w_{i,j} \cdot |w_{i,j}|^{\rho-1} \tag{9}$$

where ρ is the case amplification power, $\rho \geq 1$, and we also followed the typical choice of ρ as 2.5 (Lemire, 2005).

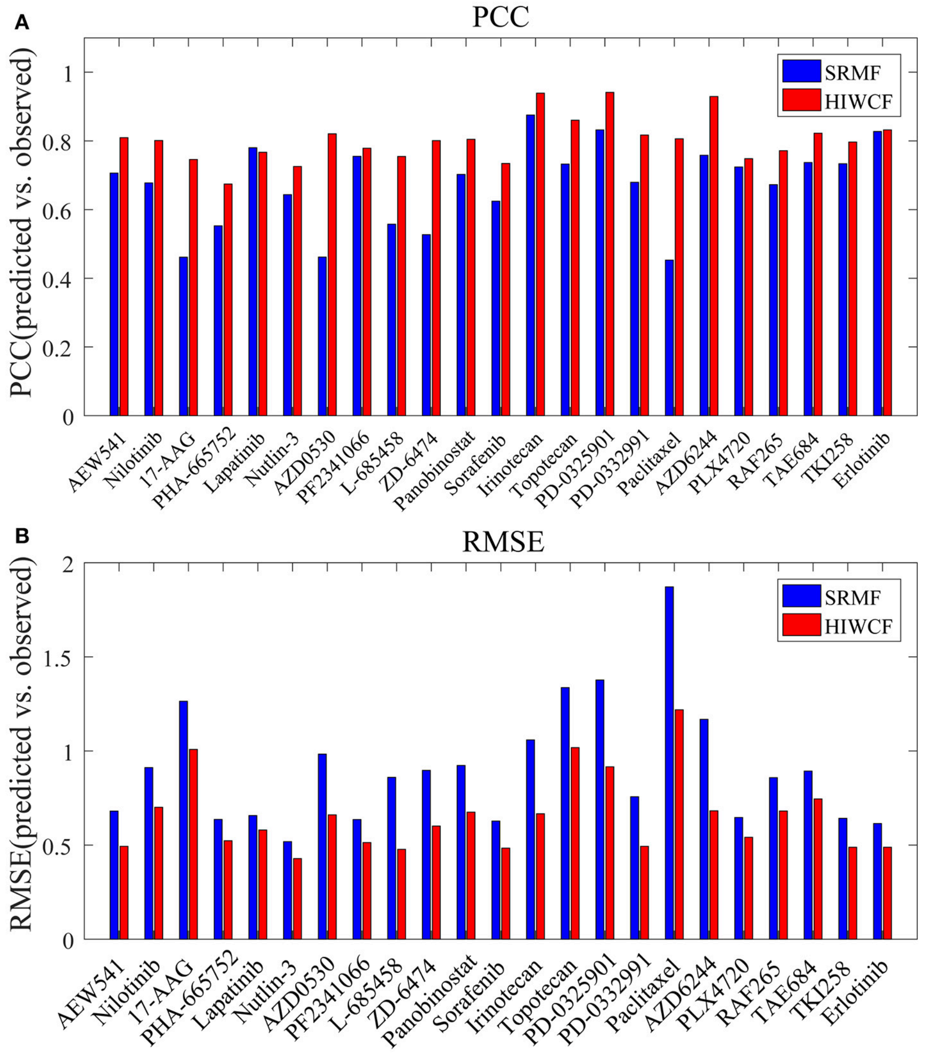


FIGURE 4 | Prediction performance of HIWCF with *MRPCC* similarity and SRMF for all 23 drugs tested in the CCLE dataset. **(A)** Bar plot shows that the prediction performance of HIWCF with *MRPCC* is better than that of SRMF in the perspective of Pearson correlations between the predicted and observed activity areas. **(B)** Bar plot shows that the prediction performance of HIWCF with *MRPCC* is better than that of SRMF in the perspective of Root Mean Square Error between the predicted and observed activity areas.

Drug Response Prediction Based on HIWCF Method

After removing the noise by baseline estimate strategy, we need to predict the unknown sensitivity for cell line u of drug i, which is $\hat{r}_{ui}$. Based on the above-mentioned similarity measure w defined in (9), we first conducted drug-oriented CF, and k drugs, which are most similar to drug i that had responses in cell line u were identified. This set of k neighboring drugs is denoted by $U(i; u)$. Then, based on w, we conducted cell line-oriented CF, and k cell lines that responded to drug i, which are most similar to cell line u were identified. This set of k neighboring cell lines is denoted by $U(u; i)$. Finally, the predicted value of $\hat{r}_{ui}$ is taken as an average of the weighted average of the response of neighboring drugs found in $U(i; u)$ and that of the response of neighboring cell lines found in$U(u; i)$, while adjusting from user and item effects through baseline estimates:

$$\hat{r}_{ui} = b_{ui} + \frac{1}{2}\Big(\frac{\sum_{j\in U(i;u)} w_{i,j}(r_{uj} - b_{uj})}{\sum_{j\in U(i;u)} w_{i,j}} + \frac{\sum_{v\in U(u;i)} w_{i,j}(r_{vi} - b_{vi})}{\sum_{v\in U(u;i)} w_{i,j}}\Big) \qquad (10)$$

RESULTS

Similarity Exhibited in Drug Response Sensitivity Shows Leading Role in Prediction

We first conducted 10-fold cross validation to evaluate the performance of different similarity definition. Incorporated with

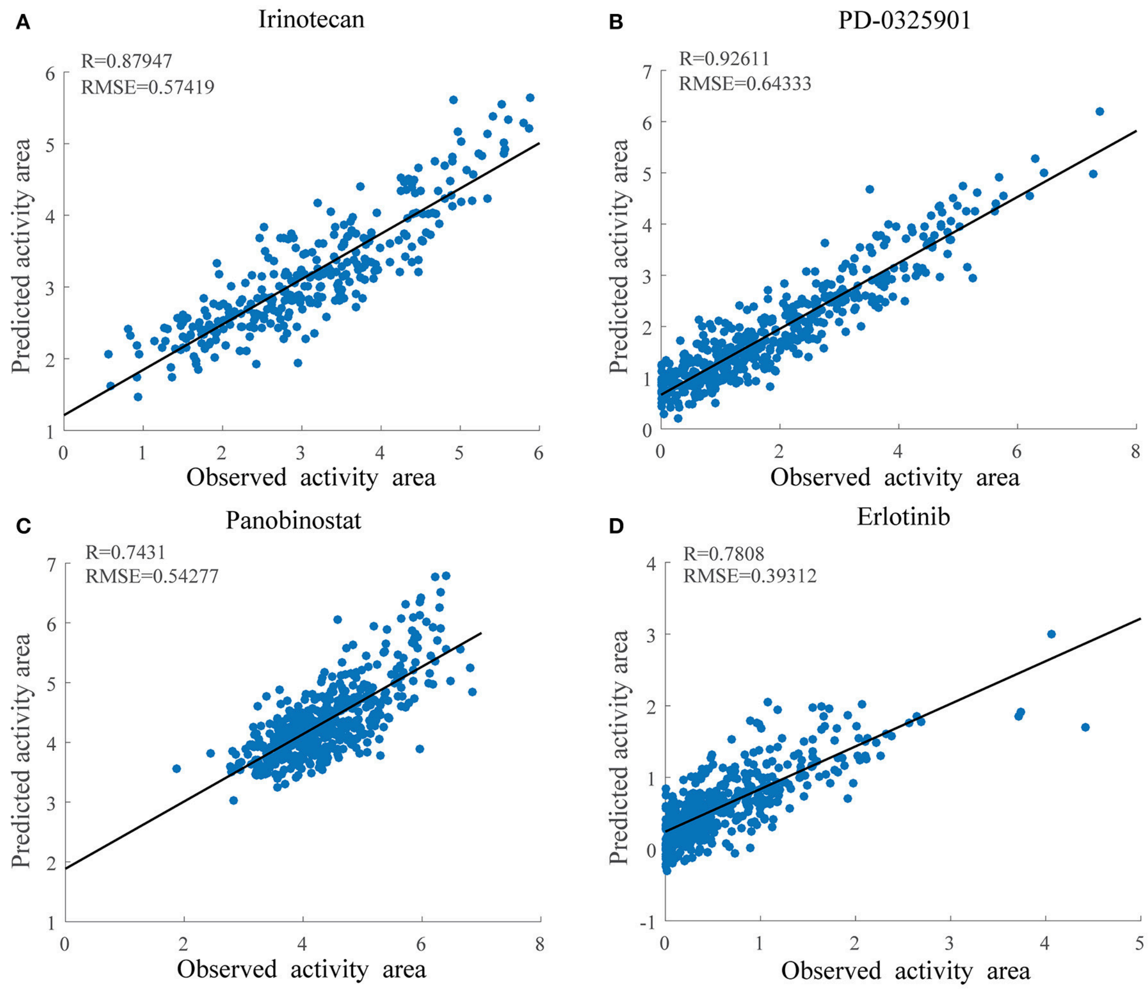


FIGURE 5 | Scatter plots of observed and predicted drug activity area for four drugs in CCLE using HIWCF with MRPCC similarity. **(A)** Scatter plot of Irinotecan. **(B)** Scatter plot of PD-0325901. **(C)** Scatter plot of Panobinostat. **(D)** Scatter plot of Erlotinib.

COEF, *RPCC* as well as *MRPCC*, drug response prediction performance of HIWCF is evaluated in both CCLE dataset and GDSC dataset with activity area or IC50 value as drug response measurement in comparison with KBMF and SRMF. The evaluation measures included average PCC, RMSE between predicted and observed drug responses through all drugs. Considering the known fact that the sensitive and resistant cell lines of each drug are more valuable to unveil mechanisms of drug actions, we also included PCC and RMSE from sensitive and resistant cell lines for each drug, which were denoted as PCC_S/R and RMSE_S/R (Wang et al., 2017).

For each dataset, the drug response entries were divided into 10-folds randomly with almost the same size. Each time, one-fold was used as the test set, while the rest nine-folds were used as the training set. The prediction was repeated 10 times such that each fold acted as a test set once. The whole cross-validation was run for 100 times for each dataset, and the prediction performance was shown in **Tables 2**, **3**.

As is shown, the prediction performance of HIWCF with *MRPCC/RPCC* similarity were far better than that with *COEF* similarity, which suggested that the similarity exhibited in drug response may lead important role than that of gene expression profiles or drug structures in the scenario of drug response prediction. Thus, we turned to use the predicted values of HIWCF with *MRPCC* similarity measure only in the rest evaluation of our paper.

In **Table 2**, we can also see that in CCLE dataset, the performance of HIWCF with *RPCC* and *MRPCC* were better than that of SRMF, without mentioning KBMF. However, as shown in **Table 3**, the performance of HIWCF with either *RPCC* or *MRPCC* were a little bit worse than that of SRMF. That may be because the similarity score of *RPCC/MRPCC* is based on the known drug response for each cell line-drug pair. Since GDSC dataset is much sparser than that of CCLE, the similarity score of *RPCC/MRPCC* of GDSC is less reliable than that of CCLE.

We further investigated the difference between *COEF* and *RPCC*. To be more specific, based CCLE dataset, we calculated the drug structure fingerprint similarity *COEF* for hierarchical clustering analysis. As shown in **Figure 1B**, it was surprising that the similarity score for most drug pairs were approaching

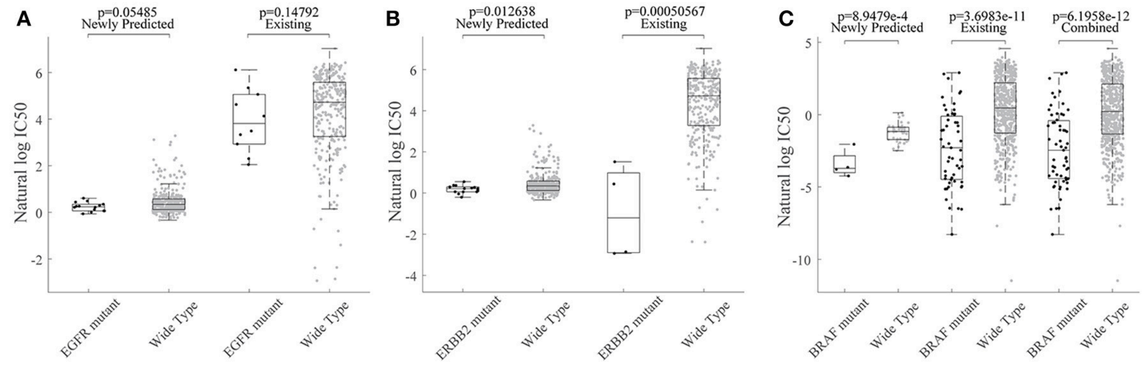


FIGURE 6 | The association of lapatinib sensitivity and cancer gene mutations were consistent for predicted response values. WT refers to the non-mutated (wild type) cell lines. **(A)** Box plot for grouped cell line response values for lapatinib based on their EGFR mutation profiles. **(B)** Box plot for grouped cell line response values for lapatinib based on their ERBB2 mutation profiles. **(C)** Box plot for grouped cell line response values for lapatinib based on their ERBB2 mutation profiles.

1, which was undistinguishable for neighbor selection. However, we can get distinguishable similarity scores from drug response similarity *RPCC*, as shown in **Figure 1A**. If we investigate the drugs that clustered into the same group, such as "Lapatinib," "AZD0530," "ZD-6474," and "Erlotinib." It is well-known that they are EGFR inhibitors, thus, they are most likely have higher similarity scores in drug response (Yuan et al., 2016). We also investigate the gene expression similarity with cell line response similarity. The cell line response similarity *RPCC* and cell line gene expression similarity *COEF* were calculated for hierarchical clustering, which were comparable with each other (**Figure 2**). The results show that cell lines collected from the same tissue type may have higher similarity score, which is consistent with previous studies. For example, most cell lines that clustered into the same group shown in **Figure 3** were collected from hematopoietic and lymphoid tissues. Hierarchical clustering was achieved in both row and column direction, with original similarity score was normalized with 0 mean.

Cross-Validation on CCLE Drug Response Datasets

We then tested the prediction performance of HIWCF for 23 drugs tested in the CCLE study, which were quantified based on PPC and RMSE between the predicted and observed activity areas.

As shown in **Figure 4**, the overall prediction performance of HIWCF throughout all the drugs was significantly higher than that of SRMF for the CCLE dataset. We believe that the improvement of HIWCF is most likely due to the involvement of similarity calculated from response matrix. The scatter plots of observed vs. predicted responses for four demonstrative drugs, Irinotecan, PD-0325901, Panobinostat, and Erlotinib are shown in **Figure 5**, which indicate the good correlations between existing response and predicted ones.

Response Data Prediction in GDSC Data

Based on the HIWCF method validated, we based on all known data to predict the unknown ones in the GDSC dataset.

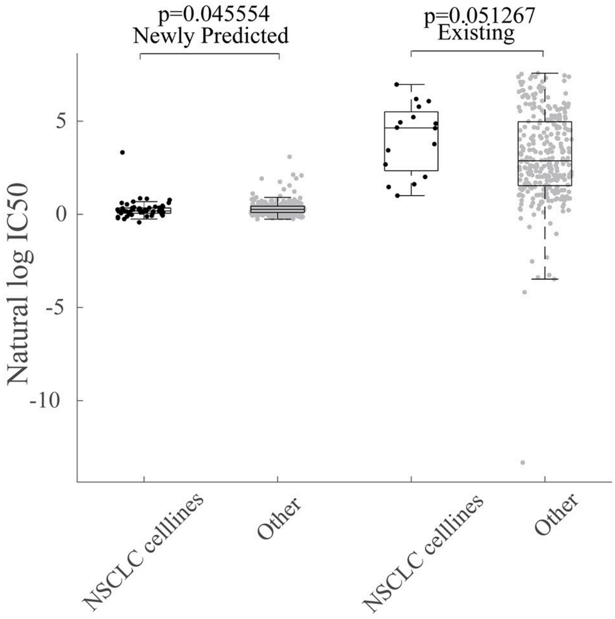


FIGURE 7 | Repositioning of sunitinib. Box plot for grouped cell line response values for Sunitinib based on their tissue type. NSCLC indicates cell lines sampled from non-small cell lung cancer tissues.

As in Wang et al. (2017), we also focused on an EGFR and ERBB2 inhibitor drug lapatinib, where more than half of response values (342/652) were unknown. Previous studies had demonstrated that EGFR and ERBB2 amplification was associated with sensitivity to lapatinib, which has been licensed for the treatment of HER2+ breast cancer clinically (Petrelli et al., 2017; Zhao et al., 2017). Thus, we tried to investigate whether the observed and predicted response of EGFR/ERBB2 mutated cell lines exhibit the sensitivity to lapatinib. All the 635 cell lines in GDSC were first grouped into mutated vs. wildtype by the total copy number variation in the exact gene (Garnett et al., 2012). Then, we found that not only EGFR mutated but also

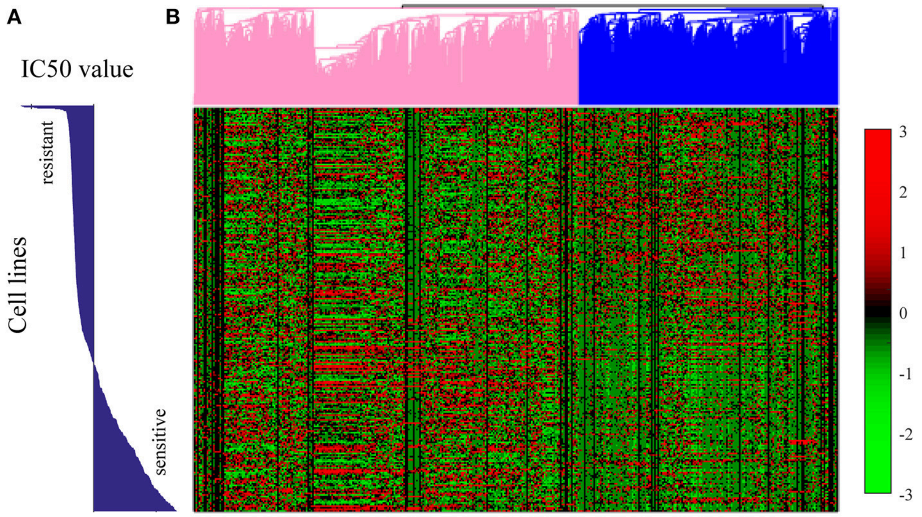


FIGURE 8 | Hierarchical clustering analysis on the gene expression profiles for all the 652 cell lines in GDSC dataset. **(A)** The bar plot of the IC^{50} values of each cell line. **(B)** The hierarchical clustering plot on the right showed the gene expression pattern for 20% most variant genes in each cell line. Each row in **(A)** corresponds to the exact row in the hierarchical clustering plot of gene expression profiles in **(B)**. The genome expression pattern was shown as some genes were up-regulated in Sunitinib resistant cell lines but down-regulated in Sunitinib sensitive cell lines, while some other genes were up-regulated in Sunitinib sensitive cell lines but down-regulated in Sunitinib resistant cell lines.

ERBB2 mutated cell lines were both significantly more sensitive to lapatinib, as shown in **Figures 6A,B**, which was consistent with previously mentioned conclusions.

We further investigated whether the newly predicted drug responses combined with known drug responses were able to detect novel drug-cancer gene association or not. To be more specific, the oncogene BRAF has been found to be significantly associated with enhanced and selective sensitivity to MEK inhibitor PD-0325901 (Solit et al., 2006) ($p = 3.70e\text{-}11$ for known drug responses; $p = 6.20e\text{-}12$ for combined response of predicted ones and known ones; **Figure 6C**).

The newly predicted drug responses of GDSC dataset may also aid in drug repositioning. For example, Sunitinib, as a kinase inhibitor targeting VEGFR2 and PDGFRβ, has been observed to be sensitive to non-small cell lung cancer (NSCLC) based on newly predicted drug responses vs. available ones, as shown in **Figure 7**.

We further conducted the hierarchical clustering analysis through genes based on the expression profile of all the 652 cell lines. Before hierarchical clustering, 80 percent genes that show less variations over all the genes were filtered out. As shown in **Figure 8**, the patterns of gene expression were shown to be related with the sensitivity of each cell line to Sunitinib. The pink marked group of genes showed higher expression in cell lines which were sensitive to Sunitinib, while the blue marked group of genes showed higher expression in cell lines which were resistant to Sunitinib.

We further conducted GO enrichment analysis for both groups of genes. For the genes that up-regulated in Sunitinib resistant cell lines were found to be related to some repair pathways, such as regulation of DNA repair ($p = 1.1e\text{-}3$), base-excision repair ($p = 0.032$), nucleotide-excision repair ($p = 6e\text{-}3$), interstrand cross-link repair ($p = 0.01$), mismatch repair ($p = 0.048$), etc., which were found to be important factors of drug resistance. For genes that were up-regulated in Sunitinib sensitive cell lines were found to be related to mTOR signaling pathway ($p = 1e\text{-}2$), NF-kappaB signaling ($p = 4.1e\text{-}10$). The inhibition of the signaling pathways help to increase drug sensitivities (Cai et al., 2014).

DISCUSSION

In this paper, we used a recommender system-based method HIWCF to predict anti-cancer drug sensitivity in GDSC and CCLE datasets respectively. The idea of the method comes from the fact that similar cell lines exhibit similar responses to the same drug, which is the exact motivation of a recommender system. This method first estimated the baseline, which helped to remove the noise in the original drug sensitivity, then shrunk the similarity measure by integration of gene expression profile, drug structure in addition to the correlation between cell lines and drugs exhibited in the drug response, which helped to weak the influence of sparseness in response matrix. Finally, it incorporated the user-orientated and item-orientated interpolation weighted collaborative filtering method to predict the unknown drug sensitivity values. Ten-fold cross validation demonstrated that the similarity drawn based on known drug response can better improve the prediction performance in comparison to the similarity drawn based on cell line gene expression profiles and drug structure only. At least, in the respective of recommender system method, it is more reliable to predict the unknown drug sensitivity based on the similarity exhibited in known drug responses. We also applied HIWCF method to predict the missing drug response values in GDSC dataset. To be more specific, we found the consistent conclusions of mutated cell lines such as EGFR/ERBB2 are more sensitive to the drug of lapatinib. We also found that the gene expression profiles showed exact pattern for Sunitinib sensitive and resistant cell lines. Genes that up-regulated in Sunitinib sensitive cell lines were subjected to repair pathways, while genes that down-regulated in Sunitinib resistant cell lines were subjected to some drug enhancement related pathways.

In comparison with existing drug response prediction methods, HIWCF follows a neighbor based collaborative filtering approach for unknown drug response prediction, which is theoretically simple and intuitive. Matrix Factorization based methods, such as SRMF model both cell lines and drugs with some latent factors for unknown drug response prediction.

However, this method has its own drawbacks. First, since HIWCF highly depends on the known drug response, the performance highly depends on the sparseness of the response matrix. The sparser the matrix is, the worse the performance it gets. Secondly, the similarity of cell lines is calculated by combining gene expression correlation coefficient and Pearson correlation coefficient exhibited in their known drug response. However, the similarity can also be improved by integrating the epigenetic, epi-transcriptomic information, etc. Furthermore, some pathway related information or other dynamic information may also help to improve the performance. Therefore, we can further work on some methods that aim in sparse issue as well as multi-omics integration one in the future.

AUTHOR CONTRIBUTIONS

LZ developed the prediction method, designed and implemented the experiments, analyzed the result, and wrote the paper. XC conceived the project, designed the experiments, analyzed the result, revised the paper, and supervised the project. N-NG prepared the data, analyzed the result, and revised the paper. HL and J-QL analyzed the result and revised the paper.

FUNDING

LZ was supported by the Fundamental Research Funds for the Central Universities and National Natural Science Foundation of China under Grant No. 2014QNB47 and 61501466. XC was supported by National Natural Science Foundation of China under Grant No. 61772531.

REFERENCES

Adomavicius, G., and Tuzhilin, A. (2005). Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Trans. Knowl. Data Eng.* 17, 734–749. doi: 10.1109/TKDE.2005.99

Ammad-ud-din, M., Georgii, E., Gonen, M., Laitinen, T., Kallioniemi, O., Wennerberg, K., et al. (2014). Integrative and personalized QSAR analysis in cancer by kernelized Bayesian matrix factorization. *J. Chem. Inf. Model.* 54, 2347–2359. doi: 10.1021/ci500152b

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., et al. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 483, 603–607. doi: 10.1038/nature11003

Boyd, M. R., and Paull, K. D. (1995). Some practical considerations and applications of the national cancer institute *in vitro* anticancer drug discovery screen. *Drug Dev. Res.* 34, 91–109. doi: 10.1002/ddr.430340203

Breese, J. S., Heckerman, D., and Kadie, C. (1998). "Empirical analysis of predictive algorithms for collaborative filtering," in *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence* (Madison: Morgan Kaufmann Publishers Inc.), 43–52.

Cai, Y., Tan, X., Liu, J., Shen, Y., Wu, D., Ren, M., et al. (2014). Inhibition of PI3K/Akt/mTOR signaling pathway enhances the sensitivity of the SKOV3/DDP ovarian cancer cell line to cisplatin *in vitro*. *Chin. J. Cancer Res.* 26, 564. doi: 10.3978/j.issn.1000-9604.2014.08.20

Chen, J., Cheng, G. H., Chen, L. P., Pang, T. Y., and Wang, X. L. (2013). Prediction of chemotherapeutic response in unresectable non-small-cell lung cancer (NSCLC) patients by 3-(4,5-dimethylthiazol-2-yl)-5-(3-carboxymethoxyphenyl)-2- (4-sulfophenyl)-2H-tetrazolium (MTS) assay. *Asian Pac. J. Cancer Prev.* 14, 3057–3062. doi: 10.7314/APJCP.2013.14.5.3057

Chen, X., Liu, M.-X., and Yan, G.-Y. (2012). Drug–target interaction prediction by random walk on the heterogeneous network. *Mol. Biosyst.* 8, 1970–1978. doi: 10.1039/c2mb00002d

Cortes-Ciriano, I., van Westen, G. J., Bouvier, G., Nilges, M., Overington, J. P., Bender, A., et al. (2016). Improved large-scale prediction of growth inhibition patterns using the NCI60 cancer cell line panel. *Bioinformatics* 32, 85–95. doi: 10.1093/bioinformatics/btv529

Costello, J. C., Heiser, L. M., Georgii, E., Gonen, M., Menden, M. P., Wang, N. J., et al. (2014). A community effort to assess and improve drug sensitivity prediction algorithms. *Nat. Biotechnol.* 32, 1202–1212. doi: 10.1038/nbt.2877

Daemen, A., Griffith, O. L., Heiser, L. M., Wang, N. J., Enache, O. M., Sanborn, Z., et al. (2013). Modeling precision treatment of breast cancer. *Genome Biol.* 14:R110. doi: 10.1186/gb-2013-14-10-r110

Garnett, M. J., Edelman, E. J., Heidorn, S. J., Greenman, C. D., Dastur, A., Lau, K. W., et al. (2012). Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* 483, 570–575. doi: 10.1038/nature11005

Heiser, L. M., Sadanandam, A., Kuo, W. L., Benz, S. C., Goldstein, T. C., Ng, S., et al. (2012). Subtype and pathway specific responses to anticancer compounds in breast cancer. *Proc. Natl. Acad. Sci. U.S.A.* 109, 2724–2729. doi: 10.1073/pnas.1018854108

Kim, S., Sundaresan, V., Zhou, L., and Kahveci, T. (2016). Integrating domain specific knowledge and network analysis to predict drug sensitivity of cancer cell lines. *PLoS ONE* 11:e0162173. doi: 10.1371/journal.pone.0162173

Koren, Y. (2010). Factor in the neighbors: scalable and accurate collaborative filtering. *ACM Trans. Knowl. Discov. Data* 4:24. doi: 10.1145/1644873.1644874

Lemire, D. (2005). Scale and translation invariant collaborative filtering systems. *Inf. Retr. Boston.* 8, 129–150. doi: 10.1023/B:INRT.0000048492.50961.a6

Menden, M. P., Iorio, F., Garnett, M., McDermott, U., Benes, C. H., Ballester, P. J., et al. (2013). Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. *PLoS ONE* 8:e61318. doi: 10.1371/journal.pone.0061318

Petrelli, F., Ghidini, M., Lonati, V., Tomasello, G., Borgonovo, K., Ghilardi, M., et al. (2017). The efficacy of lapatinib and capecitabine in HER-2 positive breast cancer with brain metastases: A systematic review and pooled analysis. *Eur. J. Cancer* 84, 141–148. doi: 10.1016/j.ejca.2017.07.024

Riddick, G., Song, H., Ahn, S., Walling, J., Borges-Rivera, D., Zhang, W., et al. (2011). Predicting *in vitro* drug sensitivity using Random Forests. *Bioinformatics* 27, 220–224. doi: 10.1093/bioinformatics/btq628

Shoemaker, R. H. (2006). The NCI60 human tumour cell line anticancer drug screen. *Nat. Rev. Cancer* 6, 813–823. doi: 10.1038/nrc1951

Shoemaker, R. H., Monks, A., Alley, M. C., Scudiero, D. A., Fine, D. L., McLemore, T. L., et al. (1988). Development of human tumor cell line panels for use in disease-oriented drug screening. *Prog. Clin. Biol. Res.* 276, 265–286.

Solit, D. B., Garraway, L. A., Pratilas, C. A., Sawai, A., Getz, G., Basso, A., et al. (2006). BRAF mutation predicts sensitivity to MEK inhibition. *Nature* 439, 358–362. doi: 10.1038/nature04304

Stanfield, Z., Coşkun, M., and Koyutürk, M. (2017). Drug response prediction as a link prediction problem. *Sci. Rep.* 7:40321. doi: 10.1145/3107411.3107459

Staunton, J. E., Slonim, D. K., Coller, H. A., Tamayo, P., Angelo, M. J., Park, J., et al. (2001). Chemosensitivity prediction by transcriptional profiling. *Proc. Natl. Acad. Sci. U.S.A.* 98, 10787–10792. doi: 10.1073/pnas.191368598

Wang, L., Li, X., Zhang, L., and Gao, Q. (2017). Improved anticancer drug response prediction in cell lines using matrix factorization with similarity regularization. *BMC Cancer* 17:513. doi: 10.1186/s12885-017-3500-5

Weinstein, J. N., Myers, T. G., O'Connor, P. M., Friend, S. H., Fornace, A. J. Jr., Kohn, K. W., et al. (1997). An information-intensive approach to the molecular pharmacology of cancer. *Science* 275, 343–349.

Yang, W., Soares, J., Greninger, P., Edelman, E. J., Lightfoot, H., Forbes, S., et al. (2013). Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.* 41, D955–D961. doi: 10.1093/nar/gks1111

Yuan, H., Paskov, I., Paskov, H., González, A. J., and Leslie, C. S. (2016). Multitask learning improves prediction of cancer drug sensitivity. *Sci. Rep.* 6:31619. doi: 10.1038/srep31619

Zhang, F., Wang, M., Xi, J., Yang, J., and Li, A. (2018). A novel heterogeneous network-based method for drug response prediction in cancer cell lines. *Sci. Rep.* 8:3355. doi: 10.1038/s41598-018-21622-4

Zhang, N., Wang, H., Fang, Y., Wang, J., Zheng, X., and Liu, X. S. (2015). Predicting anticancer drug responses using a dual-layer integrated cell line-drug network model. *PLoS Comput. Biol.* 11:e1004498. doi: 10.1371/journal.pcbi.1004498

Zhao, M., Howard, E. W., Parris, A. B., Guo, Z., Zhao, Q., Ma, Z., et al. (2017). Activation of cancerous inhibitor of PP2A (CIP2A) contributes to lapatinib resistance through induction of CIP2A-Akt feedback loop in ErbB2-positive breast cancer cells. *Oncotarget* 8, 58847–58864. doi: 10.18632/oncotarget.19375

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewers YW and CW and the handling Editor declared their shared affiliation.

ORIGINAL RESEARCH
published: 15 October 2018
doi: 10.3389/fphar.2018.01152

Prediction of Potential Small Molecule-Associated MicroRNAs Using Graphlet Interaction

Na-Na Guan[1], Ya-Zhou Sun[1], Zhong Ming[1,2], Jian-Qiang Li[1]* and Xing Chen[3]*

[1] College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China, [2] National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University, Shenzhen, China, [3] School of Information and Control Engineering, China University of Mining and Technology, Xuzhou, China

MicroRNAs (miRNAs) have been proved to be targeted by the small molecules recently, which made using small molecules to target miRNAs become a possible therapy for human diseases. Therefore, it is very meaningful to investigate the relationships between small molecules and miRNAs, which is still yet in the newly-developing stage. In this paper, we presented a prediction model of Graphlet Interaction based inference for Small Molecule-MiRNA Association prediction (GISMMA) by combining small molecule similarity network, miRNA similarity network and known small molecule-miRNA association network. This model described the complex relationship between two small molecules or between two miRNAs using graphlet interaction which consists of 28 isomers. The association score between a small molecule and a miRNA was calculated based on counting the numbers of graphlet interaction throughout the small molecule similarity network and the miRNA similarity network, respectively. Global and two types of local leave-one-out cross validation (LOOCV) as well as five-fold cross validation were implemented in two datasets to evaluate GISMMA. For Dataset 1, the AUCs are 0.9291 for global LOOCV, 0.9505, and 0.7702 for two local LOOCVs, 0.9263 ± 0.0026 for five-fold cross validation; for Dataset 2, the AUCs are 0.8203, 0.8640, 0.6591, and 0.8554 ± 0.0063, in turn. In case study for small molecules, 5-Fluorouracil, 17β-Estradiol and 5-Aza-2′-deoxycytidine, the numbers of top 50 miRNAs predicted by GISMMA and validated to be related to these three small molecules by experimental literatures are in turn 30, 29, and 25. Based on the results from cross validations and case studies, it is easy to realize the excellent performance of GISMMA.

Keywords: small molecule, microRNA, association prediction, graphlet interaction, similarity calculation

OPEN ACCESS

Edited by:
Zhi-Liang Ji,
Xiamen University, China

Reviewed by:
Quan Zou,
Tianjin University, China
Feixiong Cheng,
Cleveland Clinic, United States
Francesco Pappalardo,
Università degli Studi di Catania, Italy

Correspondence:
Jian-Qiang Li
lijq@szu.edu.cn
Xing Chen
xingchen@amss.ac.cn;
xingchen@cumt.edu.cn

Specialty section:
This article was submitted to Translational Pharmacology, a section of the journal Frontiers in Pharmacology

Received: 08 May 2018
Accepted: 24 September 2018
Published: 15 October 2018

Citation:
Guan N-N, Sun Y-Z, Ming Z, Li J-Q and Chen X (2018) Prediction of Potential Small Molecule-Associated MicroRNAs Using Graphlet Interaction. Front. Pharmacol. 9:1152. doi: 10.3389/fphar.2018.01152

INTRODUCTION

MicroRNAs (miRNAs) are a family of small non-coding RNAs, having about 22 nucleotides in length, which regulate gene expression at a post-transcriptional level (Ambros, 2003). The first miRNA was discovered over 30 years ago in the *Caenorhabditis elegans*. Subsequently, thousands of miRNAs have been discovered in many organisms, and there are currently 2588 annotated miRNAs in the human genome (Kozomara and Griffiths-Jones, 2014). MiRNAs can simultaneously regulate the expression of hundreds of genes due to the fact that their nucleotide pairing by complementarity is imperfect (He and Hannon, 2004). In this manner, they play a critical role in a variety of crucial processes such as tissue development, morphogenesis, apoptosis, signal transduction pathways, etc., (Esquela-Kerscher and Slack, 2006; Spizzo et al., 2009; Wang and Lee, 2009). This additionally implicates them in an array of disease associated processes. The development of large-expression screens has been proven useful in identifying novel miRNAs involved in diseases, which could potentially become an attractive therapeutic target (Monroig and Calin, 2013; Chen et al., 2017a, 2018a,b,c; Matsui and Corey, 2017).

Regulation of miRNAs by small molecules is an efficient mean to modulate endogenous miRNA function and to treat miRNA-related diseases (Xia et al., 2015). Small molecules have been thoroughly used with clinical applications for numerous diseases (Zhang et al., 2009). However, drug discovery and development are currently an extremely long process, which takes approximately 10–15 years (Monroig Pdel et al., 2015). Also, drug production results in an incredible economic burden and patients end up having to pay exaggerated prices for their treatments (Chen et al., 2015; Monroig Pdel et al., 2015). The use of chemical compounds that are already FDA approved to treat a specific disease would accelerate the process of completing toxicological studies and clinical trials in order to apply them to other diseases. It would shorten both money expenses and time consuming processes.

As miRNAs have been associated with many diseases (Chen et al., 2017b), the development of small-molecule drugs targeting specific miRNAs seems to be a promising approach to meet the challenge (Monroig Pdel et al., 2015). Small molecule may modulate the expression of miRNAs by either activating or repressing their transcription (Xia et al., 2015). Transcriptional inhibitors were identified by completing a small molecule screen in which a 3′ UTR complementary to miR-21 was inserted into a luciferase mRNA reporter (Gumireddy et al., 2008). This study identified a type of diazobenzene as miR-21 transcriptional inhibitors (Gumireddy et al., 2008). Small molecules were also discovered to modulate transcription of miR-122, a highly expressed and liver-specific miRNA whose aberrant expression is associated with hepatocellular carcinoma (Thomas and Deiters, 2013). Two small molecules that inhibit transcription and another small molecule that promotes transcription of pri-miR-122 were identified using a luciferase reporter system (Thomas and Deiters, 2013). The examples above show that miRNA expression can be altered with small molecules, providing promise to expand miRNAs from diagnostic signatures of disease to therapeutic targets. Therefore, the prediction of associations between small molecules and miRNAs could promote the drug repurposing for miRNA-related diseases. Besides, since the regulation of miRNA expression can be caused by targeting miRNAs directly (Zhang et al., 2010) or by targeting the relative proteins (Lim et al., 2016), identifying the small molecule-miRNA associations would be conductive to the drug discovery. However, experimental methods to study the small molecular-miRNA association are expensive and time-consuming, which makes it urgent to develop computational approaches to provide reliable predictions that can give some guidance to experiments.

Recently, several computational models have been proposed to investigate the relations between small molecules and miRNAs. For example, Jiang et al. (2012) proposed a high-throughput method to investigate the biological connections between small molecules and miRNAs in 23 human cancers based on transcriptional responses, which was the first model to systematically study the associations between bioactive small molecules and miRNAs. They constructed a complex Small molecule and MiRNA Network (SMirN) for each cancer and explored the molecular and functional features for small molecule modules, as well as miRNA modules for each cancer type. Each module of small molecular was linked to a miRNA, and each module of miRNA was connected with one small molecular. One of the advantages of this method is that it does not need to know the information of small molecule structure or miRNA structure in advance. However, the reliability of the approach was limited due to the small data of transcriptional response to genome-wide miRNA perturbations. Furthermore, Meng et al. (2014) built a bioactive Small molecule and miRNA association Network in Alzheimer's Disease (SmiRN-AD) through comparing the gene expression profiles after bioactive small molecule treating with the AD-related miRNA (ADM) regulating expressions, to get the scores of associations between small molecules and ADMs. Besides, the positive and negative associations were identified to investigate the biological insights of the SimRN-AD. Recently, Wang et al. (2016) developed another method to identify small molecule-miRNA associations based on their functional similarity. They searched the functional link of each small molecule-miRNA pair by calculating Gene Ontology enrichment after identifying differentially expressed genes for small molecules and miRNAs. Compared with previous models based on transcriptional responses, this method is more repeatable by using functional associations. Additionally, Lv et al. (2015) presented a novel computational model to predict potential associations between small molecules and miRNAs. They implemented the random walk with restart algorithm on an comprehensive network, which was established by combining small molecule similarity, miRNA similarity, as well as known small molecule-miRNA associations. Especially, this model can predict the novel related miRNAs for small molecules without any known associated miRNAs. However, it has too many adjustable parameters that need to be affirmed. Moreover, Li et al. (2016) developed a network based framework called predictive Small Molecule-miRNA Network-Based Inference (SMiR-NBI), to investigate the underlying regulations of anticancer drugs on miRNAs. This model constructed a heterogeneous network that was composed of drugs, miRNAs and genes to conduct a network based algorithm. It is mentionable that the accuracy of this method is quite high even it only depended on the network topology information. However, SMiR-NBI could not be applied to prediction of isolated miRNAs that have no interlinked small molecules. Besides, it failed to predict potential miRNAs associated with small molecules that had different dose-responses, due to lack of known data.

So far, the number of computational models is still not satisfying for the prediction of novel associations between small molecules and miRNAs. Moreover, there are still some limitations existing in the previous models. In order to predict potential small molecule-miRNA associations more effectively and reliably, in this paper, we presented the Graphlet Interaction based inference for Small Molecule-MiRNA Association prediction (GISMMA). In this model, the similarity of small molecules and the similarity of miRNAs were combined with known associations between small molecules and miRNAs in two different datasets, which were labeled with Dataset 1 and Dataset 2. In Dataset 1, only a fraction of small molecules and miRNAs were involved in known small molecule-miRNA associations, whereas in Dataset 2 all small molecules and miRNAs were implicated in known small molecule-miRNA associations. Based on the measuring of graphlet interaction between any two nodes on the network of small molecules and on the network of miRNAs, respectively, we can compute the correlation scores of small molecule-miRNA pairs. We have implemented leave-one-out cross validation (LOOCV) and five-fold cross validation to evaluate the performance of GISMMA. The AUCs of global LOOCV are 0.9291 and 0.8203 for Dataset 1 and Dataset 2, respectively; the AUCs of local LOOCV by ranking the small molecules for each fixed miRNA are, respectively 0.9505 and 0.8640 for the two datasets; the AUCs of local LOOCV by ranking the miRNAs for each fixed small molecule are, respectively 0.7702 and 0.6591 for the two datasets. And the average AUCs and standard deviations of five-fold cross validations are 0.9263 ± 0.0026 and 0.8088 ± 0.0044 for the two datasets, respectively. In case study, small molecule was set as a new one by turning all known related miRNAs into unknown ones. GISMMA was then applied to predicting latent related miRNAs for each small molecule based on the Dataset 1. For the small molecules, 5-Fluorouracil, 17β-Estradiol and 5-Aza-2′-deoxycytidine, there were in turn 30, 29, and 25 out of top 50 predicted miRNAs, which were validated to be associated with these three small molecules by experimental literatures, respectively. The results both in cross validations and case studies have suggested that GISMMA is a powerful and reliable model to predict novel associations between small molecules and miRNAs.

MATERIALS AND METHODS

Small Molecule-miRNA Associations

In this paper, we obtained the known small molecule-miRNA associations from SM2miR (Version 1) (Liu et al., 2013). The total number of known associations is 664. For comparison of model performance on different datasets, we have constructed two datasets. Dataset 1 consists of 831 small molecules extracted and integrated from SM2miR, DrugBank (Knox et al., 2011) and PubChem (Wang et al., 2009), and 541 miRNAs that were collected from SM2miR, HMDD (Lu et al., 2008), miR2Disease (Jiang et al., 2009) and PhenomiR (Jiang et al., 2009; Ruepp et al., 2010). In Dataset 1, there are only 39 small molecules and 286 miRNAs implicated in the 664 known associations, while 792 small molecules and 255 miRNAs are completely new ones without any known associations. Dataset 2 is only composed of those 39 small molecules and 286 miRNAs, which are involved in the known associations. Based on the known data, an adjacency matrix A was constructed to represent the relations between small molecules and miRNAs, in which $A(i, j)$ was set to be 1 if there is an association between small molecule $s(i)$ and miRNA $m(j)$, 0 otherwise.

Small Molecule Similarity

In this paper, according to the method proposed in (Lv et al., 2015), the small molecule similarity was calculated by integrating four usual small molecule similarities which were side effect based similarity that was computed by Jaccard score using small molecule side effect dataset (Gottlieb et al., 2011), functional consistency based similarity that was obtained by comparing the function of small molecule target genes (Lv et al., 2012), chemical structure based similarity that was calculated with the method of chemical structure comparison between any two small molecules (Hattori et al., 2003), and indication phenotype based similarity that was constructed through identifying phenotype similarity between small molecule related diseases (Gottlieb et al., 2011). Therefore, the integrated similarity of small molecules can be computed with the following formula:

$$SS = \frac{\beta_1 S_S^D + \beta_2 S_S^T + \beta_3 S_S^C + \beta_4 S_S^S}{\sum_{i=1}^{4} \beta_i} \tag{1}$$

where, $S^D{}_S$, $S^T{}_S$, $S^C{}_{S,}$ and $S^S{}_S$ denote the four different similarity types, respectively, i.e., indication phenotype based similarity, functional consistency based similarity, chemical structure based similarity and side effect based similarity, and β_i $(i = 1, 2, 3, 4)$ are the weighs used to balance the different similarity contributions, whose default values were all set as 1.

MiRNA Similarity

The miRNA similarity we used in this paper was established using the method in (Lv et al., 2015), by combining functional consistency based similarity that was calculated by comparing the function of miRNA target genes (Lv et al., 2012) and indication phenotype based similarity that was computed by measuring phenotype similarity between diseases associated with miRNAs (Gottlieb et al., 2011). Similarly, to reduce the bias of each similarity measurement, the integrated similarity of miRNAs was defined as follows:

$$SM = \frac{\alpha_1 S_M^D + \alpha_2 S_M^T}{\sum_{j=1}^{2} \alpha_j} \tag{2}$$

where, $S^D{}_M$ is the indication phenotype based similarity and $S^T{}_M$ represents the functional consistency based similarity, and α_j $(j = 1, 2)$ are the weighs of each similarity measurement, which were both set as 1.

GISMMA

In this study, by integrating small molecule similarity, miRNA similarity and known associations between small molecules and miRNAs, we developed a graphlet interaction based method to predict the potential associations between small molecules and miRNAs, which is motivated by the study of Wang et al. (2014). Prediction code of our model is available

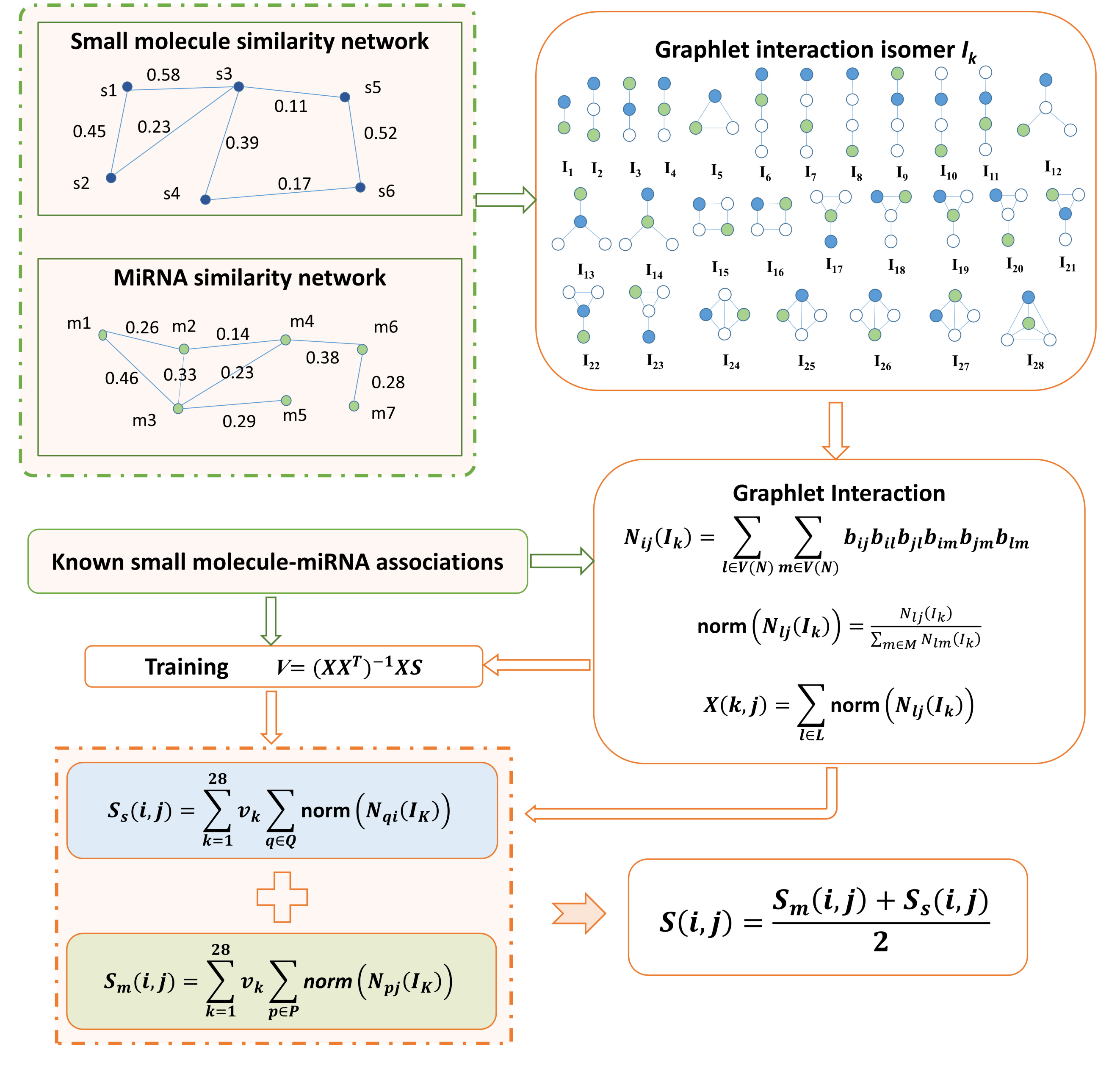

Small molecule similarity network
s1
0.58
s3
s5
0.11
0.45
0.23
0.52
0.39
s2
s4
0.17
s6
MiRNA similarity network
m1
0.26
m2
0.14
m4
m6
0.38
0.46
0.33
0.23
0.28
m3
0.29
m5
m7
Graphlet interaction isomer I_k
I_1 I_2 I_3 I_4 I_5 I_6 I_7 I_8 I_9 I_{10} I_{11} I_{12}
I_{13} I_{14} I_{15} I_{16} I_{17} I_{18} I_{19} I_{20} I_{21}
I_{22} I_{23} I_{24} I_{25} I_{26} I_{27} I_{28}
Graphlet Interaction
$N_{ij}(I_k) = \sum_{l\in V(N)} \sum_{m\in V(N)} b_{ij}b_{il}b_{jl}b_{im}b_{jm}b_{lm}$
$\text{norm}\left(N_{lj}(I_k)\right) = \frac{N_{lj}(I_k)}{\sum_{m\in M} N_{lm}(I_k)}$
$X(k,j) = \sum_{l\in L} \text{norm}\left(N_{lj}(I_k)\right)$
Known small molecule-miRNA associations
Training $V = (XX^T)^{-1}XS$
$S_s(i,j) = \sum_{k=1}^{28} v_k \sum_{q\in Q} \text{norm}\left(N_{qi}(I_K)\right)$
$S_m(i,j) = \sum_{k=1}^{28} v_k \sum_{p\in P} norm\left(N_{pj}(I_K)\right)$
$S(i,j) = \frac{S_m(i,j) + S_s(i,j)}{2}$

FIGURE 1 | Flowchart of GISMMA model based on graphlet interaction for the prediction of potential small molecule-miRNA associations.

at: https://github.com/AnnaGuan/GISMMA/tree/AnnaGuan-patch-1. The concept of graphlet interaction is traced to the definition in (Wang et al., 2014), which describes the relationship between any two nodes in a graphlet that is a type of subgraph in a large network. As was done in (Wang et al., 2014), in GISMMA only those graphlets that have 1 to 4 nodes were used, based on which 28 graphlet interaction isomers were constructed, denoted by labels I_1 to I_{28} in **Figure 1**. The graphlet interaction isomer depends on the positions of the two involved nodes, which means that the graphlet interaction between two nodes have two different set of isomers. Through counting the number of each isomer, we can represent the graphlet interaction between any two nodes in a network with a vector that contains 28 numbers (Przulj, 2007; Wang et al., 2014).

We have created a network *NS* to represent the small molecule similarity and a network *NM* to represent the miRNA similarity, where each node in the network denotes a small molecule or a miRNA. The edge with similarity value as its weight exists to link any two nodes that have similarity. The associations between small molecules and miRNAs were investigated in the two similarity networks *NS* and *NM*, respectively.

In the miRNA network *NM*, the number of isomer I_k for graphlet interaction from node $m(i)$ to node $m(j)$ can be calculated as follows (Wang et al., 2014):

$$N_{ij}\left(I_k\right) = \sum\nolimits_{l \in \mathrm{V(NM)}} \sum\nolimits_{m \in V(NM)} b_{ij} b_{il} b_{jl} b_{im} b_{jm} b_{lm} \quad (3)$$

where $V(NM)$ denotes the node set of all nodes in network NM, l, and m are two nodes different with node $m(i)$ and $m(j)$, and b is defined as:

$$b_{st} = \begin{cases} a_{st} & s \text{ and } t \text{ has a link in } I_k \\ 1 - a_{st} & s \text{ and } t \text{ has no link in } I_k \end{cases} \quad (4)$$

where, a_{st} is the edge weight assigned with the similarity value of $m(s)$ and $m(t)$. Especially, a_{st} is 0 when nodes $m(s)$ and $m(t)$ have no connection. Then we normalized the graphlet interaction as follows:

$$\mathrm{norm}\left(N_{ij}\left(I_k\right)\right) = \frac{N_{ij}\left(I_k\right)}{\sum_{m \in M} N_{im}\left(I_k\right)} \quad (5)$$

where M contains all other nodes but $m(i)$. Based on the normalized form in equation (5), we can compute the association score of a small molecule-miRNA pair as follows:

$$S_m\left(i,j\right) = \sum\nolimits_{k=1}^{28} v_k \sum_{p \in P(i)} \mathrm{norm}\left(N_{pj}\left(I_k\right)\right) \quad (6)$$

where i denotes a small molecule $s(i)$ and j denotes a miRNA $m(j)$, v_k is the weight of the kth isomer, $P(i)$ is the set of miRNAs with known associations with small molecule $s(i)$. By defining the summation of norm in equation (6) as following:

$$X_m\left(k,j\right) = \sum_{p \in P(i)} \mathrm{norm}\left(N_{pj}\left(I_k\right)\right) \quad (7)$$

we can modify equation (6) into the matrix form as following:

$$S_m = X_m^T V_m \quad (8)$$

The weight coefficients V_m can be learnt from known associations by performing a simple linear regression (Wang et al., 2014), which is given as following:

$$V_m = \left(X_m X_m^T\right)^{-1} X_m S_m \quad (9)$$

We computed the number of graphlet interaction isomer between two small molecules in the similar way as described in equations (3–5). Then the association score between small molecule $s(i)$ and miRNA $m(j)$ can be calculated in the small molecule network NS as follows:

$$S_s\left(i,j\right) = \sum\nolimits_{k=1}^{28} v_k \sum_{q \in Q(j)} \mathrm{norm}\left(N_{qi}\left(I_k\right)\right) \quad (10)$$

where $Q(j)$ is the set of small molecules that have known associations with miRNA $m(j)$. Also, the term of summation of norm in equation (10) can be defined with the matrix:

$$X_s\left(k,j\right) = \sum_{q \in Q(j)} \mathrm{norm}\left(N_{qi}\left(I_k\right)\right) \quad (11)$$

Thus equation (10) was rewritten as:$S_S = X^T{}_S V_S$, and the undetermined matrix V_s can be obtained by training the model with known association scores:

$$V_s = \left(X_s X_s^T\right)^{-1} X_s S_s \quad (12)$$

Finally, we calculated the association score between small molecule $s(i)$ and miRNA $m(j)$ by combining the scores from NM and NS in a simple average form as following:

$$S\left(i,j\right) = \frac{S_m\left(i,j\right) + S_s\left(i,j\right)}{2} \quad (13)$$

RESULTS

Performance Evaluation

In this work, two commonly used methods, LOOCV and five-fold cross validation, were implemented to evaluate the performance of GISMMA based on Dataset 1 and Dataset 2, respectively. The LOOCV has three different types including global LOOCV, local LOOCV of ranking small molecules for fixed miRNA and local LOOCV of ranking miRNAs for fixed small molecule. Each confirmed association we collected was taken as the test sample one by one and the rest of known associations were considered as the training samples in LOOCV. Candidate samples in global LOOCV consist of all the small molecule-miRNA pairs that have no known associations. In the case of local, we only consider those small molecules that do not relate to the fixed miRNA or those miRNAs unconnected to the fixed small molecule in the test sample as candidates. The scores as association probabilities were computed using the GISMMA method for both test sample and all candidate samples. Then we ranked them for the corresponding type of LOOCV. The five-fold cross validation was performed in the following steps. Firstly, all the known small molecule-miRNA associations were randomly split into five parts with equal size. Secondly, the five parts take turns to act as the test sample set one after another and the other four parts as the training sample sets; similarly, all small molecule-miRNA pairs that have no known associations play the roles of candidate samples. Thirdly, the test samples as well as the candidate samples were endowed with association scores by GISMMA. Finally, each test sample was picked out in turn to be compared with candidate samples according to their scores. The model was considered to be successfully predict the test sample only when its rank exceeded the given rank threshold.

Based on the ranking, the receiver operating characteristic (ROC) curves were used to illustrate the results of the three types of LOOCV described above, in which the abscissa axis is true positive rate (TPR, sensitivity) and the ordinate axis represents false positive rate (FPR, 1-specificity) for different thresholds given in advance. The sensitivity means the ratio that the positive samples rank above the given threshold, while the specificity is defined as the percentage of candidate samples whose ranks are below the set threshold. The area

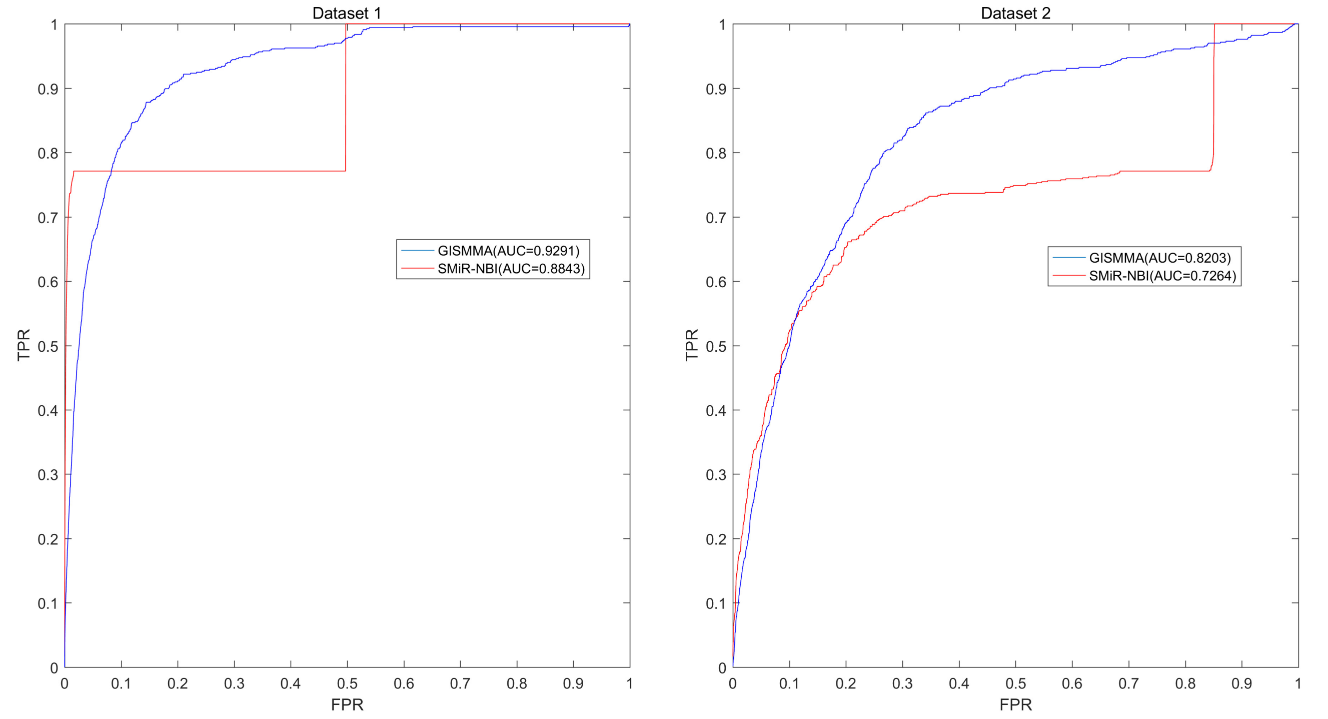


FIGURE 2 | Performance of GISMMA was compared with SMiR-NBI in terms of ROC curve and AUC of global LOOCV for Dataset 1 (left) and Dataset 2 (right). As is shown, GISMMA achieves AUCs of 0.9291 and 0.8203 for Dataset 1 and Dataset 2, respectively, significantly superior to the previous model SMiR-NBI.

under the ROC curve (AUC) was correspondingly calculated to estimate the reliability of the GISMMA. When the model correctly predicts all test samples, AUC = 1; but if the model has a random prediction, AUC = 0.5. To make comparison with previous method, we implemented SMiR-NBI (Li et al., 2016) for global and two types of local LOOCVs, 5-fold cross validation based on the same datasets. The global AUCs of GISMMA for Dataset 1 and Dataset 2 are 0.9291 and 0.8203, respectively, which are shown in **Figure 2** in comparison with previous model SMiR-NBI whose results are 0.8843 and 0.7264, respectively. In the case of local LOOCV of ranking small molecules for fixed miRNA, the AUCs of GISMMA for Dataset 1 and Dataset 2 are 0.9505 and 0.8640, respectively, compared with 0.8837 and 0.7846 of SMiR-NBI, which can be seen in **Figure 3**. The results of local LOOCV of ranking miRNAs for fixed small molecule are shown in **Figure 4**, from which we can see that the AUCs of GISMMA and SMiR-NBI are 0.7702, 0.7497 for Dataset 1, and 0.6591, 0.6100 for Dataset 2, respectively. Besides, in five-fold cross validation, the average AUCs with standard deviations of GISMMA and SMiR-NBI are 0.9263 ± 0.0026, 0.8554 ± 0.0063 for Dataset 1, and 0.8088 ± 0.0044, 0.7104 ± 0.0087 for Dataset 2. The **Table 1** lists the comparison of GISMMA and SMiR-NBI for all AUC results of the four types of cross validations on two datasets. We can make a conclusion from the comparisons that the novel method proposed in this work is more reliable and more effective in predicting potential associations between small molecules and miRNAs.

Case Study

Based on the known database and published references in PubMed database, we studied three common small molecules to further evaluate the predictive ability of GISMMA, in which the small molecule in study was set as a new one by taking away its known associations. We ulteriorly observed the number of the experimentally verified miRNAs in the top 50 ones predicted to be related to the three small molecules, respectively.

The small molecular 5-Fluorouracil (5-FU) is a widely used chemotherapeutic drug in colorectal cancer (Windle et al., 1987). For a long time, the 5-FU-induced cytotoxic effects were thought to result exclusively from its impact on DNA metabolism (Andreuccetti et al., 1996; Airley, 2009). However, several evidences indicated that the cytotoxic effect of 5-FU also results from its capacity to alter RNA metabolism and mRNA expression (Longley et al., 2003). Exposure to 5-FU promotes a profound transcriptional reprogramming leading to modification of mRNA and miRNAs expression profiles that contributes in modifying cell fate (Hernandez-Vargas et al., 2006; Rossi et al., 2007; Shah et al., 2011). After implementing GISMMA, we got the total ranking of potential miRNAs associated with 5-FU. As the result shown, among the top 10 and 50 potential 5-FU-related miRNAs, there were 8 and 30 miRNAs confirmed by experiments, respectively (See **Table 2**). For instance, miR-21 and miR-23a were predicted as the first and fifth candidates for 5-FU, respectively, which were significantly down regulated in comparison between 5-FU treated and control samples in miRNA microarray analysis of 5-FU treated MCF-7 cells (Shah et al.,

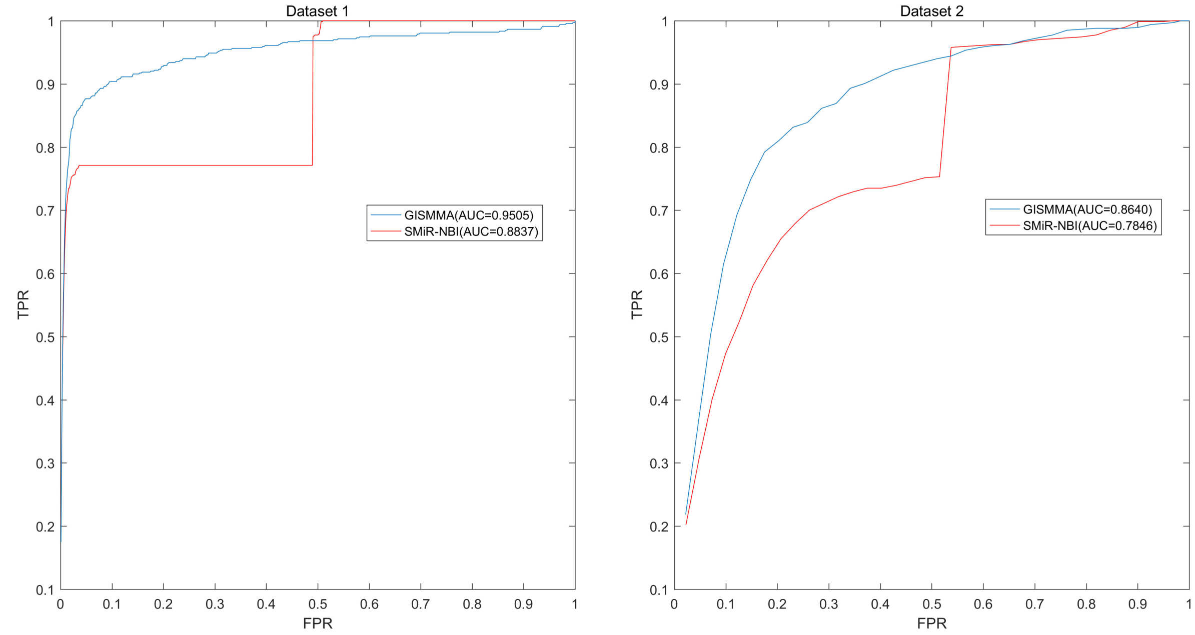


FIGURE 3 | Performance of GISMMA was compared with SMiR-NBI in terms of ROC curve and AUC of local LOOCV of ranking small molecules for fixed miRNA on Dataset 1 (left) and Dataset 2 (right). As is shown, GISMMA achieves AUCs of 0.9505 and 0.8640 for Dataset 1 and Dataset 2, respectively, significantly superior to the previous model SMiR-NBI.

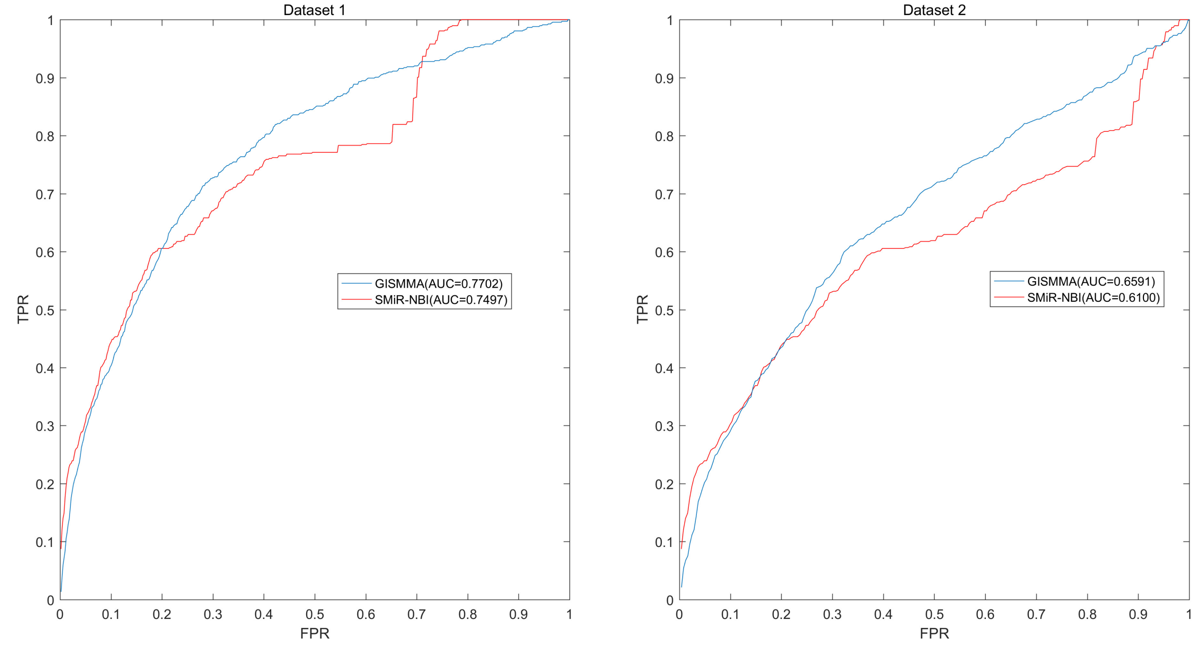


FIGURE 4 | Performance of GISMMA was compared with SMiR-NBI in terms of ROC curve and AUC of local LOOCV of ranking miRNAs for fixed small molecule on Dataset 1 (left) and Dataset 2 (right). As is shown, GISMMA achieves AUCs of 0.7702 and 0.6591 for Dataset 1 and Dataset 2, respectively, significantly superior to the previous model SMiR-NBI.

2011). Besides, miR-24-1, the third candidate in the ranking list, showed a significantly down regulation in HCT-8 colon cancer cell after exposure to 5-FU (Zhou et al., 2010). In addition, MiR-27b that ranked the fourth in the prediction list of 5-FU was found to be consistently up regulated in human colon cancer cells HC.21 following exposure to 5-FU *in vitro* (Rossi et al., 2007).

The small molecular 17β-Estradiol (E2) is the principal intracellular human estrogen that exerts important effects on

TABLE 1 | The comparison results between GISMMA and SMiR-NBI on AUC values of four cross validations based on two datasets.

DATASET	MODEL	GLOBAL LOOCV	LOCAL LOOCV (fix miRNA)	LOCAL LOOCV (fix SM)	5-FOLD CV
Dataset 1	GISMMA	0.9291	0.9505	0.7702	0.9263 ± 0.0026
	SMiR-NBI	0.8843	0.8837	0.7497	0.8554 ± 0.0063
Dataset 2	GISMMA	0.8203	0.8640	0.6591	0.8088 ± 0.0044
	SMiR-NBI	0.7264	0.7846	0.6100	0.7104 ± 0.0087

the reproductive as well as many other organ systems in both men and women (Simpson and Santen, 2015). The analogs of estradiol exhibit significant anticancer activity against human breast cancer cell lines (Sathish Kumar et al., 2014). Estrogens have associations with cancer in target tissues, which is because they have a phenolic ring structure in common with the carcinogenic hydrocarbons (Ryan, 1982). After implementing GISMMA, we got the total ranking of the E2-associated miRNAs. As the result shown, among the top 10 and 50 potential E2-related miRNAs, there were 5 and 29 miRNAs confirmed by experiments, respectively (See **Table 3**). For example, miR-21, miR-27b, and miR-23a dominated in turn the first, fourth, and fifth places of the ranking list predicted for E2, which were all down regulated after treatment of MCF-7 cells with E2 (Bhat-Nakshatri et al., 2009; Tilghman et al., 2012). Besides, E2 showed a capacity to down regulate the expression level of miR-21 in breast cancer cells (Selcuklu et al., 2012).

The small molecular 5-Aza-2′-deoxycytidine (5-Aza-CdR) is a nucleoside analog inhibitor of DNA methyltransferase (DNMT). It has been used to reverse methylation and reactivate the expression of silenced genes (Patra and Bettuzzi, 2009). 5-Aza-CdR is able to suppress the growth of various tumors *in vitro*, animal models, and clinical trials including prostate cancer (Hurtubise and Momparler, 2004; Issa et al., 2004; McCabe et al., 2006). We performed GISMMA on 5-Aza-CdR, and got the total ranking of the predicted miRNAs. As the result shown, among the top 10 and 50 potential 5-Aza-CdR related miRNAs, there were 7 and 25 miRNA-5-Aza-CdR associations confirmed by experiments (See **Table 4**). For example, in the ranking list of miRNAs predicted for 5-Aza-CdR, miR-21, and miR-27b were

TABLE 2 | Top 50 miRNAs associated with 5-Fluorouracil were predicted by GISMMA based on Dataset 1.

miRNA	Evidence	miRNA	Evidence
hsa-mir-21	26198104	hsa-mir-22	25449431
hsa-mir-324	unconfirmed	hsa-mir-409	unconfirmed
hsa-mir-24-1	26198104	hsa-mir-337	unconfirmed
hsa-mir-27b	26198104	hsa-let-7a-3	26198104
hsa-mir-23a	26198104	hsa-let-7a-2	26198104
hsa-mir-638	26198104	hsa-mir-155	28347920
hsa-mir-27a	26198104	hsa-mir-181b-2	unconfirmed
hsa-let-7b	25789066	hsa-mir-181b-1	unconfirmed
hsa-mir-181a-1	unconfirmed	hsa-mir-15b	26198104
hsa-mir-126	26062749	hsa-let-7i	unconfirmed
hsa-mir-125b-2	unconfirmed	hsa-mir-320a	26198104
hsa-mir-125b-1	unconfirmed	hsa-mir-26a-2	unconfirmed
hsa-mir-124-3	unconfirmed	hsa-mir-328	unconfirmed
hsa-mir-124-2	unconfirmed	hsa-mir-16-2	26198104
hsa-mir-124-1	unconfirmed	hsa-let-7e	26198104
hsa-let-7a-1	26198104	hsa-mir-34b	unconfirmed
hsa-mir-181a-2	24462870	hsa-mir-145	24447928
hsa-mir-24-2	26198104	hsa-mir-200b	26198104
hsa-mir-17	26198104	hsa-let-7c	25951903
hsa-mir-26a-1	unconfirmed	hsa-mir-874	27221209
hsa-mir-16-1	26198104	hsa-mir-650	unconfirmed
hsa-mir-518c	unconfirmed	hsa-mir-501	26198104
hsa-mir-99b	unconfirmed	hsa-mir-500a	unconfirmed
hsa-mir-18a	26198104	hsa-mir-1226	26198104
hsa-mir-663a	26198104	hsa-mir-200c	26198104

The top 1-25 miRNAs are shown in the first column while the top 26–50 in the second. As a result, 8 and 30 out of top 10 and top 50 were confirmed by the known experimental literatures, respectively.

TABLE 3 | Top 50 miRNAs associated with 17β-Estradiol were predicted by GISMMA based on Dataset 1.

miRNA	Evidence	miRNA	Evidence
hsa-mir-21	26198104	hsa-mir-222	24601884
hsa-mir-324	unconfirmed	hsa-mir-31	23143558
hsa-mir-24-1	unconfirmed	hsa-mir-125a	21914226
hsa-mir-27b	26198104	hsa-mir-663a	26198104
hsa-mir-23a	26198104	hsa-mir-22	24715036
hsa-mir-638	26198104	hsa-mir-132	26282993
hsa-mir-27a	26198104	hsa-mir-501	unconfirmed
hsa-mir-181a-1	unconfirmed	hsa-mir-1226	unconfirmed
hsa-mir-24-2	unconfirmed	hsa-mir-328	unconfirmed
hsa-mir-125b-2	unconfirmed	hsa-mir-155	23568502
hsa-mir-125b-1	unconfirmed	hsa-let-7a-3	26198104
hsa-mir-16-1	unconfirmed	hsa-let-7a-2	26198104
hsa-mir-124-3	26198104	hsa-mir-181b-2	unconfirmed
hsa-mir-124-2	26198104	hsa-mir-181b-1	unconfirmed
hsa-mir-124-1	26198104	hsa-mir-26a-2	unconfirmed
hsa-mir-18a	24245576	hsa-mir-15b	26198104
hsa-let-7b	26198104	hsa-mir-20a	21914226
hsa-mir-181a-2	unconfirmed	hsa-mir-29a	22334722
hsa-let-7a-1	26198104	hsa-mir-19a	unconfirmed
hsa-mir-17	26198104	hsa-mir-200b	26198104
hsa-mir-126	26198104	hsa-mir-221	21057537
hsa-mir-26a-1	unconfirmed	hsa-mir-518c	26198104
hsa-mir-320a	27965096	hsa-mir-194-2	unconfirmed
hsa-mir-16-2	unconfirmed	hsa-mir-181d	unconfirmed
hsa-mir-99b	unconfirmed	hsa-mir-197	unconfirmed

The top 1–25 miRNAs are shown in the first column while the top 26–50 in the second. As a result, 5 and 29 out of top 10 and top 50 were confirmed by the known databases or experimental literatures, respectively.

TABLE 4 | Top 50 miRNAs associated with 5-Aza-2′-deoxycytidine were predicted by GISMMA based on Dataset 1.

miRNA	Evidence	miRNA	Evidence
hsa-mir-21	26198104	hsa-mir-518c	unconfirmed
hsa-mir-324	unconfirmed	hsa-mir-200b	23626803
hsa-mir-23a	unconfirmed	hsa-let-7d	26802971
hsa-mir-24-1	26198104	hsa-mir-501	unconfirmed
hsa-mir-27b	26198104	hsa-mir-1226	unconfirmed
hsa-mir-27a	26198104	hsa-mir-200c	23626803
hsa-mir-638	26198104	hsa-mir-99b	unconfirmed
hsa-let-7a-1	unconfirmed	hsa-mir-181a-2	26198104
hsa-mir-124-3	23200812	hsa-let-7e	22053057
hsa-mir-124-2	23200812	hsa-mir-132	unconfirmed
hsa-mir-124-1	unconfirmed	hsa-mir-203a	26577858
hsa-let-7b	26708866	hsa-mir-409	unconfirmed
hsa-mir-18a	unconfirmed	hsa-mir-337	unconfirmed
hsa-mir-24-2	26198104	hsa-mir-1915	unconfirmed
hsa-mir-17	26198104	hsa-mir-128-2	unconfirmed
hsa-mir-181a-1	26198104	hsa-mir-128-1	unconfirmed
hsa-mir-663a	unconfirmed	hsa-mir-320a	26198104
hsa-let-7a-3	26227220	hsa-mir-181b-2	unconfirmed
hsa-let-7a-2	unconfirmed	hsa-mir-181b-1	unconfirmed
hsa-mir-126	26198104	hsa-mir-222	unconfirmed
hsa-mir-26a-1	unconfirmed	hsa-mir-26a-2	unconfirmed
hsa-mir-15b	unconfirmed	hsa-mir-328	unconfirmed
hsa-mir-16-1	26198104	hsa-mir-16-2	26198104
hsa-mir-125b-2	26198104	hsa-mir-29a	26198104
hsa-mir-125b-1	26198104	hsa-let-7c	unconfirmed

The top 1–25 miRNAs are shown in the first column while the top 26–50 in the second. As a result, 7 and 25 out of top 10 and top 50 were confirmed by the known databases or experimental literatures, respectively.

ranked in the first and fifth position, respectively, both of which showed significant down regulation after 5-Aza-CdR treatment in breast cancer cells (Radpour et al., 2011). Moreover, miR-24-1 was the fourth miRNA predicted to be associated with 5-Aza-CdR. Microarray analysis showed miR-24-1 were up regulated upon 5-Aza-CdR therapy in pancreatic cancer PANC-1 cells compared to control cells (Lee et al., 2009).

The whole prediction list of all candidate small molecule-miRNA pairs in Dataset 1 was provided in **Supplementary Table 1**, which was ranked in a descending order according to the association scores resulted from GISMMA. It is hoped that the ranked list can be useful in guiding biological experiments, and can be verified by more experimental results in the future.

DISCUSSION

This paper presented a graphlet interaction based method GISMMA to infer the potential associations between small molecules and miRNAs by combining small molecule similarity, miRNA similarity and known associations between small molecules and miRNAs. In GISMMA, we used a similarity network to represent the small molecules and used another similarity network to represent the miRNAs. An edge with a weight of the similarity value between two nodes was ploted when there was similarity between the two nodes, otherwise not. We utilized graphlet interaction to measure the complex relationship between two nodes in the network, where the graphlet is defined as a type of non-isomorphic subgraph (Wang et al., 2014). Then, we counted each graphlet interaction isomer in a special pattern from the node having known associations to the node which does not have known associations. Therefore, we obtained a vector to describe the graphlet interaction between the two nodes. The correlation score between a small molecule and a miRNA can be computed through summing the weighted graphlet interaction isomers, where the weighs can be learnt from the known associations. The performance of GISMMA on predicting novel small molecule-miRNA associations was evaluated with four validation approaches that were global and two types of local LOOCV, as well as five-fold cross validation. The cross validation results were compared between GISMMA and SMiR-NBI, which showed the superior performance of GISMMA over SMiR-NBI. Besides, the ROC curves of SMiR-NBI are some unusual in **Figures 2**, **3**, which may be attribute to that SMiR-NBI could not predict associated miRNAs (small molecules) for new small molecules (miRNAs). When ranking the test small molecule-miRNA pair with those candidate pairs for SMiR-NBI, we assigned fixed rank to those pairs that contain new small molecules (miRNAs) with an average number, which may cause the presence of line segments in the ROC curve. We have implemented cross validations on two datasets with different sizes. The results showed that GISMMA performed better on Dataset 1 than on Dataset 2, which could be resulted from two factors. The one is the more similarity information in Dataset 1. The other is that Dataset 1 contains those small molecules and miRNAs without any known associations, which often get lower association scores and lower rankings than the test sample. This could also make the AUCs higher. And we further executed case study for three small molecules using Dataset 1. The numbers of miRNAs that were validated to be related to these three small molecules by experimental literatures are in turn 30, 29 and 25 in top 50 miRNAs predicted by GISMMA. Via cross validations together with case study, we can see that GISMMA is well-performed and reliable in predicting new associations between small molecules and miRNAs. Furthermore, a list of all predicted small molecule-miRNA associations was provided, which would be favorable for the development of miRNA-targeted therapy and drug reposition. In detail, for a specific small molecule, we focused on the predicted miRNAs that are most possibly associated with this small molecule. These miRNAs might be related to some diseases that were not confirmed to be treated by this small molecule. Through regulating the expressions of these miRNAs, this small molecule could be used for the treatment of these diseases. Therefore, we believed that the prediction results of this work could offer some guidance for the experiment of drug reposition to some extent.

The outstanding performance of GISMMA can be attributed to several factors. Firstly, we mapped the similarity between small molecules and similarity between miRNAs into two networks, in which the similarity values were fully exploited to investigate the complex relationship between two nodes by measuring their graphlet interaction. Secondly, in GISMMA, not only direct but also indirect links were considered between the nodes in the counting of graphlet interaction isomers. Finally, the GISMMA is a bipartite method which combines miRNA network with small molecule network. It can be used to predict miRNAs associated with new small molecules without any known related miRNAs, as well as to predict small molecules associated with new miRNAs without any known related small molecules, because it computes the association score by combining the result calculated in the small molecule network with that in the miRNA network.

However, GISMMA still has some limitations. For example, the lack of the known association data, especially the presence of many new small molecules or new miRNAs that have no known associations, affected the performance to a large extent. It can be expected that the model will obtain better performance when more experimental datasets are produced in the future. Besides, the simple algorithm of averaging the scores from two networks to compute the final association score may cause bias to those pairs that can be predicted only in one network. Furthermore, GISMMA considered 4 nodes at most within a graphlet, which hindered it to contain more similarity information from more distant nodes. Finally, this model cannot be applied to the prediction of the association in which the small molecule and the miRNA are both new. We anticipate that more network-based methods could be developed to improve the prediction of novel small molecule-miRNA association. For example, Petri nets based models have been proved to be a useful tool for many prediction problems, inspired by the work in (Russo et al., 2017), we could construct algorithm using Petri nets for the inference of potential small molecule-miRNA association.

AUTHOR CONTRIBUTIONS

N-NG implemented the experiments, analyzed the result, and wrote the paper. Y-ZS analyzed the result and wrote the paper. XC conceived the project, developed the prediction method, designed the experiments, analyzed the result, and revised the paper. ZM analyzed the result. J-QL analyzed the result and revised the paper. All authors read and approved the final manuscript.

FUNDING

XC was supported by National Natural Science Foundation of China under grant nos. 61772531 and 11631014. JL was supported by National Natural Science Foundation of China under grant nos. U1713212 and 61572330, Natural Science foundation of Guangdong Province under grant no. 2014A030313554, and Technology Planning Project from Guangdong Province under grant no. 2014B010118005.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphar.2018.01152/full#supplementary-material

REFERENCES

Airley, R. (2009). Cancer chemotherapy: basic science to the clinic. *ChemMedChem* 14, 2123–2124.

Ambros, V. (2003). MicroRNA pathways in flies and worms: growth, death, fat, stress, and timing. *Cell* 113, 673–676. doi: 10.1016/S0092-8674(03)00428-8

Andreuccetti, M., Allegrini, G., Antonuzzo, A., Malvaldi, G., Conte, P. F., Danesi, R., et al. (1996). Azidothymidine in combination with 5-fluorouracil in human colorectal cell lines: in vitro synergistic cytotoxicity and DNA-induced strand-breaks. *Eur. J. Cancer* 32A, 1219–1226. doi: 10.1016/0959-8049(96)00018-4

Bhat-Nakshatri, P., Wang, G., Collins, N. R., Thomson, M. J., Geistlinger, T. R., Carroll, J. S., et al. (2009). Estradiol-regulated microRNAs control estradiol response in breast cancer cells. *Nucleic Acids Res.* 37, 4850–4861. doi: 10.1093/nar/gkp500

Chen, X., Sun, Y.-Z., Zhang, D.-H., Li, J.-Q., Yan, G.-Y., An, J.-Y., et al. (2017a). NRDTD: a database for clinically or experimentally supported non-coding RNAs and drug targets associations. *Database* 2017:bax057. doi: 10.1093/database/bax057

Chen, X., Xie, D., Zhao, Q., and You, Z.-H. (2017b). MicroRNAs and complex diseases: from experimental results to computational models. *Brief. Bioinform.* doi: 10.1093/bib/bbx130 [Epub ahead of print].

Chen, X., Wang, L., Qu, J., Guan, N.-N., and Li, J.-Q. (2018a). Predicting miRNA–disease association based on inductive matrix completion. *Bioinformatics* doi: 10.1093/bioinformatics/bty503 [Epub ahead of print].

Chen, X., Xie, D., Wang, L., Zhao, Q., You, Z.-H., and Liu, H. (2018b). BNPMDA: bipartite network projection for MiRNA–disease association prediction. *Bioinformatics* 34, 3178–3186. doi: 10.1093/bioinformatics/bty333

Chen, X., Yin, J., Qu, J., and Huang, L. (2018c). MDHGI: matrix decomposition and heterogeneous graph inference for miRNA-disease association prediction. *PLoS Comput. Biol.* 14:e1006418. doi: 10.1371/journal.pcbi.1006418

Chen, X., Yan, C. C., Zhang, X., Zhang, X., Dai, F., Yin, J., et al. (2015). Drug-target interaction prediction: databases, web servers and computational models. *Brief. Bioinform.* 17, 696–712. doi: 10.1093/bib/bbv066

Esquela-Kerscher, A., and Slack, F. J. (2006). Oncomirs - microRNAs with a role in cancer. *Nat. Rev. Cancer* 6, 259–269. doi: 10.1038/nrc1840

Gottlieb, A., Stein, G. Y., Ruppin, E., and Sharan, R. (2011). PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Mol. Syst. Biol.* 7:496. doi: 10.1038/msb.2011.26

Gumireddy, K., Young, D. D., Xiong, X., Hogenesch, J. B., Huang, Q., and Deiters, A. (2008). Small-molecule inhibitors of microrna miR-21 function. *Angew. Chem. Int. Ed. Engl.* 47, 7482–7484. doi: 10.1002/anie.200801555

Hattori, M., Okuno, Y., Goto, S., and Kanehisa, M. (2003). Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *J. Am. Chem. Soc.* 125, 11853–11865. doi: 10.1021/ja036030u

He, L., and Hannon, G. J. (2004). MicroRNAs: small RNAs with a big role in gene regulation. *Nat. Rev. Genet.* 5, 522–531. doi: 10.1038/nrg1379

Hernandez-Vargas, H., Ballestar, E., Carmona-Saez, P., Von Kobbe, C., Banon-Rodriguez, I., Esteller, M., et al. (2006). Transcriptional profiling of MCF7 breast cancer cells in response to 5-Fluorouracil: relationship with cell cycle changes and apoptosis, and identification of novel targets of p53. *Int. J. Cancer* 119, 1164–1175.

Hurtubise, A., and Momparler, R. L. (2004). Evaluation of antineoplastic action of 5-aza-2′-deoxycytidine (Dacogen) and docetaxel (Taxotere) on human breast, lung and prostate carcinoma cell lines. *Anticancer Drugs* 15, 161–167. doi: 10.1002/ijc.21938

Issa, J. P., Garcia-Manero, G., Giles, F. J., Mannari, R., Thomas, D., Faderl, S., et al. (2004). Phase 1 study of low-dose prolonged exposure schedules of the hypomethylating agent 5-aza-2′-deoxycytidine (decitabine) in hematopoietic malignancies. *Blood* 103, 1635–1640. doi: 10.1182/blood-2003-03-0687

Jiang, Q., Wang, Y., Hao, Y., Juan, L., Teng, M., Zhang, X., et al. (2009). miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic Acids Res.* 37, D98–D104. doi: 10.1093/nar/gkn714

Jiang, W., Chen, X., Liao, M., Li, W., Lian, B., Wang, L., et al. (2012). Identification of links between small molecules and miRNAs in human cancers based on transcriptional responses. *Sci. Rep.* 2:282. doi: 10.1038/srep00282

Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A., et al. (2011). DrugBank 3.0: a comprehensive resource for 'Omics' research on drugs. *Nucleic Acids Res.* 39, D1035–D1041. doi: 10.1093/nar/gkq1126

Kozomara, A., and Griffiths-Jones, S. (2014). miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.* 42, D68–D73. doi: 10.1093/nar/gkt1181

Lee, K. H., Lotterman, C., Karikari, C., Omura, N., Feldmann, G., Habbe, N., et al. (2009). Epigenetic silencing of MicroRNA miR-107 regulates cyclin-dependent kinase 6 expression in pancreatic cancer. *Pancreatology* 9, 293–301. doi: 10.1159/000186051

Li, J., Lei, K., Wu, Z., Li, W., Liu, G., Liu, J., et al. (2016). Network-based identification of microRNAs as potential pharmacogenomic biomarkers for anticancer drugs. *Oncotarget* 7, 45584–45596. doi: 10.18632/oncotarget.10052

Lim, D., Byun, W. G., Koo, J. Y., Park, H., and Park, S. B. (2016). Discovery of a small-molecule inhibitor of protein-microRNA interaction using binding assay with a site-specifically labeled Lin28. *J. Am. Chem. Soc.* doi: 10.1021/jacs.6b06965 [Epub ahead of print].

Liu, X., Wang, S., Meng, F., Wang, J., Zhang, Y., Dai, E., et al. (2013). SM2miR: a database of the experimentally validated small molecules' effects on microRNA expression. *Bioinformatics* 29, 409–411. doi: 10.1093/bioinformatics/bts698

Longley, D. B., Harkin, D. P., and Johnston, P. G. (2003). 5-fluorouracil: mechanisms of action and clinical strategies. *Nat. Rev. Cancer* 3, 330–338. doi: 10.1038/nrc1074

Lu, M., Zhang, Q., Min, D., Jing, M., Guo, Y., Wei, G., et al. (2008). An analysis of human MicroRNA and disease associations. *PLoS One* 3:e3420. doi: 10.1371/journal.pone.0003420

Lv, S., Li, Y., Wang, Q., Ning, S., Huang, T., Wang, P., et al. (2012). A novel method to quantify gene set functional association based on gene ontology. *J. R. Soc. Interface* 9, 1063–1072. doi: 10.1098/rsif.2011.0551

Lv, Y., Wang, S., Meng, F., Yang, L., Wang, Z., Wang, J., et al. (2015). Identifying novel associations between small molecules and miRNAs based on integrated molecular networks. *Bioinformatics* 31, 3638–3644. doi: 10.1093/bioinformatics/btv417

Matsui, M., and Corey, D. R. (2017). Non-coding RNAs as drug targets. *Nat. Rev. Drug Discov.* 16, 167–179. doi: 10.1038/nrd.2016.117

McCabe, M. T., Low, J. A., Daignault, S., Imperiale, M. J., Wojno, K. J., and Day, M. L. (2006). Inhibition of DNA methyltransferase activity prevents tumorigenesis in a mouse model of prostate cancer. *Cancer Res.* 66, 385–392. doi: 10.1158/0008-5472.CAN-05-2020

Meng, F., Dai, E., Yu, X., Zhang, Y., Chen, X., Liu, X., et al. (2014). Constructing and characterizing a bioactive small molecule and microRNA association network for Alzheimer's disease. *J. R. Soc. Interface* 11:20131057. doi: 10.1098/rsif.2013.1057

Monroig, P. D., and Calin, G. A. (2013). MicroRNA and Epigenetics: diagnostic and therapeutic opportunities. *Curr. Pathobiol. Rep.* 1, 43–52. doi: 10.1007/s40139-013-0008-9

Monroig Pdel, C., Chen, L., Zhang, S., and Calin, G. A. (2015). Small molecule compounds targeting miRNAs for cancer therapy. *Adv. Drug Deliv. Rev.* 81, 104–116. doi: 10.1016/j.addr.2014.09.002

Patra, S. K., and Bettuzzi, S. (2009). Epigenetic DNA-(cytosine-5-carbon) modifications: 5-aza-2′-deoxycytidine and DNA-demethylation. *Biochemistry* 74, 613–619.

Przulj, N. (2007). Biological network comparison using graphlet degree distribution. *Bioinformatics* 23, e177–e183. doi: 10.1093/bioinformatics/btl301

Radpour, R., Barekati, Z., Kohler, C., Schumacher, M. M., Grussenmeyer, T., Jenoe, P., et al. (2011). Integrated epigenetics of human breast cancer: synoptic investigation of targeted genes, microRNAs and proteins upon demethylation treatment. *PLoS One* 6:e27355. doi: 10.1371/journal.pone.0027355

Rossi, L., Bonmassar, E., and Faraoni, I. (2007). Modification of miR gene expression pattern in human colon cancer cells following exposure to 5-fluorouracil in vitro. *Pharmacol. Res.* 56, 248–253. doi: 10.1016/j.phrs.2007.07.001

Ruepp, A., Kowarsch, A., Schmidl, D., Buggenthin, F., Brauner, B., Dunger, I., et al. (2010). PhenomiR: a knowledgebase for microRNA expression in diseases and biological processes. *Genome Biol.* 11:R6. doi: 10.1186/gb-2010-11-1-r6

Russo, G., Pennisi, M., Boscarino, R., and Pappalardo, F. (2017). Continuous Petri Nets and microRNA analysis in melanoma. *IEEE/ACM Trans. Comput. Biol. Bioinform.* doi: 10.1109/TCBB.2017.2733529 [Epub ahead of print].

Ryan, K. J. (1982). Biochemistry of aromatase: significance to female reproductive physiology. *Cancer Res.* 42, 3342s–3344s.

Sathish Kumar, B., Kumar, A., Singh, J., Hasanain, M., Singh, A., Fatima, K., et al. (2014). Synthesis of 2-alkoxy and 2-benzyloxy analogues of estradiol as anti-breast cancer agents through microtubule stabilization. *Eur. J. Med. Chem.* 86, 740–751. doi: 10.1016/j.ejmech.2014.09.033

Selcuklu, S. D., Donoghue, M. T., Kerin, M. J., and Spillane, C. (2012). Regulatory interplay between miR-21, JAG1 and 17beta-estradiol (E2) in breast cancer cells. *Biochem. Biophys. Res. Commun.* 423, 234–239. doi: 10.1016/j.bbrc.2012.05.074

Shah, M. Y., Pan, X., Fix, L. N., Farwell, M. A., and Zhang, B. (2011). 5-Fluorouracil drug alters the microRNA expression profiles in MCF-7 breast cancer cells. *J. Cell. Physiol.* 226, 1868–1878. doi: 10.1002/jcp.22517

Simpson, E., and Santen, R. J. (2015). Celebrating 75 years of oestradiol. *J. Mol. Endocrinol.* 55, T1–T20. doi: 10.1530/JME-15-0128

Spizzo, R., Nicoloso, M. S., Croce, C. M., and Calin, G. A. (2009). SnapShot: microRNAs in cancer. *Cell* 137, 586.e1–586.e1. doi: 10.1016/j.cell.2009.04.040

Thomas, M., and Deiters, A. (2013). MicroRNA miR-122 as a therapeutic target for oligonucleotides and small molecules. *Curr. Med. Chem.* 20, 3629–3640. doi: 10.2174/0929867311320290009

Tilghman, S. L., Bratton, M. R., Segar, H. C., Martin, E. C., Rhodes, L. V., Li, M., et al. (2012). Endocrine disruptor regulation of microRNA expression in breast carcinoma cells. *PLoS One* 7:e32754. doi: 10.1371/journal.pone.0032754

Wang, J., Meng, F., Dai, E., Yang, F., Wang, S., Chen, X., et al. (2016). Identification of associations between small molecule drugs and miRNAs based on functional similarity. *Oncotarget* 7, 38658–38669. doi: 10.18632/oncotarget.9577

Wang, X. D., Huang, J. L., Yang, L., Wei, D. Q., Qi, Y. X., and Jiang, Z. L. (2014). Identification of human disease genes from interactome network using graphlet interaction. *PLoS One* 9:e86142. doi: 10.1371/journal.pone.0086142

Wang, Y., and Lee, C. G. (2009). MicroRNA and cancer–focus on apoptosis. *J. Cell Mol. Med.* 13, 12–23. doi: 10.1111/j.1582-4934.2008.00510.x

Wang, Y., Xiao, J., Suzek, T. O., Zhang, J., Wang, J., and Bryant, S. H. (2009). PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.* 37, W623–W633. doi: 10.1093/nar/gkp456

Windle, R., Bell, P. R., and Shaw, D. (1987). Five year results of a randomized trial of adjuvant 5-fluorouracil and levamisole in colorectal cancer. *Br. J. Surg.* 74, 569–572. doi: 10.1002/bjs.1800740707

Xia, T., Li, J., Cheng, H., Zhang, C., and Zhang, Y. (2015). Small-molecule regulators of MicroRNAs in biomedicine. *Drug Dev. Res.* 76, 375–381. doi: 10.1002/ddr.21271

Zhang, J., Yang, P. L., and Gray, N. S. (2009). Targeting cancer with small molecule kinase inhibitors. *Nat. Rev. Cancer* 9, 28–39. doi: 10.1038/nrc2559

Zhang, S., Chen, L., Jung, E. J., and Calin, G. A. (2010). Targeting microRNAs with small molecules: from dream to reality. *Clin. Pharmacol. Ther.* 87, 754–758. doi: 10.1038/clpt.2010.46

Zhou, J., Zhou, Y., Yin, B., Hao, W., Zhao, L., Ju, W., et al. (2010). 5-Fluorouracil and oxaliplatin modify the expression profiles of microRNAs in human colon cancer cells in vitro. *Oncol. Rep.* 23, 121–128.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Pharmacology

ORIGINAL RESEARCH
published: 31 October 2018
doi: 10.3389/fphar.2018.01245

Determining the Balance Between Drug Efficacy and Safety by the Network and Biological System Profile of Its Therapeutic Target

*Xiao xu Li [1,2], Jiayi Yin [1], Jing Tang [1,2], Yinghong Li [1,2], Qingxia Yang [1,2], Ziyu Xiao [1], Runyuan Zhang [1], Yunxia Wang [1], Jiajun Hong [1], Lin Tao [3], Weiwei Xue [2] and Feng Zhu [1,2]**

[1] College of Pharmaceutical Sciences, Zhejiang University, Hangzhou, China, [2] School of Pharmaceutical Sciences and Collaborative Innovation Center for Brain Science, Chongqing University, Chongqing, China, [3] Key Laboratory of Elemene Class Anti-cancer Chinese Medicine of Zhejiang Province, School of Medicine, Hangzhou Normal University, Hangzhou, China

OPEN ACCESS

Edited by:
Zhi-Liang Ji,
Xiamen University, China

Reviewed by:
Enrique Hernandez-Lemus,
Instituto Nacional de Medicina Genómica (INMEGEN), Mexico
Qing-Chuan Zheng,
Jilin University, China
Yan Li,
Army Medical University, China

***Correspondence:**
Feng Zhu
zhufeng@zju.edu.cn;
prof.zhufeng@gmail.com

Specialty section:
This article was submitted to Translational Pharmacology, a section of the journal Frontiers in Pharmacology

Received: 28 July 2018
Accepted: 12 October 2018
Published: 31 October 2018

Citation:
Li X, Yin J, Tang J, Li Y, Yang Q, Xiao Z, Zhang R, Wang Y, Hong J, Tao L, Xue W and Zhu F (2018) Determining the Balance Between Drug Efficacy and Safety by the Network and Biological System Profile of Its Therapeutic Target. Front. Pharmacol. 9:1245. doi: 10.3389/fphar.2018.01245

One of the most challenging puzzles in drug discovery is the identification and characterization of candidate drug of well-balanced profile between efficacy and safety. So far, extensive efforts have been made to evaluate this balance by estimating the quantitative structure–therapeutic relationship and exploring target profile of adverse drug reaction. Particularly, the therapeutic index (TI) has emerged as a key indicator illustrating this delicate balance, and a clinically successful agent requires a sufficient TI suitable for it corresponding indication. However, the TI information are largely unknown for most drugs, and the mechanism underlying the drugs with narrow TI (NTI drugs) is still elusive. In this study, the collective effects of human protein–protein interaction (PPI) network and biological system profile on the drugs' efficacy–safety balance were systematically evaluated. First, a comprehensive literature review of the FDA approved drugs confirmed their NTI status. Second, a popular feature selection algorithm based on *artificial intelligence (AI)* was adopted to identify key factors differencing the target mechanism between NTI and non-NTI drugs. Finally, this work revealed that the targets of NTI drugs were highly centralized and connected in human PPI network, and the number of similarity proteins and affiliated signaling pathways of the corresponding targets was much higher than those of non-NTI drugs. These findings together with the newly discovered features or feature groups clarified the key factors indicating drug's narrow TI, and could thus provide a novel direction for determining the delicate drug efficacy-safety balance.

Keywords: drug efficacy-safety balance, therapeutic index, artificial intelligence, protein-protein interaction network, biological system profile

INTRODUCTION

One of the most challenging puzzles in drug discovery is the identification and characterization of candidate drugs of well-balanced profile between efficacy and safety (Muller and Milton, 2012; Li et al., 2018; Xue et al., 2018b). In other words, apart from extensive effort made to optimize drug affinity and selectivity (Wang et al., 2017a; Zheng et al., 2017), considerable investments should be devoted to detect adverse drug reactions (Huang et al., 2018) and reveal drug likeness (Benet et al., 2016; Yang et al., 2018). So far, the identification of drug toxicities in preclinical or clinical developments has been accelerated by a variety of technological advances (Badders et al., 2018) including biomarker-guided safety assessment (Muller and Dieterle, 2009; Rzepecki et al., 2018), OMICs techniques (Iloro et al., 2013; Fu J. et al., 2018), breakthrough in computing capacity and bioinformatics method (Zhu et al., 2011; Tao et al., 2015; Chen et al., 2016), and so on. To measure the level of correlation between drug maximum efficacy and confined safety in given disorder, the therapeutic index (TI typically considered as the ratio of the highest non-toxic drug exposure to the exposure producing the desired efficacy) has emerged as a key indicator illustrating that delicate balance (Zaykov et al., 2016). The TI is essential for life-threatening diseases (such as cardiovascular and oncological disease) with limited treatment options (Zhu et al., 2008b; Kimmelman and Federico, 2017). Particularly, tiny variation in the dosage of drugs with narrow TI (NTI drugs, TI ≤ 3) may result in therapeutic failure or serious adverse drug reactions (Tao et al., 2014; Ewer and Ewer, 2015; Zheng et al., 2016), and is only acceptable for the treatment of life-threatening diseases (Yu et al., 2015). Therefore, successful therapeutic agents require sufficient TI (NNTI drugs, TI >3) suitable for it corresponding indication (Abernethy et al., 2011).

However, TI characterization is too complicated to be achieved for many drugs (Yu et al., 2015), and TI is highly susceptible to the subject variations of drug responses (Jiang et al., 2015; Yang et al., 2017). To enhance the determination and interpretation of TI, a variety of *in-silico* studies have been performed to reveal the mechanism underlying NTI drugs (Muller and Milton, 2012). In particular, the prediction models based on quantitative structure–activity (QSAR), structure–toxicity (QSTR), and structure–index (QSIR) relationship have been constructed to enable early assessment of TI (Zhu H. et al., 2008; Rodgers et al., 2010; Zhu et al., 2012a; Chen et al., 2016; Fu T. et al., 2018). These models are primarily constructed and exert their prediction capacity based on structures of the studied drugs, which thus demonstrate great limitations in coping with TI's vulnerability to the subject variation of drug responses (Jiang et al., 2015). Compared with the approaches based on drug structure, target-based approach turns out to be the one of enhanced effectiveness for characterizing confined toxicity behind the drug efficacy (Muller and Milton, 2012; Huang et al., 2018), since the population variation of drug target is capable of reflecting, to some extent, the subject variations of drug responses (Fujimoto et al., 2014; Jiang et al., 2015). But target-based method is sophisticated due to the involvement of target in complex protein–protein interaction (PPI) network (Rao et al., 2011; Li et al., 2016b; Xu et al., 2016; Wang et al., 2017b) and the necessity of considering target biological system profiles (Zhu F. et al., 2009; Xue et al., 2016).

So far, the PPI network properties (Ragusa et al., 2010; Guo et al., 2018) and biological system profiles (Zheng et al., 2006) have been adopted to analyze the drug likeness of candidate agents. On one hand, the target–protein interaction network has been constructed and the corresponding network features can be calculated for discovering the differential properties indicating disease status (Ragusa et al., 2010) and identifying candidate drug targets for a given indication (Guo et al., 2018; Xue et al., 2018a). On the other hand, the druggability of candidate target is found significantly determined by a variety of biological system profiles, which include the number of target affiliated signaling pathways (Yang et al., 2016), the number of similarity proteins outside target's protein family (Zheng et al., 2006), the number of human tissues distributed by the studied target (Zhu F. et al., 2009), and the differential level of target expression between patient and healthy individual (Ernst et al., 2017; Li et al., 2018). Since the underlying theories of network- and biological system-based approaches are distinct from each other (Guo et al., 2018; Li et al., 2018), it is essential to simultaneously consider these two types of properties for understanding drug likeness. However, these properties have not yet been collectively considered in TI-related studies, and the mechanism underlying drugs' narrow TI is still elusive.

In this study, a comprehensive analysis on the network features and biological system profiles of the primary therapeutic targets of all FDA approved drugs was conducted, and various features differentiating drugs of narrow TI (NTI drugs) from those of sufficient TI (NNTI drugs) were identified. First, due to the limited information of both NTI and NNTI drugs, a systematic literature review was conducted to collect the TI data for all approved drugs. Then, the primary therapeutic targets of these drugs were classified into four groups based on collected TI data. These four target groups include (a) targets of NTI drugs, (b) targets of both NTI and NNTI drugs, (c) targets of drugs without reported TI, and (d) targets of NNTI drugs. Third, a comparative analysis between target group (a) and (d) identified several key features able to differentiate two groups, and further study revealed three feature groups indicating the mechanisms underlying NTI drugs. In summary, these findings together with the newly discovered features or feature groups clarified key factors indicating drug's narrow TI, which gave a new direction for determining the delicate balance between drugs' maximum efficacy and confined safety.

MATERIALS AND METHODS

Systematic Collection of Drugs and Their Corresponding Targets and TI Data

The TI data of FDA approved drugs were obtained by four steps. First, FDA approved drugs were collected from the official website of FDA (Drugs@FDA), and their corresponding diseases were carefully confirmed. In total, 1,762 drugs were collected. Second, the primary therapeutic targets of these drugs were identified from the TTD database (https://db.idrblab.org/ttd/; Li et al., 2018), and 418 primary therapeutic targets of these 1,762 drugs were discovered (detail information was provided in the following paragraphs). Third, TI data of these drugs were systematically collected by a comprehensive literature review. Particularly, various keyword combinations were searched in PubMed and other academic resources, which included "drug name + therapeutic index," "drug name + therapeutic window," "drug name + critical dose," "drug name + therapeutic ranges," and "drug name + therapeutic ratio." As a result, 161 NTI and 29 NNTI drugs confirmed by the clinical evaluations or experiments were identified, which aimed at 60 and 28 human targets, respectively. **Supplementary Table S1** provided a full list of 161 NTI and 29 NNTI drugs together with their approved disease indication and corresponding targets. To the best of our knowledge, it is the first comprehensive literature review on the TI data of all drugs approved by FDA and **Supplementary Table S1** provided the most completed information of the FDA approved drugs with available TI data. Moreover, the primary therapeutic targets of all FDA approved drugs were classified into four groups based on their TI: (a) 20 targets of NTI drugs, (b) 40 targets of both NTI and NNTI drugs, (c) 339 targets of drugs without reported TI, and (d) 19 targets of NNTI drugs. Moreover, among those drugs listed in **Supplementary Table S1**, four multi-target drugs were found with NTI data available, which included *regorafenib* (hepatocellular and colorectal cancer), *sorafenib* (renal cell and hepatocellular carcinoma), *sunitinib* (gastrointestinal cancer), and *vandetanib* (medullary thyroid cancer). All these drugs are multi-kinases inhibitors for the treatment of cancer.

Identification of the Primary Therapeutic Target(S) of FDA Approved Drugs

The primary therapeutic target of each FDA approved drug was strictly determined by considering (1) the experimentally determined potency of drugs against their primary target or targets (Zhu et al., 2010), (2) the observed potency or effects of drugs against disease models (cell lines, *ex-vivo*, *in-vivo* models) linking to their primary drug targets (Zhu et al., 2012b), and (3) the observed effect of target knockout, knockdown, transgenetic, RNA interference, antibody or antisense-treated *in vivo* models (Zhu et al., 2012b). Taking the confirmation of CDK4 as the primary therapeutic target of FDA approved Palbociclib as an example, it was determined by considering: (1) experimentally defined high potency (IC50 = 11 nM) of Palbociclib against CDK4 (Fry et al., 2004), (2) the clearly observed development of multiple tumors by a point mutation (R24C) in the first coding exon of locus encoding CDK4 in the mice models (Sotillo et al., 2001), and (3) Palbociclib-induced G1-G2 arrest and apoptosis in breast tumor cell lines (IC50 <400 nM) and tumor growth reduction in human breast tumor xenograft (Lapenna and Giordano, 2009). In conclusion, only the targets with complete target determination data (including all three types of information above) were defined as the primary therapeutic targets of the corresponding FDA approved drugs.

Deriving the Human PPI Network Properties for Each Studied Target

The human protein–protein interaction (PPI) network analyzed here included 15,554 proteins and 642,304 PPIs, which was constructed using the data provided in STRING (Szklarczyk et al., 2015). In order to ensure the reliability of the analyzed data, only those PPIs with high confidence score (>0.95) were collected for the subsequent analyses (Ghosh et al., 2015; Wang S. et al., 2015). As a result, a sub-network with 8,509 proteins and 40,468 PPIs were generated and adopted for further analyses in this study. Moreover, the network properties for each studied target were generated by the PROFEAT (Zhang et al., 2017a) and the tool NetworkAnalyzer of Cytoscape (Shannon et al., 2003; Thomas and Bonchev, 2010).

In total, 32 network properties were calculated and adopted in subsequent analysis. These properties were popular for analyzing a complex biological network, which included: (1) *Average Closeness Centrality*: the average number of steps required to reach the studied node from any node in a network (Ma et al., 2016); (2) *Average Shortest Path Length*: the average length of shortest paths between the studied node and all other ones (Zhang et al., 2014); (3) *Betweenness Centrality*: the number of times the studied node serving as a linking bridge along shortest path between any two nodes (Zeidán-Chuliá et al., 2015); (4) *Bridging Centrality*: the product of the bridging coefficient and betweenness centrality (Hwang et al., 2008); (5) *Bridging Coefficient*: the extent of the studied node lying between any other densely connected nodes in the network (Paladugu et al., 2008); (6) *Closeness Centrality Sum*: the reciprocal of the sum of the shortest paths between the studied node and all other nodes in the network (Costenbader and ValenteFontanesi, 2003); (7) *Clustering Coefficient*: the number of the connected pairs between all neighbors of node (Watts and Strogatz, 1998); (8) *Current Flow Betweenness*: a centrality index measuring the level of information travels along all possible paths within network (Paladugu et al., 2008); (9) *Current Flow Closeness*: the variant of current flow betweenness (Zhang et al., 2017b); (10) *Degree*: the number of edges linked to a node (Braeuning, 2013); (11) *Degree Centrality*: the number of links incident upon a studied node (Batool and Niazi, 2014); (12) *Deviation*: the variation between sum of node distances and network unipolarity (Zhang et al., 2017a); (13) *Distance Deviation*: the absolute difference between nodes' distance sum and network's average distance (Rogelj et al., 2013); (14) *Distance Sum*: the sum of all shortest paths starting from the studied node (Bolser et al., 2003); (15) *Eccentric*: the absolute difference between nodes' eccentricities and network's average eccentricity (Zhang et al., 2017a); (16) *Eccentricity*: the maximum non-infinite shortest path length between the studied node and all other nodes in the network (Bolser et al., 2003); (17) *Eccentricity Centrality*: the largest geodesic distance between the node and any other node (Batool and Niazi, 2014); (18) *Eigenvector Centrality*: the sum of its neighbors' centrality values (Solá et al., 2013); (19) *Harmonic Closeness Centrality*: the sum of the reciprocals of the average shortest path lengths of each node in network (Zhang et al., 2017b); (20) *Interconnectivity*: a connectivity index indicating the quality of the studied nodes being connected together (Emig et al., 2013); (21) *Load Centrality*: the fraction of all the shortest paths that pass through the studied node (Kivimäki et al., 2016); (22) *Neighborhood Connectivity*: the average connectivity of all neighbors (Carson and Lu, 2015); (23) *Normalized Betweenness*: the fraction of network shortest paths that a given protein lies on (Paladugu et al., 2008); (24) *Number of Self Loops*: the number of edges starting and ending at the same node (Garlaschelli and Loffredo, 2004); (25) *Number of Triangles*: the number of triangles that include the studied node as a vertex (Rubinov and Sporns, 2010); (26) *Page Rank Centrality*: an adjustment of Katz by considering the diluted issue (Li et al., 2013); (27) *Radiality*: the level of reachability of a studied node via various shortest paths within the entire network (Koschützki and Schreiber, 2008); (28) *Residual Closeness Centrality*: the closeness measured by removing the studied node (Dangalchev, 2006); (29) *Scaled Degree*: the degree of a studied node relative to the most connected node within the same module (Sormani, 2012); (30) *Stress*: the number of shortest paths passing through a given node (Shannon et al., 2003); (31) *Topological Coefficient*: the extent to which a node in network shares interaction partners with other nodes (Zhu M. et al., 2009); (32) *Z Score*: a connectivity index based on degree distribution of a network (Rubinov and Sporns, 2010).

Assessing the Biological System Profile for Each Studied Target

The biological system profile for each studied target included: (1) the number of target-affiliated and target immediate-downstream signaling pathways in KEGG database (Kanehisa et al., 2017). The target-affiliated pathways were determined by considering that (a) the pathways of the studied target should be life-essential in both patients and healthy people and (b) the studied target should be in the pathway upstream with the capacity of regulating the biological function of the pathways. (2) The number of human tissues each target distributed in, assessed by the TissueDistributionDBs (Kogenaru et al., 2010) and Uniprot (UniProt Consortium, 2018) databases. A target was assumed to distribute in a given tissue if >5% of the total proteins are distributed in that tissue or the target concentration is higher than the average concentration of proteins in that tissue. (3) The number of human similarity proteins of a target outside the corresponding target family for probing off-target collateral effect (Zheng et al., 2006; Zhu F. et al., 2009). This was determined by BLAST similarity screening of human proteome in Uniprot database (UniProt Consortium, 2018) with a cutoff (*E-value* < 0.005; Song et al., 2006; Singh et al., 2007). (4) The differential expressions of the studied target in the disease-specific tissue between patients and healthy individuals (Li et al., 2018). The relevant data were collected directly from TTD (Li et al., 2018) and calculated based on the human gene expression raw data of *Affymetrix U133 Plus 2.0* platform in GEO (Barrett et al., 2013).

Selecting the Differential Features Indicating NTI Drugs by Artificial Intelligence

The artificial intelligence (AI) has been recently proposed as a powerful technique for drug target discovery (Xu and Wang, 2014; Zhu et al., 2018), protein function prediction (Li et al., 2016a; Seo et al., 2018; Yu et al., 2018) and biomarker identification (Li B. et al., 2016; Li et al., 2017) through mimicking the human thinking procedures, learning processes and information extractions, which included the machine learning algorithm (Zhu et al., 2008a; Wang P. et al., 2015), the deep learning method (van der Burgh et al., 2017; Seo et al., 2018), and the cognitive-computing (Krittanawong et al., 2017). As one of the most popular machine learning algorithms, the *Boruta* algorithm based on wrapper method built around a random forest classifier (Kursa, 2014) was selected and adopted in this study. It is an extension to determine the relevance via comparing the relevance of the real features to that of the random probes (Pan et al., 2018). Since *Boruta* was constructed by an AI-based technique (machine learning), it was considered to be the most powerful approach with the stability in the variable selection, especially suitable for the low-dimensional dataset among other available strategies (Degenhardt et al., 2017). In this study, the differential features between NTI and NNTI drugs were therefore identified by *R package Boruta* (Shang et al., 2017). Particularly, human PPI network properties and biological system features of each target were first calculated, and the results of feature selection were then acquired using *R package Boruta* by setting the p-value < 0.05, maxRuns = 100, and doTrace = 2. In the meantime, the getImp was set to "getImpRfZ," and the mcAdj and holdHistory were set to "TRUE."

RESULTS AND DISCUSSION

Network Properties and Biological System Profile of NTI and NNTI Drugs

As reported, the human PPI network properties and biological system profile were key factors determining efficacy-safety balance (Zheng et al., 2006; Ragusa et al., 2010; Guo et al., 2018). Network properties were inherent feature of a target in the human PPI network, while biological system profile could reflect both the on-target and off-target pharmacology (Bender et al., 2007; Han et al., 2018; Zhu et al., 2018). Herein, 32 features of human PPI network together with 4 biological system properties were therefore adopted and calculated for further analyses. To the best of our knowledge, these were the most comprehensive sets of features ever applied for TI-related analysis. **Table 1** listed the calculated values of ten properties based on the connectivity and adjacency in human PPI network. These connectivity/adjacency-based network properties were designed to describe the level of connectivity among human proteins or the neighborhood features of the studied proteins (Chen et al., 2016). The properties included bridging coefficient, clustering coefficient, degree, degree centrality, interconnectivity, neighbor connectivity, number of triangles, scaled degree, topological coefficient, and Z-score (corresponding definitions were provided in section Materials and Methods). As shown in **Table 1**, 8 (80.0%) out of 10 properties were significantly different (p-value < 0.05, highlighted by bold font) between the targets of NTI and NNTI drugs, and half of those 10 properties were with the most significant differences (p-value < 0.01, highlighted by bold-underline).

Similar to the connectivity/adjacency-based network property, the calculated values of 16 properties based on the shortest path length in the human PPI network were provided in **Table 2** (corresponding definitions of these properties were provided in section Materials and Methods). As shown in

TABLE 1 | The calculated values of 10 properties based on the connectivity and adjacency in the human PPI network.

Connectivity/Adjacency based properties	Targets of the NTI drugs		Targets of the NNTI drugs		*p*-values
	Mean ± SD	**Median**	**Mean ± SD**	**Median**	
Bridging coefficient	5.62E–01 ± 1.44E+00	7.10E-02	3.72E+00 ± 9.02E+00	7.47E-01	2.15E-01
Clustering coefficient	1.07E–01 ± 1.67E–01	1.82E-02	4.06E–01 ± 4.06E–01	3.33E-01	**1.40E-02**
Degree	1.04E+01 ± 4.14E+00	1.10E+01	4.53E+00 ± 3.89E+00	3.00E+00	**3.56E-05**
Degree centrality	1.09E–03 ± 5.70E–04	1.00E-03	5.71E–04 ± 7.56E–04	0.00E+00	**2.90E-02**
Interconnectivity	2.59E–01 ± 1.04E–01	1.86E-01	5.89E–01 ± 1.44E–01	6.18E-01	**3.21E-07**
Neighbor connectivity	3.33E+01 ± 2.50E+01	2.79E+01	1.25E+01 ± 8.45E+00	1.13E+01	**1.57E-05**
Number of triangles	5.28E+00 ± 8.06E+00	1.00E+00	5.29E+00 ± 7.47E+00	3.00E+00	9.98E-01
Scaled degree	1.41E–02 ± 5.44E–03	1.50E-02	6.36E–03 ± 5.17E–03	4.00E-03	**7.01E-05**
Topological coefficient	1.67E–01 ± 1.53E–01	1.07E-01	3.41E–01 ± 2.42E–01	3.60E-01	**1.76E-02**
Z score	1.23E–03 ± 1.27E–02	3.00E-03	–1.63E–02 ± 1.22E–02	–2.20E-02	**1.17E-04**

The mean values (together with standard deviation) and median values of these properties between the targets of NTI and NNTI drugs were provided, and the statistical difference (p-value) for each property between targets of NTI and NNTI drugs were also calculated (p-values <0.05 and <0.01 were highlighted by bold and bold-underline, respectively).

TABLE 2 | The calculated values of 16 properties based on the shortest path length in human PPI network.

Shortest path length-based properties	Targets of the NTI drugs		Targets of the NNTI drugs		*p*-values
	Mean ± SD	**Median**	**Mean ± SD**	**Median**	
Average shortest path length	4.06E+00 ± 2.90E–01	3.95E+00	4.88E+00 ± 1.08E+00	5.09E+00	**1.06E-02**
Betweenness centrality	1.26E–03 ± 6.77E–04	1.77E-03	2.54E–04 ± 3.94E–04	1.09E-05	**1.59E-08**
Average closeness centrality	2.47E–01 ± 1.63E–02	2.53E-01	1.97E–01 ± 2.26E–02	1.92E-01	**5.31E-07**
Current flow betweenness	3.07E–03 ± 1.35E–03	4.00E-03	8.57E–04 ± 1.17E–03	5.00E-04	**3.38E-06**
Deviation	1.11E+04 ± 2.31E+03	1.03E+04	1.96E+04 ± 4.39E+03	2.03E+04	**4.24E-06**
Distance deviation	6.17E+03 ± 2.02E+03	6.93E+03	4.30E+03 ± 2.37E+03	4.16E+03	**1.57E-02**
Distance sum	3.23E+04 ± 2.31E+03	3.14E+04	4.08E+04 ± 4.39E+03	4.15E+04	**4.24E-06**
Eccentric	1.11E+00 ± 4.27E–01	1.34E+00	5.97E–01 ± 4.14E–01	3.40E-01	**6.09E-04**
Eccentricity	1.02E+01 ± 4.27E–01	1.00E+01	1.14E+01 ± 7.45E–01	1.10E+01	**6.44E-05**
Eccentricity centrality	9.79E–02 ± 3.85E–03	1.00E-01	8.84E–02 ± 5.79E–03	9.10E-02	**2.30E-05**
Harmonic closeness centrality	2.10E+03 ± 1.53E+02	2.14E+03	1.64E+03 ± 2.03E+04	1.59E+03	**3.95E-07**
Load centrality	1.35E–03 ± 7.83E–04	2.00E-03	2.86E–04 ± 4.69E–04	0.00E+00	**3.75E-07**
Normalized betweenness	2.81E–03 ± 1.53E–03	4.00E-03	5.71E–04 ± 8.52E–04	0.00E+00	**2.53E-08**
Residual closeness centrality	6.00E+02 ± 1.05E+02	6.29E+02	3.07E + 02 ± 1.23E+02	2.74E+02	**1.32E-07**
Radiality	8.09E–01 ± 1.81E–02	8.16E-01	7.43E–01 ± 3.33E–02	7.44E-01	**1.21E-06**
Stress	1.56E+06 ± 9.11E+05	2.24E+06	3.04E+05 ± 4.82E+05	4.69E+03	**2.13E-08**

Mean values (together with standard deviation) and median values of these properties between the targets of NTI and NNTI drugs were provided, and the statistical difference (p-value) for each property between targets of NTI and NNTI drugs were also calculated (p-values <0.05 and <0.01 were highlighted by bold and bold-underline, respectively).

Table 2, all properties were found to be significantly different (p-values < 0.05, in bold font) between the targets of NTI and NNTI drug, and 14 (87.5%) of the 16 properties were with the most significant difference (p-value < 0.01, bold-underline). Moreover, the calculated values of 4 human biological system properties were shown in **Table 3** (definition of these properties was given in section Materials and Methods). As reported, these properties were frequently adopted to analyze the druggability of therapeutic targets for not only approved drugs but also the drugs in clinical trial development or withdrawn from market (Li et al., 2018). Herein, two properties were identified as significantly different (p-value < 0.01, bold-underline) between targets of NTI and NNTI drugs, which included the number of pathways affiliated by the targets of the studied drugs and the number of similarity proteins outside target's functional family. One thing needed to be emphasized was that the standard deviation of many properties was even larger than their mean value (such as bridging coefficient, clustering coefficient, and Z-score). These deviations indicated that the corresponding p-value may not be enough to measure the difference between the targets of NTI and NNTI drug. Moreover, any of the individual feature (p-value < 0.05 shown in **Tables 1–3**) could not be used to satisfactorily differentiate the targets of NTI drugs from that of the NNTI ones. Thus, this finding inspired us to discover the differential features using more advanced computational algorithm and collectively considering multiple properties.

TABLE 3 | The calculated values of four human biological system properties.

Human biological system properties	Targets of the NTI drugs		Targets of the NNTI drugs		*p*-values
	Mean ± SD	Median	Mean ± SD	Median	
No. of pathways affiliated by the primary therapeutic target	6.10 ± 1.80	7.00	1.14 ± 0.38	1.00	**2.50E-15**
No. of similarity proteins outside the target family	24.4 ± 15.22	29.00	11.79 ± 6.21	11.00	**1.46E-05**
Differential expression levels between patients and healthy individuals	0.42 ± 0.35	0.56	0.33 ± 0.32	0.20	3.86E-01
No. of tissues distributed by the primary therapeutic target	3.38 ± 0.81	3.00	3.61 ± 1.82	3.00	6.06E-01

The mean values (together with standard deviation) and median values of these properties between the targets of NTI and NNTI drugs were provided, and the statistical difference (p-value) for each property between targets of NTI and NNTI drugs were also calculated (p-values <0.05 and <0.01 were highlighted by bold and bold-underline, respectively).

TABLE 4 | 19 substantially overlapped network properties grouped into 5 property groups based on their innate mutual dependence.

Property group	Original property	Equation of the property	Description of the property
Average closeness centrality	Average closeness centrality	$1/(\frac{1}{N}\sum_{j=1}^{N} D_{ij})$	The average number of steps required to reach the studied node from any node in the network
	Harmonic closeness centrality	$\sum_{j=1}^{N} \frac{1}{D_{ij}}$	The sum of the reciprocals of the average shortest path lengths of each node in the network
	Residual closeness centrality	$\sum_{j=1}^{N} \frac{1}{2^{D_{ij}}}$	The closeness measured by removing the studied node
	Sum closeness centrality	$1/\sum_{j=1}^{N} D_{ij}$	The reciprocal of the sum of the shortest paths between the studied node and all other nodes in the network
Average shortest path length	Average shortest path length	$\frac{1}{N-1}\sum_{j=1}^{N} D_{ij}$	The average length of the shortest paths between the studied node and all other nodes in network
	Deviation	$distSum_i - unipolarity_i$	The variation between the total sum of node distances and the network unipolarity
	Distance sum	$\sum_{j=1}^{N} D_{ij}$	The sum of all shortest paths starting from the studied node
Betweenness centrality	Betweenness centrality	$\sum_{s \neq i \neq t} \frac{\sigma_{st}(i)}{\sigma_{st}}$	The number of times the studied node serving as a linking bridge along the shortest paths between any two nodes
	Current flow betweenness	$\frac{1}{N_b}\sum_{s,t \in V} \tau_{st}(i)$	A centrality index measuring the level of information travels along all possible paths within the network
	Current flow closeness	$\frac{N_C}{\sum_{s \neq t} p_{st}(s) - p_{st}(t)}$	The variant of current flow betweenness
	Load centrality	$\sum_{s \neq i \neq t} \sigma_{st}(i)$	The fraction of all the shortest paths that pass through the studied node
	Normalized betweenness centrality	$\frac{cenBtw_i - min(cenBtw_G)}{max(cenBtw_G) - min(cenBtw_G)}$	The fraction of network shortest paths that the studied protein lies on
Degree	Degree	$Degree_i$	The total number of edges linked to a node
	Degree centrality	$\frac{deg_i}{N-1}$	The number of links incident upon the studied node
	Number of self-loops	$Selfloop_i$	The number of edges starting and ending at the same node
	Scaled degree	$\frac{deg_i}{max(deg_G)}$	The degree of the studied node relative to the most connected node within the same module
	Z score	$[deg_i - avg(deg_G)]/dev(deg_G)$	A connectivity index based on the degree distribution of network
Eccentricity	Eccentricity	$max(D_{ij})$	The maximum non-infinite shortest path length between the studied node and all other nodes in the network
	Eccentricity centrality	$1/max(D_{ij})$	The largest geodesic distance between the node and any other nodes

Discovering the Key Features of NTI Drug Targets by Artificial Intelligence

Based on the in-depth investigation of 36 properties in **Tables 1–3**, several properties were found to be not fully independent or even duplicate in their descriptions (like degree vs. scaled degree). In this study, all 36 properties were systematically reviewed, and 19 of these 36 were identified to be substantially overlapped with some other properties (**Table 4**). Since there was significant dependence among the 19 properties, the use of all 36 properties for statistical feature selection may introduce strong biases. Thus, the 19 properties were grouped based on their innate mutual dependence. As shown in **Table 4**,

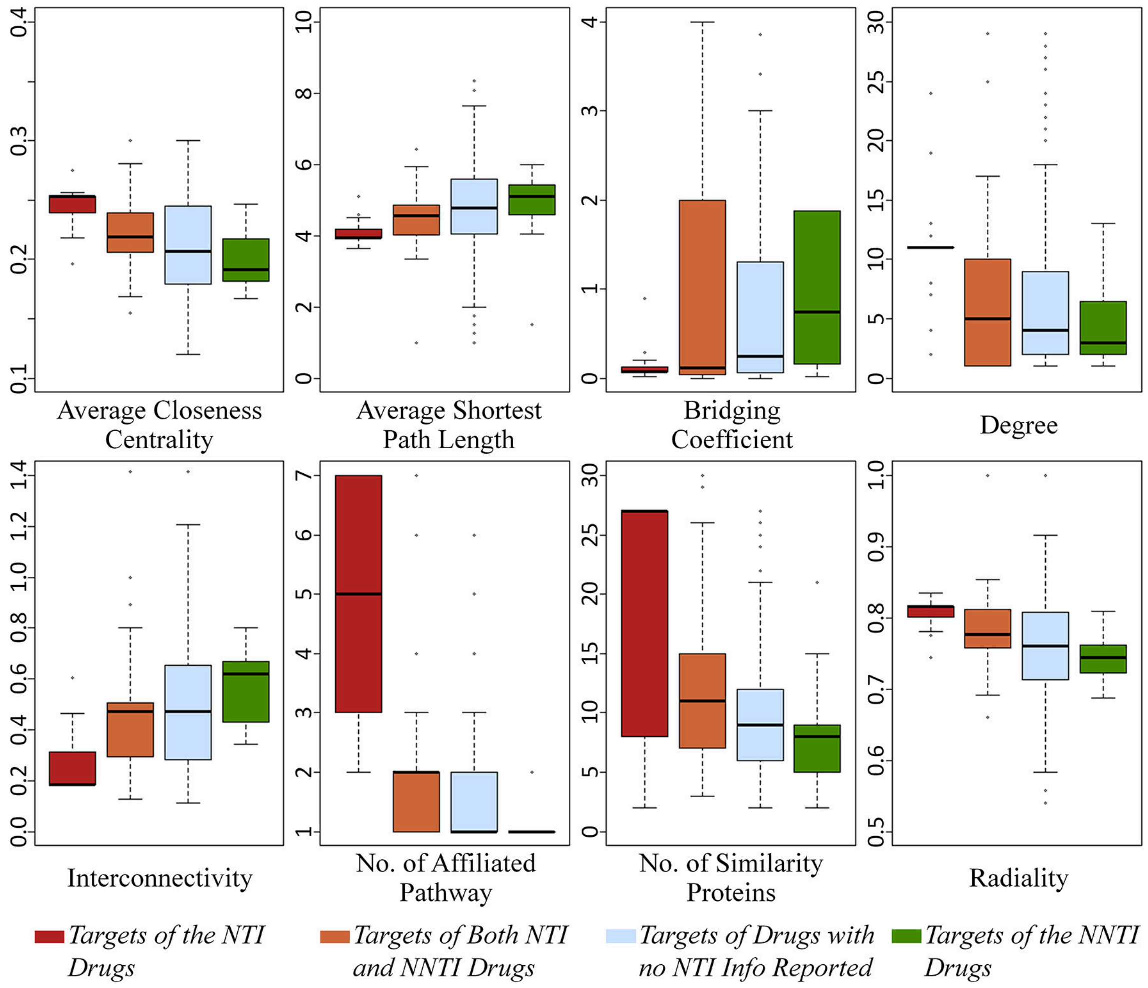


FIGURE 1 | Boxplots of eight key features identified in this study. For each feature, there were four plots colored in red, orange, light blue and green which indicated the targets of NTI drugs, both NTI and NNTI drugs, drugs with no NTI data reported and NNTI drugs, respectively.

five property groups were generated by considering equation and description of these 19 properties, and each group was named by the first property (ordered alphabetically) in the corresponding group. As a result, these five groups included: the *average closeness centrality*, *average shortest path length*, *betweenness centrality*, *degree*, *eccentricity*. To minimize the possible bias induced by the innate mutual dependence among properties, only these five properties were considered in subsequent feature selection analysis, instead of investigating all 19 properties. Taking the remaining 17 relatively independent properties into consideration, 22 properties in total of each target were selected for subsequent feature selection.

As one of the most popular feature selection strategies based on AI, the *Boruta* algorithm based on a wrapper method built around a random forest classifier (Kursa, 2014) was adopted in this study. *Boruta* was considered the most powerful method with the stability in variable selection, especially suitable for the low-dimensional dataset among other reported strategies (Degenhardt et al., 2017). In this study, the key differential features were thus selected from 22 properties using *R package Boruta* by setting the p-value < 0.05. As a result, eight properties were selected as able to collectively reflect the target's mechanism underlying NTI drugs. As illustrated in **Figure 1**, the boxplots colored in red and green referred to the targets of NTI and NNTI drugs, respectively. Some key features increased from the targets of NTI drug to that of NNTI one (such as *average shortest path length*), while others demonstrated a decrease (such as *average closeness centrality*). Based on the comprehensive literature review, some of those 8 key features had been reported to be indirectly relevant to drugs' efficacy-safety balances. For example, the lower value of *average closeness centrality* of target was reported to demonstrate a less lethality risk (Chen et al., 2011), which was consistent with the findings of this study (a much higher *average closeness centrality* of the targets of NTI drugs was observed compared with that of NNTI ones, shown in **Figure 1**). Moreover, the higher level (lower value) of *interconnectivity* was frequently observed in lethal diseases such as cardiovascular disorder and cancer (Muhammd et al., 2018).

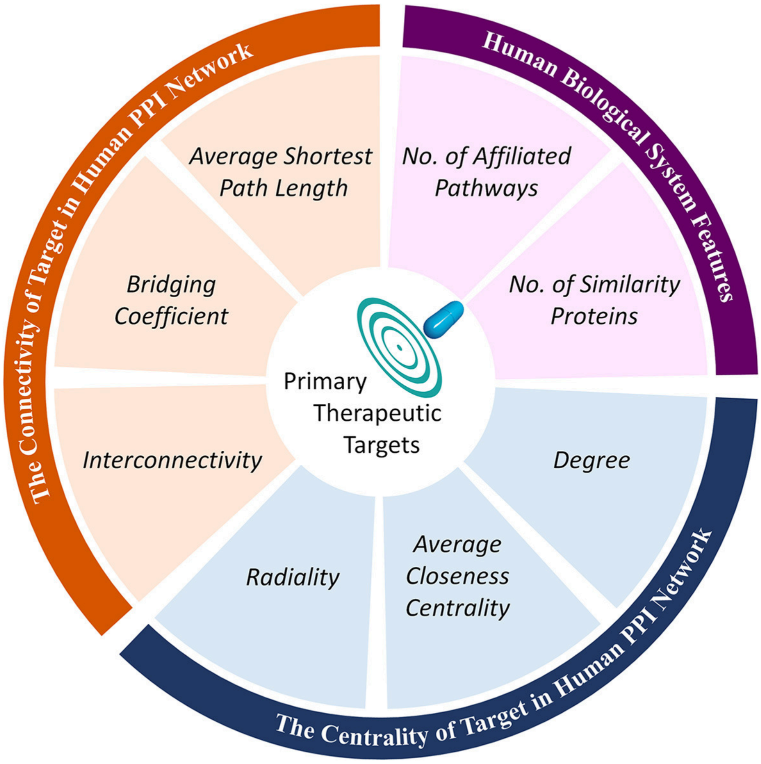

The Connectivity of Target in Human PPI Network
Human Biological System Features
The Centrality of Target in Human PPI Network
Average Shortest Path Length
No. of Affiliated Pathways
Bridging Coefficient
No. of Similarity Proteins
Primary Therapeutic Targets
Interconnectivity
Degree
Radiality
Average Closeness Centrality

FIGURE 2 | Classification of eight key features identified in this study into three feature groups.

Oncological and cardiovascular disorder had been recognized as life-threatening diseases, and the majority of their drugs were reported to be NTI ones (Muller and Milton, 2012; Yu et al., 2015). Thus, the result of *interconnectivity* in **Figure 1** was consistent with these previous reports, which further validated the effectiveness of applied algorithm in identifying key target features underlying NTI drugs.

Moreover, there were four groups of targets as defined in section Materials and Methods: (a) targets of NTI drugs, (b) targets of both NTI and NNTI drugs, (c) targets of drugs without reported TI, and (d) targets of NNTI drugs. Apart from the target groups (a) and (d), the remaining groups provided more complicated and informative data for illustrating the mechanism underlying NTI drugs. On one hand, the targets in group (b) were affected by both NTI and NNTI drugs, which might reflect properties from both sides, but might also be significantly affected by the properties of confirmed NTI drugs. On the other hand, no TI data of the group (c) targets was reported based on literature review. It was possible that some NTI drugs were not discovered for those targets. But considering the large number of group (c) targets (339 in total), it was highly possible that most of those group (c) targets were only aimed by NNTI drugs, and just a small fraction of which could find new NTI drug in the future. The value of 8 properties of those 4 target groups were illustrated in **Figure 1**. It was interesting that all properties followed a clear descending/ascending trend from the targets of group (a) to (d), which was in accordance with the analyses provided above. Thus, these findings could be another line of evidence that validated the effectiveness of the feature identification algorithm applied in this study.

Target Mechanism Underlying NTI Drugs Collectively Determined by Multiple Profiles

By collectively considering **Figure 1** and **Tables 1–3**, seven out of those eight selected key features showed significant difference (p-value < 0.05), but it was clear that these significant differences did not guarantee the corresponding feature as the key differential one (57.7% of the features with significant difference (p-value < 0.05) were not selected as key differential ones). Moreover, significant difference was not observed for the selected key feature *bridging coefficient* (p-value $= 0.22$). This finding indicated that those eight features collectively determined the target mechanism of NTI drugs, and the TI-related mechanism might be the result of the synergistical effects among those features. Moreover, the majority of these eight key features were identified for the first time by this study, and this work was also the first analysis on the collective effects of both PPI network properties and biological system profile on the drug efficacy-safety balance.

Further analysis on these eight identified key features (shown in **Figure 1**) revealed that these key features were found to belong to three feature groups. These feature groups were connectivity and centrality of targets in human PPI network together with human biological system features. By combining the data in **Figure 1**, the key features within the same feature group (illustrated in **Figure 2**) followed the same ascending/descending trends, which were colored by the same background. As shown in **Figure 2**, the targets of NTI drugs were highly centralized and connected, and the number of similarity proteins and the number of affiliated pathways were substantially higher than those of NNTI drug. Since the number of similarity proteins and affiliated pathways was reported to be good indicator of target druggability (Zhu F. et al., 2009; Li et al., 2018), the NTI profile identified in this study was in accordance with that of reported target druggability.

CONCLUSION

This work is the first study conducting comprehensive review on the TI data of all FDA approved drugs (**Supplementary Table S1**) and revealing the collective effects of both human PPI network properties and biological system profiles on drug efficacy-safety balance. Eight key features were identified here as collectively differentiating the target mechanisms between NTI and NNTI drugs. These features revealed that the targets of NTI drugs were highly centralized and connected in human PPI network, and the numbers of similarity proteins and target-affiliated pathways were both much higher than those of NNTI drugs. These findings together with the newly discovered features/feature groups clarified the key factors indicating drug's narrow TI and could therefore provide a novel direction for determining the delicate drug efficacy-safety balance.

AUTHOR CONTRIBUTIONS

FZ conceived the idea and supervised the work. XL, JY, and JT performed the research. XL, JY, JT, YL, QY, ZX, RZ, YW, JH, LT, and WX prepared and analyzed the data. FZ wrote the manuscript. All authors have read and approved this manuscript.

FUNDING

This work was funded by National Natural Science Foundation of China (81872798); Innovation Project on Industrial Generic Key Technologies of Chongqing (cstc2015zdcy-ztzx120003); and Fundamental Research Funds for Central Universities (10611CDJXZ238826, CDJZR14468801, CDJKXB14011, 2015CDJXY).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphar.2018.01245/full#supplementary-material

REFERENCES

Abernethy, D. R., Woodcock, J., and Lesko, L. J. (2011). Pharmacological mechanism-based drug safety assessment and prediction. *Clin. Pharmacol. Ther.* 89, 793–797. doi: 10.1038/clpt.2011.55

Badders, N. M., Korff, A., Miranda, H. C., Vuppala, P. K., Smith, R. B., Winborn, B. J., et al. (2018). Selective modulation of the androgen receptor AF2 domain rescues degeneration in spinal bulbar muscular atrophy. *Nat. Med.* 24, 427–437. doi: 10.1038/nm.4500

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., et al. (2013). NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.* 41, D991–D995. doi: 10.1093/nar/gks1193

Batool, K., and Niazi, M. A. (2014). Towards a methodology for validation of centrality measures in complex networks. *PLoS ONE* 9:e90283. doi: 10.1371/journal.pone.0090283

Bender, A., Scheiber, J., Glick, M., Davies, J. W., Azzaoui, K., Hamon, J., et al. (2007). Analysis of pharmacology data and the prediction of adverse drug reactions and off-target effects from chemical structure. *ChemMedChem* 2, 861–873. doi: 10.1002/cmdc.200700026

Benet, L. Z., Hosey, C. M., Ursu, O., and Oprea, T. I. (2016). BDDCS, the rule of 5 and drugability. *Adv. Drug Deliv. Rev.* 101, 89–98. doi: 10.1016/j.addr.2016.05.007

Bolser, D., Dafas, P., Harrington, R., Park, J., and Schroeder, M. (2003). Visualisation and graph-theoretic analysis of a large-scale protein structural interactome. *BMC Bioinformatics* 4:45. doi: 10.1186/1471-2105-4-45

Braeuning, A. (2013). The connection of beta-catenin and phenobarbital in murine hepatocarcinogenesis: a critical discussion of Awuah et al. *Arch. Toxicol.* 87, 401–402. doi: 10.1007/s00204-012-1002-4

Carson, M. B., and Lu, H. (2015). Network-based prediction and knowledge mining of disease genes. *BMC Med. Genomics* 8(Suppl. 2):S9. doi: 10.1186/1755-8794-8-S2-S9

Chen, L., Wang, Q., Zhang, L., Tai, J., Wang, H., Li, W., et al. (2011). A novel paradigm for potential drug-targets discovery: quantifying relationships of enzymes and cascade interactions of neighboring biological processes to identify drug-targets. *Mol. Biosyst.* 7, 1033–1041. doi: 10.1039/c0mb00249f

Chen, S., Zhang, P., Liu, X., Qin, C., Tao, L., Zhang, C., et al. (2016). Towards cheminformatics-based estimation of drug therapeutic index: predicting the protective index of anticonvulsants using a new quantitative structure-index relationship approach. *J. Mol. Graph. Model.* 67, 102–110. doi: 10.1016/j.jmgm.2016.05.006

Costenbader, E., and ValenteFontanesi, T. W. (2003). The stability of centrality measures when networks are sampled. *Soc. Netw.* 25, 283–307. doi: 10.1016/S0378-8733(03)00012-1

Dangalchev, C. (2006). Residual closeness in networks. *Phys. A* 365, 556–564. doi: 10.1016/j.biortech.2018.08.122

Degenhardt, F., Seifert, S., and Szymczak, S. (2017). Evaluation of variable selection methods for random forests and omics data sets. *Brief. Bioinform.* doi: 10.1093/bib/bbx124. [Epub ahead of print].

Emig, D., Ivliev, A., Pustovalova, O., Lancashire, L., Bureeva, S., Nikolsky, Y., et al. (2013). Drug target prediction and repositioning using an integrated network-based approach. *PLoS ONE* 8:e60618. doi: 10.1371/journal.pone.0060618

Ernst, M., Du, Y., Warsow, G., Hamed, M., Endlich, N., Endlich, K., et al. (2017). FocusHeuristics - expression-data-driven network optimization and disease gene prediction. *Sci. Rep.* 7:42638. doi: 10.1038/srep42638

Ewer, M. S., and Ewer, S. M. (2015). Cardiotoxicity of anticancer treatments. *Nat. Rev. Cardiol.* 12, 547–558. doi: 10.1038/nrcardio.2015.65

Fry, D. W., Harvey, P. J., Keller, P. R., Elliott, W. L., Meade, M., Trachet, E., et al. (2004). Specific inhibition of cyclin-dependent kinase 4/6 by PD 0332991 and associated antitumor activity in human tumor xenografts. *Mol. Cancer. Ther.* 3, 1427–1438.

Fu, J., Tang, J., Wang, Y., Cui, X., Yang, Q., Hong, J., et al. (2018). Discovery of the consistently well-performed analysis chain for SWATH-MS based pharmacoproteomic quantification. *Front. Pharmacol.* 9:681. doi: 10.3389/fphar.2018.00681

Fu, T., Zheng, G., Tu, G., Yang, F., Chen, Y., Yao, X., et al. (2018). Exploring the binding mechanism of metabotropic glutamate receptor 5 negative allosteric modulators in clinical trials by molecular dynamics simulations. *ACS Chem. Neurosci.* 9, 1492–1502. doi: 10.1021/acschemneuro.8b00059

Fujimoto, G. M., Monroe, M. E., Rodriguez, L., Wu, C., MacLean, B., Smith, R. D., et al. (2014). Accounting for population variation in targeted proteomics. *J. Proteome Res.* 13, 321–323. doi: 10.1021/pr4011052

Garlaschelli, D., and Loffredo, M. I. (2004). Patterns of link reciprocity in directed networks. *Phys. Rev. Lett.* 93:268701. doi: 10.1103/PhysRevLett.93.268701

Ghosh, S., Kumar, G. V., Basu, A., and Banerjee, A. (2015). Graph theoretic network analysis reveals protein pathways underlying cell death following neurotropic viral infection. *Sci. Rep.* 5:14438. doi: 10.1038/srep14438

Guo, R., Zhang, X., Su, J., Xu, H., Zhang, Y., Zhang, F., et al. (2018). Identifying potential quality markers of Xin-Su-Ning capsules acting on arrhythmia by integrating UHPLC-LTQ-Orbitrap, ADME prediction and network target analysis. *Phytomedicine* 44, 117–128. doi: 10.1016/j.phymed.2018.01.019

Han, Z. J., Xue, W. W., Tao, L., and Zhu, F. (2018). Identification of novel immune-relevant drug target genes for Alzheimer's disease by combining ontology inference with network analysis. *CNS Neurosci. Ther.* doi: 10.1111/cns.13051. [Epub ahead of print].

Huang, L. H., He, Q. S., Liu, K., Cheng, J., Zhong, M. D., Chen, L. S., et al. (2018). ADReCS-Target: target profiles for aiding drug safety research and application. *Nucleic Acids Res.* 46, D911–D917. doi: 10.1093/nar/gkx899

Hwang, W. C., Zhang, A., and Ramanathan, M. (2008). Identification of information flow-modulating drug targets: a novel bridging paradigm for drug discovery. *Clin. Pharmacol. Ther.* 84, 563–572. doi: 10.1038/clpt.2008.129

Iloro, I., Gonzalez, E., Gutierrez-de Juan, V., Mato, J. M., Falcon-Perez, J. M., and Elortza, F. (2013). Non-invasive detection of drug toxicity in rats by solid-phase extraction and MALDI-TOF analysis of urine samples. *Anal. Bioanal. Chem.* 405, 2311–2320. doi: 10.1007/s00216-012-6644-9

Jiang, X. L., Samant, S., Lesko, L. J., and Schmidt, S. (2015). Clinical pharmacokinetics and pharmacodynamics of clopidogrel. *Clin. Pharmacokinet.* 54, 147–166. doi: 10.1007/s40262-014-0230-6

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 45, D353–D361. doi: 10.1093/nar/gkw1092

Kimmelman, J., and Federico, C. (2017). Consider drug efficacy before first-in-human trials. *Nature* 542, 25–27. doi: 10.1038/542025a

Kivimäki, I., Lebichot, B., Saramaki, J., and Saerens, M. (2016). Two betweenness centrality measures based on randomized shortest paths. *Sci. Rep.* 6:19668. doi: 10.1038/srep19668

Kogenaru, S., del Val, C., Hotz-Wagenblatt, A., and Glatting, K. H. (2010). TissueDistributionDBs: a repository of organism-specific tissue-distribution profiles. *Theor. Chem. Acc.* 125:9. doi: 10.1186/s40199-014-0080-7

Koschützki, D., and Schreiber, F. (2008). Centrality analysis methods for biological networks and their application to gene regulatory networks. *Gene Regul. Syst. Biol.* 2, 193–201.

Krittanawong, C., Zhang, H., Wang, Z., Aydar, M., and Kitai, T. (2017). Artificial intelligence in precision cardiovascular medicine. *J. Am. Coll. Cardiol.* 69, 2657–2664. doi: 10.1016/j.jacc.2017.03.571

Kursa, M. B. (2014). Robustness of random forest-based gene selection methods. *BMC Bioinformatics* 15:8. doi: 10.1186/1471-2105-15-8

Lapenna, S., and Giordano, A. (2009). Cell cycle kinases as therapeutic targets for cancer. *Nat. Rev. Drug Discov.* 8, 547–566. doi: 10.1038/nrd2907

Li, B., Tang, J., Yang, Q., Cui, X., Li, S., Chen, S., et al. (2016). Performance evaluation and online realization of data-driven normalization methods used in LC/MS based untargeted metabolomics analysis. *Sci. Rep.* 6:38881. doi: 10.1038/srep38881

Li, B., Tang, J., Yang, Q., Li, S., Cui, X., Li, Y., et al. (2017). NOREVA: normalization and evaluation of MS-based metabolomics data. *Nucleic Acids Res.* 45, W162–W170. doi: 10.1093/nar/gkx449

Li, W., Chen, L., Li, X., Jia, X., Feng, C., Zhang, L., et al. (2013). Cancer-related marketing centrality motifs acting as pivot units in the human signaling network and mediating cross-talk between biological pathways. *Mol. Biosyst.* 9, 3026–3035. doi: 10.1039/c3mb70289h

Li, Y. H., Wang, P. P., Li, X. X., Yu, C. Y., Yang, H., Zhou, J., et al. (2016b). The human kinome targeted by FDA approved multi-target drugs and combination products: a comparative study from the drug-target interaction network perspective. *PLoS ONE* 11:e0165737. doi: 10.1371/journal.pone.0165737

Li, Y. H., Xu, J. Y., Tao, L., Li, X. F., Li, S., Zeng, X., et al. (2016a). SVM-prot 2016: a web-server for machine learning prediction of protein functional families from sequence irrespective of similarity. *PLoS ONE* 11:e0155290. doi: 10.1371/journal.pone.0155290

Li, Y. H., Yu, C. Y., Li, X. X., Zhang, P., Tang, J., Yang, Q., et al. (2018). Therapeutic target database update 2018: enriched resource for facilitating bench-to-clinic research of targeted therapeutics. *Nucleic Acids Res.* 46, D1121–D1127. doi: 10.1093/nar/gkx1076

Ma, B., Wang, H., Dsouza, M., Lou, J., He, Y., Dai, Z., et al. (2016). Geographic patterns of co-occurrence network topological features for soil microbiota at continental scale in eastern China. *ISME J.* 10, 1891–1901. doi: 10.1038/ismej.2015.261

Muhammd, J., Khan, A., Ali, A., Fang, L., Yanjing, W., Xu, Q., et al. (2018). Network pharmacology: exploring the resources and methodologies. *Curr. Top. Med. Chem.* 18, 949–964. doi: 10.2174/156802661866618030141351

Muller, P. Y., and Dieterle, F. (2009). Tissue-specific, non-invasive toxicity biomarkers: translation from preclinical safety assessment to clinical safety monitoring. *Expert Opin. Drug Metab. Toxicol.* 5, 1023–1038. doi: 10.1517/17425250903114174

Muller, P. Y., and Milton, M. N. (2012). The determination and interpretation of the therapeutic index in drug development. *Nat. Rev. Drug Discov.* 11, 751–761. doi: 10.1038/nrd3801

Paladugu, S. R., Zhao, S., Ray, A., and Raval, A. (2008). Mining protein networks for synthetic genetic interactions. *BMC Bioinformatics* 9:426. doi: 10.1186/1471-2105-9-426

Pan, Y., Wang, Z., Zhan, W., and Deng, L. (2018). Computational identification of binding energy hot spots in protein-RNA complexes using an ensemble approach. *Bioinformatics* 34, 1473–1480. doi: 10.1093/bioinformatics/btx822

Ragusa, M., Avola, G., Angelica, R., Barbagallo, D., Guglielmino, M. R., Duro, L. R., et al. (2010). Expression profile and specific network features of the apoptotic machinery explain relapse of acute myeloid leukemia after chemotherapy. *BMC Cancer* 10:377. doi: 10.1186/1471-2407-10-377

Rao, H. B., Zhu, F., Yang, G. B., Li, Z. R., and Chen, Y. Z. (2011). Update of PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Res.* 39, W385–W390. doi: 10.1093/nar/gkr284

Rodgers, A. D., Zhu, H., Fourches, D., Rusyn, I., and Tropsha, A. (2010). Modeling liver-related adverse effects of drugs using knearest neighbor quantitative structure-activity relationship method. *Chem. Res. Toxicol.* 23, 724–732. doi: 10.1021/tx900451r

Rogelj, P., Hudej, R., and Petric, P. (2013). Distance deviation measure of contouring variability. *Radiol. Oncol.* 47, 86–96. doi: 10.2478/raon-2013-0005

Rubinov, M., and Sporns, O. (2010). Complex network measures of brain connectivity: uses and interpretations. *Neuroimage* 52, 1059–1069. doi: 10.1016/j.neuroimage.2009.10.003

Rzepecki, A. K., Cheng, H., and McLellan, B. N. (2018). Cutaneous toxicity as a predictive biomarker for clinical outcome in patients receiving anticancer therapy. *J. Am. Acad. Dermatol.* 79, 545–555. doi: 10.1016/j.jaad.2018.04.046

Seo, S., Oh, M., Park, Y., and Kim, S. (2018). DeepFam: deep learning based alignment-free method for protein family modeling and prediction. *Bioinformatics* 34, i254–i262. doi: 10.1093/bioinformatics/bty275

Shang, L., Liu, C., Tomiura, Y., and Hayashi, K. (2017). Machine-learning-based olfactometer: prediction of odor perception from physicochemical features of odorant molecules. *Anal. Chem.* 89, 11999–12005. doi: 10.1021/acs.analchem.7b02389

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303

Singh, S., Malik, B. K., and Sharma, D. K. (2007). Choke point analysis of metabolic pathways in E.histolytica: a computational approach for drug target identification. *Bioinformation* 2, 68–72. doi: 10.6026/97320630002068

Solá, L., Romance, M., Criado, R., Flores, J., Garcia del Amo, A., and Boccaletti, S. (2013). Eigenvector centrality of nodes in multiplex networks. *Chaos* 23:033131. doi: 10.1063/1.4818544

Song, L., Xu, W., Li, C., Li, H., Wu, L., Xiang, J., et al. (2006). Development of expressed sequence tags from the bay scallop, *Argopecten irradians irradians*. *Mar. Biotechnol.* 8, 161–169. doi: 10.1007/s10126-005-0126-4

Sormani, M. P. (2012). Modeling the distribution of new MRI cortical lesions in multiple sclerosis longitudinal studies. *Mult. Scler. Relat. Disord.* 1:108. doi: 10.1016/j.msard.2012.01.001

Sotillo, R., Dubus, P., Martin, J., de la Cueva, E., Ortega, S., Malumbres, M., et al. (2001). Wide spectrum of tumors in knock-in mice carrying a Cdk4 protein insensitive to INK4 inhibitors. *EMBO J.* 20, 6637–6647. doi: 10.1093/emboj/20.23.6637

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., et al. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* 43, D447–D452. doi: 10.1093/nar/gku1003

Tao, L., Zhu, F., Qin, C., Zhang, C., Xu, F., Tan, C. Y., et al. (2014). Nature's contribution to today's pharmacopeia. *Nat. Biotechnol.* 32, 979–980. doi: 10.1038/nbt.3034

Tao, L., Zhu, F., Xu, F., Chen, Z., Jiang, Y. Y., and Chen, Y. Z. (2015). Co-targeting cancer drug escape pathways confers clinical advantage for multi-target anticancer drugs. *Pharmacol. Res.* 102, 123–131. doi: 10.1016/j.phrs.2015.09.019

Thomas, S., and Bonchev, D. (2010). A survey of current software for network analysis in molecular biology. *Hum. Genomics* 4, 353–360. doi: 10.1186/1479-7364-4-5-353

UniProt Consortium, T. (2018). UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* 46:2699. doi: 10.1093/nar/gky092

van der Burgh, H. K., Schmidt, R., Westeneng, H. J., de Reus, M. A., van den Berg, L. H., and van den Heuvel, M. P. (2017). Deep learning predictions of survival based on MRI in amyotrophic lateral sclerosis. *Neuroimage Clin.* 13, 361–369. doi: 10.1016/j.nicl.2016.10.008

Wang, P., Fu, T., Zhang, X., Yang, F., Zheng, G., Xue, W., et al. (2017a). Differentiating physicochemical properties between NDRIs and sNRIs clinically important for the treatment of ADHD. *Biochim. Biophys. Acta* 1861, 2766–2777. doi: 10.1016/j.bbagen.2017.07.022

Wang, P., Yang, F., Yang, H., Xu, X., Liu, D., Xue, W., et al. (2015). Identification of dual active agents targeting 5-HT1A and SERT by combinatorial virtual screening methods. *Biomed. Mater. Eng.* 26(Suppl. 1), S2233–2239. doi: 10.3233/BME-151529

Wang, P., Zhang, X., Fu, T., Li, S., Li, B., Xue, W., et al. (2017b). Differentiating physicochemical properties between addictive and nonaddictive ADHD drugs revealed by molecular dynamics simulation studies. *ACS Chem. Neurosci.* 8, 1416–1428. doi: 10.1021/acschemneuro.7b00173

Wang, S., Tong, Y., Ng, T. B., Lao, L., Lam, J. K., Zhang, K. Y., et al. (2015). Network pharmacological identification of active compounds and potential actions of Erxian decoction in alleviating menopause-related symptoms. *Chin. Med.* 10:19. doi: 10.1186/s13020-015-0051-z

Watts, D. J., and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature* 393, 440–442. doi: 10.1038/30918

Xu, J., Wang, P., Yang, H., Zhou, J., Li, Y., Li, X., et al. (2016). Comparison of FDA approved kinase targets to clinical trial ones: insights from their system profiles and drug-target interaction networks. *Biomed. Res. Int.* 2016:2509385. doi: 10.1155/2016/2509385

Xu, R., and Wang, Q. (2014). Automatic construction of a large-scale and accurate drug-side-effect association knowledge base from biomedical literature. *J. Biomed. Inform.* 51, 191–199. doi: 10.1016/j.jbi.2014.05.013

Xue, W., Wang, P., Li, B., Li, Y., Xu, X., Yang, F., et al. (2016). Identification of the inhibitory mechanism of FDA approved selective serotonin reuptake inhibitors: an insight from molecular dynamics simulation study. *Phys. Chem. Chem. Phys.* 18, 3260–3271. doi: 10.1039/c5cp05771j

Xue, W., Wang, P., Tu, G., Yang, F., Zheng, G., Li, X., et al. (2018a). Computational identification of the binding mechanism of a triple reuptake inhibitor amitifadine for the treatment of major depressive disorder. *Phys. Chem. Chem. Phys.* 20, 6606–6616. doi: 10.1039/c7cp07869b

Xue, W., Yang, F., Wang, P., Zheng, G., Chen, Y., Yao, X., et al. (2018b). What contributes to serotonin-norepinephrine reuptake inhibitors' dual-targeting mechanism? The key role of transmembrane domain 6 in human serotonin and norepinephrine transporters revealed by molecular dynamics simulation. *ACS Chem. Neurosci.* 9, 1128–1140. doi: 10.1021/acschemneuro.7b00490

Yang, F., Zheng, G., Fu, T., Li, X., Tu, G., Li, Y. H., et al. (2018). Prediction of the binding mode and resistance profile for a dual-target pyrrolyl diketo acid scaffold against HIV-1 integrase and reverse-transcriptase-associated ribonuclease H. *Phys. Chem. Chem. Phys.* 20, 23873–23884. doi: 10.1039/c8cp01843j

Yang, F. Y., Fu, T. T., Zhang, X. Y., Hu, J., Xue, W. W., Zheng, G. X., et al. (2017). Comparison of computational model and X-ray crystal structure of human serotonin transporter: potential application for the pharmacology of human monoamine transporters. *Mol. Simul.* 43, 1089–1098. doi: 10.1080/08927022.2017.1309653

Yang, H., Qin, C., Li, Y. H., Tao, L., Zhou, J., Yu, C. Y., et al. (2016). Therapeutic target database update 2016: enriched resource for bench to clinical drug target and targeted pathway information. *Nucleic Acids Res.* 44, D1069–D1074. doi: 10.1093/nar/gkv1230

Yu, C. Y., Li, X. X., Yang, H., Li, Y. H., Xue, W. W., Chen, Y. Z., et al. (2018). Assessing the performances of protein function prediction algorithms from the perspectives of identification accuracy and false discovery rate. *Int. J. Mol. Sci.* 19:19010183. doi: 10.3390/ijms19010183

Yu, L. X., Jiang, W., Zhang, X., Lionberger, R., Makhlouf, F., Schuirmann, D. J., et al. (2015). Novel bioequivalence approach for narrow therapeutic index drugs. *Clin. Pharmacol. Ther.* 97, 286–291. doi: 10.1002/cpt.28

Zaykov, A. N., Mayer, J. P., and DiMarchi, R. D. (2016). Pursuit of a perfect insulin. *Nat. Rev. Drug Discov.* 15, 425–439. doi: 10.1038/nrd.2015.36

Zeidán-Chuliá, F., Gursoy, M., Neves de Oliveira, B. H., Ozdemir, V., Kononen, E., and Gursoy, U. K. (2015). A systems biology approach to reveal putative host-derived biomarkers of periodontitis by network topology characterization of MMP-REDOX/NO and apoptosis integrated pathways. *Front. Cell. Infect. Microbiol.* 5:102. doi: 10.3389/fcimb.2015.00102

Zhang, P., Tao, L., Zeng, X., Qin, C., Chen, S., Zhu, F., et al. (2017a). A protein network descriptor server and its use in studying protein, disease, metabolic and drug targeted networks. *Brief. Bioinform.* 18, 1057–1070. doi: 10.1093/bib/bbw071

Zhang, P., Tao, L., Zeng, X., Qin, C., Chen, S. Y., Zhu, F., et al. (2017b). PROFEAT update: a protein features web server with added facility to compute network descriptors for studying omics-derived networks. *J. Mol. Biol.* 429, 416–425. doi: 10.1016/j.jmb.2016.10.013

Zhang, W., Zhang, Q., Zhang, M., Zhang, Y., Li, F., and Lei, P. (2014). Network analysis in the identification of special mechanisms between small cell lung cancer and non-small cell lung cancer. *Thorac. Cancer* 5, 556–564. doi: 10.1111/1759-7714.12134

Zheng, C. J., Han, L. Y., Yap, C. W., Ji, Z. L., Cao, Z. W., and Chen, Y. Z. (2006). Therapeutic targets: progress of their exploration and investigation of their characteristics. *Pharmacol. Rev.* 58, 259–279. doi: 10.1124/pr.58.2.4

Zheng, G., Xue, W., Wang, P., Yang, F., Li, B., Li, X., et al. (2016). Exploring the inhibitory mechanism of approved selective norepinephrine reuptake inhibitors and reboxetine enantiomers by molecular dynamics study. *Sci. Rep.* 6:26883. doi: 10.1038/srep26883

Zheng, G., Xue, W., Yang, F., Zhang, Y., Chen, Y., Yao, X., et al. (2017). Revealing vilazodone's binding mechanism underlying its partial agonism to the 5-HT1A receptor in the treatment of major depressive disorder. *Phys. Chem. Chem. Phys.* 19, 28885–28896. doi: 10.1039/c7cp05688e

Zhu, F., Han, B., Kumar, P., Liu, X., Ma, X., Wei, X., et al. (2010). Update of TTD: therapeutic target database. *Nucleic Acids Res.* 38, D787–D791. doi: 10.1093/nar/gkp1014

Zhu, F., Han, L., Zheng, C., Xie, B., Tammi, M. T., Yang, S., et al. (2009). What are next generation innovative therapeutic targets? Clues from genetic, structural, physicochemical, and systems profiles of successful targets. *J. Pharmacol. Exp. Ther.* 330, 304–315. doi: 10.1124/jpet.108.149955

Zhu, F., Han, L. Y., Chen, X., Lin, H. H., Ong, S., Xie, B., et al. (2008a). Homology-free prediction of functional class of proteins and peptides by support vector machines. *Curr. Protein Pept. Sci.* 9, 70–95. doi: 10.2174/138920308783565697

Zhu, F., Li, X. X., Yang, S. Y., and Chen, Y. Z. (2018). Clinical success of drug targets prospectively predicted by *in silico* study. *Trends Pharmacol. Sci.* 39, 229–231. doi: 10.1016/j.tips.2017.12.002

Zhu, F., Ma, X. H., Qin, C., Tao, L., Liu, X., Shi, Z., et al. (2012a). Drug discovery prospect from untapped species: indications from approved natural product drugs. *PLoS ONE* 7:e39782. doi: 10.1371/journal.pone.0039782

Zhu, F., Qin, C., Tao, L., Liu, X., Shi, Z., Ma, X., et al. (2011). Clustered patterns of species origins of nature-derived drugs and clues for future bioprospecting. *Proc. Natl. Acad. Sci. U. S. A.* 108, 12943–12948. doi: 10.1073/pnas.1107336108

Zhu, F., Shi, Z., Qin, C., Tao, L., Liu, X., Xu, F., et al. (2012b). Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery. *Nucleic Acids Res.* 40, D1128–D1136. doi: 10.1093/nar/gkr797

Zhu, F., Zheng, C. J., Han, L. Y., Xie, B., Jia, J., Liu, X., et al. (2008b). Trends in the exploration of anticancer targets and strategies in enhancing the efficacy of drug targeting. *Curr. Mol. Pharmacol.* 1, 213–232. doi: 10.2174/1874467210801030213

Zhu, H., Tropsha, A., Fourches, D., Varnek, A., Papa, E., Gramatica, P., et al. (2008). Combinatorial QSAR modeling of chemical toxicants tested against Tetrahymena pyriformis. *J. Chem. Inf. Model.* 48, 766–784. doi: 10.1021/ci700443v

Zhu, M., Gao, L., Li, X., Liu, Z., Xu, C., Yan, Y., et al. (2009). The analysis of the drug-targets based on the topological properties in the human protein-protein interaction network. *J. Drug Target.* 17, 524–532. doi: 10.1080/10611860903046610

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Pharmacology

ORIGINAL RESEARCH
published: 08 March 2019
doi: 10.3389/fphar.2019.00191

Quantitative Systems Pharmacological Analysis of Drugs of Abuse Reveals the Pleiotropy of Their Targets and the Effector Role of mTORC1

Fen Pei[†], *Hongchun Li*[†], *Bing Liu** *and Ivet Bahar**

Department of Computational and Systems Biology, School of Medicine, University of Pittsburgh, Pittsburgh, PA, United States

Existing treatments against drug addiction are often ineffective due to the complexity of the networks of protein-drug and protein-protein interactions (PPIs) that mediate the development of drug addiction and related neurobiological disorders. There is an urgent need for understanding the molecular mechanisms that underlie drug addiction toward designing novel preventive or therapeutic strategies. The rapidly accumulating data on addictive drugs and their targets as well as advances in machine learning methods and computing technology now present an opportunity to systematically mine existing data and draw inferences on potential new strategies. To this aim, we carried out a comprehensive analysis of cellular pathways implicated in a diverse set of 50 drugs of abuse using quantitative systems pharmacology methods. The analysis of the drug/ligand-target interactions compiled in DrugBank and STITCH databases revealed 142 known and 48 newly predicted targets, which have been further analyzed to identify the KEGG pathways enriched at different stages of drug addiction cycle, as well as those implicated in cell signaling and regulation events associated with drug abuse. Apart from synaptic neurotransmission pathways detected as upstream signaling modules that "sense" the early effects of drugs of abuse, pathways involved in neuroplasticity are distinguished as determinants of neuronal morphological changes. Notably, many signaling pathways converge on important targets such as mTORC1. The latter emerges as a universal effector of the persistent restructuring of neurons in response to continued use of drugs of abuse.

Keywords: drug abuse, quantitative systems pharmacology, pleiotropic proteins, mTOR complex 1, drug-target interactions, neurotransmission, machine learning, cellular pathways

OPEN ACCESS

Edited by:
Zhi-Liang Ji,
Xiamen University, China

Reviewed by:
Amit P. Bhavsar,
University of Alberta, Canada
Huanmei Wu,
Indiana University, Purdue University Indianapolis, United States

****Correspondence:***
Bing Liu
liubing@pitt.edu
Ivet Bahar
bahar@pitt.edu

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Translational Pharmacology, a section of the journal Frontiers in Pharmacology

Received: *08 November 2018*
Accepted: *14 February 2019*
Published: *08 March 2019*

Citation:
Pei F, Li H, Liu B and Bahar I (2019) Quantitative Systems Pharmacological Analysis of Drugs of Abuse Reveals the Pleiotropy of Their Targets and the Effector Role of mTORC1. Front. Pharmacol. 10:191. doi: 10.3389/fphar.2019.00191

INTRODUCTION

Drug addiction is a chronic relapsing disorder characterized by compulsive, excessive, and self-damaging use of drugs of abuse. It is a debilitating condition that potentially leads to serious physiological injury, mental disorder and death, resulting in major health, and social economic impacts worldwide (Nestler, 2013; Koob and Volkow, 2016). Substances with diverse chemical structures and mechanisms of action are known to cause addiction. Except for alcohol and tobacco, substances of abuse are commonly classified into six groups based on their primary targets or effects: cannabinoids (e.g., cannabis), opioids (e.g., morphine, heroin, fentanyl), central nervous system (CNS) depressants (e.g., pentobarbital, diazepam), CNS stimulants (e.g., cocaine, amphetamine), hallucinogens (e.g., ketamine, lysergic acid diethylamide), and anabolic steroids (e.g., nandrolone, oxymetholone).

The primary actions of drugs of abuse have been well studied. In spite of the pleiotropy and heterogeneity of drugs of abuse, they share similar phenotypes: from acute intoxication to chronic dependence (Taylor et al., 2013), the reinforcement shift from positive to negative through a three-stage cycle involving binge/intoxication, withdrawal/negative effect, and preoccupation/anticipation (Koob and Volkow, 2016). Notably, virtually all drugs of abuse augment dopaminergic transmission in the reward system (Wise, 1996). However, the detailed cellular pathways of addiction processes are still far from known. For example, cocaine acts primarily as an inhibitor of dopamine (DA) transporter (DAT) and results in DA accumulation in the synapses of DA neurons (Shimada et al., 1991; Volkow et al., 1997). However, it has been shown that DA accumulation *per se* is not sufficient to account for the rewarding process associated with cocaine addiction; serotonin (5-HT) and noradrenaline (or norepinephrine, NE) also play important roles (Rocha et al., 1998; Sora et al., 1998). Another example is ketamine, a non-selective antagonist for N-methyl-d-aspartate (NMDA) receptor (NMDAR), notably most effective in the amygdala and hippocampal regions of neurons (Collingridge et al., 1983). In addition to its primary action, ketamine affects a number of other neurotransmitter receptors, including sigma-1 (Mendelsohn et al., 1985), substance P (Okamoto et al., 2003), opioid (Hustveit et al., 1995), muscarinic acetylcholine (mACh) (Hirota et al., 2002), nicotinic acetylcholine (nACh) (Coates and Flood, 2001), serotonin (Kapur and Seeman, 2002), and γ-aminobutyric acid (GABA) receptors (Hevers et al., 2008). The promiscuity of drugs of abuse brings an additional layer of complexity, which prevents the development of efficient treatment against drug addiction.

In recent years, there has been significant progress in the characterization of drug/target/pathway relations driven by the accumulation of drug-target interactions and pathways data, as well as the development of machine learning, *in silico* genomics, chemogenomics, and quantitative systems pharmacology (QSP) tools. Several innovative studies started to provide valuable information on substance abuse targets and pathways. For example, Li et al. curated 396 drug abuse related genes from the literature and identified five common pathways underlying the reward and addiction actions of cocaine, alcohol, opioids, and nicotine (Li et al., 2008). Hu et al. analyzed the genes related to nicotine addiction via a pathway and network-based approach (Hu et al., 2018). Biernacka et al. performed genome-wide analysis on 1,165 alcohol-dependence cases and identified two pathways associated with alcohol dependence (Biernacka et al., 2013). Xie et al. generated chemogenomics knowledgebases focused on G-protein coupled receptors (GPCRs) related to drugs of abuse in general (Xie et al., 2014), and cannabinoids in particular (Xie et al., 2016). Notably, these studies have shed light on selected categories or subgroups of drugs. There is a need to understand the intricate couplings between multiple pathways implicated in the cellular response to drugs of abuse, identify mechanisms common to various categories of drugs while distinguishing those unique to selected categories.

We undertake here such a systems-level approach using a dataset composed of six different categories of drugs of abuse. Following a QSP approach proposed earlier (Stern et al., 2016), we provide a comprehensive, unbiased glimpse of the complex mechanisms implicated in addiction. Specifically, as shown in **Figure 1**, a set of 50 drugs of abuse with a diversity of chemical structures (**Supplementary Figure 1**) and pharmacological actions were collected as probes, and the known targets of these drugs as well as the targets predicted using our probabilistic matrix factorization (PMF) method (Cobanoglu et al., 2013) were analyzed to infer biological pathways associated with drug addiction. Our analysis yielded 142 known and 48 predicted targets and 173 pathways permitting us to identify both generic mechanisms regulating the responses to drug abuse as well as specific mechanisms associated with selected categories, which could facilitate the development of auxiliary agents for treatment of addiction.

A key step in our approach is to identify the targets for drugs of abuse. There exists various drug-target interaction databases (DBs), web servers and computational models, as summarized recently (Chen et al., 2016). The DBs utilized in this work are the drug-target database DrugBank (Wishart et al., 2018) and the protein-chemical database STITCH (Szklarczyk et al., 2016). DrugBank is a bioinformatics and cheminformatics resource that combines drug data with comprehensive target information. It is frequently updated, with the current version containing 10,562 drugs, 4,493 targets and corresponding 16,959 interactions. Since most of drugs of abuse are approved or withdrawn drugs, DrugBank is a good source for obtaining information on their interactions. STITCH, on the other hand, is much more extensive. It integrates chemical-protein interactions from experiments, other DBs, literature and predictions, resulting in data on 430,000 chemicals and 9,643,763 proteins across 2,031 genomes. We have used in the present analysis the subset of human protein-chemicals data supported by experimental evidence. The method of approach adopted here is an important advance over our original PMF-based machine learning methodology for predicting drug-target interactions (Cobanoglu et al., 2013). First, the approach originally developed for mining DrugBank has been extended to analyzing the STITCH DB, the content of which is 2–3 orders of magnitude larger than DrugBank (based on the respective numbers of interactions). Second, the information on predicted drug-target associations is complemented by pathway data on *humans* inferred from the KEGG pathway DB (December 2017 version; Kanehisa et al., 2017) upon pathway enrichment analysis of known and predicted targets. Third, the outputs are subjected to extensive analyses to detect recurrent patterns and formulate new hypotheses for preventive or therapeutic strategies against drug abuse.

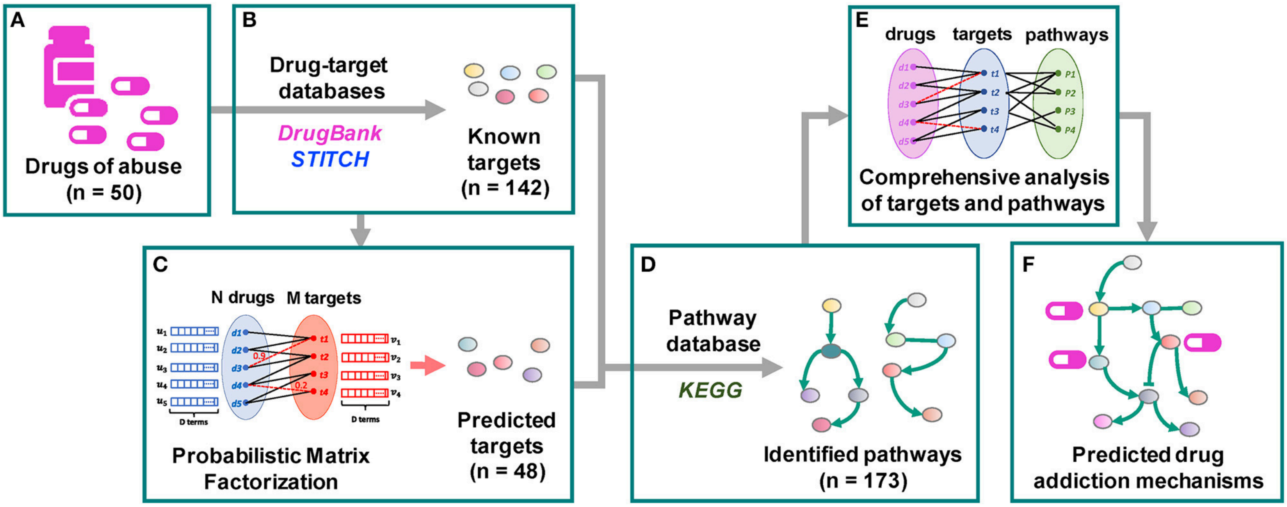


FIGURE 1 | Workflow of the quantitative systems pharmacological analysis. **(A)** 50 drugs of abuse with a diversity of chemical structures and pharmacological actions were collected as probes. **(B)** 142 known targets of these drugs were identified through drug-target interaction database DrugBank and chemical-protein interaction database STITCH. **(C)** 48 predicted targets were predicted using our probabilistic matrix factorization (PMF) method (Cobanoglu et al., 2013). **(D)** 173 human pathways were inferred from the KEGG pathways database by mapping the known and predicted targets. **(E,F)** The pathways were grouped into 5 clusters. The functioning of identified targets and pathways and their involvement in drug addiction were comprehensively examined.

MATERIALS AND METHODS

Selection of Drugs of Abuse and Their Known Targets

We selected as input 50 drugs commonly known as drugs of abuse using two basic criteria: (i) diversity in terms of structure and mode of action, and (ii) availability of information on at least one human target protein in DrugBank v5 (Wishart et al., 2018) or STITCH v5 (Szklarczyk et al., 2016). The selected drugs represent six different categories: CNS stimulants, CNS depressants, opioids, cannabinoids, anabolic steroids, and hallucinogens (see **Supplementary Table 1** and **Supplementary Figure 1**).

A dataset of 142 known targets, listed in **Supplementary Table 2**, were retrieved from DrugBank and STITCH DBs for these 50 drugs. The list includes all targets reported for these drugs in DrugBank, and those with high confidence score, based on experiments, reported in STITCH. Each chemical-target interaction is annotated with five confidence scores in STITCH: experimental, DB, text-mining, prediction, and a combination score of the previous four, each ranging from 0 to 1. We selected the human protein targets with experimental confidence scores of 0.4 or higher. **Supplementary Table 2** summarizes the 142 targets we identified as well as the associated 445 drug-target interactions.

Structure-based and interaction-pattern-based similarities between pairs of drugs were evaluated using two different criteria. The former was based on *structure-based distance* calculated as the Tanimoto distance between their 2D structure fingerprints. Tanimoto distances were evaluated using Python RDKit suite (RDKit: Open-Source Cheminformatics Software. https://www.rdkit.org/). Similarities based on their interactions patterns with known targets were evaluated by evaluating *target-based distances*. To this aim, we represented each drug i by a 142-dimensional "target vector" $\boldsymbol{d}_i$, the entries of which represent the known targets and are assigned values of 0 or 1, depending on the existence/observation of an interaction between the corresponding target and drug i. Interaction-pattern similarities between drug pairs i and j were evaluated by calculating the correlation cosine $\cos(\boldsymbol{d}_i \, . \, \boldsymbol{d}_j) = (\boldsymbol{d}_i \, . \, \boldsymbol{d}_j)/(|\boldsymbol{d}_i| \; |\boldsymbol{d}_j|)$ between these vectors, and the corresponding cosine distance is $[1–\cos(\boldsymbol{d}_i \, . \, \boldsymbol{d}_j)]$. Likewise, *ligand-based distances* between target pairs i and j were evaluated as the cosine distance between the 50-dimensional vectors $\boldsymbol{t}_i$ and $\boldsymbol{t}_j$ corresponding to the two targets, the entries of which are 0 or 1 depending on absence or existence of an interaction between the target and the corresponding drug of abuse.

Probabilistic Matrix Factorization (PMF) Based Drug-Target Interaction Prediction

Novel targets for each drug were predicted using our probabilistic matrix factorization (PMF) based machine learning approach (Cobanoglu et al., 2013, 2015). Briefly, we start with a sparse matrix $\boldsymbol{R}$ representing the known interactions between N drugs and M targets. Using the PMF algorithm, we decomposed $\boldsymbol{R}$ into a drug matrix $\boldsymbol{U}$ and a target matrix $\boldsymbol{V}$, by learning the optimal D latent variables to represent each drug and each target. The product of $\boldsymbol{U}^T$ and $\boldsymbol{V}$ assigns values to the unknown (experimentally not characterized) entries of the reconstructed $\boldsymbol{R}$, each value representing the *confidence score* for a novel drug-target interaction prediction

$$R_{N\times M} = U^T_{N\times D} V_{D\times M}$$

Using this method, we trained two PMF models, one based on 11,681 drug-target interactions between 6,640 drugs and 2,255 targets from DrugBank v5, and the other based on 8,579,843 chemical-target interactions for 311,507 chemicals and 9,457 targets from STITCH v5 human experimentally confirmed subset, respectively. We evaluated the confidence scores in the range [0, 1] for each predicted drug-target interaction, in both cases. We selected the interactions with confidence scores higher than 0.7 within the top 10 predicted targets for each input drug. This led to 161 novel interactions identified between 27 out of the 50 input drugs and 89 targets (composed of 41 known and 48 novel targets; **Supplementary Table 3**).

Pathway Enrichment Analysis

We mapped the 50 drugs with 142 known and 48 predicted targets to the KEGG pathways (version December 2017, *homo sapiens*) (Kanehisa et al., 2017). 114 and 173 pathways were mapped by 142 known targets and all targets (both known and predicted) respectively (see **Supplementary Table 4**). In order to prioritize enriched pathways, we calculated the hypergeometric *p*-values based on the targets as the enrichment score as follows. Given a list of targets, the enrichment *p*-value for pathway *A* (P^A) is the probability of randomly drawing k_0 or more targets that belong to pathway *A*:

$$P^A = \sum_{k_0 \le k \le m} \frac{\binom{K}{k}\binom{M-K}{m-k}}{\binom{M}{m}}$$

where *M* is the total number of human proteins in the KEGG Pathway, *m* is the total number of proteins/targets we identified, and *K* is the number of proteins that belong to pathway *A,* while k_0 is the number of targets we identified that belong to pathway *A*. The obtained *p*-values are adjusted by a False Discovery Rate (FDR) correction to account for multiple testing, using the widely used Benjamini-Hochberg method (Benjamini and Yekutieli, 2001). The cutoff of the adjusted *p*-values gives us an upper bound of the false discovery rate. The false discovery rate is the fraction of false significant pathways maximally expected from the significant pathways identified in our case. We sort *p*-values from smallest to largest, with *m* being the total number of pathways. The adjusted *p*-value, p_i^*, corresponding to the i*th* pathway is:

$$p_i{}^* = min_{k=i\ldots m}\{\min(p_k m/i, 1)\}$$

Supplementary Table 4 lists these *p*-values for pathway enrichments based on both known and predicted targets.

The source code used for generating the results reported in this study is available at https://github.com/Fengithub/DA.

RESULTS

Functional Similarity of Drugs of Abuse Does Not Imply Structural Similarity, Consistent With the Multiplicity of Their Actions

Figure 2 presents a quantitative analysis of the functional and structural diversity of the examined $n = 50$ drugs of abuse, and the similarities among the $m = 142$ known targets of these addictive drugs. The $n \times n$ maps in **Figures 2A,B** display the drug-drug pairwise distances/dissimilarities based on their 2D fingerprints (**Figure 2A**), and their interaction patterns with their targets. **Figures 2C,D** display the corresponding dendrograms. The drugs are indexed and color-coded as in **Supplementary Table 1** and **Supplementary Figure 1**. As expected, drugs belonging to the same functional category (*same color*) exhibit more similar interaction patterns (**Figure 2D**). However, we also note outliers, such as cocaine lying among opioids, as opposed to its categorization as a CNS stimulant, or promethazine, a CNS depressant, lying among hallucinogens (shown by *arrows*). The peculiar behavior of cocaine is consistent with its high promiscuity (see **Figure 3A** for the number of targets associated with each examined drug). This type of promiscuity becomes even more apparent when the drugs are organized based on their structure (or 2D fingerprints; see section Materials and Methods) as may be seen in **Figure 2A**. For example, opioids (*cyan labels/arc*; clustered together in **Figures 2B,D** based on their interactions) are now distributed in two or more branches of the structure-based dendrogram in **Figure 2C**; likewise, CNS depressants (*blue*) and cannabinoids (*light brown*), grouped each as a single cluster in target-based dendrograms in **Figure 2D**, are now distributed into two or more clusters in **Figure 2C**.

Overall these results suggest that the functional categorization of the drugs does not necessarily comply with their structural characteristics. The similar functionality presumably originates from targeting similar pathways, but the difference in the structure suggests that either their targets, or the binding sites on the same target, are different; or the binding is not selective enough such that multiple drugs can bind the same site. Consequently, a diversity of pathways or a multiplicity of cellular responses are triggered by the use and abuse of these drugs.

The Selected Drugs and Identified Targets Are Highly Diverse and Promiscuous

We evaluated the similarities between proteins targeted by drugs of abuse, based on their interaction patterns with the studied drugs of abuse. **Figures 2E,F** display the respective target-target distances, and corresponding dendrogram. **Supplementary Table 2** lists the full names of these targets, organized in the same order as the **Figure 2E** axes. We discern several groups of targets clustered together in consistency with their biological functions. For example, practically all GABA receptor subtypes (*brown*) are clustered together. This large cluster also includes the riboflavin transporter 2A (SLC52A2), which may be required for GABA release (Tritsch et al., 2012).

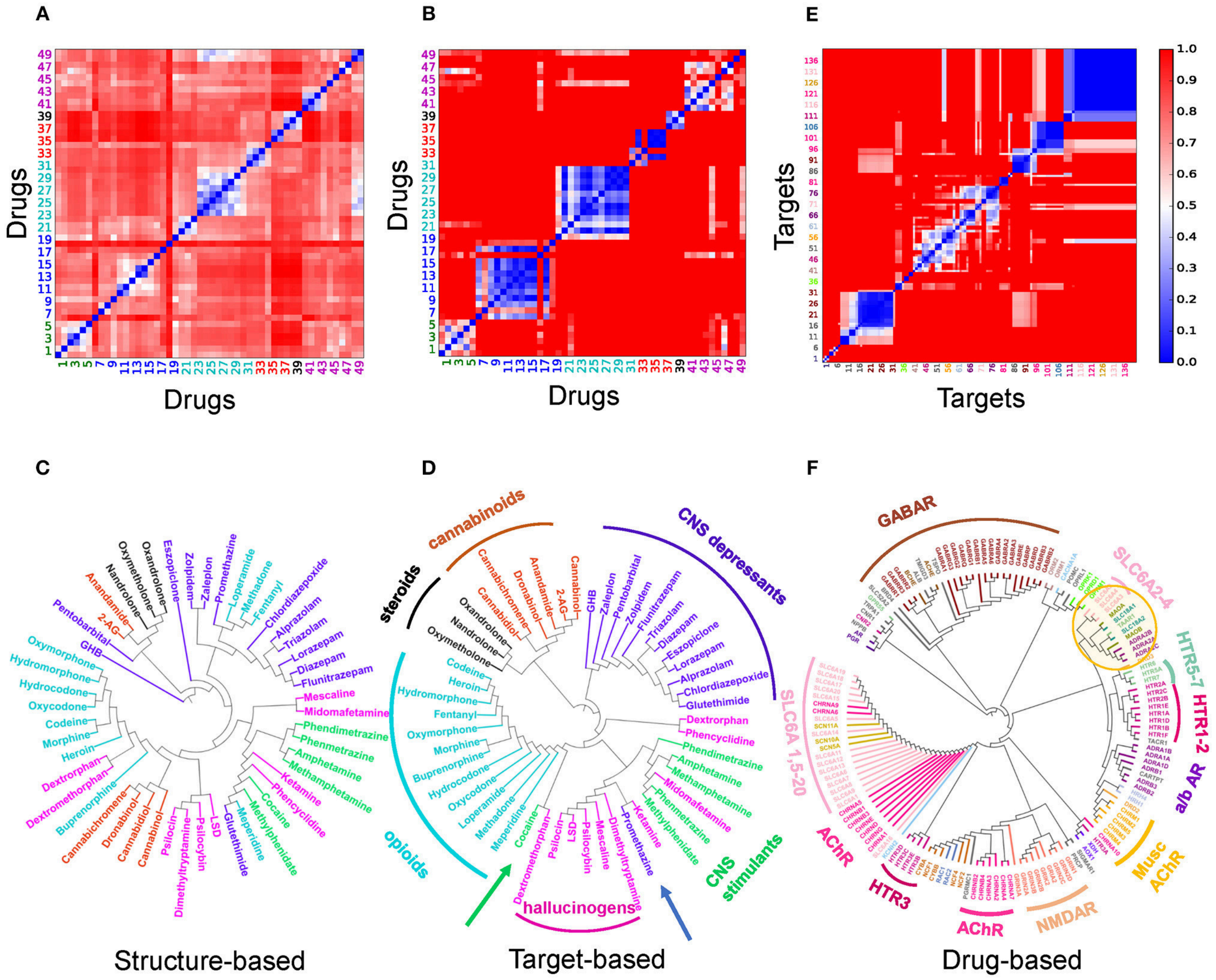


FIGURE 2 | Distribution of the dataset of 50 drugs of abuse based on their structure and interaction (with targets) similarities **(A–D)**, and pairwise similarities and classification of the corresponding targets based on their interaction patterns with the drugs of abuse. **(A–D)** Drug-drug distance maps for the studied 50 addictive drugs based on **(A)** 2D structure fingerprints and **(B)** interaction patterns with targets using the correlation cosines between their target vectors (see *Materials and Methods*), and corresponding dendrograms **(C,D)**. The indices of drugs of abuse in **(A,B)** follow the same order as those used in **Supplementary Table 1**. The drug labels in **(C,D)** are color-coded based on their categories: CNS stimulants (*green*), CNS depressants (*blue*), opioids (*cyan*), cannabinoids (*light brown*), anabolic steroids (*black*) and hallucinogens (*magenta*). Note that the drugs of abuse in the same category do not necessarily show structural similarities nor similar interaction pattern with targets. **(E)** Pairwise distance map for the 142 known targets based on their interaction patterns with the 50 drugs. The indices in **(E)** follows the same order as those listed clockwise in the dendrogram **(F).** The tree maps in **(C,D,F)** are generated based on the respective distances values in the **(A,B,E)**.

On the other hand, the different subtypes of serotonin (or 5-hydroxytryptamine, 5-HT) receptors (5HTRs) participate in distinct clusters pointing to the specificity of different subtypes vis-à-vis different drugs of abuse (*labeled* in **Figure 2F**).

The large majority of neurotransmitter transporters, such as Na^+/Cl^--dependent GABA transporters (SLC6A1) and glycine transporter (SLC6A9) are in the same cluster (*pink*, *labeled*). Acetylcholine receptors also lie close to (or are even interspersed among) Na^+/Cl^--dependent neurotransmitter transporters, presumably due to shared drugs such as cocaine. However, the three transporters playing a crucial role in developing drug addiction, DAT, NE transporter (NET) and serotonin transporter (SERT) (*labeled* SLC6A2: NET, SLC6A3: DAT, SLC6A4: SERT) are distinguished by from all other neurotransmitter transporters as a completely disjoint group. The corresponding branch of the dendrogram (*highlighted by the yellow circle*) also includes vesicular amino acid transporters and trace amine-associated receptor 1 (TAAR1) known to interact with these transporters (Miller, 2011). We also note in the same branch two seemingly unrelated targets: flavin monoamine oxidase which draws attention to the role of oxidative events; and α2-adrenergic receptor subtypes A-C, which uses NE as a chemical messenger for mediating stimulant effects such as sensitization and reinstatement of drug seeking, and adenylate

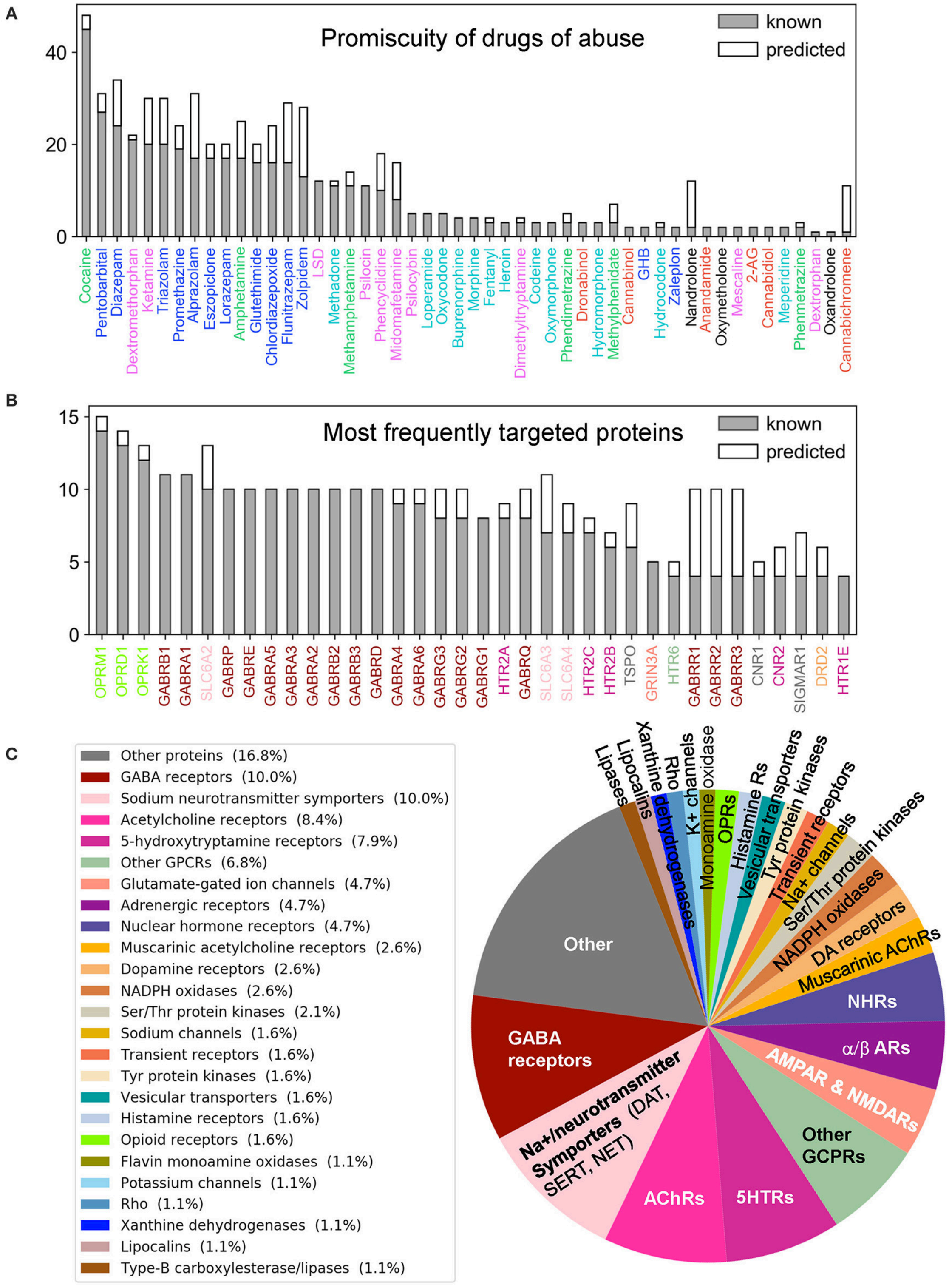


FIGURE 3 | Promiscuity of drugs of abuse and their targets, and major families of proteins targeted by drugs of abuse. Number of known (*gray*) and predicted (*white*) interactions are shown by bars for **(A)** drugs of abuse and **(B)** their targets. The examined set consists of 50 drugs of abuse and a total of 142 known and 48 predicted targets, involved in 445 (known) and 161 (predicted) interactions. **(A)** Displays the number of interactions known or predicted for all 50 drugs. **(B)** Displays the results for the targets that interact with at least 4 known drugs (36 targets). The colors used for names of drugs and targets are same as those used in **Figure 2**. **(C)** Displays the distribution of families of proteins targeted by drugs of abuse.

cyclase as another messenger to regulate cAMP levels (Sofuoglu and Sewell, 2009).

Supplementary Table 2 summarizes the 445 known interactions between these 50 drugs and 142 targets. We observe an average of 8.9 interactions per drug and 3.1 interactions per target. There are 23 promiscuous drugs that target at least 10 proteins as shown in **Figure 3A**. Cocaine, the most promiscuous psychostimulant, interacts with 45 known, and 3 predicted targets. It is known that cocaine binds DAT to lock it in the outward-facing state (OFS) and block the reuptake of DA. It similarly antagonizes SERT and NET (Heikkila et al., 1975; Sora et al., 1998), and also affects muscarinic acetylcholine receptors (mAChRs) M1 and M2 (Williams and Adinoff, 2008). Our PMF model also predicted a potential interaction between cocaine and M5. While this interaction is not listed in current DBs, there is experimental evidence suggesting that muscarinic AChR M5 plays an important role in reinforcing the effects of cocaine (Fink-Jensen et al., 2003), in support of the PMF model prediction.

The PMF model enables us to predict novel targets. For example, anabolic steroid nandrolone has only two known interactions, and cannabinoid cannabichromene has one. However, 10 new targets were predicted with high confidence scores for each of them (**Supplementary Table 3** and **Supplementary Figure 2A**). This is due to the data available in STITCH DB, which offers a large training dataset that enhances the performance of our machine learning approach. Overall, 89 new interactions were predicted for known targets, and 42 novel targets were predicted with 72 interactions. **Figure 3C** displays the distribution of all targets among different protein families. As will be further elaborated below, among the newly identified drug-target pairs, nandrolone-MAPK14 (mitogen-activated protein kinase 14, also known as p38α) and canabichromene-IKBKB (inhibitor of NFκ-B kinase subunit β) play a role in regulating mTORC1 signaling, which will be shown to be a potential effector of drug addiction.

Turning to targets, three opioid receptors (OPRM1, OPRD1, and OPRL1) exhibit the highest level of promiscuity (**Supplementary Figure 2B**). The μ-type opioid receptor (OPRM1) interacts with 14 known drugs including all opioids as well as ketamine and dextromethorphan. We also predicted a novel interaction between OPRM1 and the CNS stimulant methylphenidate. This is consistent with experimental observations that methylphenidate upregulates OPRM1's activity in the reward circuitry in a mouse model (Zhu et al., 2011). Furthermore, tissue-based transcriptome analysis (Uhlén et al., 2015) shows that 69% of our 190 targets are expressed in the brain, and 49 of them show elevated expression levels in the brain compared to other tissue types (**Supplementary Table 5**). Among all the targets, NMDA receptor 1 (GRIN1) shows the highest elevated expression. It is also one of the top 5 enriched genes overall in the brain (Uhlén et al., 2015).

Taken together, the 50 selected drugs of abuse and the 142 known and 48 novel targets we identified cover a diversity of biological functions, are involved in many cellular pathways, and are generally promiscuous. In order to reveal the common mechanisms that underlie the development and escalation of drug addiction and also distinguish the effects specific to selected drugs, we proceed now to a detailed pathway analysis, presented next.

Pathway Enrichment Analysis Reveals the Major Pathways Implicated in Various Stages of Addiction Development

Our QSP analysis yielded a total of 173 pathways, including 114 associated with the known targets of the examined dataset of drugs of abuse, and 59 associated with the predicted targets. The detailed pathway enrichment results can be found in **Supplementary Table 4**. These pathways can be grouped in five categories (**Figure 4**; **Supplementary Figures 3**, **4**, and **Supplementary Table 4**):

Synaptic Neurotransmission (NT)

Six significantly enriched (with adjusted p-value < 0.05) pathways are associated with synaptic neurotransmission: dopaminergic, serotonergic, glutamatergic, synaptic vesicle cycle, cholinergic, and GABAergic synapses pathways. Sixty-eight known targets and 7 predicted targets are involved in these pathways. This is consistent with the fact that neurotransmission plays a dominant role in the rewarding system and is key to drug addiction (Volkow and Morales, 2015).

Signal Transduction (SG)

Forty-six intracellular signaling pathways were mapped by 92 targets comprised of 66 known and 25 predicted targets. Notably, many of these pathways have been reported to play a role in mediating the effects of drugs of abuse. These include the top five [calcium signaling (Li et al., 2008), retrograde endocannabinoid signaling (Mechoulam and Parker, 2013), cGMP-PKG signaling (Shen et al., 2016), cAMP signaling (Philibin et al., 2011), and Rap1 signaling (Cahill et al., 2016)] as well as some pathways with relatively low enrichment score (i.e., 0.2 < adjusted p-value), such as TNF signaling (Zhu et al., 2018), MAPK signaling (Sun et al., 2016), PI3K-Akt signaling (Neasta et al., 2011), NF-κB signaling (Nennig and Schank, 2017), and mTOR signaling (Neasta et al., 2014). We note that many receptors targeted by drugs of abuse take part in the KEGG neuroactive ligand-receptor interaction pathway. In the interest of focusing on intracellular signaling effects, we have not included these in the SG category; they are listed in the "Other Pathways" in **Supplementary Table 4**.

Autonomic Nervous System (ANS)-Innervation (ANS)

We also identified 10 pathways regulating ANS-innervated systems such as endocrine secretion, taste transduction, and circadian entrainment. Recent evidences suggested drugs of abuse such as morphine (Al-Hasani and Bruchas, 2011) and cocaine (Moeller et al., 1997; Prosser et al., 2014) can influence ANS-innervated systems and may contribute to the withdrawn symptoms associated with drug addiction. Thirty-seven known and 9 predicted targets take part in these pathways.

Neuroplasticity (NP). Eight enriched pathways with potential to alter the morphology of neurons, were found to be related to drug addiction. Among them, long-term potentiation (LTP) and

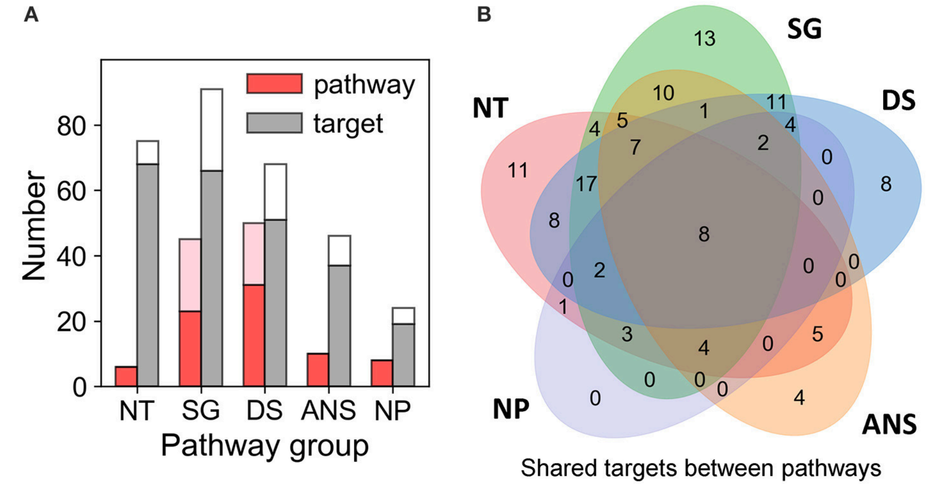


FIGURE 4 | Results from pathway and target enrichments analysis. Five broad categories of pathways are distinguished among those involving the targets of drug abuse: NT, synaptic neurotransmission pathways; SG, signal transduction pathways; DS, disease-associated pathways; ANS, autonomic nervous system-innervation pathways; and NP, neuroplasticity related pathways. **(A)** Numbers of pathways (*red bars*) and targets (*gray bars*) of drug abuse lying in the five categories, based on data available in DrugBank and STITCH. The *pink* and *white* stacked bars are the corresponding numbers for pathways and targets additionally predicted by PMF. **(B)** Overlaps between the target content of the five pathway categories. Note that all targets belonging to the NP category pathways are represented in the other four categories. See the complete list of pathways and targets in **Supplementary Table 4**.

long-term depression (LTD) are key to reward-related learning and addiction by modifying the fine tuning of dopaminergic firing (Jones and Bonci, 2005). Axon guidance pathway regulates the growth direction of neuron cells (Bahi and Dreyer, 2005). Regulation of actin cytoskeleton plays important role in morphological development and structural changes of neurons (Luo, 2002). Gap junctions connect neighboring neurons via intercellular channels that allow direct electrical communication (Belousov and Fontes, 2013) and regulate the efficiency of communication between electrical synapses (Belousov and Fontes, 2013). Nineteen known targets and 5 predicted targets are involved in these pathways. Insulin-like growth factor 1 receptor (IGF1R) is predicted as a target of drug triazolam (**Supplementary Table 4**). IGF1R is involved in LTP, adherens junction and focal adhesion pathways. It functions via canonical signaling pathways noted above in the SG category, such as the PI3K-Akt-mTOR and Ras-Raf-MAPK pathways (Lee et al., 2016) and it plays important role in neuroplasticity (Lee et al., 2016). We note that the NP group involves many pathways directly relevant to drug addiction (Bahi and Dreyer, 2005; Kalivas and Volkow, 2011; Moradi et al., 2013; Rothenfluh and Cowan, 2013). There is no target unique to this particular group of pathways (**Figure 4B**). However, the fact that the targets belonging to the NP group are also shared by other groups consolidates the significance of these targets.

Disease-Associated Pathways (DS)

Fifty enriched pathways mapped by 51 known and 17 predicted targets are associated with diverse diseases in different organs such as brain, liver, and lung. They also cover various drug addiction mechanisms including: nicotine addiction, morphine addiction, cocaine addiction, amphetamine addiction, and alcoholism. Additionally, there are "other pathways" such as those involved in cell migration, differentiation, immune responses, and metabolic events, which can be seen in **Supplementary Table 4**.

Taken together, the enrichment analysis reveals five major categories of pathways that regulate the three stages of drug addiction cycle: (1) binge and intoxication, (2) withdrawal and negative affect, and (3) preoccupation and anticipation (or craving) (Koob and Volkow, 2010). Drugs of abuse directly affect neurotransmission pathways: they increase the accumulation of DA and other neurotransmitters in the synaptic and extrasynaptic regions, which in turn results in the hedonic feeling (stage 1) and triggers the DA reward system. Dysregulation of ANS-innervation pathways may cause negative effects and feelings (stage 2) and feedback to the CNS. Addictive drugs impair executive processes by disrupting the reward system (neurotransmission pathways) and imparting morphological changes via neuroplasticity pathways (e.g., LTD and LTP), which then result in craving (stage 3). Below, we present an in-depth analysis of the role of these pathways or their shared targets in drug addiction.

Selected Targets Shared by Dominant Pathways Emerge as Common Mediators of Drug Addiction

We next analyzed the overlapping targets between the pathways in different functional categories.

First, we note that eight pleiotropic proteins are shared by all five categories (at the intersection of the five Venn diagrams in **Figure 4B**): AMPA receptor (subtype GluA2; GRIA2), NMDA receptors 1 and 2A-D (designated as GRIN1, GRIN2A, GRIN2B, GRIN2C, and GRIN2D) and voltage-dependent calcium channel $Ca_v2.1$ (or CACNA1A) as well as the predicted target phosphatidylinositol 3-kinase class 1A catalytic subunit α (PIK3CA) (**Supplementary Table 4**).

Second, 15 proteins are distinguished as targets of four of these major pathways: Serotonin receptors 5HTR2-A, -B and -C), $GABA_A$ receptors 1-6 (GABRA1- GABRA6), β-1 adrenergic receptor 1 (ADRB1), Ras-related C3 botulinum toxin substrate 1 (RAC1; member of Rho family of GTPases), mAChR M_3 (CHRM3) and DA receptor D_2 (DRD2), and two predicted targets - p38α (MAPK14) and DA receptor D_1 (DRD1).

AMPA receptor plays a crucial role in LTP and LTD, which are vital to neuroplasticity, memory and learning (Volkow et al., 2016). Serotonin receptors, expressed in both the CNS and the peripheral nervous system (e.g., gastrointestinal tract), are responsible for anxiety, impulsivity, memory, mood, sleep, thermoregulation, blood pressure, gastrointestinal motility, and nausea (Pytliak et al., 2011). They have been proposed to be therapeutic targets for treating cocaine use disorder (Howell and Cunningham, 2015). RAC1 is involved in five neuroplasticity pathways, including axon guidance, adherens junction and tight junction pathways (**Supplementary Table 4**), and 13 intracellular signal transduction pathways. It regulates neuroplasticity, as well as apoptosis and autophagy (Natsvlishvili et al., 2015). DA receptor D_2 is a target of 28 drugs of abuse (out of 50 examined here) and is involved in cAMP signaling, and gap junction pathways, in addition to dopaminergic signaling. It is implicated in reward mechanisms in the brain (Blum et al., 1996) and the regulation of drug-seeking behaviors (Edwards et al., 2006). Finally, PI3K turns out to be the most pleiotropic target among those targeted by drugs of abuse, being involved in 61 pathways identified here, including neuroplasticity pathways such as axon guidance, and several downstream signaling pathways such as PI3K-Akt, mTOR, Ras and Jak-STAT pathways.

Overall, the above listed 23 proteins shared by at least four different groups of pathways are distinguished here as highly pleiotropic proteins involved in the large majority of pathway categories implicated in drug abuse. Most of them are ligand- or voltage-gated ion channels or neurotransmitter receptors, mainly AMPAR, NMDAR, Cav2.1, mAChR, and serotonin and DA receptors. However, it is interesting to note the targets PI3K and p38α, not currently reported in DrugBank and STITCH, emerge as highly pleiotropic targets of the drugs of abuse. These are suggested by the current analysis to directly or indirectly affect addiction development and await future experimental validation. Finally, a number of proteins take part in specific drug-abuse-related pathways and might serve as targets for selective treatments. **Supplementary Table 6** provides a list of such targets uniquely implicated in distinctive pathways.

Pathway Enrichment Highlights the Interference of Drugs of Abuse With Synaptic Neurotransmission

It is broadly known that neurotransmitters such as DA, 5-HT, NE, endogenous opioids, ACh, endogenous cannabinoids, Glu, and GABA are implicated in drug addiction (Tomkins and Sellers, 2001; Everitt and Robbins, 2005; Parolaro and Rubino, 2008; Benarroch, 2012). Our analysis also showed that the serotonergic synapse (adjusted *p*-value $p_i^* = 2.01E\text{-}18$), GABAergic synapse ($p_i^* = 1.19E\text{-}17$), cholinergic synapse ($p_i^* = 2.36E\text{-}07$), dopaminergic synapse ($p_i^* = 1.66E\text{-}06$) and glutamatergic synapse ($p_i^* = 1.86E\text{-}03$) pathways were significantly enriched (**Supplementary Table 4**). A total number of 34 drugs (across six different groups) target at least one of these pathways. However, the identification of a pathway does not necessarily mean that the drug directly affects that particular neurotransmitter transport/signaling. There may be indirect effects due to the crosstalks between synaptic signaling pathways. For example, the ionotropic glutamate receptors NMDAR and AMPAR are also the downstream mediators in the dopaminergic synapse pathway. Likewise, GABARs are downstream mediators in the serotonergic synapse pathway.

In **Figure 5**, we highlight five major neurotransmission events that directly mediate addiction, and illustrate how eight drugs of abuse interfere with them. Despite the promiscuity of the drugs of abuse, some selectively map onto a single synaptic neurotransmission pathway. For example, psilocin [a hallucinogen whose structure is similar to 5HT (Diaz, 1997)] interacts with several types of 5HTRs, regulating serotonergic synapse exclusively (see **Figure 5** and **Supplementary Table 4**). In contract, loperamide (not shown) affects all neurotransmission pathways by interacting with the voltage-dependent P/Q-type calcium channel (VGCC), regulating calcium flux on synapses. Cocaine targets four of these synaptic neurotransmission events (serotonergic, GABAergic, cholinergic, and dopaminergic synapses), through its interactions with 5-HT3R, sodium- and chloride-dependent GABA transporter (GAT), muscarinic (M1 and M2) and nicotinic AChRs, and DAT, respectively. Methadone affects three synaptic neurotransmissions, including serotonergic synapse, dopaminergic synapse, and glutamatergic synapse through the interactions with SERT, DAT, and glutamate receptors (NMDAR), respectively.

It is worth noting that the current analysis helps us generate new hypotheses, yet to be experimentally validated, on the ways drugs of abuse affect neurotransmission. In addition to the new role of the muscarinic AChR M5 suggested by the current analysis in section the selected drugs and identified targets are highly diverse and promiscuous, our PMF model suggested that cannabichromene, a cannabinoid whose primary target is the transient receptor (TRPA1), could interact with DAT and thus regulate dopaminergic transmission, which will require further examination.

The above synaptic neurotransmission events act as upstream signaling modules that "sense" the early effects of drug abuse. In the next section, we focus on the downstream signaling events elicited by drug abuse.

mTORC1 Emerges as a Potential Downstream-Effector Activated by Drugs Abuse

The calcium-, cAMP-, Rap1-, Ras-, AMPK-, ErbB-, MAPK-, and PI3K-Akt-signaling pathways in the SG category (**Supplementary Table 4**) crosstalk with each other and form a unified signaling network. As shown in **Figure 6**,

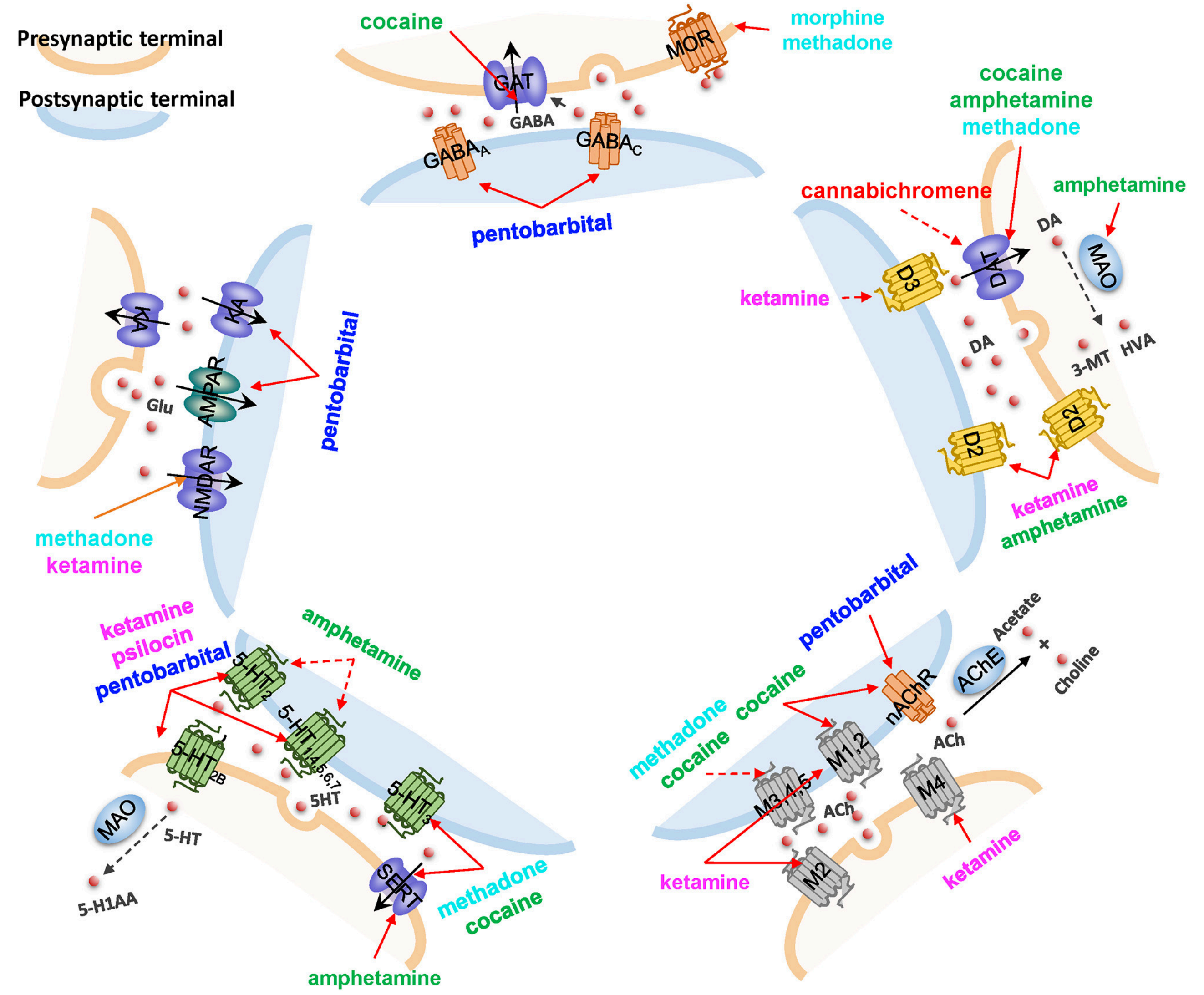


FIGURE 5 | The impact of drugs of abuse on synaptic neurotransmission. Five major neurotransmission events are highlighted, mediated by (*counterclockwise, starting from top*): GABA receptors and transporters, ionotropic glutamate receptors (NMDAR and AMPAR) and cation channels, serotonin (5HT) receptors (5-HTR) and transporters (SERT), muscarinic or nicotinic AChRs, and dopamine (DA) receptors and transporters. Vesicular monoamine transporters (VMAT) that translocate DA are also shown. Drugs affecting the different pathways are listed, color coded with their categories, as presented in **Figure 2**. *Solid red arrows* indicate a known drug-target interaction, *dashed red arrows* indicate predicted drug-target interactions. Other molecules shown in the diagram are: KA, kainate receptor; MAO, monoamine oxidase; HVA, homovanillate; 3-MT, 3-methoxytyramine; MOR, mu-type opioid receptor; AChE, acetylcholinesterase; and 5-H1AA, 5-hydroxyindoleacetate.

ligand-binding to GPCRs modulates the production of cAMP, which leads to the activation of Rap1. Activated Rap1 modules the Ca^{2+} signaling by inducing the production of inositol triphosphate (IP_3) and also activates the PI3K-Akt signaling cascade. Stimulations of ErbB family of receptor tyrosine kinases (related to epidermal growth factor receptor EGFR) as well as insulin-like growth factor receptor IGF1R trigger both PI3K-Akt and MAPK signaling cascades (proteins colored *blue* in **Figure 6**). Notably all these pathways merge and regulate a group of downstream proteins (shown in *dark yellow* in **Figure 6**); and at the center of this cluster lies the mammalian target of rapamycin (mTOR) complex 1 (mTORC1) which is likely to be synergistically regulated by all these merging pathways.

mTORC1 is not only a master regulator of autophagy (Rabanal-Ruiz et al., 2017), but also controls protein synthesis and transcription (Ma and Blenis, 2009). It has been reported to promote neuroadaptation following exposure to drugs of abuse including cocaine, alcohol, morphine and Δ^9-tetrahydrocannabinol (THC) (Neasta et al., 2014). Our results lead to the hypothesis that mTORC1 may act as a universal effector of the cellular response to drug abuse at an advanced

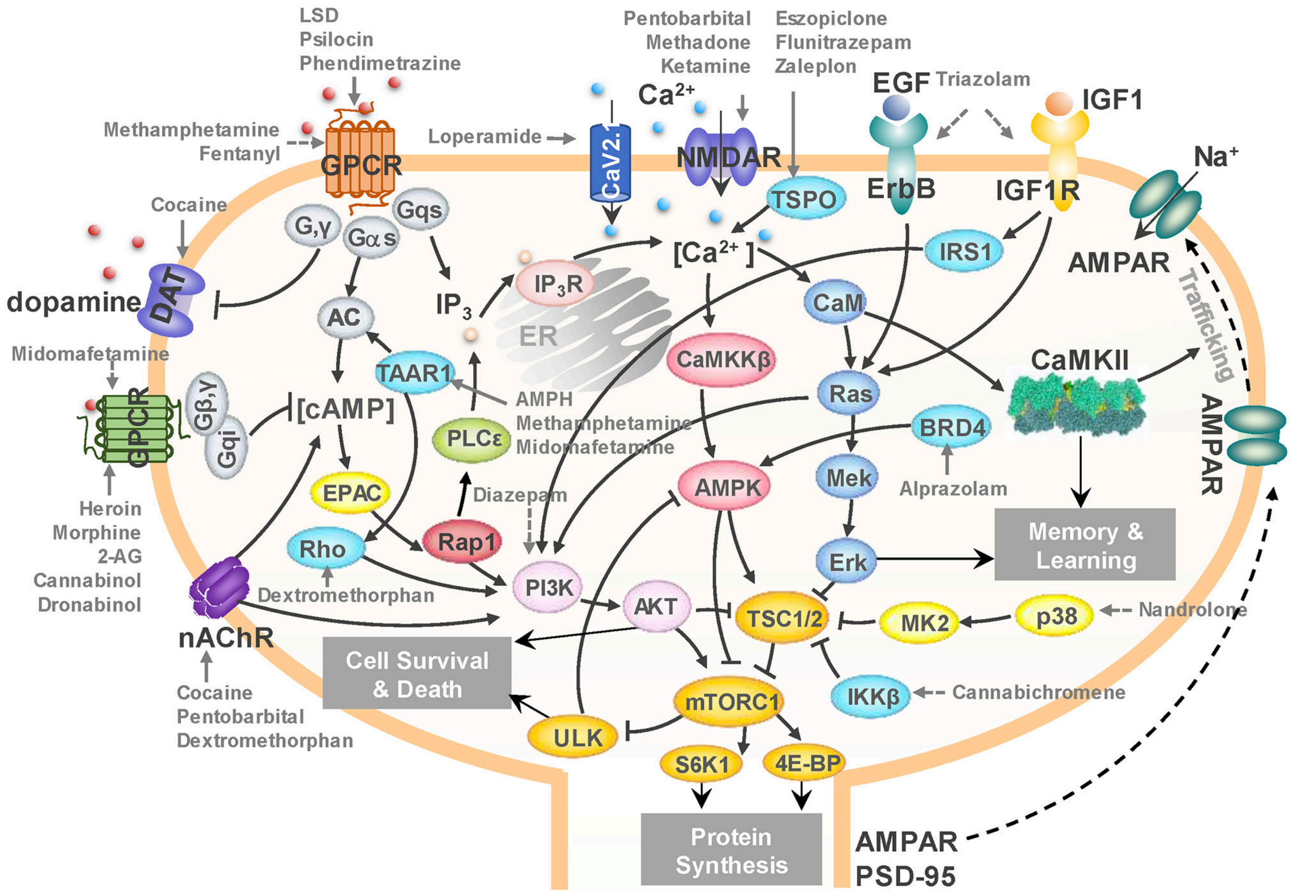


FIGURE 6 | A unified signaling network mediates the effects of drugs of abuse. *Black arrows* represent the activation, inhibition, and translocation events during signal transduction. *Solid gray arrows* represent the known drug-target interactions. *Dashed gray arrows* represent predicted drug-target interactions. The diagram illustrates the targets of several drugs of abuse belonging to different categories: loperamide, fentanyl, heroin, morphine, and methadone from opioids; midomafetamine, ketamine, dextromethorphan, LSD, and psilocin from hallucinogens; triazolam, diazepam, alprazolam, pentobarbital, eszopiclone, flunitrazepam, and zaleplon from CNS depressants; cannabichromene, 2-AG, cannabinol, and dronabinol from cannabinoids; methamphetamine, cocaine, AMPH, and phendimetrazine from CNS stimulants; and nandrolone from anabolic steroids. mTORC1 emerges as a hub where the effects on several targets of addictive drugs appear to be consolidated to lead to cell death and/or protein synthesis in the CNS, and in particular, to AMPAR/PSD95 synthesis that induces morphological changes in the dendrites.

(preoccupation and anticipation, or craving) stage, controlling the synthesis of selected proteins and ensuing cell growth, which may result in persistent alterations in the dendritic morphology and neuronal circuitry.

In **Figure 6**, selected interactions between drugs from different substance groups and their targets are highlighted using *gray* arrows. The figure illustrates that not only many known drug-target interactions, but also predicted ones involved in the unified signaling network. For example, our PMF model predicted that diazepam would interact with PI3K to influence mTORC1 signaling (*dashed gray arrows* denote predictions). It has been reported that Ro5-4864, a benzodiazepine derivative of diazepam suppresses activation of PI3K (Yousefi et al., 2013), which corroborates our prediction. We further predicted that cannabichromene may interact with IκB kinase β (IKKβ) to regulate mTORC1 by inhibiting TSC1/2. Interestingly, another cannabinoid, arachidonoyl ethanolamine, is known to directly inhibits IKKβ (Sancho et al., 2003). Taken together, our results suggest a unified network that underlies the development of drugs addiction, in which mTORC1 appears to play a key effector role.

DISCUSSION

In the present study we focused on the targets and pathways affected by drugs of abuse, toward gaining a systems-level understanding of key players and dominant interactions that control the response to drug abuse and the development of drug addiction. Using machine learning methods, we focused on 50 drugs of abuse that form a chemically and functionally diverse set, and analyzed their 142 targets as well as the corresponding cellular pathways and their crosstalk. Our analysis identified:

(i) 48 additional proteins targeted by drugs of abuse, including PIK3CA, IKBKB, EGFR, and IGF1R, are shown to be key mediators of downstream effects of drug abuse.

(ii) 161 new interactions between the drugs of abuse and the known and predicted targets, including those between cocaine and M5, methylphenidate and OPRM1, and diazepam and PI3K, not reported in existing DBs, but supported by prior experiments, and others (e.g., the interactions of cannabichromene with IKBKB and DAT) that await experimental validation.

(iii) A dataset of 70 pathways, composed of 6 neurotransmission pathways, 46 signal transduction pathways, 8 neuroplasticity pathways and 10 autonomic nervous system innervation pathways which are proposed to govern different stages of the molecular, cellular and tissue level responses to drug abuse and in addiction development.

Overall, our comprehensive analysis led to new hypotheses on drug-target interactions and signaling and regulation mechanism elicited by drugs of abuse in general, along with those on selected targets and pathways for specific drugs. Below we elaborate on the biological and biomedical implications of these findings.

Persistent Restructuring in Neuronal Systems as a Feature Underlying Drug Addiction

Enriched pathways in the neuroplasticity category include gap junction, LTP, LDP, adherens junction, regulation of actin cytoskeleton, focal adhesion, axon guidance, and tight junction (**Supplementary Table 4**). These are responsible for the changes in the morphology of dendrites. For instance, DA regulates excitatory synaptic plasticity by modulating the strength and size of synapses through LTP and LTD (De Roo et al., 2008; Volkow and Morales, 2015). The restructuring of dendritic spines involves the rearrangements of cytoskeleton and actin-myosin (Volkow and Morales, 2015). The axon guidance molecules guide the direction of neuronal growth.

Drugs of abuse can induce the changes in CNS through these pathways. For example, chronic exposure to cocaine increases dendritic spine density in medium spiny neurons (Russo et al., 2010). The disruption in axon guidance pathway and alteration in synaptic geometry can result in drug-related plasticity (Bahi and Dreyer, 2005). The persistent restructuring in the CNS caused by drugs of abuse is responsible for long-term behavioral plasticity driving addiction (Volkow et al., 2003; Russo et al., 2010; Volkow and Morales, 2015). As will be further discussed below, mTORC1 plays a central role in the synthesis of new proteins (e.g., AMPARs) and thereby neuronal (dendrites) growth, alteration of the synaptic geometry and therefore rewiring of the neuronal circuitry.

ANS May Mediate the Negative-Reinforcement of Drug Addiction

The current study further points to pathways regulating the ANS-innervated systems. As the NP pathways influence the neuroplasticity in the ANS, we hypothesize that drugs of abuse might induce a persistent restructuring in the ANS as well. The drug-related plasticity in ANS may lead to the dysregulation of ANS-innervated systems and cause negative effects and feelings during the second stage of drug addiction. Drug addiction is well known as a brain disease (Volkow and Morales, 2015). However, many drugs of abuse can disrupt the activity of ANS and cause disorders in ANS-innervated systems (Al-Hasani and Bruchas, 2011; Huang, 2017). For example, opioids (e.g., morphine) alter neuronal excitability and neurotransmission in the ANS (Wood and Galligan, 2004), and induce disorders in gastrointestinal system, smooth muscle, skin, cardiovascular, and immune system (Al-Hasani and Bruchas, 2011). Cannabinoids (e.g., THC) modulate the exocytotic NE release in ANS-innervated organs through presynaptic cannabinoid receptors (Ishac et al., 1996).

The pathways we identified in the ANS category regulate insulin secretion, gastric acid secretion, vascular smooth muscle contraction, pancreatic secretion, salivary secretion, and renin secretion (**Supplementary Table 4**). Their dysfunction may be associated with the autonomic withdrawal syndrome, such as thermoregulatory disorder (chills and sweats) and gastrointestinal upset (abdominal cramps and diarrhea), which has been observed in drug/substance users (Wise and Koob, 2014). In addition, the stress and depression caused by these negative effects may be part of the negative reinforcement of drug addiction (Self and Nestler, 1995; Koob and Le Moal, 2001). In other words, the drug induced ANS disorders can feedback to CNS and mediate the negative reinforcement. Compared to the structural changes in CNS, the disorder and persistent restructuring in ANS is less studied and it could be a future direction in the study of development of drug addiction and related diseases.

mTORC1 Appears as a Key Mediator of Cellular Morphological Changes Elicited in Response to Continued Drug Abuse

The functioning and regulation of mTOR signaling has been elucidated over the past two decades. It became clear that mTORC1 plays a crucial role in regulating diverse cellular processes including protein synthesis, autophagy, lipid metabolism, and mitochondrial biogenesis (Saxton and Sabatini, 2017). In the brain, mTORC1 coordinates neural development, circuit formation, synaptic plasticity, and long-term memory (Lipton and Sahin, 2014). The dysregulation of mTORC1 pathway is associated with many neurodevelopmental and neurodegenerative diseases such as Parkinson's disease and Alzheimer's disease. mTORC1 has been noted to be an important mediator of the development of drug addiction and relapse vulnerability (Dayas et al., 2012). Accumulating evidences show that pharmacological inhibition of mTORC1 (often through rapamycin treatment) can prevent sensitization of methamphetamine-induced place preference (Narita et al., 2005), reduce craving in heroin addicts (Shi et al., 2009), attenuate the expression of alcohol-induced locomotor sensitization (Neasta et al., 2010), suppress the expression of cocaine-induced place preference (Bailey et al., 2012), protect against the expression of drug-seeking and relapse by reducing AMPAR (GluA1) and CaMKII levels (James et al., 2014), and inhibit reconsolidation of morphine-associated memories (Lin et al., 2014).

Our unbiased computational analysis based on a diverse set of 50 drugs of abuse supports the hypothesis that mTORC1 may act as a universal effector or controller of neuroadaptations induced by drugs of abuse (Neasta et al., 2014). The major signal transduction pathways we identified that involve targets of drugs of abuse interconnect and converge to the mTORC1 signaling cascade (**Figure 6**). Most drugs of abuse in our list target upstream regulators of mTORC1, including membrane receptors (e.g., GPCRs, RTKs and NMDAR), kinases (e.g., PI3K, p38α, and IKKβ), and ion channels (e.g., $Ca_V2.1$ and TRPV2). Notably, the impact of some of these known or predicted targets has been experimentally confirmed. For example, blockade of the known target NMDAR using MK801 reduces the amnesic-like effects of cannabinoid THC (Puighermanal et al., 2009). Likewise, inhibition of PI3K (a predicted target) by LY294002 suppresses morphine-induced place preference in rats (Cui et al., 2010) and the expression of cocaine-sensitization (Izzo et al., 2002). Our results thus provide a pool of candidate targets implicated in cellular responses to addictive drugs, which await to be consolidated by further tests.

The downstream effectors of mTORC1, which specifically mediate drug behavioral plasticity is far from known. mTORC1 can mediate the activation of S6Ks and 4E-BPs, which leads to increased production of proteins required for synaptic plasticity including AMPAR and PSD-95 (Dayas et al., 2012). EM reconstruction of hippocampal neuropil showed the variability in the size and shape of dendrites depending on synaptic activity (Bartol Jr et al., 2015), which in turn correlates with information storage. Recently studies have revealed that Atg5- and Atg7-dependent autophagy in dopaminergic neurons regulates cellular and behavioral responses to morphine (Su et al., 2017). Cocaine exposure results in ER stress-induced and mTORC1-dependent autophagy (Guo et al., 2015). Fentanyl induces autophagy via activation of ROS/MAPK pathway (Yao et al., 2016). Methamphetamine induces autophagy through the κ-opioid receptor (Ma et al., 2014). These observations are consistent with the currently inferred role of mTORC1 as a downstream effector of cellular responses to drug addiction.

Drug Repurposing Opportunities for Combatting Drug Addiction

Autophagy modulating drugs have been shown to have therapeutic effects against liver and lung diseases. The signaling network presented in **Figure 6** involves many targets of such drugs. For instance, carbamazepine affects IP_3 production and enhances autophagy via calcium-AMPK-mTORC1 pathway (Hidvegi et al., 2010). It has been identified as a potential drug for treating α1-antitrypsin deficiency, hepatic fibrosis, and lung proteinopathy (Hidvegi et al., 2010, 2015). Rapamycin is a potential drug for lung disease such as fibrosis (Abdulrahman et al., 2011; Patel et al., 2012). Other liver and lung drugs which facilitate the removal of aggregates by promoting autophagy may also affect drug-related neurodegenerative disorders. **Supplementary Table 7** summarizes 15 autophagy-modulating drugs for liver and lung diseases. Target identification and pathway analysis of this subset of drugs using the same protocol as those adopted for the 50 drugs of abuse indeed confirmed that drugs of abuse and liver/lung drugs share many common pathways (**Supplementary Figure 5**). Notably, among those pathways, neuroactive ligand-receptor interactions, calcium signaling, and serotonergic synapse pathways are among the top 10 enriched pathways of both drugs of abuse and liver/lung drugs. Amphetamine addiction and alcoholism are also enriched by targets of liver/lung drugs. Thus, an interesting future direction is to examine whether autophagy modulating drugs for liver and lung diseases could be repurposed, if necessary by suitable refinements to increase their selectivity, for treating drug addiction.

In summary, our results invite attention to new targets of addictive drugs and pathways implicated in the development of addiction, as well as new therapeutic opportunities. Recent studies support the utility of such computationally-driven QSP predictions. The validation of these predictions requires comprehensive wet-lab bioactivity assays (Pahikkala et al., 2015). In particular, the establishment of the proposed role of mTORC1 would require *in vitro* and *in vivo* longitudinal studies given that our current study points to the involvement of mTORC1 at later stages of drug addiction. In a recent study, we identified the role of protein kinase A (PKA) pathway in Huntington's disease using a QSP approach and verified experimentally (Pei et al., 2017). A similar combined computational-experimental framework could be adopted to extend the current study and establish new strategies. Though these experiments are beyond the scope of the current paper, our unbiased computational study provides insights into the pleiotropy of the targets of addictive drugs as well as the common signaling platforms that may serve as mediators of drug addiction.

Knowledge of pathways implicated in drug addiction may be used, as a next step, to construct kinetic models to quantitatively assess the orchestration of signals induced by pathway crosstalks. Our previous studies on Toll-like receptors (Liu et al., 2016) and cell fate decision processes (Liu et al., 2014, 2017) have demonstrated the utility of identifying such crosstalks for detecting synergistic response mechanisms and designing polypharmacological strategies. Therefore, the computational data presented here presents a milestone toward developing new therapies against drug addiction by identifying new targets beyond those usually investigated by focused studies. Finally, our analysis framework is generic and could be adopted for characterizing the targets and pathways of other complex disorders by suitable redefinition of the input set of drugs of interest.

DATA AVAILABILITY

All datasets generated for this study are included in the manuscript and/or the supplementary files.

AUTHOR CONTRIBUTIONS

FP, HL, and IB conceived and designed the research. FP and HL performed the research. FP, HL, BL, and IB analyzed the results and wrote the manuscript.

FUNDING

This work was supported by the National Institutes of Health awards P30DA035778, and P41GM103712.

ACKNOWLEDGMENTS

FP wish to thank Dr. D. Lansing Taylor for his mentorship.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphar.2019.00191/full#supplementary-material

REFERENCES

Abdulrahman, B. A., Khweek, A. A., Akhter, A., Caution, K., Kotrange, S., Abdelaziz, D. H., et al. (2011). Autophagy stimulation by rapamycin suppresses lung inflammation and infection by *Burkholderia cenocepacia* in a model of cystic fibrosis. *Autophagy* 7, 1359–1370. doi: 10.4161/auto.7.11.17660

Al-Hasani, R., and Bruchas, M. R. (2011). Molecular mechanisms of opioid receptor-dependent signaling and behavior. *Anesthesiology* 115, 1363–1381. doi: 10.1097/ALN.0b013e318238bba6

Bahi, A., and Dreyer, J. L. (2005). Cocaine-induced expression changes of axon guidance molecules in the adult rat brain. *Mol. Cell. Neurosci.* 28, 275–291. doi: 10.1016/j.mcn.2004.09.011

Bailey, J., Ma, D., and Szumlinski, K. K. (2012). Rapamycin attenuates the expression of cocaine-induced place preference and behavioral sensitization. *Addict. Biol.* 17, 248–258. doi: 10.1111/j.1369-1600.2010.00311.x

Bartol Jr, T. M., Bromer, C., Kinney, J., Chirillo, M. A., Bourne, J. N., Harris, K. M., et al. (2015). Nanoconnectomic upper bound on the variability of synaptic plasticity. *Elife* 4:e10778. doi: 10.7554/eLife.10778

Belousov, A. B., and Fontes, J. D. (2013). Neuronal gap junctions: making and breaking connections during development and injury. *Trends Neurosci.* 36, 227–236. doi: 10.1016/j.tins.2012.11.001

Benarroch, E. E. (2012). Endogenous opioid systems: current concepts and clinical correlations. *Neurology* 79, 807–814. doi: 10.1212/WNL.0b013e3182662098

Benjamini, Y., and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* 29, 1165–1188. doi: 10.2307/2674075

Biernacka, J. M., Geske, J., Jenkins, G. D., Colby, C., Rider, D. N., Karpyak, V. M., et al. (2013). Genome-wide gene-set analysis for identification of pathways associated with alcohol dependence. *Int. J. Neuropsychopharmacol.* 16, 271–278. doi: 10.1017/S1461145712000375

Blum, K., Sheridan, P. J., Wood, R. C., Braverman, E. R., Chen, T. J., et al. (1996). The D2 dopamine receptor gene as a determinant of reward deficiency syndrome. *J. R Soc. Med.* 89, 396–400. doi: 10.1177/014107689608900711

Cahill, M. E., Bagot, R. C., Gancarz, A. M., Walker, D. M., Sun, H., Wang, Z. J., et al. (2016). Bidirectional synaptic structural plasticity after chronic cocaine administration occurs through Rap1 small GTPase signaling. *Neuron* 89, 566–582. doi: 10.1016/j.neuron.2016.01.031

Chen, X., Yan, C. C., Zhang, X., Zhang, X., Dai, F., Yin, J., et al. (2016). Drug-target interaction prediction: databases, web servers and computational models. *Brief. Bioinform.* 17, 696–712. doi: 10.1093/bib/bbv066

Coates, K. M., and Flood, P. (2001). Ketamine and its preservative, benzethonium chloride, both inhibit human recombinant α7 and α4β2 neuronal nicotinic acetylcholine receptors in Xenopus oocytes. *Br. J. Pharmacol.* 134, 871–879. doi: 10.1038/sj.bjp.0704315

Cobanoglu, M. C., Liu, C., Hu, F., Oltvai, Z. N., and Bahar, I. (2013). Predicting drug-target interactions using probabilistic matrix factorization. *J. Chem. Inf. Model.* 53, 3399–3409. doi: 10.1021/ci400219z

Cobanoglu, M. C., Oltvai, Z. N., Taylor, D. L., and Bahar, I. (2015). BalestraWeb: efficient online evaluation of drug-target interactions. *Bioinformatics* 31, 131–133. doi: 10.1093/bioinformatics/btu599

Collingridge, G. L., Kehl, S. J., and McLennan, H. (1983). Excitatory amino acids in synaptic transmission in the schaffer collateral-commissural pathway of the rat hippocampus. *J. Physiol.* 334, 33–46. doi: 10.1113/jphysiol.1983.sp014478

Cui, Y., Zhang, X., Cui, Y., Xin, W., Jing, J., and Liu, X. (2010). Activation of phosphatidylinositol 3-kinase/Akt-mammalian target of rapamycin signaling pathway in the hippocampus is essential for the acquisition of morphine-induced place preference in rats. *Neuroscience* 171, 134–143. doi: 10.1016/j.neuroscience.2010.08.064

Dayas, C. V., Smith, D. W., and Dunkley, P. R. (2012). An emerging role for the mammalian target of rapamycin in "pathological" protein translation: relevance to cocaine addiction. *Front. Pharmacol.* 3:13. doi: 10.3389/fphar.2012.00013

De Roo, M., Klauser, P., Garcia, P. M., Poglia, L., and Muller, D. (2008). Spine dynamics and synapse remodeling during LTP and memory processes. *Prog. Brain Res.* 169, 199–207. doi: 10.1016/S0079-6123(07)00011-8

Diaz, J. (1997). *How Drugs Influence Behavior: A Neuro-Behavioral Approach*. Upper Saddle River, NJ: Prentice Hall.

Edwards, S., Whisler, K. N., Fuller, D. C., Orsulak, P. J., and Self, D. W. (2006). Addiction-related alterations in D1 and D2 Dopamine receptor behavioral responses following chronic cocaine self-administration. *Neuropsychopharmacology* 32, 354–366. doi: 10.1038/sj.npp.1301062

Everitt, B. J., and Robbins, T. W. (2005). Neural systems of reinforcement for drug addiction: from actions to habits to compulsion. *Nat. Neurosci.* 8, 1481–1489. doi: 10.1038/nn1579

Fink-Jensen, A., Fedorova, I., Wortwein, G., Woldbye, D. P., Rasmussen, T., Thomsen, M., et al. (2003). Role for M5 muscarinic acetylcholine receptors in cocaine addiction. *J. Neurosci. Res.* 74, 91–96. doi: 10.1002/jnr.10728

Guo, M. L., Liao, K., Periyasamy, P., Yang, L., Cai, Y., Callen, S. E., et al. (2015). Cocaine-mediated microglial activation involves the ER stress-autophagy axis. *Autophagy* 11, 995–1009. doi: 10.1080/15548627.2015.1052205

Heikkila, R. E., Orlansky, H., and Cohen, G. (1975). Studies on the distinction between uptake inhibition and release of [3H] dopamine in rat brain tissue slices. *Biochem. Pharmacol.* 24, 847–852. doi: 10.1016/0006-2952(75)90152-5

Hevers, W., Hadley, S. H., Lüddens, H., and Amin, J. (2008). Ketamine, but not phencyclidine, selectively modulates cerebellar GABAA receptors containing α6 and δ subunits. *J. Neurosci.* 28, 5383–5393. doi: 10.1523/JNEUROSCI.5443-07.2008

Hidvegi, T., Ewing, M., Hale, P., Dippold, C., Beckett, C., Kemp, C., et al. (2010). An autophagy-enhancing drug promotes degradation of mutant alpha1-antitrypsin Z and reduces hepatic fibrosis. *Science* 329, 229–232. doi: 10.1126/science.1190354

Hidvegi, T., Stolz, D. B., Alcorn, J. F., Yousem, S. A., Wang, J., Leme, A. S., et al. (2015). Enhancing autophagy with drugs or lung-directed gene therapy reverses the pathological effects of respiratory epithelial cell proteinopathy. *J. Biol. Chem.* 290, 29742–29757. doi: 10.1074/jbc.M115.691253

Hirota, K., Hashimoto, Y., and Lambert, D. G. (2002). Interaction of intravenous anesthetics with recombinant human M1-M3 muscarinic receptors expressed in chinese hamster ovary cells. *Anesth. Analg.* 95, 1607–1610. doi: 10.1097/00000539-200212000-00025

Howell, L. L., and Cunningham, K. A. (2015). Serotonin 5-HT2 receptor interactions with dopamine function: implications for therapeutics in cocaine use disorder. *Pharmacol. Rev.* 67, 176–197. doi: 10.1124/pr.114.009514

Hu, Y., Fang, Z., Yang, Y., Rohlsen-Neal, D., Cheng, F., and Wang, J. (2018). Analyzing the genes related to nicotine addiction or schizophrenia via a pathway and network based approach. *Sci. Rep.* 8:2894. doi: 10.1038/s41598-018-21297-x

Huang, A. C. W. (2017). "Autonomic nervous system and brain circuitry for internet addiction," in *Internet Addiction*, eds C. Montag and M. Reuter (Cham: Springer), 161–180. doi: 10.1007/978-3-319-46276-9_10

Hustveit, O., Maurset, A., and Øye, I. (1995). Interaction of the chiral forms of ketamine with opioid, phencyclidine, σ and muscarinic receptors. *Pharmacol. Toxicol.* 77, 355–359. doi: 10.1111/j.1600-0773.1995.tb01041.x

Ishac, E. J., Jiang, L., Lake, K. D., Varga, K., Abood, M. E., and Kunos, G. (1996). Inhibition of exocytotic noradrenaline release by presynaptic cannabinoid CB 1 receptors on peripheral sympathetic nerves. *Br. J. Pharmacol.* 118, 2023–2028. doi: 10.1111/j.1476-5381.1996.tb15639.x

Izzo, E., Martin-Fardon, R., Koob, G. F., Weiss, F., and Sanna, P. P. (2002). Neural plasticity and addiction: PI3-kinase and cocaine behavioral sensitization. *Nat. Neurosci.* 5, 1263–1264. doi: 10.1038/nn977

James, M. H., Quinn, R. K., Ong, L. K., Levi, E. M., Charnley, J. L., Smith, D. W., et al. (2014). mTORC1 inhibition in the nucleus accumbens 'protects' against the expression of drug seeking and 'relapse'and is associated with reductions in GluA1 AMPAR and CAMKIIα levels. *Neuropsychopharmacology* 39, 1694–1702. doi: 10.1038/npp.2014.16

Jones, S., and Bonci, A. (2005). Synaptic plasticity and drug addiction. *Curr. Opin. Pharmacol.* 5, 20–25. doi: 10.1016/j.coph.2004.08.011

Kalivas, P. W., and Volkow, N. D. (2011). New medications for drug addiction hiding in glutamatergic neuroplasticity. *Mol. Psychiatry* 16, 974–986. doi: 10.1038/mp.2011.46

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 45, D353–D361. doi: 10.1093/nar/gkw1092

Kapur, S., and Seeman, P. (2002). NMDA receptor antagonists ketamine and PCP have direct effects on the dopamine D 2 and serotonin 5-HT 2 receptors—implications for models of schizophrenia. *Mol. Psychiatry* 7, 837–844. doi: 10.1038/sj.mp.4001093

Koob, G. F., and Le Moal, M. (2001). Drug addiction, dysregulation of reward, and allostasis. *Neuropsychopharmacology* 24, 97–129. doi: 10.1016/S0893-133X(00)00195-0

Koob, G. F., and Volkow, N. D. (2010). Neurocircuitry of addiction. *Neuropsychopharmacology* 35, 217–238. doi: 10.1038/npp.2009.110

Koob, G. F., and Volkow, N. D. (2016). Neurobiology of addiction: a neurocircuitry analysis. *Lancet* 3, 760–773. doi: 10.1016/S2215-0366(16)00104-8

Lee, H. T., Chang, H. T., Lee, S., Lin, C. H., Fan, J. R., Lin, S. Z., et al. (2016). Role of IGF1R(+) MSCs in modulating neuroplasticity via CXCR4 cross-interaction. *Sci Rep* 6:32595. doi: 10.1038/srep32595

Li, C. Y., Mao, X., and Wei, L. (2008). Genes and (common) pathways underlying drug addiction. *PLoS Comput. Biol.* 4:e2. doi: 10.1371/journal.pcbi.0040002

Lin, J., Liu, L., Wen, Q., Zheng, C., Gao, Y., Peng, S., et al. (2014). Rapamycin prevents drug seeking via disrupting reconsolidation of reward memory in rats. *Int. J. Neuropsychopharmacol.* 17, 127–136. doi: 10.1017/S1461145713001156

Lipton, J. O., and Sahin, M. (2014). The neurology of mTOR. *Neuron* 84, 275–291. doi: 10.1016/j.neuron.2014.09.034

Liu, B., Bhatt, D., Oltvai, Z. N., Greenberger, J. S., and Bahar, I. (2014). Significance of p53 dynamics in regulating apoptosis in response to ionizing radiation, and polypharmacological strategies. *Sci Rep.* 4:6245. doi: 10.1038/srep06245

Liu, B., Liu, Q., Yang, L., Palaniappan, S. K., Bahar, I., Thiagarajan, P. S., et al. (2016). Innate immune memory and homeostasis may be conferred through crosstalk between the TLR3 and TLR7 pathways. *Sci. Signal.* 9:ra70. doi: 10.1126/scisignal.aac9340

Liu, B., Oltvai, Z. N., Bayir, H., Silverman, G. A., Pak, S. C., Perlmutter, D. H., et al. (2017). Quantitative assessment of cell fate decision between autophagy and apoptosis. *Sci. Rep.* 7:17605. doi: 10.1038/s41598-017-18001-w

Luo, L. (2002). Actin cytoskeleton regulation in neuronal morphogenesis and structural plasticity. *Annu. Rev. Cell Dev. Biol.* 18, 601–635. doi: 10.1146/annurev.cellbio.18.031802.150501

Ma, J., Wan, J., Meng, J., Banerjee, S., Ramakrishnan, S., and Roy, S. (2014). Methamphetamine induces autophagy as a pro-survival response against apoptotic endothelial cell death through the Kappa opioid receptor. *Cell Death Dis.* 5:e1099. doi: 10.1038/cddis.2014.64

Ma, X. M., and Blenis, J. (2009). Molecular mechanisms of mTOR-mediated translational control. *Nat. Rev. Mol. Cell Biol.* 10, 307–318. doi: 10.1038/nrm2672

Mechoulam, R., and Parker, L. A. (2013). The endocannabinoid system and the brain. *Annu. Rev. Psychol.* 64, 21–47. doi: 10.1146/annurev-psych-113011-143739

Mendelsohn, L. G., Kalra, V., Johnson, B. G., and Kerchner, G. A. (1985). Sigma opioid receptor: characterization and co-identity with the phencyclidine receptor. *J. Pharmacol. Exp. Ther.* 233, 597–602.

Miller, G. M. (2011). The emerging role of trace amine-associated receptor 1 in the functional regulation of monoamine transporters and dopaminergic activity. *J. Neurochem.* 116, 164–176. doi: 10.1111/j.1471-4159.2010.07109.x

Moeller, F. G., Dougherty, D. M., Rustin, T., Swann, A. C., Allen, T. J., Shah, N., et al. (1997). Antisocial personality disorder and aggression in recently abstinent cocaine dependent subjects. *Drug Alcohol Depend.* 44, 175–182. doi: 10.1016/S0376-8716(96)01335-X

Moradi, S., Charkhpour, M., Ghavimi, H., Motahari, R., Ghaderi, M., and Hassanzadeh, K. (2013). Gap junction blockers: a potential approach to attenuate morphine withdrawal symptoms. *J. Biomed. Sci.* 20:77. doi: 10.1186/1423-0127-20-77

Narita, M., Akai, H., Kita, T., Nagumo, Y., Narita, M., Sunagawa, N., et al. (2005). Involvement of mitogen-stimulated p70-S6 kinase in the development of sensitization to the methamphetamine-induced rewarding effect in rats. *Neuroscience* 132, 553–560. doi: 10.1016/j.neuroscience.2004.12.050

Natsvlishvili, N., Goguadze, N., Zhuravliova, E., and Mikeladze, D. (2015). Sigma-1 receptor directly interacts with Rac1-GTPase in the brain mitochondria. *BMC Biochem.* 16:11. doi: 10.1186/s12858-015-0040-y

Neasta, J., Barak, S., Ben Hamida, S., and Ron, D. (2014). mTOR complex 1: a key player in neuroadaptations induced by drugs of abuse. *J. Neurochem.* 130, 172–184. doi: 10.1111/jnc.12725

Neasta, J., Ben Hamida, S., Yowell, Q., Carnicella, S., and Ron, D. (2010). Role for mammalian target of rapamycin complex 1 signaling in neuroadaptations underlying alcohol-related disorders. *Proc. Natl. Acad. Sci. U.S.A.* 107, 20093–20098. doi: 10.1073/pnas.1005554107

Neasta, J., Hamida, S. B., Yowell, Q. V., Carnicella, S., and Ron, D. (2011). AKT signaling pathway in the nucleus accumbens mediates excessive alcohol drinking behaviors. *Biol. Psychiatry* 70, 575–582. doi: 10.1016/j.biopsych.2011.03.019

Nennig, S., and Schank, J. (2017). The role of NFkB in drug addiction: beyond inflammation. *Alcohol Alcohol.* 52, 172–179. doi: 10.1093/alcalc/agw098

Nestler, E. J. (2013). Cellular basis of memory for addiction. *Dialogues Clin. Neurosci.* 15, 431–443.

Okamoto, T., Minami, K., Uezono, Y., Ogata, J., Shiraishi, M., Shigematsu, A., et al. (2003). The inhibitory effects of ketamine and pentobarbital on substance p receptors expressed in Xenopus oocytes. *Anesth. Analg.* 97, 104–110. doi: 10.1213/01.ANE.0000066260.99680.11

Pahikkala, T., Airola, A., Pietila, S., Shakyawar, S., Szwajda, A., Tang, J., et al. (2015). Toward more realistic drug-target interaction predictions. *Brief. Bioinform.* 16, 325–337. doi: 10.1093/bib/bbu010

Parolaro, D., and Rubino, T. (2008). The role of the endogenous cannabinoid system in drug addiction. *Drug News Perspect.* 21, 149–157.

Patel, A. S., Lin, L., Geyer, A., Haspel, J. A., An, C. H., Cao, J., et al. (2012). Autophagy in idiopathic pulmonary fibrosis. *PLoS ONE* 7:e41394. doi: 10.1371/journal.pone.0041394

Pei, F., Li, H., Henderson, M. J., Titus, S. A., Jadhav, A., Simeonov, A., et al. (2017). Connecting neuronal cell protective pathways and drug combinations in a Huntington's disease model through the application of quantitative systems pharmacology. *Sci. Rep.* 7:17803. doi: 10.1038/s41598-017-17378-y

Philibin, S. D., Cortes, A., Self, D. W., and Bibb, J. A. (2011). Striatal signal transduction and drug addiction. *Front. Neuroanat.* 5:60. doi: 10.3389/fnana.2011.00060

Prosser, R. A., Stowie, A., Amicarelli, M., Nackenoff, A. G., Blakely, R. D., and Glass, J. D. (2014). Cocaine modulates mammalian circadian clock timing by decreasing serotonin transport in the SCN. *Neuroscience* 275, 184–193. doi: 10.1016/j.neuroscience.2014.06.012

Puighermanal, E., Marsicano, G., Busquets-Garcia, A., Lutz, B., Maldonado, R., and Ozaita, A. (2009). Cannabinoid modulation of hippocampal long-term memory is mediated by mTOR signaling. *Nat. Neurosci.* 12, 1152–1158. doi: 10.1038/nn.2369

Pytliak, M., Vargov,á, V., Mechírov,á, V., and Felsöci, M. (2011). Serotonin receptors-from molecular biology to clinical applications. *Physiol. Res.* 60, 15–25.

Rabanal-Ruiz, Y., Otten, E. G., and Korolchuk, V. I. (2017). mTORC1 as the main gateway to autophagy. *Essays Biochem*. 61, 565–584. doi: 10.1042/EBC20170027

Rocha, B. A., Fumagalli, F., Gainetdinov, R. R., Jones, S. R., Ator, R., Giros, B., et al. (1998). Cocaine self-administration in dopamine-transporter knockout mice. *Nat. Neurosci.* 1, 132–137. doi: 10.1038/381

Rothenfluh, A., and Cowan, C. W. (2013). Emerging roles of actin cytoskeleton regulating enzymes in drug addiction: actin or reactin'? *Curr. Opin. Neurobiol.* 23, 507–512. doi: 10.1016/j.conb.2013.01.027

Russo, S. J., Dietz, D. M., Dumitriu, D., Morrison, J. H., Malenka, R. C., and Nestler, E. J. (2010). The addicted synapse: mechanisms of synaptic and structural plasticity in nucleus accumbens. *Trends Neurosci.* 33, 267–276. doi: 10.1016/j.tins.2010.02.002

Sancho, R.,o., Calzado, M. A., Di Marzo, V., Appendino, G., and Muñoz, E. (2003). Anandamide inhibits nuclear factor-κB activation through a cannabinoid receptor-independent pathway. *Mol. Pharmacol.* 63, 429–438. doi: 10.1124/mol.63.2.429

Saxton, R. A., and Sabatini, D. M. (2017). mTOR signaling in growth, metabolism, and disease. *Cell* 168, 960–976. doi: 10.1016/j.cell.2017.02.004

Self, D. W., and Nestler, E. J. (1995). Molecular mechanisms of drug reinforcement and addiction. *Annu. Rev. Neurosci.* 18, 463–495. doi: 10.1146/annurev.ne.18.030195.002335

Shen, F., Wang, X.-W., Ge, F.-F., Li, Y.-J., and Cui, C.-L. (2016). Essential role of the NO signaling pathway in the hippocampal CA1 in morphine-associated memory depends on glutaminergic receptors. *Neuropharmacology* 102, 216–228. doi: 10.1016/j.neuropharm.2015.11.008

Shi, J., Jun, W., Zhao, L. Y., Xue, Y. X., Zhang, X. Y., Kosten, T. R., et al. (2009). Effect of rapamycin on cue-induced drug craving in abstinent heroin addicts. *Eur. J. Pharmacol.* 615, 108–112. doi: 10.1016/j.ejphar.2009.05.011

Shimada, S., Kitayama, S., Lin, C. L., Patel, A., Nanthakumar, E., Gregor, P., et al. (1991). Cloning and expression of a cocaine-sensitive dopamine transporter complementary DNA. *Science* 254, 576–578. doi: 10.1126/science.1948034

Sofuoglu, M., and Sewell, R. A. (2009). Norepinephrine and stimulant addiction. *Addict. Biol.* 14, 119–129. doi: 10.1111/j.1369-1600.2008.00138.x

Sora, I., Wichems, C., Takahashi, N., Li, X. F., Zeng, Z., Revay, R., et al. (1998). Cocaine reward models: conditioned place preference can be established in dopamine- and in serotonin-transporter knockout mice. *Proc. Natl. Acad. Sci. U.S.A.* 95, 7699–7704. doi: 10.1073/pnas.95.13.7699

Stern, A. M., Schurdak, M. E., Bahar, I., Berg, J. M., and Taylor, D. L. (2016). A perspective on implementing a quantitative systems pharmacology platform for drug discovery and the advancement of personalized medicine. *J. Biomol. Screen.* 21, 521–534. doi: 10.1177/1087057116635818

Su, L. Y., Luo, R., Liu, Q., Su, J. R., Yang, L. X., Ding, Y. Q., et al. (2017). Atg5- and Atg7-dependent autophagy in dopaminergic neurons regulates cellular and behavioral responses to morphine. *Autophagy* 13, 1496–1511. doi: 10.1080/15548627.2017.1332549

Sun, W. L., Quizon, P. M., and Zhu, J. (2016). Molecular mechanism: ERK signaling, drug addiction, and behavioral effects. *Prog. Mol. Biol. Transl. Sci.* 137, 1–40. doi: 10.1016/bs.pmbts.2015.10.017

Szklarczyk, D., Santos, A., von Mering, C., Jensen, L. J., Bork, P., and Kuhn, M. (2016). STITCH 5: augmenting protein-chemical interaction networks with tissue and affinity data. *Nucleic Acids Res.* 44, D380–384. doi: 10.1093/nar/gkv1277

Taylor, S. B., Lewis, C. R., and Olive, M. F. (2013). The neurocircuitry of illicit psychostimulant addiction: acute and chronic effects in humans. *Subst. Abuse Rehabil.* 4, 29–43. doi: 10.2147/SAR.S39684

Tomkins, D. M., and Sellers, E. M. (2001). Addiction and the brain: the role of neurotransmitters in the cause and treatment of drug dependence. *CMAJ* 164, 817–821.

Tritsch, N. X., Ding, J. B., and Sabatini, B. L. (2012). Dopaminergic neurons inhibit striatal output through non-canonical release of GABA. *Nature* 490, 262–266. doi: 10.1038/nature11466

Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., et al. (2015). Tissue-based map of the human proteome. *Science* 347:1260419. doi: 10.1126/science.1260419

Volkow, N., and Morales, M. (2015). The brain on drugs: from reward to addiction. *Cell* 162, 712–725. doi: 10.1016/j.cell.2015.07.046

Volkow, N. D., Fowler, J. S., and Wang, G.-J. (2003). The addicted human brain: insights from imaging studies. *J. Clin. Investig.* 111, 1444–1451. doi: 10.1172/JCI18533

Volkow, N. D., Koob, G. F., and McLellan, A. T. (2016). Neurobiologic advances from the brain disease model of addiction. *N. Engl. J. Med.* 374, 363–371. doi: 10.1056/NEJMra1511480

Volkow, N. D., Wang, G.-J., Fischman, M., Foltin, R., Fowler, J., Abumrad, N., et al. (1997). Relationship between subjective effects of cocaine and dopamine transporter occupancy. *Nature* 386, 827–830. doi: 10.1038/386827a0

Williams, M. J., and Adinoff, B. (2008). The role of acetylcholine in cocaine addiction. *Neuropsychopharmacology* 33, 1779–1797. doi: 10.1038/sj.npp.1301585

Wise, R. A. (1996). Addictive drugs and brain stimulation reward. *Annu. Rev. Neurosci.* 19, 319–340. doi: 10.1146/annurev.ne.19.030196.001535

Wise, R. A., and Koob, G. F. (2014). The development and maintenance of drug addiction. *Neuropsychopharmacology* 39, 254–262. doi: 10.1038/npp.2013.261

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., et al. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* 46, D1074–D1082. doi: 10.1093/nar/gkx1037

Wood, J. D., and Galligan, J. (2004). Function of opioids in the enteric nervous system. *Neurogastroenterol. Motil.* 16, 17–28. doi: 10.1111/j.1743-3150.2004.00554.x

Xie, X.-Q., Wang, L., Wang, J., Xie, Z., Yang, P., and Ouyang, Q. (2016). "*In silico* chemogenomics knowledgebase and computational system neuropharmacology approach for cannabinoid drug research," in *Neuropathology of Drug Addictions and Substance Misuse*, ed V. Preedy (Cambridge, MA: Elsevier), 183–195. doi: 10.1016/B978-0-12-800634-4.00019-6

Xie, X. Q., Wang, L., Liu, H., Ouyang, Q., Fang, C., and Su, W. (2014). Chemogenomics knowledgebased polypharmacology analyses of drug abuse related G-protein coupled receptors and their ligands. *Front. Pharmacol.* 5:3. doi: 10.3389/fphar.2014.00003

Yao, J., Ma, C., Gao, W., Liang, J., Liu, C., Yang, H., et al. (2016). Fentanyl induces autophagy via activation of the ROS/MAPK pathway and reduces the sensitivity of cisplatin in lung cancer cells. *Oncol. Rep.* 36, 3363–3370. doi: 10.3892/or.2016.5183

Yousefi, O. S., Wilhelm, T., Maschke-Neu,ß, K., Kuhny, M., Martin, C., Molderings, G. J., et al. (2013). The 1, 4-benzodiazepine Ro5-4864 (4-chlorodiazepam) suppresses multiple pro-inflammatory mast cell effector functions. *Cell Commun. Signal.* 11:13. doi: 10.1186/1478-811X-11-13

Zhu, J., Spencer, T. J., Liu-Chen, L.-Y., Biederman, J., and Bhide, P. G. (2011). Methylphenidate and μ opioid receptor interactions: a pharmacological target for prevention of stimulant abuse. *Neuropharmacology* 61, 283–292. doi: 10.1016/j.neuropharm.2011.04.015

Zhu, M., Xu, Y., Wang, H., Shen, Z., Xie, Z., Chen, F., et al. (2018). Heroin abuse results in shifted RNA expression to neurodegenerative diseases and attenuation of TNFα signaling pathway. *Sci. Rep.* 8:9231. doi: 10.1038/s41598-018-27419-9

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Pharmacology

ORIGINAL RESEARCH
published: 10 September 2019
doi: 10.3389/fphar.2019.00975

Ontological and Non-Ontological Resources for Associating Medical Dictionary for Regulatory Activities Terms to SNOMED Clinical Terms With Semantic Properties

Cédric Bousquet[1,2]*, *Julien Souvignet*[1,2], *Éric Sadou*[1], *Marie-Christine Jaulent*[1] *and Gunnar Declerck*[3]

[1] Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances en e-Santé, LIMICS, Sorbonne Université, Inserm, Université Paris 13, Paris, France, [2] Unit of Public Health and Medical Informatics, University of Saint Etienne, Saint Etienne, France, [3] EA 2223 Costech (Connaissance, Organisation et Systèmes Techniques), Centre de Recherche, Sorbonne Universités, Université de technologie de Compiègne, Compiègne, France

OPEN ACCESS

Edited by:
Lixia Yao,
Mayo Clinic, United States

Reviewed by:
Yongqun Oliver He,
University of Michigan Health System,
United States
Zhe He,
Florida State University,
United States

****Correspondence:***
Cédric Bousquet
cedric.bousquet@chu-st-etienne.fr

Specialty section:
This article was submitted to
Translational Pharmacology,
a section of the journal
Frontiers in Pharmacology

Received: *18 July 2018*
Accepted: *31 July 2019*
Published: *10 September 2019*

Citation:
Bousquet C, Souvignet J, Sadou É, Jaulent M-C and Declerck G (2019) Ontological and Non-Ontological Resources for Associating Medical Dictionary for Regulatory Activities Terms to SNOMED Clinical Terms With Semantic Properties. Front. Pharmacol. 10:975. doi: 10.3389/fphar.2019.00975

Background: Formal definitions allow selecting terms (e.g., identifying all terms related to "Infectious disease" using the query "has causative agent organism") and terminological reasoning (e.g., "hepatitis B" is a "hepatitis" and is an "infectious disease"). However, the standard international terminology Medical Dictionary for Regulatory Activities (MedDRA) used for coding adverse drug reactions in pharmacovigilance databases does not beneficiate from such formal definitions. Our objective was to evaluate the potential of reuse of ontological and non-ontological resources for generating such definitions for MedDRA.

Methods: We developed several methods that collectively allow a semiautomatic semantic enrichment of MedDRA: 1) using MedDRA-to-SNOMED Clinical Terms (SNOMED CT) mappings (available in the Unified Medical Language System metathesaurus or other mapping resources, e.g., the MedDRA preferred term "hepatitis B" is associated to the SNOMED CT concept "type B viral hepatitis") to extract term definitions (e.g., "hepatitis B" is associated with the following properties: has finding site liver structure, has associated morphology inflammation morphology, and has causative agent hepatitis B virus); 2) using MedDRA labels and lexical/syntactic methods for automatic decomposition of complex MedDRA terms (e.g., the MedDRA systems organ class "blood and lymphatic system disorders" is decomposed in blood system disorders and lymphatic system disorders) or automatic suggestions of properties (e.g., the string "cyclic" in preferred term "cyclic neutropenia" leads to the property has clinical course cyclic).

Results: The Unified Medical Language System metathesaurus was the main ontological resource reusable for generating formal definitions for MedDRA terms. The non-ontological resources (another mapping resource provided by Nadkarni and Darer in 2010 and MedDRA labels) allowed defining few additional preferred terms. While the Ci4SeR tool helped the curator to define 1,935 terms by suggesting potential supplemental relations based on the parents' and siblings' semantic definition, defining manually all MedDRA terms remains expensive in time.

Discussion: Several ontological and non-ontological resources are available for associating MedDRA terms to SNOMED CT concepts with semantic properties, but providing manual definitions is still necessary. The ontology of adverse events is a possible alternative but does not cover all MedDRA terms either. Perspectives are to implement more efficient techniques to find more logical relations between SNOMED CT and MedDRA in an automated way.

Keywords: adverse drug reaction, Medical Dictionary for Regulatory Activities, SNOMED Clinical Terms, ontology, clinical terminology, pharmacovigilance

INTRODUCTION

Formal representation of semantics as provided by computational ontologies and associated semantic Web techniques have been extensively used in medical data integration systems in the last decade (Sheth et al., 2005), and they now tend to be acknowledged as a powerful means to improve the quality of the processing chain of medical data, process automatic extraction of information and knowledge from large databases or ensure semantic interoperability between disparate data processing systems (Park and Hardiker, 2009; Schriml et al., 2012; Schulz and Jansen, 2013).

In the medical domain, classic terminologies are gradually giving way to clinical terminologies, in which terms are defined using knowledge representation languages (Rossi Mori et al., 1998). An example is SNOMED Clinical Terms (SNOMED CT), a general clinical terminology whose objective is to represent all possible terms required for coding the patient record and other applications for representation of biomedical information (Khorrami et al., 2018). SNOMED CT presents several advantages compared with classic terminologies, especially the ability to apply techniques of semantic reasoning in order to build new groups of terms, whereas classic terminologies are limited to default groupings (generally made manually by experts) that are already specified as part of the terminology (Bousquet et al., 2005).

Medical Dictionary for Regulatory Activities (MedDRA) is a classic terminology used by regulatory authorities and pharmaceutical companies for coding adverse drug reactions (ADR) in pharmacovigilance databases (Brown et al., 1999). MedDRA terms are not formally defined and search is therefore limited to existing categories (Bousquet et al., 2005). It is frequently difficult to identify the exact MedDRA category that represents a given medical condition under investigation in a sufficiently specific and exhaustive way, for example, during a pharmacovigilance database search (Brown, 2003).

Since several years, we have performed studies that showed that a knowledge-based approach is efficient for building new groups of ADR terms with World Health Organization Adverse Reaction Terminology (WHO ART) (Alecu et al., 2006) (Iavindrasana et al., 2006) and with MedDRA (Henegar et al., 2006; Declerck et al., 2012; Asfari et al., 2016; Souvignet et al., 2016a) in an automated way. This means that starting from a resource containing formal definitions of ADR terms, it is possible to make queries that correspond to a case definition in order to retrieve the related set of terms. This strategy was applied in Pharmacovigilance Adverse Reaction Terminology Server (Alecu et al., 2007) where building a knowledge base for all WHO ART terms was a challenge. Indeed, all definitions were to be set manually, and we therefore focused on automated ways to enrich WHO ART (Iavindrasana et al., 2006). We found that mapping of WHO ART with SNOMED CT by means of the Unified Medical Language System (UMLS) metathesaurus proved to be a very efficient method to build formal definitions of WHO ART terms in an automated way (Alecu et al., 2008).

Difficulties we encountered for enriching WHO ART now appear at a larger scale in MedDRA due to a growing number of terms and a more complex organization of MedDRA. Indeed, only about 50% of MedDRA terms [excluding lowest level term (LLT)] were associated with a SNOMED CT concept in UMLS (Bodenreider, 2009). Therefore, the mapping method we applied to WHO ART was a fair starting point but proved to be insufficient for obtaining an exhaustive enrichment of MedDRA.

Our objective was to evaluate the potential of reuse of ontological and non-ontological resources for defining and/or enriching definitions of MedDRA terms. We present in this article

Abbreviations: ADR, Adverse drug reactions; AERS, Adverse Events Reporting System; Ci4SeR, Curation Interface for Semantic Resources; HLGT, Higher Level Group Term; HLT, High Level Term; ICD-10, International classification of diseases, 10th edition; LALR, Lexically assign logically refine; LLT, Lowest Level Term; LOINC, Logical Observation Identifiers Names & Codes; MedDRA, Medical dictionary for drug regulatory activities; NEC, Not elsewhere classified; NOS, Not otherwise specified; OAE, Ontology of Adverse Events; OWL, Web Ontology Language; PT, Preferred Term; SMQ, Standardized MedDRA Queries; SNOMED CT, SNOMED Clinical Terms; SOC, System Organ Class; UMLS, Unified Medical Language System; WHO ART, World Health Organization-Adverse Reaction Terminology.

several complementary methods that may benefit different levels of automation and could also be reused in order to semantically enrich other terminologies. These include methods such as i) extracting SNOMED CT definitions based on MedDRA-to-SNOMED CT mappings available in the UMLS metathesaurus or other mapping resources and ii) developing lexical and syntactic methods using MedDRA term label information. The ability to reuse the selected ontological and non-ontological resources was measured by comparing the number of MedDRA terms associated with a formal definition after processing of these resources with the number of MedDRA terms to define. Additionally, we manually curated some term definitions using expert knowledge that allowed us to evaluate the time necessary and validated a sample of the formal definitions provided by the previous methods. We stored formal definitions of MedDRA terms in a semantic resource named OntoADR (Bousquet et al., 2014).

The organization of OntoADR and results of semantic queries on OntoADR have been already published (Bousquet et al., 2014; Souvignet et al., 2016b). This article presents methods we implemented for reusing ontological and non-ontological resources to enable the formalization of the semantics and results obtained with each of these methods, how they automate the development of formal representations of MedDRA terms, the limits related to these methods, and additional developments that would be required for a more complete semantic enrichment of MedDRA.

BACKGROUND

Hierarchical Organization of Medical Dictionary for Regulatory Activities

The MedDRA hierarchy consists of five levels (from broad to narrow), among which four are depicted in **Figure 1**: System Organ Class (SOC), e.g., hepatobiliary disorders; higher level group term (HLGT), e.g., hepatic and hepatobiliary disorders; high level term (HLT), e.g., hepatic viral infections; preferred term (PT), e.g., hepatitis B; and LLT not shown on the figure. The PT level is preferred for data analysis and retrieval. MedDRA was defined as multi-axial because one PT may be present in one primary SOC and also in several secondary SOC. However, one PT may exist only within one single HLT within a SOC. As HLT within a SOC constitutes disjoint classes, it is seldom reliable to consider only one HLT or higher level category when searching for MedDRA terms related to a pharmacovigilance safety topic (Bousquet et al., 2005; Asfari et al., 2016).

Moreover, it was recognized that HLT are not always sufficient to represent clinical conditions involving several organs (e.g., anaphylactic shock involving the kidney, liver, cardiovascular, and respiratory systems) because they only group together terms belonging to the same SOC. When searching for signals associated with a drug, MedDRA terms representing the suspected ADR must thus be identified prior to the running of signal detection algorithms Souvignet et al. (2012). For instance, if one suspects a given drug to cause acute renal failure, using the MedDRA term "renal failure acute" is generally not sufficient for the algorithms to extract a signal because the acute renal failure condition can be coded with several related MedDRA terms by health professionals (e.g., "renal impairment," "blood creatinine abnormal," or "dialysis"). Identifying clinically related terms in MedDRA is not an easy task, as those terms might exist in different locations of the MedDRA hierarchy.

Since several years, the Maintenance and Support Services Organization, which is responsible for MedDRA maintenance and diffusion, builds standardized MedDRA queries (SMQ) to address these issues (Mozzicato, 2007). SMQs consist of sets of PT from different branches of MedDRA that allow describing a particular medical condition and are intended to aid in case identification. SMQs are a way to describe safety topics relevant

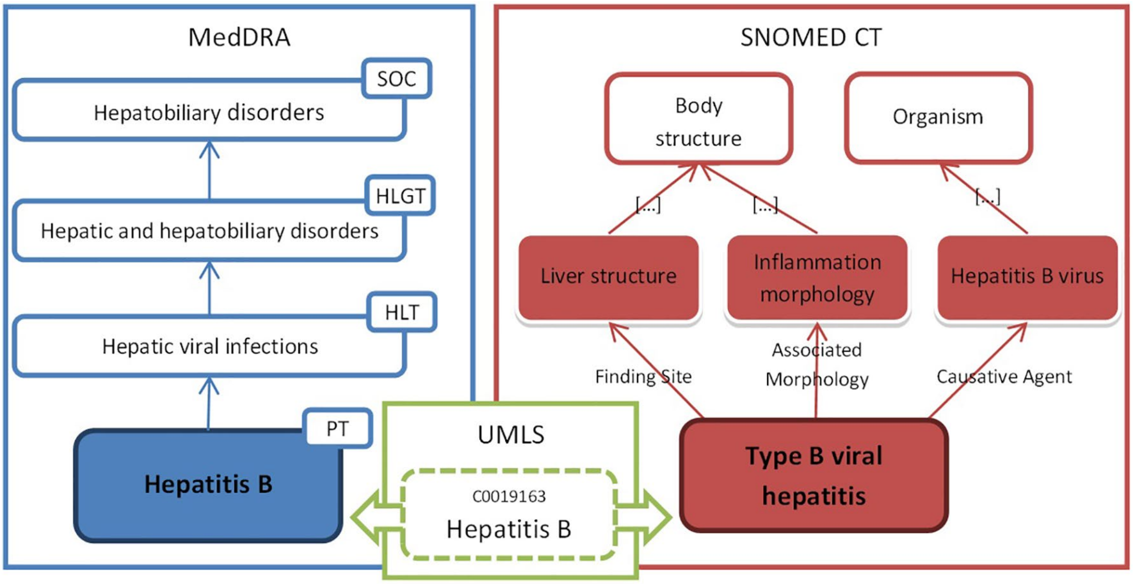


FIGURE 1 | Example of the formal definition of "hepatitis B" in OntoADR.

for pharmacovigilance that are not covered by the HLT and HLGT present in the MedDRA hierarchy. However, the achievement of SMQ raises important difficulties.

SMQs are currently developed manually by experts from the Maintenance and Support Services Organization, which is time consuming. Furthermore, once defined, they should not be modified or customized, and the existing SMQs do not cover all issues possible with drugs. Because experts have (even slightly) different understandings of the medical condition targeted by an SMQ, the kind of terms and the rationale for their selection may differ from an SMQ to another. This means that from one group of experts to another, for the same safety topic, the list of MedDRA terms selected could be different. Because SMQs are manually implemented, they could also miss important MedDRA terms. For those different reasons, the development of methods for automated selection of MedDRA terms on the basis of semantic information is desirable. Indeed, an automation of the process of PT selection in SMQ, even partial, could increase the quality and reproducibility of the SMQ and allow an important saving of time.

Difficulties for Searching Terms in Medical Dictionary for Regulatory Activities

The performance of pharmacovigilance systems based on spontaneous reporting is dependent on the information systems in which case reports are stored. In particular, these systems are subordinated to the ability of users to retrieve and exploit case reports in order to 1) reinforce existing knowledge on drug safety, 2) make assumptions about the existence of a causal relationship between a drug and an adverse event, and 3) evaluate the available information to implement regulatory measures to secure drug therapies. The search for pharmacovigilance case reports is difficult because it is necessary to identify the medical terms indicating the safety topic that one wishes to evaluate. In general, a term is not sufficient to designate this safety topic, and it is preferable to look for all case reports in relation to a set of terms (Hauben et al., 2006; Hansen et al., 2007). According to the MedDRA® Data Retrieval and Presentation: Points To Consider (ICH Working Group, 2018), "clinically related PTs might be overlooked or not recognized as belonging together because they might be in different groupings within a single SOC or they may be located in more than one SOC."

Figure 2 shows the problem of finding terms in MedDRA. Terms associated with a green tick are related to valvulopathy, while terms marked with a red cross do not correspond to valvulopathy. It is observed that the search terms are located in different branches of the terminology, which requires the pharmacovigilant specialist more time and effort to carry out his query. In addition, several HLT or HLGT must be combined to arrive at the final result, and irrelevant terms are present in these groups, which means that a search based on HLT or HLGT groupings will be associated with a large number of irrelevant PT. Another method for searching MedDRA terms is the textual query, but the terminology seems complex and does not reveal discriminating strings in the search for valvulopathies. For example "stenos * aort *" gives as results "stenosis of the aortic valve," "congenital aortic stenosis of the

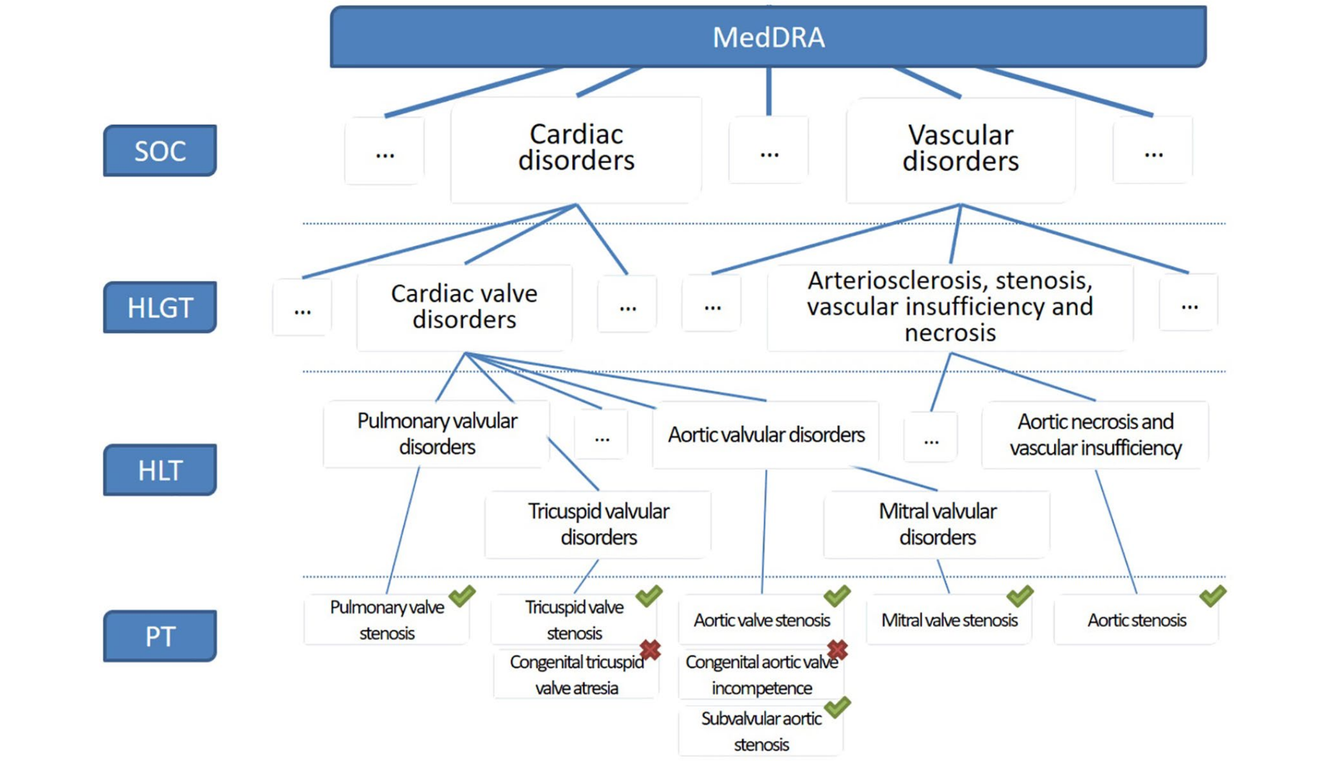

MedDRA
SOC
HLGT
HLT
PT
...
Cardiac disorders
...
Vascular disorders
...
...
Cardiac valve disorders
...
...
Arteriosclerosis, stenosis, vascular insufficiency and necrosis
...
Pulmonary valvular disorders
...
Aortic valvular disorders
...
Aortic necrosis and vascular insufficiency
Tricuspid valvular disorders
Mitral valvular disorders
Pulmonary valve stenosis ✓
Tricuspid valve stenosis ✓
Congenital tricuspid valve atresia ✗
Aortic valve stenosis ✓
Congenital aortic valve incompetence ✗
Subvalvular aortic stenosis ✓
Mitral valve stenosis ✓
Aortic stenosis ✓

FIGURE 2 | Extract of the MedDRA hierarchy showing preferred terms related or non-related to valvulopathies.

valve," and "stenosis of the mitral valve and insufficiency of the aortic valve," but it is necessary to perform additional text searches corresponding to findings for valvular involvement such as calcification or insufficiency.

Interface, Aggregation, and Reference Terminologies

Schulz et al. (2017) recently evaluated how interface, aggregation, and reference terminologies may interact in the context of the new 11th version of the International Classification of Diseases and its relations with SNOMED CT. Interface terminology was defined by Rosenbloom et al. (2006) as a "systematic collection of health care–related phrases (terms) that supports clinicians' entry of patient-related information into computer programs […]." According to Spackman (1997), the "main purpose of a reference terminology […] is the retrieval and analysis of data." The term "aggregation terminology" was first introduced by Rogers (2005) to designate a classification systems in which its main purpose is to enable "statistical aggregation," further defined by Schulz et al. (2017) as consisting of single hierarchies and disjoint classes.

We consider that such clarification would be useful in our case study, to better explain what we intend to do with MedDRA and SNOMED CT. Within our approach, MedDRA is the aggregation terminology, and SNOMED CT is the reference terminology. As our purpose is to improve retrieval and not coding of MedDRA terms, we did not work on building an interface terminology. Such interface terminology would be desirable to facilitate data entry in pharmacovigilance databases but is outside of our scope. In the following paragraph, we show how a graphical user interface implementing SNOMED CT as a reference terminology could help users' experience in selecting MedDRA terms and potentially improving search in pharmacovigilance databases.

Rationale for Supplementing Medical Dictionary for Regulatory Activities With Formal Definitions

We consider it is possible to overcome the limitations associated with the organization of MedDRA terminology and the difficulty to identify related MedDRA terms, by proposing an alternative method for the grouping of PTs based on their medical meaning rather than their position in the hierarchy. This new method is based on PT modeling in a form that allows logical inferences by a computer. From a technical point of view, the implementation of this method is based on knowledge engineering, a branch of artificial intelligence in which it is possible to describe MedDRA terms using a formal language (Bousquet et al., 2014). In the field of knowledge engineering, we define "ontology" as the set of objects of a domain and relations between these objects.

While McKnight (1999) recognized the need for user-directed composition of controlled health terminologies, and the required improvement of the user interface in the context of data entry, we believe that such user interface is also of great importance to enable composition for data retrieval. In a previous work, we implemented OntoADR query tools, a graphical user interface that relies on OntoADR, and compared the performances of eight users in selecting MedDRA PTs with the MedDRA web browser in a pilot study on five medical conditions (Souvignet et al., 2019). Although the number of medical conditions was low, we observed a statistically significant improvement by using OntoADR query tools compared with the MedDRA web browser for selecting MedDRA PTs (+27% precision and +34% recall). Similar to Maedche and Staab (2001), we consider that the target application may serve as a measure for validating the implemented ontology and believe that such criteria is more important than criteria relying only on the evaluation of the ontology without taking into account the context where it is used. This pilot study confirmed the validity of our approach and justifies that we continue the implementation of our mappings between MedDRA and SNOMED CT.

In order to address a safety issue, it is necessary to identify case reports in pharmacovigilance databases relative to this issue. A safety issue may concern the causality assessment of drug D in the occurrence of medical condition C. **Figure 3** shows the example of three use cases where one is evaluating the causal role of suspected drug D in three medical conditions: a) upper gastrointestinal hemorrhage, b) medical conditions with symptom of erythema, and c) fungal infectious disorders. A single MedDRA term is usually not sufficient to characterize a given medical condition. OntoADR is intended to support selection of MedDRA terms according to different criteria. **Figure 4** depicts the query performed in OntoADR to retrieve MedDRA terms associated to these medical conditions, e.g., "finding site: upper gastrointestinal tract," and "associated morphology: hemorrhage," and the 10 first MedDRA terms retrieved by this query, e.g., anastomotic ulcer hemorrhage, aorto-esophageal fistula, chronic gastrointestinal bleeding, etc.

Table 1 shows parts of formal definitions for 10 MedDRA terms associated with upper gastrointestinal hemorrhage among 27, in particular, SNOMED CT concepts that are filler of relations "finding site" and "associated morphology." Fillers that are not relevant to upper gastrointestinal hemorrhage are in italic, e.g., *Stenosis* for the MedDRA PT "anastomotic ulcer hemorrhage." In case the filler of the "associated morphology" relation is "hemorrhage," the query is immediately satisfied for this condition, e.g., three MedDRA terms (aorto-esophageal fistula, gastric antral vascular ectasia, and gastric hemorrhage) are defined as having "hemorrhage" as their associated morphology. When "hemorrhage" is not the filler of the "associated morphology" relation, other relevant fillers may be retrieved thanks to the subsumption mechanism that establishes hierarchical relations between a parent concept and its children concept. For example, "hemorrhage" subsumes "acute bleeding ulcer," 'bleeding varices," "chronic hemorrhage," and "hemorrhagic inflammation." "Upper gastro intestinal tract structure" subsumes "duodenal structure," "esophageal structure," "gastrojejunal junction structure," "pyloric antrum structure," and "stomach structure."

Semantic Enrichment

The traditional process of domain ontology construction is based on expert intervention (Bedini and Nguyen, 2007). Although this manual procedure guarantees a fair quality of the generated

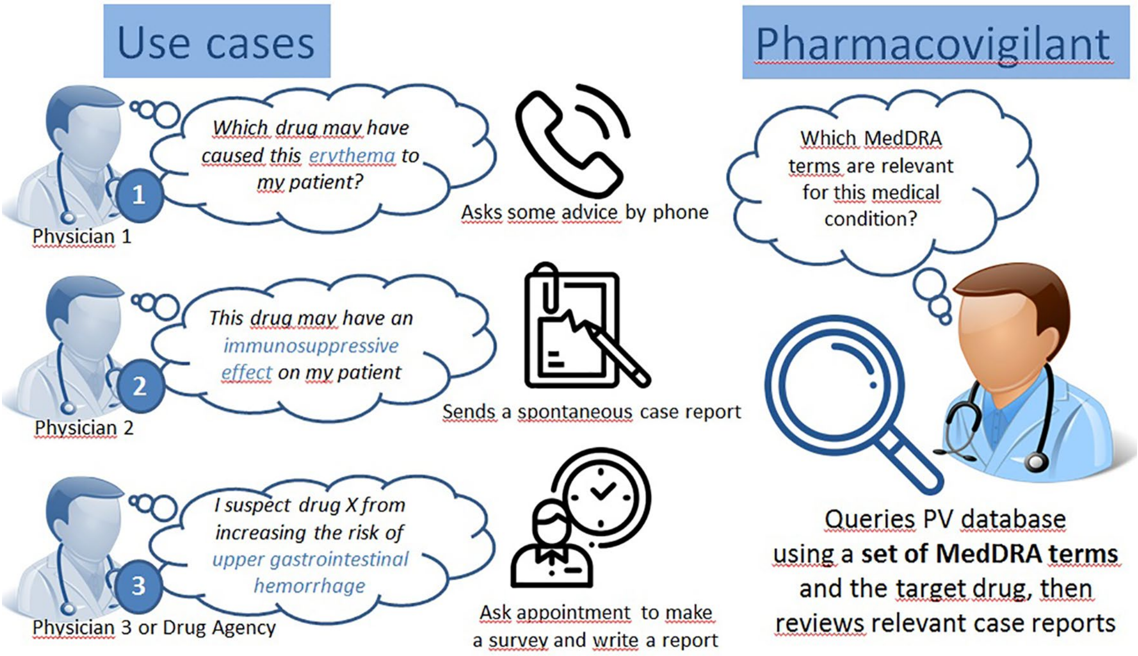
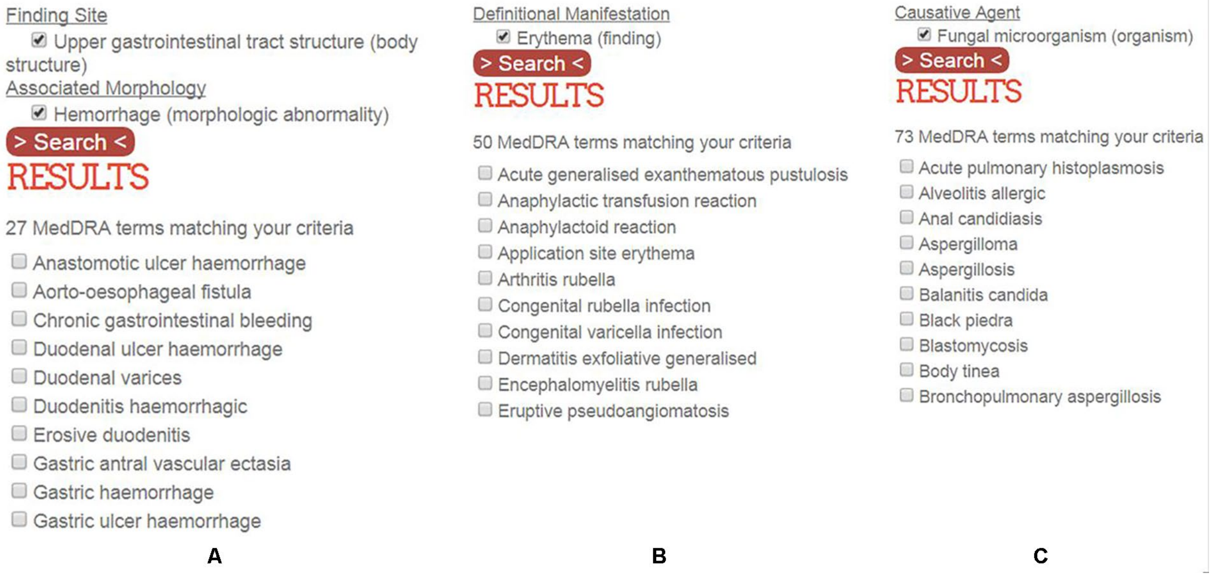

Use cases
Pharmacovigilant
Which drug may have caused this erythema to my patient?
1
Physician 1
Asks some advice by phone
Which MedDRA terms are relevant for this medical condition?
This drug may have an immunosuppressive effect on my patient
2
Physician 2
Sends a spontaneous case report
I suspect drug X from increasing the risk of upper gastrointestinal hemorrhage
3
Physician 3 or Drug Agency
Ask appointment to make a survey and write a report
Queries PV database using a set of MedDRA terms and the target drug, then reviews relevant case reports

FIGURE 3 | Three use case where data retrieval in a pharmacovigilance database is necessary.



Finding Site
☑ Upper gastrointestinal tract structure (body structure)
Associated Morphology
☑ Hemorrhage (morphologic abnormality)
> Search <
RESULTS
27 MedDRA terms matching your criteria
□ Anastomotic ulcer haemorrhage
□ Aorto-oesophageal fistula
□ Chronic gastrointestinal bleeding
□ Duodenal ulcer haemorrhage
□ Duodenal varices
□ Duodenitis haemorrhagic
□ Erosive duodenitis
□ Gastric antral vascular ectasia
□ Gastric haemorrhage
□ Gastric ulcer haemorrhage

A

Definitional Manifestation
☑ Erythema (finding)
> Search <
RESULTS
50 MedDRA terms matching your criteria
□ Acute generalised exanthematous pustulosis
□ Anaphylactic transfusion reaction
□ Anaphylactoid reaction
□ Application site erythema
□ Arthritis rubella
□ Congenital rubella infection
□ Congenital varicella infection
□ Dermatitis exfoliative generalised
□ Encephalomyelitis rubella
□ Eruptive pseudoangiomatosis

B

Causative Agent
☑ Fungal microorganism (organism)
> Search <
RESULTS
73 MedDRA terms matching your criteria
□ Acute pulmonary histoplasmosis
□ Alveolitis allergic
□ Anal candidiasis
□ Aspergilloma
□ Aspergillosis
□ Balanitis candida
□ Black piedra
□ Blastomycosis
□ Body tinea
□ Bronchopulmonary aspergillosis

C

FIGURE 4 | MedDRA terms associated to a formal definition (only the 10 first MedDRA terms retrieved by the query are depicted). **(A)** MedDRA terms associated to upper gastrointestinal hemorrhage, **(B)** MedDRA terms describing medical conditions associated with erythema, **(C)** MedDRA terms describing infectious diseases induced by fungi.

resource, it suffers from several difficulties; among those are the cold start problem (starting from scratch) and the lack of availability of domain experts (Qawasmeh et al., 2018). In fact, the high cost of experts' interventions is the major bottleneck identified early in the state of the art of ontology construction (Cullen and Bryman, 1988; Simperl et al., 2006; Balakrishna et al., 2010). This bottleneck justifies reusing and linking existing resources, when available, to create new ontologies (Alani, 2006). Reuse is not always possible because ontologies may not exist in the field of interest. For example, Mazo et al. (2017) describe a histological ontology of the human cardiovascular system and report that they "did not find in the State-of-the-Art an ontology of

TABLE 1 | Finding site and associated morphology of 10 MedDRA terms describing upper gastrointestinal hemorrhage among 27.

	hasFindingSite	hasAssociatedMorphology
Anastomotic ulcer haemorrhage	Gastrojejunal junction structure	Acute bleeding ulcer Stenosis
Aorto-esophageal fistula	Aortic structure Esophageal structure	Hemorrhage
Chronic gastrointestinal bleeding	Stomach structure	Chronic hemorrhage
Duodenal ulcer haemorrhage	Duodenal structure	Acute bleeding ulcer
Duodenal varices	Portal vein structure Duodenal structure	Bleeding varices
Duodenitis hemorrhagic	Duodenal structure	Hemorrhagic inflammation
Erosive duodenitis	Duodenal structure	Hemorrhagic inflammation
Gastric antral vascular ectasia	Pyloric antrum structure	Hemorrhage Angiectasia
Gastric hemorrhage	Stomach structure	Hemorrhage
Gastric ulcer hemorrhage	Stomach structure	Acute bleeding ulcer

histology neither a similar organization of hierarchies of histology terms that [they] may be able to reuse." At the opposite, when all the ontologies that are needed are available, it is sufficient to reuse and assemble them, such as in the example of the development of the orthology ontology (Fernández-Breis et al., 2016).

Ontology enrichment is the task of extending an existing ontology with additional concepts and semantic relations and placing them at the correct position in the ontology (Petasis et al., 2011). Automatic ontological construction is often based on learning (Maedche and Staab, 2001; Buitelaar et al., 2005). Such approach can be based on unstructured texts (Asim, 2018; Cimiano et al., 2006; Emani et al., 2015; Costa et al., 2016; Dasgupta et al., 2018), informal ontologies (Astrakhantsev and Turdakov, 2013), or linked data (Gavankar et al., 2012; Tiddi et al., 2012; Riga et al., 2017). A particular case of unstructured data corresponds to the labels of ontology identifiers that may be very dense with information. Such information described as "hidden semantics" by Third (2012) received little attention, at the exception of the gene ontology identifiers (e.g., Quesada-Martínez et al., 2015). SNOMED CT identifiers may also benefit from such approach, but this was limited to taking into account the "acute" and "chronic" qualifiers for evolution of diseases (Rector and Iannone, 2012) and the occurrence of congenital diseases (van Damme et al., 2018). Such "hidden semantics" were also detected by Nadkarni and Darer (2010) to build correspondences between MedDRA and SNOMED CT, but these correspondences were not associated with relations, which makes this work interesting for reuse but requires reengineering to transform the correspondences into semantic relations.

One of the major difficulties of these approaches is the extraction of non-taxonomic relationships (Dahab et al., 2008; Sánchez and Moreno, 2008; Villaverde et al., 2009; Petasis et al., 2011; Serra et al., 2014). Furthermore, several automatic approaches for ontology reusing and engineering still require domain experts and knowledge engineers (Bobed et al., 2012). Thus, semiautomatic approaches could be a good alternative (Balakrishna et al., 2010).

Semiautomatic approaches employ intelligent methods to significantly reduce, without completely replacing, human efforts (Huang et al., 2014). In such approaches, the role of experts could be limited to validating final automatic learning results (Wächter and Schroeder, 2010) or suggesting improvements at the end of ontology life cycle (Alobaidi, 2018). Expert intervention can be achieved with the help of graphical user interface (Wächter and Schroeder, 2010), spreadsheets (Blfgeh et al., 2017; Judkins et al., 2018), or specified pipelines such as the eXtensible ontology development (He et al., 2018).

NeON Methodology

After comparing several methods for ontology development, we selected the NeON methodology (Suárez-Figueroa et al., 2012) because it was the most appropriate to illustrate the strategy that we followed for designing OntoADR. While other methodologies may also be considered for knowledge engineering and may be more relevant in other contexts, we considered that dimensions of reuse were the most important features when selecting NeON.

While other approaches for ontology engineering provide methodological guidance, the NeON Methodology does not prescribe a rigid workflow. It suggests a variety of pathways based on nine flexible scenarios that address common issues, such as reusing, reengineering, and merging ontological resources. These ontological resources also comprise ontology design patterns (Aranguren, 2008; Gangemi, 2005; Blomqvist, 2008; Presutti and Gangemi, 2008), which are generic templates or abstract descriptions proposed to enforce best practices in ontology implementation. One particularity of NeON matching well with our specific approach is that it also takes into account reusing and reengineering of non-ontological resources, which is not the case of other methodologies such as METHONTOLOGY (Fernández-López et al., 1997) and On-To-Knowledge (Sure et al., 2004). These non-ontological resources may consist of structured data such as terminologies (Jimeno-Yepes et al., 2009) or databases, unstructured data (e.g., articles), or semi-structured data (e.g., XML, JSON) (Qawasmeh et al., 2018). In addition to these nine scenarios, NeON also integrates support activities such as knowledge acquisition, documentation, and evaluation that should be carried out during the whole ontology development cycle.

MATERIAL AND METHODS

Summary of the Method

Application of the NeON methodology

We applied the following scenarios of the NeON methodology for implementing OntoADR (**Figure 5**).

- Scenario 1. From Specification to Implementation: this includes four steps (specification, conceptualization, formalization, and implementation). We previously presented these steps in previous work [(Bousquet et al., 2014)] and limit here the

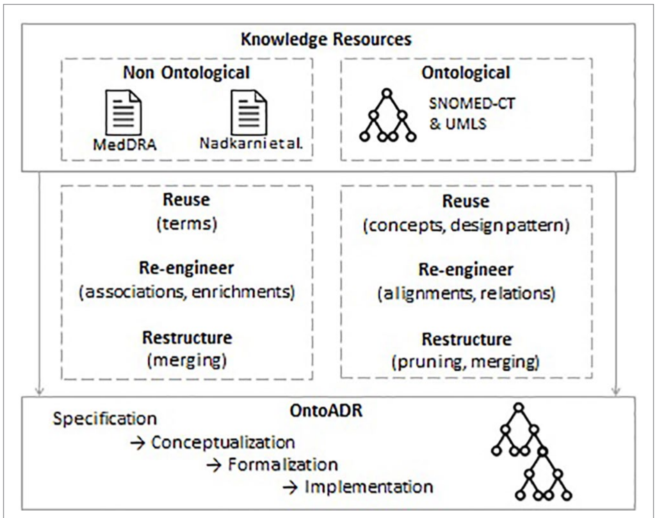


FIGURE 5 | Scenarios of the NeON methodology for implementing OntoADR.

scope of this presentation to scenarios that emphasize the reuse of ontological and non-ontological resources.

- Scenario 2. Reusing and Re-engineering Non-Ontological Resources: We transformed MedDRA into a subsumption tree (see [(Bousquet et al., 2014)], and retrieved a non-ontological resource provided by Nadkarni & Darer (2010). We analyzed this non-ontological resource in order to establish the correspondence between its content (mappings between MedDRA terms and SNOMED CT) and formal definitions that benefit from explicit relations derived from the SNOMED CT concept model. Then, we generated the formal definitions based on this non-ontological resource for inclusion in OntoADR.
- Scenario 6. Reusing, Merging, and Re-engineering Ontological Resources: We reused available ontological resources (UMLS, SNOMED CT) and merged SNOMED CT with MedDRA using mappings available in UMLS. We also reused the SNOMED CT concept model and had to reengineer it to keep only classes and relations that are relevant to formally define MedDRA terms. While the concept model may be considered as a building block that enforces best practices for ontological design in SNOMED CT, it does not strictly correspond to the description of content ontology design patterns, which explains why we did not implement Scenario 7: Reusing ontology design patterns (ODPs).
- Scenario 8. Restructuring Ontological Resources: This consists in pruning parts of the SNOMED CT that are not relevant for the description that was previously described by Souvignet et al., (2016b) and enriching the ontology by adding supplementary concepts and axioms.

In addition to these different scenarios, we also implemented "ontology support activities" for "knowledge acquisition" that comprises activities for (1) capturing knowledge from the MedDRA labels and work by a domain expert for adding formal definition using the Ci4SeR tool [(Souvignet et al., 2014)] and (2) "ontology validation" that consists in checking that the meaning of the ontology definitions are compliant with the definitions we intended the MedDRA terms to convey.

Flow chart of the method

Figure 6 depicts a flow chart representing an overall representation of the several steps and tasks proposed in the article to get an overview of the algorithm at a glance. While this diagram could make readers believe that all these steps were conducted in parallel, it is proposed only as a convenient way to apprehend the method as a whole. The previous paragraph where we applied the NeON methodology shows a different perspective where different scenarios were applied at different time. In the flow chart, each MedDRA term is considered one after the other and can go through several parallel paths according to different conditions.

- If a MedDRA term is associated to a SNOMED CT concept in the UMLS metathesaurus, one mapping is manually selected: the use of the mapping information is described in the section *Using MedDRA-to-SNOMED CT Mappings From UMLS Metathesaurus*.
- If a MedDRA term is present in another mapping source than the UMLS (e.g., the Nadkarni and Darer's proposal), then this mapping is used for the definition in a way that is described in the following section: *Using Another MedDRA-to-SNOMED CT Mapping Resource*.
- If a MedDRA term is composed of several distinct terms by a conjunction (e.g., "acute and chronic thyroiditis" that consists of two medical conditions "acute thyroiditis" and "chronic thyroiditis"), then the MedDRA term is decomposed according to the algorithm described in the section *Using a Syntactic Decomposition Algorithm on Complex MedDRA Terms*. Then, MedDRA subterms are considered as additional MedDRA terms that can be used as a new input for the whole algorithm.
- If a MedDRA term contains a substring with associated meaning (e.g., medical words ending in -algia that indicate pain), this MedDRA term may benefit from a potential syntactic enrichment: the generation of a partial definition is described in the section *Automatic Lexical Enrichment Methods*.

Some manual definitions can optionally be added (see the section *NeON Methodology*). All partial definitions acquired with the different algorithms are then automatically combined into a merged definition of the MedDRA term. We used MedDRA version 17 that consists of 26 SOC, 334 HLGT, 1,720 HLT, 20.559 PT, and 72,637 LLT, SNOMED CT version March 2015 and UMLS version 2014AB. SNOMED CT concepts were extracted from the Concepts_Core_INT file (Release Format 1) and the hierarchy and semantic properties from the Relationships_Core_INT file. This version of MedDRA was applied to the following paths: "using UMLS metathesaurus mappings," "automatic enrichment methods," and "manual definition of concepts." MedDRA 13 that consists of 26 SOC, 335 HLGT, 1,709 HLT, 18,786 PT, and 68,258 LLT was applied to the following paths: "Using other mapping resources" and "Using a decomposition algorithm and Metamap software to map complex MedDRA terms."

Problems With Mapping Other Layers Than the PT Level

We have tried to map other layers (SOC, HLT, and HLGT). For instance, the cardiac disorders SOC concept has for formal definition hasFindingSite some "Heart Structure." While

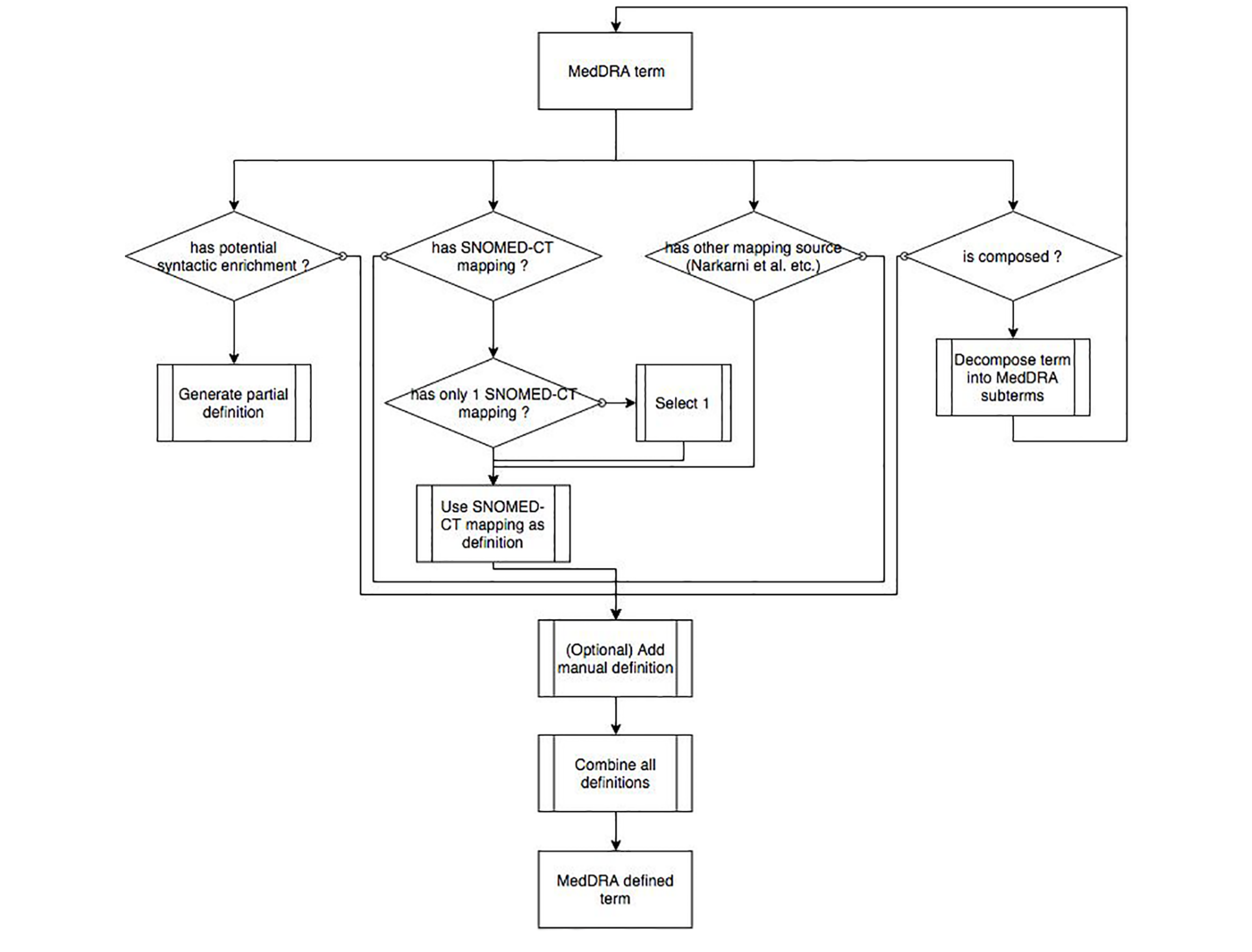

MedDRA term
has potential syntactic enrichment ?
has SNOMED-CT mapping ?
has other mapping source (Narkarni et al. etc.)
is composed ?
Generate partial definition
has only 1 SNOMED-CT mapping ?
Select 1
Decompose term into MedDRA subterms
Use SNOMED-CT mapping as definition
(Optional) Add manual definition
Combine all definitions
MedDRA defined term

FIGURE 6 | Flow chart representing an overall representation of the several steps and tasks.

some SOC, HLGT, and HLT were accurately defined, essentially thanks to mappings in UMLS, it was decided not to present the results in this article. We used formal definitions associated with the three MedDRA higher levels only in the Ci4SeR tool (Souvignet et al., 2014), where the curator could use them if she considered them as relevant. We decided not to map the LLT because PT is the preferred level for case report analysis and search in pharmacovigilance databases.

We explained in previous work why the MedDRA hierarchy cannot be converted into a subsumption tree because this sometimes causes semantic inconsistencies (Bousquet et al., 2014). The reason is that most high level categories in MedDRA are intended to reflect the domain actors' practices (i.e., following the different medical specialties) and are not necessarily organized according to different semantic criteria as one would expect in a well-formed ontology. A first example concerns groups of symptoms (HLGT or HLT) that are placed under the general categories of disorders that they are the symptom of (SOC or HLGT). Such hierarchical organization would not be authorized in an ontology as the relation being-a-symptom-of does not imply an is-a relation. For instance, the PT "dyspnoea" and "dizziness or syncope" belong to the HLGT "cardiac disorder signs and symptoms" that is under the SOC cardiac disorders. While dyspnea for instance refers to conditions that may be associated to cardiac disorders, such symptom cannot be considered as a cardiac disorder. A second example is the MedDRA PT sudden death that belongs to the following hierarchies: 1) ventricular arrhythmias and cardiac arrest (HLT)/cardiac arrhythmias (HLGT), and 2) death and sudden death (HLT)/fatal outcomes (HLGT). While "cardiac arrhythmias" is defined in OntoADR with the HASFINDINGSITE SOME "HEART STRUCTURE" property, the sudden death PT should not inherit from such property because sudden death could be the consequence of death that is not of cardiac origin.

Using Medical Dictionary for Regulatory Activities-to-SNOMED Clinical Terms Mappings From UMLS Metathesaurus

The UMLS may be used as a source of knowledge for adding formal definitions to medical terminologies (Schulz and Hahn, 2001). Based on our initial experience (Alecu et al., 2008), we assume that SNOMED CT is currently the best candidate for providing formal definitions to MedDRA. SNOMED CT terms are defined using description logic (DL) formalism, and a fair number of alignments between MedDRA and SNOMED CT are present in the UMLS metathesaurus. Therefore, most of the formal definitions attributed to MedDRA terms in OntoADR are based on semantic information extracted from the SNOMED CT clinical terminology. When a MedDRA term is mapped to a SNOMED CT concept, we reused the semantic information within SNOMED CT in order to build the formal definition of the MedDRA term. Identifying reliable MedDRA-to-SNOMED CT mappings is thus an essential step in our methodology to define MedDRA term semantics.

The UMLS (Lindberg et al., 1993) consists of a semantic network and a metathesaurus developed by the US National Library of Medicine to link terms from more than a hundred controlled vocabularies, including SNOMED CT and MedDRA. Terms from the different vocabularies are linked together by association to a unique UMLS concept defined by a concept unique identifier, e.g., "C0019163" that is mapped to both MedDRA term "hepatitis B" and SNOMED CT concept "type B viral hepatitis."

In OntoADR, MedDRA term "hepatitis B" has the formal definition: hasFindingSite some "Liver Structure," hasAssociatedMorphology some "Inflammation Morphology," and hasCausativeAgent some "Hepatitis B Virus" where has FindingSite, hasAssociatedMorphology, and hasCausativeAgent are OntoADR semantic relations inspired from SNOMED CT, and "liver structure," "inflammation morphology," and "hepatitis B virus" are SNOMED CT concepts we imported in OntoADR (**Figure 1**).

In order to map MedDRA and SNOMED CT terms, we developed an algorithm following these steps: i) search for the MedDRA PT in the UMLS using the MedDRA identifier; ii) for MedDRA PT without SNOMED CT mappings in the UMLS, if the PT has one or more related LLT considered as synonymous, then the LLT identifier is used for a new UMLS search. iii) If this second search is unsuccessful, the algorithm performs a last UMLS search using PT and LLT labels, seeking to pair these labels with SNOMED CT concepts by string matching.

All mapping propositions selected from the UMLS metathesaurus were validated, modified, or completed by knowledge engineers and pharmacovigilance experts of our team. i) All one-to-one mappings we decided to use were first validated by checking manually the correspondence between the meanings of terms.

Several SNOMED CT concepts may be proposed as synonyms (Fung et al., 2005) of the same MedDRA concept, although they have different meanings. Each MedDRA concept in OntoADR can have only one equivalent SNOMED CT concept. When several SNOMED CT concepts are proposed in UMLS as synonym of a MedDRA term, only one was selected by an expert for building the formal definition. Such selection should be based on synonymy between a SNOMED CT concept and a MedDRA term. According to Fung, synonymy between term X and Y may be defined according to linguistic criteria (Fung et al., 2005) such as enforcing that it is possible to replace X by Y in any sentence without modifying the meaning. Examples of such synonyms are "celiac disease" and "gluten enteropathy," or "kidney stone" and "renal calculus." Selection of a SNOMED CT concept was performed first by comparing its label with the MedDRA term label. In case both were identical, which occurred most of the time, it was obvious to select this mapping, but in other cases, we took into account the medical relevance of the mapping and had to rely on expert evaluation. For example, three SNOMED CT concepts are mapped to the MedDRA term "Spondylitis": "inflammatory spondylopathy," "undifferentiated spondylitis," and "spondylitis," and the later appears as a perfect match according to label comparison.

Such a validation process was necessary because UMLS mapping propositions are not always semantically valid. MedDRA terms and SNOMED CT concepts mapped together in UMLS can refer to different medical entities, even if they are homonyms. For instance, the MedDRA term "vascular disorders" and its SNOMED CT homonym "vascular disorder" are mapped together in UMLS; however, the former refers to disorders of blood and lymphatic systems and the later only to disorders of blood vessels (lymphatic system disorders are caught by the concept "disorder of lymphatic system" in SNOMED CT). ii) In case of one-to-n mappings, a manual expert choice was made to select the SNOMED CT concept whose definition best fitted the meaning of the correspondent MedDRA concept. iii) When no SNOMED CT concept among the ones suggested by UMLS was satisfactory, the definition of the correspondent MedDRA concept was made manually. iv) Mapping a MedDRA term with a SNOMED CT concept does not ensure that the former gets a complete (or even a satisfying) formal definition of its semantics: the formal definition can be incomplete or even be literally absent: it is common to find SNOMED CT concepts, for instance psychiatric concepts, that have no definitional properties. When necessary, the semantic properties from SNOMED CT attributed to MedDRA concepts through the mapping procedure were thus completed manually by additional assertions.

Using Another Medical Dictionary for Regulatory Activities-to-SNOMED Clinical Terms Mapping Resources

To complete the mappings selected from UMLS, we also made use of Nadkarni and Darer's propositions of mappings (Nadkarni and Darer, 2010). Using one year of data (recorded between July 1, 2008 and April 30, 2009) from the US Food and Drug Administration Adverse Events Reporting System (AERS) pharmacovigilance database, the authors identified 3,705 MedDRA PT that collectively accounted for 95% of case reports. The 3,705 selected MedDRA terms correspond to high-frequency terms in the US Food and Drug Administration database and potentially have a great added value. After eliminating terms already mapped to SNOMED CT concepts in UMLS, they attempted to map manually the remaining terms (786 in total) with software assistance. Most of those terms (733) could be mapped by Nadkarni and Darer with SNOMED CT concepts *via* one-to-one or one-to-n mappings.

Several problems have been encountered when trying to reuse Nadkarni and Darer's propositions of mappings

(Nadkarni and Darer, 2010). i) First, in the case of one-to-n mappings, the authors broke down a MedDRA term in such a way to associate it to several SNOMED CT concepts but did not specify which semantic relation was relevant. For example, they mapped the MedDRA concept "tongue discoloration" with SNOMED CT concepts "abnormal color" and "entire tongue" but did not specify which semantic relation interconnected the first to the others. Obviously, it cannot be here an equivalence or synonymy (Fung et al., 2005) ("same as") relation as in the case of one-to-one mapping. When SNOMED CT concepts belong to branches such as *body structure* or *morphologic abnormality*, the relationship to use is easy to deduce, and its creation can be automated: it will be in the first case the HASFINDINGSITE relationship and, in the second, the HASASSOCIATEDMORPHOLOGY relationship. However, when it comes to SNOMED CT concepts from branches *finding*, *qualifier value*, or *disorder*, the relationship to use is not obvious, and only a human expert can decide. A major part of our recovery work was to specify these relationships by making use of the set of relationships available in OntoADR.

ii) We have also occasionally been forced to revise the proposed Nadkarni and Darer's mappings, partly for reasons of pure semantic accuracy and partly because of the purpose of OntoADR, which we illustrate here using three examples:

- For the MedDRA term "feeding disorder neonatal," the authors propose a mapping with SNOMED CT concepts "feeding disorder of infancy OR early childhood" and "neonatal." The second mapping is correct, and we have included "neonatal" in OntoADR to define the MedDRA concept "feeding disorder neonatal" with the HASOCCURRENCE relationship (HASOCCURRENCE SOME NEONATAL). We did not, however, reuse the first mapping. The concept "feeding disorder of infancy OR early childhood" is broader than the MedDRA concept "feeding disorder neonatal," which concerns only the newborn. A mapping with SNOMED CT concept "feeding problems in newborn" would have been more accurate. To complete the formal definition of "feeding disorder neonatal," we added the properties INTERPRETS SOME "FEEDING PATTERN" and HASINTERPRETATION SOME "ABNORMAL."
- The MedDRA concept "azotaemia" is mapped by Nadkarni and Darer to SNOMED CT concepts "blood urea nitrogen measurement" and "increased." We used the first by giving the MedDRA concept "azotaemia" the property INTERPRETS SOME BLOOD UREA NITROGEN MEASUREMENT in OntoADR. However, the use of the second to describe the relationship HASINTERPRETATION appeared problematic. Indeed, azotemia is characterized not so much by an increased concentration of nitrogen compounds in the blood but as a concentration above a certain reference threshold. We thus opted for the creation of the property HASINTERPRETATION SOME ABOVE REFERENCE RANGE, more accurate in this context.
- For the MedDRA term "anorectal discomfort," the authors propose a mapping with SNOMED CT concepts "discomfort" and "anus and rectum (combined site)." However, the problem here is that the SNOMED CT concept "anus and rectum (combined site)" is set in an isolated portion of the SNOMED CT branch "body structure" (branch called "group of anatomical entities"). Nothing connects it to the concepts of the digestive system structures (e.g., no relationship PART-OF). Due to the SEP decomposition (structure, entire, part) of the anatomical branch of SNOMED CT, the concept "anal structure" has no relation to the concept "anus and rectum." It would have been impossible to use this localization by semantic reasoning, for example, to identify concepts located on part of the anorectal system (principle of subsumption reasoning: concepts that have a relationship of location on parts of the anorectal structure are considered by inference as siblings of the concept of diseases that are located on the whole anorectal structure). We therefore preferred to use the SNOMED CT concept "anorectal structure" to define the relationship HASFINDINGSITE of the MedDRA concept "anorectal discomfort" in OntoADR. This SNOMED CT concept allows the semantic reasoning operation described previously. Moreover, we can assume that the MedDRA term "anorectal discomfort" is sometimes used to encode ADRs in a non-specified way that may be anal or rectal, and not both, as is implied by the use of the SNOMED CT concept "anus and rectum (combined site)." It is therefore important to locate by subsumption concepts that are located within a substructure of the whole anorectal structure.

Using a Syntactic Decomposition Algorithm on Complex Medical Dictionary for Regulatory Activities Terms

Among MedDRA terms that are not mapped with SNOMED CT terms in UMLS, there are many complex terms, i.e., corresponding to composed expressions. MedDRA complex terms are of several kinds: a) expressions composed with an AND logical operator or commas (e.g., "acute and chronic thyroiditis" or "pregnancy, labour, delivery, and postpartum conditions"); b) expressions composed with "NEC" (not elsewhere classified), "unspecified," or with a text between brackets, usually to specify exclusion clauses [e.g., "autoimmune disorders NEC," "laryngeal neoplasms malignancy unspecified," or "ocular neoplasms malignant (excl. melanomas)"]; c) they can also combine these different kinds of complexity [e.g., "gastrointestinal and abdominal pains (excl. oral and throat)" or "ocular structural change, deposit, and degeneration NEC"]. These terms are usually terms of level HLT, HLGT, and SOC in the MedDRA hierarchy. However, their definitions have a great added value because some terms they subsume may inherit their properties. Indeed, defining one high level term with a morphology property may amount to defining all child terms with this property within the limits of what we have indicated in the section *Problems With Mapping Other Layers Than the PT Level*.

The complex MedDRA terms present two kinds of difficulties: i) the difficulty of mapping with SNOMED CT that tends to favor simple concepts probably because most complex concepts correspond to pure classifying artifacts, e.g., "not elsewhere classified," without real counterpart in the phenomena that are part of medicine; ii) difficulties for formalization of meaning: representing in OWL (Web Ontology Language) the meaning of a compound concept containing exclusions with logical operators is constrained by the expressiveness of the DL language used. In OntoADR, it is not possible to describe the exact same MedDRA semantics due to computability constraints. To date, we have developed a technical solution for the first point, but no satisfactory solution of conceptualization (especially in terms of human cost modeling) has yet been developed to meet the second point. It should be noted that this issue is regarding mainly terms of high levels and does not affect the progress of definitions for PT terms in OntoADR.

In order to map complex MedDRA terms, we developed an algorithm for syntactic decomposition. It consists of three routines: 1) a routine for "cleaning" terms; 2) a routine for identification of complex expressions; and 3) a routine for decomposition of an expression from a set of formal rules. Routine 1 begins by suppressing from the MedDRA labels unnecessary characters or characters that cannot be supported by the decomposition routine [stop words, content between brackets, terms as "unspecified," "NOS" (not otherwise specified), etc.]. Routine 2 identifies decomposable expressions: it searches for keywords that indicate a probable composition of the expression ("AND," "OR," "WITH," ";", etc.). Finally, routine 3 decomposes the complex expression in a set of simpler expressions (cf. **Table 2**), by applying different rules, for example:

$$\text{(A AND B).q} \rightarrow \text{A.q} + \text{B.q}$$
$$\text{q.(A AND B)} \rightarrow \text{q.A} + \text{q.B}$$

We then used the MetaMap software (Aronson, 2001) to map all new decomposed concepts to existing SNOMED CT concepts.

TABLE 2 | Examples of parsing of complex MedDRA terms using different rules.

Input : MedDRA complex term	Selected Rule	Output : result of the decomposition
"Ear and labyrinth disorders"	(A AND B).q	"ear disorders" "labyrinth disorders"
"Manic and bipolar mood disorders and disturbances"	((A AND B).q).(C AND D)	"manic mood disorders" "bipolar mood disorders" "manic mood disturbances" "bipolar mood disturbances"
"Blood and lymphatic system disorders"	(A AND B).q	"blood system disorders" "lymphatic system disorders"

Automatic Lexical Enrichment Methods

We have used a rule-based algorithm for automatic suggestion of properties from the MedDRA label to enrich the formal definition of concepts. Two key procedures have been implemented in the algorithm:

1. When the algorithm detects a given string S_x in a MedDRA label, it automatically adds a corresponding property P_x in the OWL concept definition. For example: if the string "pain" or "algia" is found in a MedDRA concept's label, the semantic property hasDefinitionalManifestation some Pain is automatically added to the concept's definition. Similarly, if the string "perforation" is found, the formal definition hasAssociatedMorphology some Perforation is suggested. All created properties are then validated by an expert. Illegitimate properties are rejected. For example, the algorithm proposed to add the formal property hasAssociatedMorphology some Hernia to the MedDRA concept "hernia repair," as the string "hernia" was found in the label. Semantically, this assignment is obviously illegitimate: the "hernia repair" is not a type of hernia and cannot be defined by this morphological property. The property has therefore been rejected. The expert, however, took advantage of this suggestion to correct it in: OccursAfter some Hernia. This validation step is also necessary due to the occasionally polysemic expressions used for automatic generation of properties. For example, the automatic generation of the hasClinicalCourse some Cyclic property, when the algorithm detects the "cyclic" string in a MedDRA label is valid for concepts such as "cyclic neutropenia" or "cyclic vomiting syndrome," where the term "cyclic" indicates the clinical course of the disease. However, it is not valid for the concept "cyclic AMP," which refers to a clinical test (a measure of the presence or amount of cyclic adenosine monophosphate, e.g., in urine). We could have improved the automatic processing in order to detect these problematic cases, but the formalization of these exceptions would have taken longer time than using a manual approval process.

A restriction is applied to prevent the property P_x to be duplicated when it already exists in a MedDRA term definition, e.g., the detection of the "perforation" string in the label of a MedDRA concept C_{Med} only results in the creation of the property hasAssociatedMorphology some Perforation if C_{Med} does not already own a property hasAssociatedMorphology some <Morphology>. If it is the case, we assume that the relation R_x (in this case hasAssociatedMorphology) has already been filled in correctly.

2. A second procedure is implemented by the algorithm to automatically generate properties. Based on the same principles, but working with more complex patterns of recognition, it was designed to complete definitions of MedDRA concepts referring to investigations and their results (SOC « Investigations »).

Two relationships are available in SNOMED CT to define the examination results (whether clinical observations or investigations): interprets, which refers to "the entity being evaluated or interpreted, when an evaluation, interpretation, or "judgment" is intrinsic to the meaning of a concept"; and hasInterpretation, which, grouped with the attribute Interprets, "designates the judgment aspect being evaluated or interpreted for a concept (e.g., presence, absence, degree, normality, abnormality, etc.)" (Rector and Brandt, 2008). It is important that these two relationships are filled in OntoADR in order to apply semantic reasoning not only to ADR concepts as such, but also to concepts referring to abnormal results of investigations that are the consequence of an ADR (for instance, "neutrophil count decreased" for the neutropenia condition), as such results are frequently used to describe ADRs in pharmacovigilance databases. However, it turned out that very few MedDRA concepts located in the investigations branch could be identified through the procedures described in the previous sections, in particular the mapping from UMLS. A large majority of MedDRA concepts in SOC investigations thus remained undefined in OntoADR.

To remedy this situation, we have integrated into the algorithm a module supporting the properties INTERPRETS and HASINTERPRETATION for MedDRA concepts from SOC "investigations." Results of investigations are usually expressed in MedDRA using the following adjectives: abnormal, normal, absent, present, increased, decreased, positive, and negative. All these qualifiers are also used in SNOMED CT to fill the property HASINTERPRETATION. The procedure followed by the algorithm was therefore as follows:

When the string S_x corresponding to one of these adjectives is detected in the label < LAB1> of a MedDRA concept C_{Med1} from SOC "Investigations":

1. Create in the definition of C_{Med1} the property HASINTERPRETATION SOME S_x.
2. Find if it exists in the investigations branch a concept C_{Med2}, whose label < LAB2> corresponds to (< LAB1> minus S_x). If C_{Med2} exists, create in the definition of C_{Med1} the property INTERPRETS SOME C_{MED2}. This second phase of the procedure is used to connect *via* the property INTERPRETS the results to the related investigations.

In the example of the concept C_{Med1} « Alpha hydroxybutyrate dehydrogenase decreased », this procedure gives the following results:

1. Creation of the property HASINTERPRETATION SOME DECREASED in the definition of C_{Med1}
2. There is a concept C_{Med2}: "alpha hydroxybutyrate dehydrogenase". The property INTERPRETS SOME 'ALPHA HYDROXYBUTYRATE DEHYDROGENASE' is thus created in the definition of C_{Med1}.

Once again, all of the created properties were reviewed and validated by an expert.

Manual Definition

Besides these semiautomatic methods for defining MedDRA concepts in OntoADR, we also performed the manual definition of about 1,935 concepts (Souvignet et al., 2016b). We had insufficient human resources to carry out the manual definition of all MedDRA terms that previous methods had failed to define. So, we decided to focus on high value-added terms for pharmacovigilance. In the EU-ADR project, Trifirò et al. (2009) developed a ranked list of 23 first importance adverse drug events (e.g., cardiac valve fibrosis) based on a review of scientific literature, medical textbooks, and websites of regulatory agencies. To identify which MedDRA terms are related to those 23 topics, pharmacovigilance experts familiar with MedDRA have chosen for each topic an SMQ and/or MedDRA hierarchy-based grouping (HLT or HLGT) or a custom set of preferred terms (PT) fitting the definition of the targeted topics (see Declerck et al., 2012 for details). When no existing MedDRA groupings could be identified to fit the safety topic, *ad hoc* manual groupings of MedDRA PT were proposed by the experts. This work benefited from using a dedicated tool we implemented, Ci4SeR (curation interface for semantic resources) (Souvignet et al., 2014).

RESULTS

Using Other Medical Dictionary for Regulatory Activities-to-SNOMED Clinical Terms Mapping Resources

"Once the Nadkarni and Darer's mapping propositions were validated, modified or completed, we applied the same procedure as described in the section *Using MedDRA-to-SNOMED CT Mappings From UMLS Metathesaurus* to pick up information from SNOMED CT and define the MedDRA concepts of OntoADR. Using the set of SNOMED CT relations available in OntoADR, we also realized manually the definition of those MedDRA terms (53 in total) for which no mapping could be found by Nadkarni and Darer. The use, after verification and eventually correction and complementation, of mappings proposed by Nadkarni and Darer, allowed us to complete the definition of 786 supplementary MedDRA PTs in OntoADR.

Using a Syntactic Decomposition Algorithm on Complex Medical Dictionary for Regulatory Activities Terms

Among the 2,070 HLT, HLGT, and SOC in MedDRA 13.0, a total of 1,011 terms was decomposed by the algorithm generating an average of 2.7 terms by decomposition. The consistency of automatic decomposition was checked by an expert. The errors were corrected through a progressive adjustment of the decomposition algorithm. Only the decomposition of 30 complex terms that were not supported by the algorithm was done manually. Once the decomposition was performed, we used the UMLS MetaMap 2010 AB mapping software, which returns from a given string (in our case, a part of the decomposition), the UMLS concept unique identifier of the nearest syntactically SNOMED CT concepts (fuzzy match). With this method, a total of 638 MedDRA concepts (9 SOCs, 131 HLGTs, and 498 HLTs) could be mapped to the SNOMED CT concepts (mappings one-to-one or one-to-n).

This additional mapping method has the advantage of enabling the definition of high level terms in MedDRA. These definitions may then be inherited by subsumed low level terms. However, the definitions have also the disadvantage of being broad and thus potentially insufficiently precise for specific preferred terms.

Automatic Lexical Enrichment Methods

This procedure was applied to 11 of the 25 SNOMED CT properties used in OntoADR, using 82 different matching strings. In total, this procedure has led to the creation of 8,194 properties, among which 7,691 were validated (i.e., 93.9%). A sample of the strings detected by the algorithm and properties created is shown in **Table 3**.

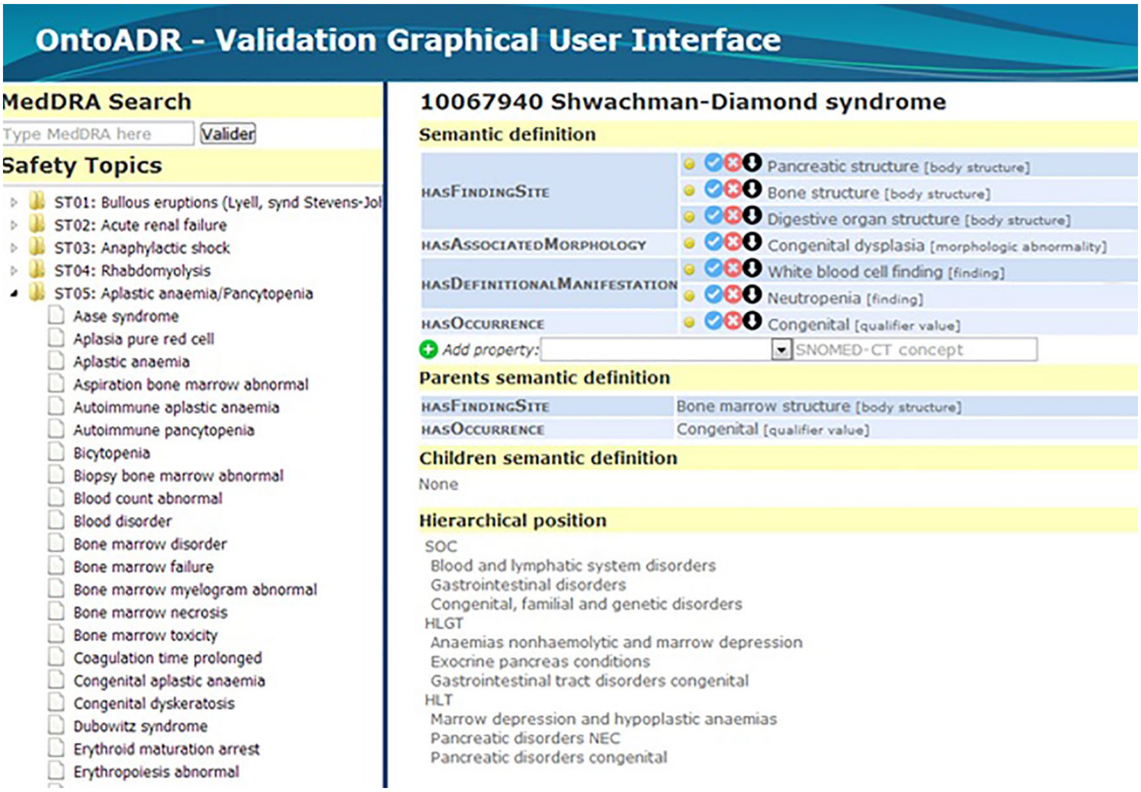
Manual Definition of Concepts

Figure 7 depicts as an example the formal definition associated to the term "Shwachman-Diamond syndrome," as it was described in OntoADR after application of the different algorithms that precede manual refinement.

The curation, which took approximately 750 h, allowed refining the definition of 1,935 MedDRA terms to validate and fully define these terms (Souvignet et al., 2016b). Among the 3,482 properties available in OntoADR for these terms, the curator validated 2,636 properties (76%), proposed 350 (10%) more precise terms (i.e., narrower terms in the SNOMED CT hierarchy), and removed 496 properties (14%). The curator also proposed 13,675 additional properties, but these should not be considered as errors related to missing properties but rather

TABLE 3 | Sample of the properties created automatically from the MedDRA label to enrich the formal definitions of MedDRA concepts in OntoADR.

Relation	Matching strings	Value of the property	Nb properties created	% properties validated
hasClinicalCourse	acute	Sudden onset and/or short duration	84	95.1%
	cyclic	Cyclic	10	20%
	recurrent	Recurrent	121	100%
hasCausativeAgent	bacteria	Bacteria	83	100%
	viral	Virus	157	88.5%
hasAssociatedMorphology	abscess	Abscess morphology	121	100%
	hernia	Hernia	70	82.9%
	haemorrhage, haemorrhagic, bleeding	Hemorrhage	272	86.4%
hasPathologicalProcess	infection, infections, infectious, infective	Infectious process	726	89.4%
	parasitic	Parasitic process	12	91.7%
Interprets	motor, movement, kines	Motor function behaviour	66	59.1%
DueTo	allergic	Hypersensitive reaction	32	93.8%



OntoADR - Validation Graphical User Interface

MedDRA Search

Type MedDRA here [Valider]

Safety Topics

- ST01: Bullous eruptions (Lyell, synd Stevens-Jol
- ST02: Acute renal failure
- ST03: Anaphylactic shock
- ST04: Rhabdomyolysis
- ST05: Aplastic anaemia/Pancytopenia
  - Aase syndrome
  - Aplasia pure red cell
  - Aplastic anaemia
  - Aspiration bone marrow abnormal
  - Autoimmune aplastic anaemia
  - Autoimmune pancytopenia
  - Bicytopenia
  - Biopsy bone marrow abnormal
  - Blood count abnormal
  - Blood disorder
  - Bone marrow disorder
  - Bone marrow failure
  - Bone marrow myelogram abnormal
  - Bone marrow necrosis
  - Bone marrow toxicity
  - Coagulation time prolonged
  - Congenital aplastic anaemia
  - Congenital dyskeratosis
  - Dubowitz syndrome
  - Erythroid maturation arrest
  - Erythropoiesis abnormal

10067940 Shwachman-Diamond syndrome

Semantic definition

hasFindingSite	Pancreatic structure [body structure] Bone structure [body structure] Digestive organ structure [body structure]
hasAssociatedMorphology	Congenital dysplasia [morphologic abnormality]
hasDefinitionalManifestation	White blood cell finding [finding] Neutropenia [finding]
hasOccurrence	Congenital [qualifier value]

Add property: SNOMED-CT concept

Parents semantic definition

hasFindingSite	Bone marrow structure [body structure]
hasOccurrence	Congenital [qualifier value]

Children semantic definition

None

Hierarchical position

SOC
Blood and lymphatic system disorders
Gastrointestinal disorders
Congenital, familial and genetic disorders
HLGT
Anaemias nonhaemolytic and marrow depression
Exocrine pancreas conditions
Gastrointestinal tract disorders congenital
HLT
Marrow depression and hypoplastic anaemias
Pancreatic disorders NEC
Pancreatic disorders congenital

FIGURE 7 | Formal definition associated to the preferred term "Shwachman-Diamond syndrome" before manual refinement.

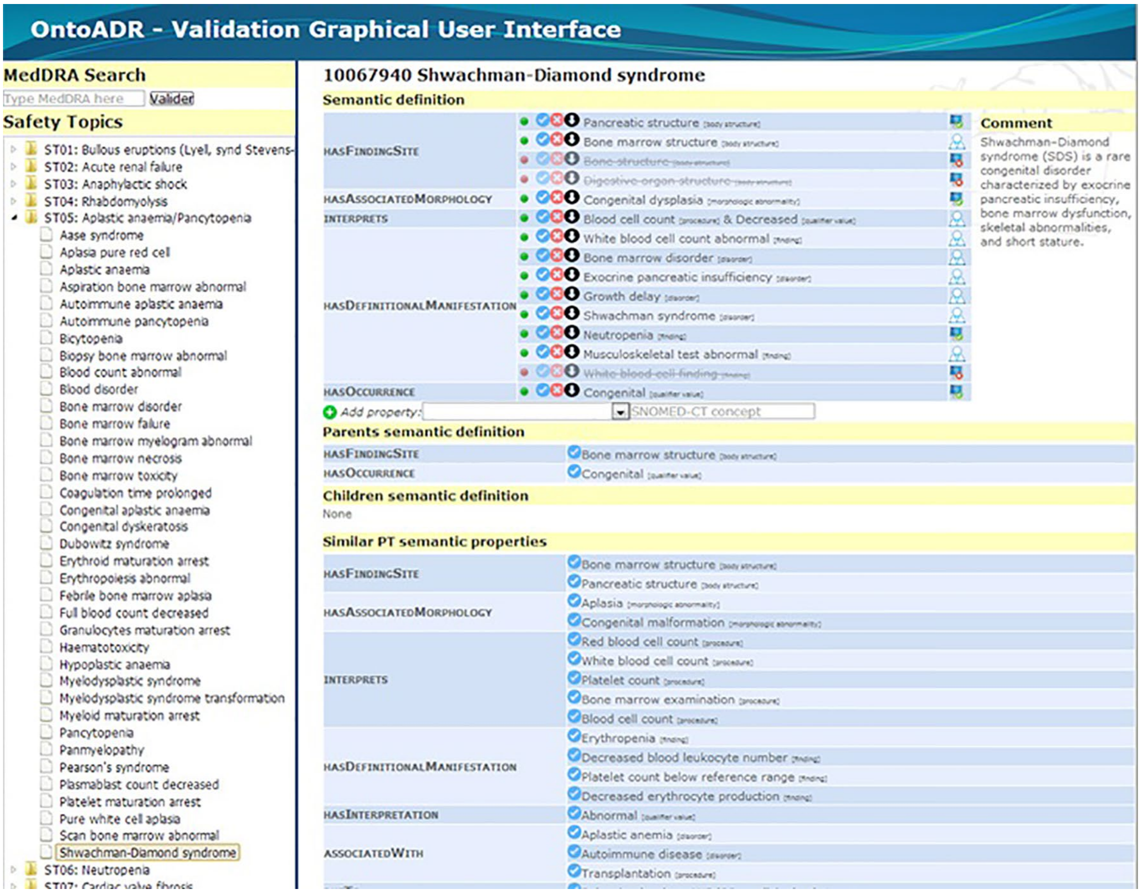
as the curator's desire to better document diagnoses with signs and symptoms and investigations that may be associated to a given disease but are not specific, as they may be absent in some occurrences of this disease.

Figure 8 shows how the "Shwachman-Diamond syndrome" PT's formal definition was modified by the curator in the Ci4SeR tool. The lowest part of the screenshot contains the properties that were automatically proposed considering the parent's and siblings' formal definitions. **Table 4** depicts the results using each method.

DISCUSSION

Summary

We have described in this article several methods that allow collectively a better semantic enrichment of MedDRA. **Table 4** shows that using UMLS metathesaurus is the method that was the most efficient considering the number of mappings and helped to add formal definitions for about half MedDRA terms. As other mapping resources than UMLS are rare and concern



OntoADR - Validation Graphical User Interface

MedDRA Search

Type MedDRA here [Valider]

Safety Topics

- ST01: Bullous eruptions (Lyell, synd Stevens-
- ST02: Acute renal failure
- ST03: Anaphylactic shock
- ST04: Rhabdomyolysis
- ST05: Aplastic anaemia/Pancytopenia
  - Aase syndrome
  - Aplasia pure red cell
  - Aplastic anaemia
  - Aspiration bone marrow abnormal
  - Autoimmune aplastic anaemia
  - Autoimmune pancytopenia
  - Bicytopenia
  - Biopsy bone marrow abnormal
  - Blood count abnormal
  - Blood disorder
  - Bone marrow disorder
  - Bone marrow failure
  - Bone marrow myelogram abnormal
  - Bone marrow necrosis
  - Bone marrow toxicity
  - Coagulation time prolonged
  - Congenital aplastic anaemia
  - Congenital dyskeratosis
  - Dubowitz syndrome
  - Erythroid maturation arrest
  - Erythropoiesis abnormal
  - Febrile bone marrow aplasia
  - Full blood count decreased
  - Granulocytes maturation arrest
  - Haematotoxicity
  - Hypoplastic anaemia
  - Myelodysplastic syndrome
  - Myelodysplastic syndrome transformation
  - Myeloid maturation arrest
  - Pancytopenia
  - Panmyelopathy
  - Pearson's syndrome
  - Plasmablast count decreased
  - Platelet maturation arrest
  - Pure white cell aplasia
  - Scan bone marrow abnormal
  - Shwachman-Diamond syndrome
- ST06: Neutropenia
- ST07: Cardiac valve fibrosis

10067940 Shwachman-Diamond syndrome

Semantic definition

HASFINDINGSITE	Pancreatic structure (body structure) Bone marrow structure (body structure) ~~Bone structure (body structure)~~ ~~Digestive organ structure (body structure)~~
HASASSOCIATEDMORPHOLOGY	Congenital dysplasia (morphologic abnormality)
INTERPRETS	Blood cell count (procedure) & Decreased (qualifier value)
HASDEFINITIONALMANIFESTATION	White blood cell count abnormal (finding) Bone marrow disorder (disorder) Exocrine pancreatic insufficiency (disorder) Growth delay (disorder) Shwachman syndrome (disorder) Neutropenia (finding) Musculoskeletal test abnormal (finding) ~~White blood cell finding (finding)~~
HASOCCURRENCE	Congenital (qualifier value)

Add property: SNOMED-CT concept

Comment

Shwachman-Diamond syndrome (SDS) is a rare congenital disorder characterized by exocrine pancreatic insufficiency, bone marrow dysfunction, skeletal abnormalities, and short stature.

Parents semantic definition

HASFINDINGSITE	Bone marrow structure (body structure)
HASOCCURRENCE	Congenital (qualifier value)

Children semantic definition

None

Similar PT semantic properties

HASFINDINGSITE	Bone marrow structure (body structure) Pancreatic structure (body structure)
HASASSOCIATEDMORPHOLOGY	Aplasia (morphologic abnormality) Congenital malformation (morphologic abnormality)
INTERPRETS	Red blood cell count (procedure) White blood cell count (procedure) Platelet count (procedure) Bone marrow examination (procedure) Blood cell count (procedure)
HASDEFINITIONALMANIFESTATION	Erythropenia (finding) Decreased blood leukocyte number (finding) Platelet count below reference range (finding) Decreased erythrocyte production (finding)
HASINTERPRETATION	Abnormal (qualifier value)
ASSOCIATEDWITH	Aplastic anemia (disorder) Autoimmune disease (disorder) Transplantation (procedure)

FIGURE 8 | Formal definition associated to the term "Shwachman-Diamond syndrome" after manual refinement.

TABLE 4 | Synthesis of mappings and properties found using all previously described methods.

Source/Method	MedDRA version	Comparator	Number of mappings	Number of properties
Using UMLS Metathesaurus mappings	v17	20,599 PT	11,281 PT (54.8%)	74,598
Using other mapping resources	v13	18,786 PT	455[a] PT (2.4%)	469[a]
Using a decomposition algorithm and Metamap software to map complex MedDRA terms	v13	2,070 HLT, HLGT, and SOC	638[a,b] HLT, HLGT, and SOC (30.8%)	–
Automatic enrichment methods	v17	–	–	7,691[a]
Manual definition of concepts	v17	20,599 PT	1,935 PT (9.1%)	13,675

[a] *Represent only the number of mapping/properties that was not found by other methods.*

[b] *Limited to SOC, HLGT, and HLT.*

TABLE 5 | Summary of inconveniences and advantages of the different methods.

Algorithms	What is already available?	Characteristics of the algorithm
Using MedDRA-to-SNOMED CT mappings from UMLS Metathesaurus	MedDRA-to-SNOMED CT mappings in UMLS Metathesaurus	Available mappings are used to retrieve SNOMED CT concepts associated to a MedDRA term. Properties are added to the formal definition according to the SNOMED CT concept position in the hierarchy. It illustrates Scenario 6. “Reusing, Merging and Re-engineering Ontological Resources” of the NeON methodology. The algorithm is automatic, except when several SNOMED CT concepts are available which requires expert selection.
Using other MedDRA-to-SNOMED CT mapping resources	Nadkarni and Darer’s propositions of mappings	This is an expert-based process entirely manual of validation and refinement that illustrates Scenario 2. “Reusing and Re-engineering Non-Ontological Resources” of the NeON methodology but benefits from a non-ontological resource that expedites formal definition of MedDRA terms compared with manual definitions.
Using a syntactic decomposition algorithm on complex MedDRA terms	–	This algorithm is automated and developed ad hoc. It illustrates one of the ontology support activities, “knowledge acquisition,” and exploits hidden semantics as proposed by Third (2012). It is limited to complex MedDRA terms.
Automatic lexical enrichment methods	–	This algorithm is based on substring search. It necessitates defining beforehand substrings that may be associated to a SNOMED CT concept. Review of the algorithm’s proposal is mandatory in order to check that substrings allow associating with relevant SNOMED CT concepts. It also illustrates Ontology support activity “knowledge acquisition” and exploits hidden semantics.
Manual definition	–	This process is manual and expert based, but the Ci4SeR tool suggests definitions on the basis of definitions already available for siblings and parents of a MedDRA PT. Such approach addresses both “knowledge acquisition” and “ontology validation” within the Ontology support activities.

only few MedDRA terms, the Nadkarni and Darer’s resource allowed to add properties to 4.2% of MedDRA terms but only to 2.4% of MedDRA terms that were not associated with mappings to SNOMED CT in UMLS.

Our proposal to decompose complex MedDRA terms was applied only to SOC, HLGT, and HLT levels and accounted for 30.8% of these MedDRA terms above the PT level. Manual definitions and refinements of definitions obtained with other methods allowed to process 9.1% of MedDRA terms, which is more than the proportion of terms that were defined using Nadkarni and Darer’s mapping resource. However, it was associated with high time-consuming effort by the domain expert that confirms previous work, e.g., Giannangelo and Millar (2012) who observed that “map specialists on average mapped 6.5 SNOMED CT concepts an hour.” **Table 5** summarizes the main characteristics of each method and indicates if the proposed method reuses existing knowledge, if it requires manual adaptations or may be performed in an automated way.

Related Work in Medical Informatics

He et al. (2014) have introduced the Ontology of Adverse Events (OAE). OAE was originally targeted for vaccine adverse events (Marcos et al., 2013) and now also includes adverse drug events. In practice, using OAE to select case reports in the Vaccine Adverse Event Reporting System proved difficult: “AE data stored in Vaccine Adverse Event Reporting System are annotated using MedDRA” (Marcos et al., 2013). Authors complained that “many disadvantages of MedDRA, including the lack of term definitions and a well-defined hierarchical and logical structure, prevent its effective usage in VAE (vaccine adverse event) term classification.” Therefore, for an efficient analysis, they performed a mapping between MedDRA and OAE (Sarntivijai et al., 2012).

OAE contains about 2,300 AE entities but only 1,900 MedDRA mappings (9% of all MedDRA PT). For example, there is a single term for upper gastrointestinal hemorrhage in OAE (He et al., 2014), whereas one can cite several in MedDRA (see the section *Rationale for Supplementing MedDRA With Formal Definitions* where we identified 27 using OntoADR). Furthermore, OAE formal definitions are limited to anatomical and physiopathological descriptions. He and colleagues proposed extensions to OAE such as the Ontology of Drug Neuropathy Adverse Events (Guo et al., 2016), which suggests that providing supplementary MedDRA mappings is possible using the same methodology. One advantage of OAE is the possibility to use it in open access, which allows wide dissemination to users, while legal issues related to ownership of MedDRA and SNOMED CT should be solved before we can make OntoADR available.

Adverse Events Reporting Ontology aims to allow storing of pharmacovigilance data related to anaphylaxis according to guidelines defined by the Brighton collaboration (Courtot et al., 2014) but may also be extended to other safety topics, e.g., malaria (Courtot et al., 2013). Nevertheless, ADRs are not formally defined in Adverse Events Reporting Ontology.

While we did not find any resource available providing definitions for every ADR in MedDRA, there are more general resources with formal representation of clinical terms. In order not to start from scratch the definitions of ADRs, we needed a trustworthy formal resource, standardized and reliable. We chose SNOMED CT for three main reasons: first, pharmacovigilance concepts generally do not differ from those used in other medical fields. Second, SNOMED CT is the most complete and most detailed terminology of medicine with a formal semantic foundation currently available (Elkin et al., 2006) sharing common fields with MedDRA (medical pathologies in all medical specialties, signs and symptoms, laboratory tests results, some diagnostic and therapeutic procedures). Finally, SNOMED CT has the advantage of covering to a large extent, if not entirely, other standard medical terminologies such as International Classification of Diseases, 10th edition (ICD-10), and especially more than 50% of MedDRA terms (excluding LLT) are associated with a SNOMED CT concept

(Bodenreider, 2009) in UMLS, a degree of coverage that, to our knowledge, no other current medical ontology was able to match.

We found in the literature several examples of mappings from a terminology to SNOMED CT (Vikström et al., 2007; Merabti et al., 2009; Nyström et al., 2010; Dhombres and Bodenreider, 2016; Fung et al., 2017). However, the objective was usually to integrate a terminology in SNOMED CT or to map this terminology to SNOMED CT but not to enrich this terminology by the means of formal definitions. The lexically assign logically refine method is an example of an automated method in which logical observation identifiers names and codes (LOINC) and SNOMED terms are first decomposed, then refined by the means of knowledge-based methods that allowed to map LOINC and SNOMED together (Dolin et al., 1998). In another work, Adamusiak and Adamusiak and Bodenreider (2012) developed an OWL version of both LOINC and SNOMED CT and made use of mappings between SNOMED CT terms to identify redundancy and inconsistencies in LOINC multi-axial hierarchy. Roldán-García et al. (2016) implemented Dione, an OWL representation of ICD-10-CM where formal definitions were obtained thanks to mappings between ICD-10-CM and SNOMED CT available in UMLS and the Bioportal. More recently, Nikiema et al. (2017) benefited from SNOMED CT logical definitions to find mappings between ICD-10 and ICD-O3 concepts in the domain of cancer diagnosis terminologies.

It is usually recommended to build medical terminologies following the model of clinical terminologies that obey to Cimino's desiderata (Cimino, 1998; Bales et al., 2006). Such model brings several advantages such as improving the maintenance of large terminologies (Cimino et al., 1994), and formal definitions were implemented in several terminologies such as the NCI-Thesaurus (Hartel et al., 2005). Our approach is more in line with what is recommended by Ingenerf and Giere (1998), that is to say, to keep terminologies with disjoint classes required for statistics (in a clinical terminology, the same term may be present in several separate categories because of multiple inheritance and be counted more than once) and instead implement a mapping of terms of first-generation system to a formal system. This allows keeping the MedDRA terminology in its current format, counting ADRs according to predefined categories that are standardized and replicable at the international level with MedDRA and building new categories on demand by using knowledge engineering methods. This is what we have done in our implementation of OntoADR (Bousquet et al., 2014) in the form of an OWL-DL file and in the form of a database (Souvignet et al., 2016b).

We have no knowledge of other works in which the formalization of complex terms involving AND/OR relations has been performed in an automated way. We have not proposed formal definitions of LLT because this level is reserved for the coding of case reports, in order to improve the accuracy of coding, but it is not useful for grouping data for analysis (which is performed at the PT level). Although the analysis of pharmacovigilance databases is performed preferentially at the PT level, it could be important to also define the upper levels: SOC, HLGT, and HLT. This formalization would bring several advantages: i) preferred terms may inherit properties from their parents that allows to give them a formal definition in case the synonymous SNOMED CT concept has no definition, or there is no SNOMED CT concept mapped to this PT in UMLS; ii) This would allow to calculate by the means of terminological reasoning high level MedDRA categories in which PTs should be included and therefore restore multiple inheritance that does not exist in MedDRA. However, it is advisable to remain modest insofar as the relations between a PT and the higher hierarchical levels to which it is attached are not always of a taxonomic nature.

Perspectives

Our perspectives are to add formal definitions to a larger number of MedDRA terms. Our approach may be improved using more advanced natural language processing techniques (Iavindrasana et al., 2006; Deléger et al., 2009; Liu et al., 2011; Dupuch et al., 2014) compared with the basic semantic enrichment we performed considering MedDRA labels. We estimate that the methods proposed here can be reused for other first-generation terminologies provided that these terminologies have a mapping with SNOMED CT with fair coverage and that this mapping is available in accessible sources of knowledge such as the UMLS. The terminology can also be treated using methods of natural language processing as was done for example with LOINC in the lexically assign logically refine method (Dolin et al., 1998). One can also consider cases in which the terminology would be normally defined by mapping to another clinical terminology than SNOMED CT. This may be the case in other areas of application in which SNOMED CT is not the best choice.

As the manual approach was time consuming and necessitates human resources we do not have, we plan to rely on the development of complementary automated approaches. First, formal definitions could be extracted from textual definitions (Petrova et al., 2015) or directly using morphosemantic analysis on the term label, e.g., blepharitis where "itis" stands for "inflammation," and "blephar" stands for "eyelid." Such approach is limited to terms containing "compound forms" that have a medical meaning (Deléger et al., 2009). Second, formal definitions could be based on ontology design patterns, such as implemented in tools like Ontorat (Xiang et al., 2015) or TermGenie (Dietze et al., 2014), which partially automate the process, as they still rely on expert curation. Third, additional mappings between MedDRA and other terminologies could be obtained *via* improved mappings in the UMLS metathesaurus (Bodenreider et al., 1998; Fung et al., 2007; Diallo, 2014). Fourth, semantic definitions may be audited by comparing definitions associated to terms that present lexical similarities (Agrawal and Elhanan, 2014). However, this presents an intrinsic limit: terms to compare should consist of at least three words that constraints this method mainly to MedDRA procedures.

Fifth, we plan to extract knowledge using additional sources than SNOMED CT such as NCI Thesaurus (Sioutos et al., 2007) that could be useful to build definitions for MedDRA terms that describe cancer-related adverse reactions. A recent work by Oliveira and Pesquita, (2018) reports that current ontology matching techniques and systems are mostly devoted to finding links between two equivalent entities from two distinct ontologies. However, different domains may be involved that requires the implementation of matching techniques that allow linking more than two ontologies through more complex relations. An example is "aortic valve stenosis" (from human phenotype ontology) that is equivalent to the combination of "aortic valve" (from the Foundational Model of Anatomy) and "constricted" (from Phenotype And Trait Ontology).

CONCLUSION

The possibility of selecting terms using formal definitions and terminological reasoning are major advantages of clinical terminologies with formal semantics such as SNOMED CT, which present several advantages compared with classic terminologies. MedDRA, as a standard international terminology for the coding of ADRs in pharmacovigilance databases, could beneficiate from these knowledge engineering techniques, but MedDRA terms have to be defined using formal languages first. As defining manually MedDRA terms takes much time, it is important to reuse as much as possible ontological and non-ontological resources available to expedite the generation of formal definitions. The collection of methods we present can collectively support a semiautomatic semantic enrichment of MedDRA. Perspectives are to implement more efficient techniques to find more logical relations between SNOMED CT and MedDRA in an automated way.

AUTHOR CONTRIBUTIONS

GD adapted Nadkarni and Darer's definitions to OntoADR, performed automatic lexical enrichment methods, and wrote the first draft of the manuscript. M-CJ provided significant advice on the design of the study and contributed to the evaluation. ES first, then JS, performed mappings between MedDRA and SNOMED CT using the UMLS metathesaurus and developed new versions of OntoADR. JS performed the syntactic decomposition algorithm on complex MedDRA terms and wrote the corresponding section of the manuscript. CB conducted the study, contributed to the evaluation, reviewed state-of-the-art related work, and wrote the final article. CB and M-CJ were responsible for submitting the PROTECT project to the IMI requests for proposal and for submitting the PEGASE project to the ANR request for proposal. All authors have made substantial contributions and approved the final manuscript and agreed to be accountable for all aspects of the work.

FUNDING

This work was funded by the grant N° ANR-16-CE23-0011-01 from the ANR, the French Agence nationale de la Recherche through the PEGASE (*Pharmacovigilance enrichie par des Groupements Améliorant la detection des Signaux Emergents*). Some results described in this article were obtained as part of the PROTECT consortium (Pharmacoepidemiological Research on Outcomes of Therapeutics by a European Consortium, www.imi-protect.eu) which is a public–private partnership coordinated by the European Medicines Agency. The PROTECT project has received support from the Innovative Medicine Initiative Joint Undertaking (www.imi.europa.eu) under grant agreement no. 115004, resources of which are composed of financial contribution from the European Union's Seventh Framework Program (FP7/2007-2013) and EFPIA companies' in kind contribution.

ACKNOWLEDGMENTS

The authors express their gratitude to Adrien Fanet for his technical contribution to the conception of OntoADR, and to Anne Jamet and Hadyl Asfari for their medical expertise.

REFERENCES

Adamusiak, T., and Bodenreider, O. (2012). Quality assurance in LOINC using Description Logic. *AMIA Annu. Symp. Proc.* 2012, 1099–1108.

Agrawal, A., and Elhanan, G. (2014). Contrasting lexical similarity and formal definitions in SNOMED CT: consistency and implications. *J. Biomed. Inform.* 47, 192–198. doi: 10.1016/j.jbi.2013.11.003

Alani, H. (2006). Ontology construction from online ontologies. In: 15th International World Wide Web Conference, 23-26 May 2006, Edinburgh, Scotland. 491–495. doi: 10.1145/1135777.1135849

Alecu, I., Bousquet, C., Mougin, F., and Jaulent, M. C. (2006). Mapping of the WHO-ART terminology on Snomed CT to improve grouping of related adverse drug reactions. *Stud. Health Technol. Inform.* 124, 833–838. doi: 10.3233/978-1-58603-647-8-833

Alecu, I., Bousquet, C., Degoulet, P., and Jaulent, M. C. (2007). PharmARTS: terminology web services for drug safety data coding and retrieval. *Stud. Health Technol. Inform.* 129 (Pt 1), 699–704. doi: 10.3233/978-1-58603-774-1-699

Alecu, I., Bousquet, C., and Jaulent, M. C. (2008). A case report: using SNOMED CT for grouping Adverse Drug Reactions Terms. *BMC Med. Inform. Decis. Mak.* 8Suppl 1, S4. doi: 10.1186/1472-6947-8-S1-S4

Alobaidi, M., Malik, K. M., and Sabra, S. (2018). Linked open data-based framework for automatic biomedical ontology generation. *BMC Bioinformatics* 19 (1), 319. doi: 10.1186/s12859-018-2339-3

Aranguren, M. E., Antezana, E., Kuiper, M., and Stevens, R. (2008). Ontology Design Patterns for bio-ontologies: a case study on the Cell Cycle Ontology. *BMC Bioinformatics*. 9 Suppl 5:S1. doi: 10.1186/1471-2105-9-S5-S1

Aronson, A. R. (2001). "Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program," in *Proceedings of the AMIA Symposium* (American Medical Informatics Association), 17.

Asim, M. N., Wasim, M., Khan, M. U. G., Mahmood, W., and Abbasi, H. M. (2018). A survey of ontology learning techniques and applications. Database, 2018, bay101. doi: 10.1093/database/bay101

Asfari, H., Souvignet, J., Lillo-Le Louët, A., Trombert, B., Jaulent, M. C., and Bousquet, C. (2016). Automated grouping of terms associated to cardiac valve fibrosis in MedDRA. *Therapie.* 71 (6), 541–552. doi: 10.1016/j.therap.2016.06.003

Astrakhantsev, N. A., and Turdakov, D. Y. (2013). Automatic construction and enrichment of informal ontologies: A survey. *Program Com. Software* 39 (1), 34–42. doi: 10.1134/S0361768813010039

Balakrishna, M., Moldovan, D. I., Tatu, M., and Olteanu, M. (2010). Semi-automatic domain ontology creation from text resources In: Proc. 7th International Conference on Language Resources and Evaluation (LREC 10). Valletta, Malta, 19–21.

Bales, M. E., Kukafka, R., Burkhardt, A., and Friedman, C. (2006). Qualitative assessment of the International Classification of Functioning, Disability, and Health with respect to the desiderata for controlled medical vocabularies. *Int. J. Med. Inform.* 75 (5), 384–395. doi: 10.1016/j.ijmedinf.2005.07.026

Bedini, I., and Nguyen, B. (2007). *Automatic ontology generation: State of the art. PRiSM Laboratory Technical Report*. University of Versailles.

Blfgeh, A., Warrender, J., Hilkens, C. M. U., and Lord, P. (2017). A document-centric approach for developing the tolAPC ontology. *J. Biomed. Semantics* 8 (1), 54. doi: 10.1186/s13326-017-0159-4

Blomqvist, E. (2008). Pattern ranking for semi-automatic ontology construction. In Proceedings of the 2008 ACM symposium on Applied computing pp. 2248–2255.

Bobed, C., Mena, E., and Trillo, R., (2012). "FirstOnt: Automatic Construction of Ontologies out of Multiple Ontological Resources," in *Proceedings of the 16th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems* (Spain: KES'12, IOS Press), 1909–1919.

Bodenreider, O., Nelson, S. J., Hole, W. T., and Chang, H. F., (1998). Beyond synonymy: exploiting the UMLS semantics in mapping vocabularies. *Proceedings/AMIA Annual Symposium*. 815–819.

Bodenreider, O. (2009). Using SNOMED CT in combination with MedDRA for reporting signal detection and adverse drug reactions reporting. *AMIA*. *Annu. Symp. Proc.* 14, 45–49.

Bousquet, C., Lagier, G., Lillo-Le Louët, A., Le Beller, C., Venot, A., and Jaulent, M. C. (2005). Appraisal of the MedDRA conceptual structure for describing and grouping adverse drug reactions. *Drug Saf.* 28 (1), 19–34. doi: 10.2165/00002018-200528010-00002

Bousquet, C., Sadou, E., Souvignet, J., Jaulent, M. C., and Declerck, G. (2014). Formalizing MedDRA to support semantic reasoning on adverse drug reaction terms. *J. Biomed. Inform.* 49,282–291 pii: S1532–0464(14)00079-3. doi: 10.1016/j.jbi.2014.03.012

Brown, E. G., Wood, L., and Wood, S. (1999). The medical dictionary for regulatory activities (MedDRA). *Drug Saf.* 20 (2), 109–117. doi: 10.2165/00002018-199920020-00002

Brown, E. G. (2003). Methods and pitfalls in searching drug safety databases utilising the Medical Dictionary for Regulatory Activities (MedDRA). *Drug Saf.* 26 (3), 145–158. doi: 10.2165/00002018-200326030-00002

Buitelaar, P., Cimiano, P., and Magnini, B., (2005). Ontology learning from text: an overview. Ontology learning from text: methods, evaluation and applications 123, 3–12.

Cimino, J. J., Clayton, P. D., Hripcsak, G., and Johnson, S. B. (1994). Knowledge-based approaches to the maintenance of a large controlled medical terminology. *J. Am. Med. Inform. Assoc.* 1 (1), 35–50. doi: 10.1136/jamia.1994.95236135

Cimiano, P., Völker, J., and Studer, R. (2006). Ontologies on demand?-a description of the state-of-the-art, applications, challenges and trends for ontology learning from text. *Information, Wissenschaft und Praxis* 57 (6-7), 315–320.

Cimino, J. J. (1998). Desiderata for controlled medical vocabularies in the twenty-first century. *Methods Inf. Med.* 37 (4-5), 394–403. doi: 10.1055/s-0038-1634558

Costa, R., Lima, C., Sarraipa, J., and Jardim-Gonçalves, R. (2016). Facilitating knowledge sharing and reuse in building and construction domain: an ontology-based approach. *J. Intell. Manuf.* 27 (1), 263–282. doi: 10.1007/s10845-013-0856-5

Courtot, M., Zheng, J., Stoeckert, C. J., Jr., Brinkman, R. R., and Ruttenberg, A. (2013). Diagnostic criteria and clinical guidelines standardization to automate case classification. *In ICBO*, 101–104.

Courtot, M., Brinkman, R. R., and Ruttenberg, A. (2014). The logic of surveillance guidelines: an analysis of vaccine adverse event reports from an ontological perspective. *PLoS One* 9 (3), e92632. doi: 10.1371/journal.pone.0092632

Cullen, J., and Bryman, A. (1988). The knowledge acquisition bottleneck: time for reassessment? *Expert Syst.* 5 (3), 216–225. doi: 10.1111/j.1468-0394.1988.tb00065.x

Dahab, M. Y., Hassan, H. A., and Rafea, A. (2008). TextOntoEx: automatic ontology construction from natural English text. *Expert Syst. Appl.* 34 (2), 1474–1480. doi: 10.1016/j.eswa.2007.01.043

Dasgupta, S., Padia, A., Maheshwari, G., Trivedi, P., and Lehmann, J., (2018). Formal Ontology Learning from English IS-A Sentences. *arXiv preprint arXiv*:1802.03701.

Declerck, G., Bousquet, C., and Jaulent, M. C. (2012). Automatic generation of MedDRA terms groupings using an ontology. *Stud. Health Technol. Inform.* 180, 73–77. doi: 10.3233/978-1-61499-101-4-73

Deléger, L., Namer, F., and Zweigenbaum, P. (2009). Morphosemantic parsing of medical compound words: transferring a French analyzer to English. *Int. J. Med. Inform.* 78 Suppl 1, S48–S55. doi: 10.1016/j.ijmedinf.2008.07.016

Dhombres, F., and Bodenreider, O. (2016). Interoperability between phenotypes in research and healthcare terminologies–Investigating partial mappings between HPO and SNOMED CT. *J. Biomed. Semantics* 7, 3. doi: 10.1186/s13326-016-0047-3

Diallo, G. (2014). An effective method of large scale ontology matching. *J. Biomed. Semantics* 5 (1), 44. doi: 10.1186/2041-1480-5-44

Dietze, H., Berardini, T. Z., Foulger, R. E., Hill, D. P., Lomax, J., Osumi-Sutherland, D., et al. (2014). TermGenie - a web-application for pattern-based ontology class generation. *J. Biomed. Semantics* 5, 48. doi: 10.1186/2041-1480-5-48

Dolin, R. H., Huff, S. M., Rocha, R. A., Spackman, K. A., and Campbell, K. E. (1998). Evaluation of a "lexically assign, logically refine" strategy for semi-automated integration of overlapping terminologies. *J. Am. Med. Inform. Assoc.* 5 (2), 203–213. doi: 10.1136/jamia.1998.0050203

Dupuch, M., Dupuch, L., Hamon, T., and Grabar, N. (2014). Exploitation of semantic methods to cluster pharmacovigilance terms. *J. Biomed. Semantics* 5 (1), 18. doi: 10.1186/2041-1480-5-18

Elkin, P. L., Brown, S. H., Husser, C. S., Bauer, B. A., Wahner-Roedler, D., Rosenbloom, S. T., et al., (2006). "Evaluation of the content coverage of SNOMED CT: ability of SNOMED clinical terms to represent clinical problem lists. ," in *Mayo Clinic Proceedings*, vol. 81. (Elsevier), 741–748. doi: 10.4065/81.6.741

Emani, C. K., Da Silva, C. F., Fiès, B., Ghodous, P., and Bourdeau, M., (2015). "Automated Semantic Enrichment of Ontologies in the Construction Domain," in *Proc. of the 32nd CIB W78 Conference 2015, 27th-29th,* Eindhoven, Netherlands.

Fernández-Breis, J. T., Chiba, H., Legaz-García Mdel, C., and Uchiyama, I. (2016). The Orthology Ontology: development and applications. *J. Biomed. Semantics* 7 (1), 34. doi: 10.1186/s13326-016-0077-x

Fernández-López, M., Gómez-Pérez, A., and Juristo, N., (1997). Methontology: from ontological art towards ontological engineering. Proc. AAAI Spring Symp. Series, AAAI Press, Menlo Park, Calif., 33–40.

Fung, K. W., Hole, W. T., Nelson, S. J., Srinivasan, S., Powell, T., and Roth, L. (2005). Integrating SNOMED CT into the UMLS: an exploration of different views of synonymy and quality of editing. *J. Am. Med. Inform. Assoc.* 12 (4), 486–494. doi: 10.1197/jamia.M1767

Fung, K. W., Bodenreider, O., Aronson, A. R., Hole, W. T., and Srinivasan, S. (2007). Combining lexical and semantic methods of inter-terminology mapping using the UMLS. *Stud. Health Technol. Inform.* 129 (Pt 1), 605–609. doi: 10.3233/978-1-58603-774-1-605

Fung, K. W., Xu, J., Ameye, F., Gutiérrez, A. R., and D'Havé, A. (2017). Leveraging Lexical Matching and Ontological Alignment to Map SNOMED CT Surgical Procedures to ICD-10-PCS. *AMIA Annu. Symp. Proc.* 2016, 570–579. eCollection 2016.

Gangemi, A. (2005). Ontology design patterns for semantic web content. In International semantic web conference pp. 262–276. Springer, Berlin, Heidelberg.

Gavankar, C., Kulkarni, A., Fang Li, Y., and Ramakrishnan. (2012). Enriching an academic knowledge base using linked open data In: Proceedings of Workshop on Speech and Language Processing Tools in Education in 24th International Conference on Computational Linguistics, 51–60.

Giannangelo, K., and Millar, J. (2012). Mapping SNOMED CT to ICD-10. *Stud. Health Technol. Inform.* 180, 83–87. doi: 10.3233/978-1-61499-101-4-83

Guo, A., Racz, R., Hur, J., Lin, Y., Xiang, Z., Zhao, L., et al. (2016). Ontology-based collection, representation and analysis of drug-associated neuropathy adverse events. *J. Biomed. Semantics* May 217, 29. doi: 10.1186/s13326-016-0069-x

Hansen, R. A., Gartlehner, G., Powell, G. E., and Sandler, R. S. (2007). Serious adverse events with infliximab: analysis of spontaneously reported adverse events. *Clin. Gastroenterol. Hepatol.* 5 (6), 729–735. doi: 10.1016/j.cgh.2007.02.016

Hartel, F. W., de Coronado, S., Dionne, R., Fragoso, G., and Golbeck, J. (2005). Modeling a description logic vocabulary for cancer research. *J. Biomed. Inform.* 38 (2), 114–129. doi: 10.1016/j.jbi.2004.09.001

Hauben, M., Patadia, V. K., and Goldsmith, D. (2006). What counts in data mining? *Drug Saf.* 29 (10), 827–832. doi: 10.2165/00002018-200629100-00001

He, Y., Sarntivijai, S., Lin, Y., Xiang, Z., Guo, A., Zhang, S., et al. (2014). OAE: the ontology of adverse events. *J. Biomed. Semantics* 2014, 5:29. doi: 10.1186/2041-1480-5-29

He, Y., Xiang, Z., Zheng, J., Lin, Y., Overton, J. A., and Ong, E. (2018). The eXtensible ontology development (XOD) principles and tool implementation to support ontology interoperability. *J. Biomed. Semantics* 9 (1), 3. doi: 10.1186/s13326-017-0169-2

Henegar, C., Bousquet, C., Lillo-Le Louët, A., Degoulet, P., and Jaulent, M. C. (2006). Building an ontology of adverse drug reactions for automated signal generation in pharmacovigilance. *Comput. Biol. Med.* 36 (7-8), 748–767. doi: 10.1016/j.compbiomed.2005.04.009

Huang, J., Dang, J., Borchert, G. M., Eilbeck, K., Zhang, H., Xiong, M., et al. (2014). OMIT: dynamic, semi-automated ontology development for the microRNA domain. *PLoS One* 9 (7), e100855. doi: 10.1371/journal.pone.0100855

Iavindrasana, J., Bousquet, C., and Jaulent, M. C. (2006). Knowledge acquisition for computation of semantic distance between WHO-ART terms. *Stud. Health Technol. Inform.* 124, 839–844. doi: 10.3233/978-1-58603-647-8-839

ICH Working Group. (2018). MedDRA® data retrieval and presentation: points to consider. Release 4.15 Based on MedDRA Version 21.0 MedDRA® TERM SELECTION: POINTS TO CONSIDER. ICH-Endorsed Guide for MedDRA Users. https://www.meddra.org/sites/default/files/guidance/file/000157_datretptc_r3_15_mar2018.pdf.

Ingenerf, J., and Giere, W. (1998). Concept-oriented standardization and statistics-oriented classification: continuing the classification versus nomenclature controversy. *Methods Inf. Med.* 37 (4-5), 527–539. doi: 10.1055/s-0038-1634544

Jimeno-Yepes, A., Jiménez-Ruiz, E., Berlanga-Llavori, R., and Rebholz-Schuhmann, D. (2009). Reuse of terminological resources for efficient ontological engineering in Life Sciences. *BMC Bioinformatics* 10 (10), S4. doi: 10.1186/1471-2105-10-S10-S4

Judkins, J., Tay-Sontheimer, J., Boyce, R. D., and Brochhausen, M. (2018). Extending the DIDEO ontology to include entities from the natural product drug interaction domain of discourse. *J. Biomed. Semantics* 9 (1), 15. doi: 10.1186/s13326-018-0183-z

Khorrami, F., Ahmadi, M., and Sheikhtaheri, A. (2018). Evaluation of SNOMED CT Content Coverage: a systematic literature review. *Stud. Health Technol. Inform.* 248, 212–219. doi: 10.3233/978-1-61499-858-7-212

Lindberg, D. A., Humphreys, B. L., and McCray, A. T. (1993). The unified medical language system. *Methods Inf. Med.* 32 (4), 281–291. doi: 10.1055/s-0038-1634945

Liu, K., Hogan, W. R., and Crowley, R. S. (2011). Natural Language Processing methods and systems for biomedical ontology learning. *J. Biomed. Inform.* 44 (1), 163–179. doi: 10.1016/j.jbi.2010.07.006

Maedche, A., and Staab, S. (2001). Ontology learning for the semantic web. *IEEE Intell. Syst.* 16 (2), 72–79. doi: 10.1109/5254.920602

Marcos, E., Zhao, B., and He, Y. (2013). The Ontology of Vaccine Adverse Events (OVAE) and its usage in representing and analyzing adverse events associated with US-licensed human vaccines. *J. Biomed. Semantics* 4, 40. doi: 10.1186/2041-1480-4-40

Mazo, C., Salazar, L., Corcho, O., Trujillo, M., and Alegre, E. (2017). A histological ontology of the human cardiovascular system. *J. Biomed. Semantics* 8 (1), 47. doi: 10.1186/s13326-017-0158-5

McKnight, L. K., Elkin, P. L., Ogren, P. V., and Chute, C. G. (1999). Barriers to the clinical implementation of compositionality. *Proc. AMIA Symp.*, 320–324.

Merabti, T., Letord, C., Abdoune, H., Lecroq, T., Joubert, M., and Darmoni, S. J. (2009). Projection and inheritance of SNOMED CT relations between MeSH terms. *Stud. Health Technol. Inform.* 150, 233–237. doi: 10.3233/978-1-60750-044-5-233

Mozzicato, P. (2007). StandardisedMedDRA Queries. *Drug Saf.* 30 (7), 617–619. doi: 10.2165/00002018-200730070-00009

Nadkarni, P. M., and Darer, J. A. (2010). Determining correspondences between high frequency MedDRA concepts and SNOMED: a case study. *BMC Med. Inform. Deci. Mak.* 10. doi: 10.1186/1472-6947-10-66

Nikiema, J. N., Jouhet, V., and Mougin, F. (2017). Integrating cancer diagnosis terminologies based on logical definitions of SNOMED CT concepts. *J. Biomed. Inform.* 74, 46–58. doi: 10.1016/j.jbi.2017.08.013

Nyström, M., Vikström, A., Nilsson, G. H., Ahlfeldt, H., and Orman, H. (2010). Enriching a primary health care version of ICD-10 using SNOMED CT mapping. *J. Biomed. Semantics* 1 (1), 7. doi: 10.1186/2041-1480-1-7

Oliveira, D., and Pesquita, C. (2018). Improving the interoperability of biomedical ontologies with compound alignments. *J. Biomed. Semantics* 9 (1), 1. doi: 10.1186/s13326-017-0171-8

Park, H., and Hardiker, N. (2009). Clinical terminologies: a solution for semantic interoperability. *J. Korean Soc. Med. Inform.* 15 (1), 1–11. doi: 10.4258/jksmi.2009.15.1.1

Petasis, G., Karkaletsis, V., Paliouras, G., Krithara, A., and Zavitsanos, E., (2011). "Ontology population and enrichment: State of the art," in *Knowledge-driven multimedia information extraction and ontology evolution*, Springer-Verlag, 134–166. doi: 10.1007/978-3-642-20795-2_6

Petrova, A., Ma, Y., Tsatsaronis, G., Kissa, M., Distel, F., Baader, F., et al. (2015). Formalizing biomedical concepts from textual definitions. *J. Biomed. Semantics* 6, 22. doi: 10.1186/s13326-015-0015-3

Presutti, V., and Gangemi, A. (2008). Content ontology design patterns as practical building blocks for web ontologies. In International Conference on Conceptual Modeling pp. 128–141. Springer, Berlin, Heidelberg.

Quesada-Martínez, M., Mikroyannidi, E., Fernández-Breis, J. T., and Stevens, R. (2015). Approaching the axiomatic enrichment of the Gene Ontology from a lexical perspective. *Artif Intell Med.* Sep65 (1), 35–48. doi: 10.1016/j.artmed.2014.09.003

Roldán-García, M. D., García-Godoy, M. J., and Aldana-Montes, J. F. (2016). Dione: an OWL representation of ICD-10-CM for classifying patients' diseases. *J. Biomed. Semantics* 7 (1), 62. doi: 10.1186/s13326-016-0105-x

Qawasmeh, O., Lefrançois, M., Zimmermann, A., and Maret, P., (2018). Improved categorization of computer-assisted ontology construction systems: focus on bootstrapping capabilities. In Extended semantic web conference (ESWC2018). doi: 10.1007/978-3-319-98192-5_12

Rector, A., and Brandt, S. (2008). Why Do It the Hard Way? The Case for an Expressive Description Logic for SNOMED. *J. Am. Med. Inform. Assoc.* 15 (6), 744–751. doi: 10.1197/jamia.M2797

Rector, A., and Iannone, L. (2012). Lexically suggest, logically define: quality assurance of the use of qualifiers and expected results of post-coordination in SNOMED CT. *J. Biomed. Inform.* 45 (2), 199–209. doi: 10.1016/j.jbi.2011.10.002

Riga, M., Mitzias, P., Kontopoulos, E., and Kompatsiaris, I., (2017). PROPheT–Ontology Population and Semantic Enrichment from Linked Data Sources. In International Conference on Data Analytics and Management in Data Intensive Domains, Springer, Cham 157–168. doi: 10.1007/978-3-319-96553-6_12

Rogers. (2005). Using Medical Terminologies. http://www.cs.man.ac.uk/~jeremy/HealthInf/RCSEd/terminology-using.htm.

Rossi Mori, A., Consorti, F., and Galeazzi, E. (1998). Standards to support development of terminological systems for healthcare telematics. *Methods Inf. Med.* 37 (4-5), 551–563. doi: 10.1055/s-0038-1634542

Rosenbloom, S. T., Miller, R. A., Johnson, K. B., Elkin, P. L., and Brown, S. H. (2006). Interface terminologies: facilitating direct entry of clinical data into electronic health record systems. *J. Am. Med. Inform. Assoc.* 13 (3), 277–288. doi: 10.1197/jamia.M1957

Sánchez, D., and Moreno, A. (2008). Learning non-taxonomic relationships from web documents for domain ontology construction. *Data Knowl. Eng.* 64 (3), 600–623. doi: 10.1016/j.datak.2007.10.001

Sarntivijai, S., Xiang, Z., Shedden, K. A., Markel, H., Omenn, G. S., Athey, B. D., et al., (2012). Ontology-based combinatorial comparative analysis of adverse events associated with killed and live influenza vaccines. doi: 10.1371/journal.pone.0049941

Schriml, L. M., Arze, C., Nadendla, S., Chang, Y. W. W., Mazaitis, M., Felix, V., et al. (2012). Disease Ontology: a backbone for disease semantic integration. *Nucleic Acids Res.* 40 (D1), D940–D946. doi: 10.1093/nar/gkr972

Schulz, S., and Jansen, L. (2013). Formal ontologies in biomedical knowledge representation. *Yearb Med. Inform.* 8 (1), 132–146. doi: 10.1055/s-0038-1638845

Schulz, S., and Hahn, U. (2001). Medical knowledge reengineering–converting major portions of the UMLS into a terminological knowledge base. *Int. J. Med. Inform.* 64(2–3):207-21. doi: 10.1016/S1386-5056(01)00201-5

Schulz, S., Rodrigues, J. M., Rector, A., and Chute, C. G. (2017). Interface Terminologies, Reference Terminologies and Aggregation Terminologies: A Strategy for Better Integration. *Stud. Health Technol. Inform*. 245,940–944. doi 10.3233/978-1-61499-830-3-940

Serra, I., Girardi, R., and Novais, P. (2014). Evaluating techniques for learning non-taxonomic relationships of ontologies from text. *Expert Syst. Appl.* 41 (11), 5201–5211. doi: 10.1016/j.eswa.2014.02.042

Sheth, A., Ramakrishnan, C., and Thomas, C. (2005). Semantics for the semantic web: The implicit, the formal and the powerful. *Int. J. Semant Web Inf. Syst.* 1 (1), 1–18. doi: 10.4018/jswis.2005010101

Simperl, E. P. B., Tempich, C., and Sure, Y., (2006). "Ontocom: a cost estimation model for ontology engineering," in *International Semantic Web Conference* (Berlin, Heidelberg: Springer), 625–639. doi: 10.1007/11926078_45

Sioutos, N., de Coronado, S., Haber, M. W., Hartel, F. W., Shaiu, W. L., and Wright, L. W. (2007). NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. *J. Biomed. Inform.* 40 (1), 30–43. doi: 10.1016/j.jbi.2006.02.013

Souvignet, J., Declerck, G., Trombert, B., Rodrigues, J. M., Jaulent, M. C., and Bousquet, C. (2012). Evaluation of automated term groupings for detecting anaphylactic shock signals for drugs. *AMIA Annu. Symp. Proc.* 2012, 882–890.

Souvignet, J., Asfari, H., Declerck, G., Lardon, J., Trombert-Paviot, B., Jaulent, M. C., et al. (2014). Ci4SeR–Curation Interface for Semantic Resources–Evaluation with Adverse Drug Reactions. *EHealth-For Continuity of Care: Proceedings of MIE2014* 205, 116.

Souvignet, J., Asfari, H., Lardon, J., Del Tedesco, E., Declerck, G., and Bousquet, C. (2016a). MedDRA automated term groupings using OntoADR: evaluation with upper gastrointestinal bleedings. *Expert Opin. Drug Saf.* 15 (9), 1153–1161. doi: 10.1080/14740338.2016.1206075

Souvignet, J., Declerck, G., Asfari, H., Jaulent, M. C., and Bousquet, C. (2016b). OntoADR a semantic resource describing adverse drug reactions to support searching, coding, and information retrieval. *J. Biomed. Inform.* 63, 100–107. doi: 10.1016/j.jbi.2016.06.010

Souvignet, J., Declerck, G., Trombert-Paviot, B., Asfari, H., Jaulent, M. C., and Bousquet, C. (2019). Semantic Queries Expedite MedDRA Terms Selection Thanks to a Dedicated User Interface: A Pilot Study on Five Medical Conditions. *Front. Pharmacol.* 10, 50. doi: 10.3389/fphar.2019.00050

Spackman, K. A., Campbell, K. E., and Côté, R. A. (1997). SNOMED RT: a reference terminology for health care. *Proc AMIA Annu Fall Symp*. 640–4.

Suárez-Figueroa, M. C., Gómez-Pérez, A., and Fernández-López, M., (2012). "The NeOn methodology for ontology engineering," in *Ontology engineering in a networked world* (Berlin, Heidelberg: Springer), 9–34. doi: 10.1007/978-3-642-24794-1_2

Sure, Y., Staab, S., and Studer, R., (2004). "On-to-knowledge methodology (OTKM)," in *Handbook on ontologies* (Berlin, Heidelberg: Springer), 117–132. doi: 10.1007/978-3-540-24750-0_6

Third, A. (2012). "Hidden semantics: what can we learn from the names in an ontology?," in *Proceedings of the Seventh International Natural Language Generation Conference, Association for Computational Linguistics*, 67–75.

Tiddi, I., Mustapha, N. B., Vanrompay, Y., and Aufaure, M. A., (2012). "Ontology learning from open linked data and web snippets," in *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"* (Berlin, Heidelberg: Springer), 434–443. doi: 10.1007/978-3-642-33618-8_59

Trifirò, G., Pariente, A., Coloma, P. M., Kors, J. A., Polimeni, G., Miremont-Salamé, G., et al. (2009). EU-ADR group. Data mining on electronic health record databases for signal detection in pharmacovigilance: which events to monitor? *Pharmacoepidemiol Drug Saf.* 18 (12), 1176–1184. doi: 10.1002/pds.1836

van Damme, P., Quesada-Martínez, M., Cornet, R., and Fernández-Breis, J. T. (2018). From lexical regularities to axiomatic patterns for the quality assurance of biomedical terminologies and ontologies. *J. Biomed. Inform.* 84, 59–74. doi: 10.1016/j.jbi.2018.06.008

Vikström, A., Skånér, Y., Strender, L. E., and Nilsson, G. H. (2007). Mapping the categories of the Swedish primary health care version of ICD-10 to SNOMED CT concepts: rule development and intercoder reliability in a mapping trial. *BMC Med. Inform. Decis. Mak.* 2 (7), 9. doi: 10.1186/1472-6947-7-9

Villaverde, J., Persson, A., Godoy, D., and Amandi, A. (2009). Supporting the discovery and labeling of non-taxonomic relationships in ontology learning. *Expert Syst. Appl.* 36 (7), 10288–10294. doi: 10.1016/j.eswa.2009.01.048

Wächter, T., and Schroeder, M. (2010). Semi-automated ontology generation within OBO-Edit. *Bioinformatics* 26 (12), i88–i96. doi: 10.1093/bioinformatics/btq188

Xiang, Z., Zheng, J., Lin, Y., and He, Y. (2015). Ontorat: automatic generation of new ontology terms, annotations, and axioms based on ontology design patterns. *J. Biomed. Semantics* 6, 4. doi: 10.1186/2041-1480-6-4

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: info@frontiersin.org | +41 21 510 17 00

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership