Integrative cancer pharmacogenomics to infer large-scale drug taxonomy

Nehme El-Hachem[1,2,‡], Deena M.A. Gendoo[3,4,‡], Laleh Soltan Ghoraie[3,4,‡], Zhaleh Safikhani[3,4], Petr Smirnov[3], Christina Chung[5], Kenan Deng[5], Ailsa Fang[5], Erin Birkwood[6], Chantal Ho[5], Ruth Isserlin[5], Gary D. Bader[5,7,8], Anna Goldenberg[5,9], Benjamin Haibe-Kains[3,4,5,10,*]

[1] Integrative Computational Systems Biology, Institut de Recherches Cliniques de Montréal, Montreal, Quebec, Canada
[2] Department of Biomedical Sciences. Université de Montréal, Montreal, Quebec, Canada
[3] Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, Canada
[4] Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada
[5] Department of Computer Science, University of Toronto, Toronto, Ontario, Canada
[6] School of Computer Science, McGill University, Montreal, Quebec, Canada
[7] The Donnelly Centre, Toronto, Ontario, Canada
[8] The Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, Ontario, Canada
[9] Hospital for Sick Children, Toronto, Ontario, Canada
[10] Ontario Institute of Cancer Research, Toronto, Ontario, Canada

‡ Co-first authors
* Corresponding author: Benjamin Haibe-Kains, Princess Margaret Cancer Centre, 101 College Street, Toronto, M5G1L7, Ontario, Canada

Running Title: Integrative drug taxonomy using pharmacogenomics

Keywords: Anticancer Therapies, Pharmacogenomics, Drug Taxonomy, Bioinformatics, Machine Learning, Integrative Analysis

Abbreviations: APC: Affinity Propagation Clustering; AUC: Area under the receiver operating characteristics curve; AUC_d: Area under the drug dose-response curve (drug sensitivity); CTRP: Cancer Therapeutics Portal; DNF: Drug Network Fusion; FDR: False Discovery Rate; GDSC:

Genomics of Drug Sensitivity in Cancer; MoA: Mechanism of Action; PR: Precision-Recall; SMILES: Simplified Molecular Input Line Entry Specification.

Financial support. D.A.M. Gendoo was supported by the Ontario Institute for Cancer Research through funding provided by the Government of Ontario. Z. Safikhani was supported by The Cancer Research Society (Canada). N. El-Hachem was supported by the Ministry of Economic Development, Employment and Infrastructure and the Ministry of Innovation of the Government of Ontario. B. Haibe-Kains was supported by the Gattuso-Slaight Personalized Cancer Medicine Fund at Princess Margaret Cancer Centre and the Canadian Institutes of Health Research.

Disclosure of Potential Conflicts of Interest. No potential conflicts of interest were disclosed.

ABSTRACT

Identification of drug targets and mechanism of action (MoA) for new and uncharacterized anticancer drugs is important for optimization of treatment efficacy. Current MoA prediction largely relies on prior information including side effects, therapeutic indication, and chemo-informatics. Such information is not transferable or applicable for newly identified, previously uncharacterized small molecules. Therefore, a shift in the paradigm of MoA predictions is necessary towards development of unbiased approaches that can elucidate drug relationships and efficiently classify new compounds with basic input data. We propose here a new integrative computational pharmacogenomic approach, referred to as Drug Network Fusion (DNF), to infer scalable drug taxonomies that relies only on basic drug characteristics towards elucidating drug-drug relationships. DNF is the first framework to integrate drug structural information, high-throughput drug perturbation, and drug sensitivity profiles, enabling drug classification of new experimental compounds with minimal prior information. DNF taxonomy succeeded in identifying pertinent and novel drug-drug relationships, making it suitable for investigating experimental drugs with potential new targets or MoA. The scalability of DNF facilitated identification of key drug relationships across different drug categories and poses as a flexible tool for potential clinical applications in precision medicine. Our results support DNF as a valuable resource to the cancer research community by providing new hypotheses on compound MoA and potential insights for drug repurposing.

INTRODUCTION

Continuous growth and ongoing deployment of large-scale pharmacogenomic datasets has opened new avenues of research for the prediction of biochemical interactions of small molecules with their respective targets and therapeutic effects, also referred to as drug mechanisms of action (MoA). Development of computational methods to predict MoA of new compounds is an active field of research in the past decade (1). Despite major advancements in this field, key challenges still remain in the (*i*) design of classification approaches that rely on minimal drug characteristics to classify drugs, and (*ii*) selection and integration of complementary datasets to best characterize drugs' effects on biological systems.

The notion of a 'minimalist' approach to represent similarities among drug compounds has been extensively explored, with varying results. Several computational strategies have solely relied on chemical structure similarity to infer drug-target interactions (2,3), based on the assumption that structurally-similar drugs share similar targets, and ultimately, similar pharmacological and biological activity (4). However, sole dependence on chemical structure information fails to consider drug-induced genomic and phenotypic perturbations, which directly connect with biological pathways and molecular disease mechanisms (5,6). Seminal approaches by Iorio et al. (7) and Iskar et al. (8) leveraged drug-induced transcriptional profiles from Connectivity Map (CMAP) (9) towards identification of drug-drug similarities and MoA solely based on gene expression profiles (7). The major limitation of CMAP however is the lack of global scope, as only 1,309 drugs are characterized across 5 cancer cell lines (9). Other methods have integrated prior knowledge such as adverse effects annotations (10,11) and recent approaches showed that integrating multiple layers of information had improved Anatomical Therapeutic Classification system (ATC) prediction for FDA-approved drugs (12). While these initiatives constitutes major advances towards characterizing drug MoA, comparing the consistency of such efforts towards prediction of new, uncharacterized small molecules remains a challenge.

Computational approaches designed to characterize drug MoA have yet to capitalize on newly generated high-throughput data types. The published CMAP has recently been superseded by the L1000 dataset from the NIH Library of Integrated Network-based Cellular Signatures (LINCS) consortium, which contains over 1.8 million gene expression profiles spanning 20,413 chemical perturbations. A recent study of the LINCS data showed that structural similarity are significantly associated with similar transcriptional changes (6). While the L1000 dataset provides an unprecedented compendium of both transcriptomic drug data, its integration with other pharmacogenomics data types has not been explored extensively.

The advent of high-throughput *in vitro* drug screening promises to provide additional insights into drug MoA. The pioneering initiative of the NCI60 panel provided an assembly of tumour cell lines that have been treated against a panel of over 100,000 small molecules (13), and served as the first large-scale resource enabling identification of lineage-selective small molecule sensitivities (13). However, its relatively small number of cancer cell lines (n=59) restricted the relevance of these data for prediction of drug MoA. To address this issue, the Cancer Therapeutics Response Portal (CTRP) initiative screened 860 cancer cell lines against a set of 481 small molecule compounds (14), which makes it the largest repository of *in vitro* drug sensitivity measurements to date. Individual assessment of these *in vitro* sensitivity datasets have highlighted their relevance for inference of MoA of approved and experimental compounds. It remains to be demonstrated, however, whether integration of drug sensitivity data with other drug-related data, such as drug structures and drug-induced transcriptional signatures, can be used to systematically infer drug MoA.

To efficiently harness these recent high-throughput datasets, we have developed a scalable approach that maximizes complementarity between different data types to provide a complete landscape of drug-drug relationships and similarities. We leveraged our recent

Similarity Network Fusion algorithm (15) to integrate drug structure, sensitivity, and perturbation data in a large-scale molecular drug taxonomy, called Drug Network Fusion (DNF) (Fig. 1). We demonstrate how the resulting integrative drug taxonomies improve characterization of drug target and anatomical classifications compared to taxonomies based on single data types. Importantly, we show how the DNF taxonomy can be used to infer MoA for new compounds that lack deep pharmacological and biochemical characterization.

MATERIAL AND METHODS

A schematic overview of the analysis design is presented in Supplementary Fig. 1.

Processing of drug-related data and identification of drug similarity

Cancer Cell lines: We refer to the original publications of the NCI60 (16), CTRPv2 (14) and L1000 (17) datasets for the source, number of passages and authentication of the cancer cell lines used in this study.

Drug structure annotations: Canonical SMILES strings for the small molecules were extracted from PubChem (18), a database of more than 60 millions unique structures. Tanimoto similarity measures (19) between drugs were calculated by first parsing annotated SMILES strings for existing drugs through the *parse.smiles* function of the *rcdk* package (version 3.3.2). Extended connectivity fingerprints (hash-based fingerprints, default length 1,024) across all drugs was subsequently calculated using the *rcdk::get.fingerprints* function (20).

Drug perturbation signatures: We obtained transcriptional profiles of cancer lines treated with drugs from the L1000 dataset recently released by the Broad Institute, which contains over 1.8

million gene expression profiles of 1000 ‘landmark’ genes across 20,413 drugs. We used our PharmacoGx package (version 1.3.4) (21) to compute signatures for the effect of drug concentration on the transcriptional state of a cell, using a linear regression model adjusted for treatment duration, cell line identity, and batch to identify the genes whose expression is significantly perturbed by drug treatment:

$$G = \beta_0 + \beta_i C_i + \beta_t T + \beta_d D + \beta_b B$$

where

G = *molecular feature expression (gene)*

C_i = *concentration of the compound applied*

T = *cell line identity*

D = *experiment duration*

B = *experimental batch*

β = *regression coefficients.*

The strength of the feature response is quantified by β_i. The transcriptional changes induced by drugs on cancer cell lines are subsequently referred to throughout the text as *drug perturbation signatures*. Similarity between estimated coefficients of drug perturbation signatures was computed using the Pearson correlation coefficient, with the assumption that drugs similarly perturbing the same set of genes might have similar mechanisms of action.

Drug sensitivity profiles: We obtained summarized dose-response curves from the published drug sensitivity data of the NCI60 (13) and CTRPv2 (14) datasets integrated in the PharmacoGx package. We relied on the Z-score and area under the curve (AUC_d) metrics from drug-dose response curves of the NCI60 and CTRPv2, respectively. Drug similarity was defined as the Pearson correlation of drug sensitivity profiles.

Development of a drug network fusion (DNF) taxonomy

We used our Similarity Network Fusion algorithm (15) to identify drugs that have similar mechanisms of actions by integrating three data types representing drug structure, drug perturbation, and drug sensitivity profiles. Drug structure and drug perturbation taxonomies were based on drug-drug similarity matrices computed from the PubChem SMILES and the L1000 dataset, respectively. To test the robustness of the fusion algorithm with respect to the scale of the drug sensitivity profiles, we also applied our methodology on both the CTRPv2 and NCI60 datasets. The NCI60 dataset comprises a much smaller panel of cell lines (60 vs. 860 for NCI60 and CTRPv2, respectively). The NCI60 panel compensates for its small cell line panel by the large number of screened drugs (>40,000 drugs tested on the full panel; Supplementary Fig. 2). Accordingly, the drug sensitivity taxonomy was composed of the drug-drug similarity matrix of the sensitivity profiles extracted from either the NCI60 or CTRPv2 datasets. For each dataset, an affinity matrix was first calculated using the *affinityMatrix* function as described in the *SNFtool* package (version 2.2), using default parameters. We combined the three affinity matrices of the structure, perturbation, and sensitivity taxonomies into a Drug Network Fusion (DNF) matrix using the *SNFtool::SNF* function (Supplementary Fig. 1). Two separate DNF matrices were generated dependant on the sensitivity layer used (either CTRPv2 or NCI60). The developed DNF taxonomies, as well as the single data type taxonomies, were subsequently tested against benchmark datasets to validate their drug MoA.

Assessment of drug mode of action across drug taxonomies

Drug-target associations. Known target associations for drugs pertaining to the NCI60 dataset were downloaded from DrugBank (version 5.0) (22). Drug-target associations for drugs of the CTRPv2 dataset were obtained from the CTRPv2 website (http://www.broadinstitute.org/ctrp.v2/?page=#ctd2Target). Drugs with annotated targets were filtered to retain only targets with at least two drugs.

Anatomical therapeutic classification system (ATC). ATC annotations (23) for the drugs common to the NCI60 and CTRPv2 datasets were downloaded from ChEMBL (file version 16-5-10-02) (24). These ATC codes were filtered to retain only those categories with at least one pair of drugs sharing a pharmacological indication. The drugs with known ATC annotations from the NCI60 and CTRPv2 datasets were subsequently used as a validation benchmark against singular drug taxonomies and the DNF taxonomy.

Evaluation of drug mechanism of action across taxonomies

We assessed the predictive value of our developed taxonomies against drug-target and ATC benchmark datasets to determine the extent to which single data type taxonomies and the DNF taxonomy recapitulate known drug MoA (Supplementary Fig. 3). We adapted the method from Cheng et al (25) to compare benchmarked datasets against singular drug taxonomies (drug perturbation, drug structure, or drug sensitivity) as well as the integrated DNF taxonomy. This method is further detailed below for the benchmark datasets used in our study. First, we created adjacency matrices that indicate whether each pair of drugs share a target molecule or ATC annotation. The drug-target and ATC adjacency matrices were then converted into a vector of similarities between every possible pair of drugs where the value ‘1’ was assigned in the vector if the paired drugs were observed the same target/ATC set, and ‘0’ otherwise. Similarly, the affinity matrices of singular drug taxonomies as well as the DNF taxonomy matrix were converted into vectors of drug pairs, with the similarity value of the drug pairs retained from their original corresponding matrix. Binary vectors of the benchmarks were compared to the four continuous vectors of the drug taxonomies by computing the receiver-operating curves (ROC) and the area under the curve (AUROC) using the *ROCR* package (version 1.0.7). Concordance indices were calculated using the *rcorr.cens* function of the *Hmisc* package (version 1.18.0). The AUROC estimates the probability that, for two pairs of drugs, drugs that are part of the same drug set (same therapeutic targets or ATC functional annotations) have higher similarity than drugs that do not belong to the same drug set. AUROC calculations for each of the four taxonomies were statistically compared against each other using the *survcomp::compare.cindex* function (26). Precision-Recall (PR) curve is an alternative to ROC curves for measuring an algorithm's performance, especially in classification problems with highly skewed class distributions. We used the PRROC package (version 1.1) to compute PR curves. This package does not implement functions for statistical comparison of PR curves. For these types of classification tasks, algorithms that optimize AUROC do not necessarily optimize the area under the PR curves (AUPRC). Therefore, computing both curves brings more insight to measuring performance and comparing multiple algorithms for the same prediction task.

Detection of drug communities and visualization

Clusters of drug communities were determined from the DNF taxonomy using the affinity propagation algorithm (APC) from the *apcluster* package (version 1.4.2). The APC algorithm generates non-redundant drug communities, with each community represented by an exemplar drug. An elevated *q* value parameter, which determines the quantiles of similarities to be used as input preference to generate small or large number of clusters, was set at *q*=0.9 within the *apcluster* function to produce a large number of communities. Networks of exemplar drugs were rendered in *Cytoscape* (version 3.3.0) (27). Drug structures were rendered using the *chemViz* plugin (version 1.0.3) for cytoscape. A minimal spanning tree of the exemplar drugs was determined using Kruskal's algorithm as part of the *CySpanningTree* plugin (version 1.1) for cytoscape.

Assessment of Drug Community Enrichment

We tested for the enrichment of drug communities determined from the DNF taxonomy against all drugs attributed to a specific target or ATC class. We first generated lists of drug targets and ATC classes from our benchmarks, filtered to retain only drug targets or ATC classes with two or more drugs. Each of the 53 and 52 communities from the DNF taxonomy (using CTRPv2 and NCI60 drug sensitivity data, respectively) was subsequently compared against the lists using a Fisher's exact test, followed by multiple testing (FDR) correction.

DNF web-application

We developed the DNF web-application (dnf.pmgenomics.ca) to facilitate the exploration of the drug communities and the full drug similarity networks built using the CTRPv2 and NCI60 drug sensitivity datasets. The application was implemented using JavaScript and AngularJS for its frontend and Node.js for its backend. Drug network information is stored on the server as JSON and rendered as graphs with the Cytoscape.js library (28). As drug clusters are fully connected, showing all edges in the web application will overwhelm the browser. Thus, the graphs shown in the web application are thresholded to display the top one thousand edges between drugs that have the greatest similarity. Users can click on drug communities to display the full set of drugs and their similarity edges. Clicking on an edge reports the similarity score and its relative contribution to the fused similarity score. Clicking on drugs provides basic descriptors of the compound and its pubchem link.

Research reproducibility

The code and data links required to reproduce this analysis is open source and publicly available on github.com/bhklab/DNF. All software dependencies are available on the Comprehensive Repository R Archive Network (CRAN) or Bioconductor (BioC). A detailed tutorial describing how to setup the software environment, and run our analysis pipeline to generate the figures and tables are provided in the DNF GitHub repository. The code and documentation of the DNF web-application is open source and publicly available on github.com/bhklab/DNF-webapp. This work complies with the guidelines proposed by Sandve et al. (29) in terms of code availability and replicability of results.

RESULTS

We developed the DNF approach to generate a large-scale molecular taxonomy by integrating drug similarity matrices from structural information, transcriptomic perturbation, and sensitivity profiles (Supplementary Fig. 1). Drug structure profiles (SMILES representations) were extracted from the PubChem database. Drug perturbation signatures, representing drug-induced gene expression changes, were extracted from the recent LINCS L1000 dataset. Drug sensitivity profiles representing cell line viability across cancer cell lines were extracted from our PharmacoGx platform (21), which contains pharmacological profiles of several hundred cell lines generated by the CTRP (14) and NCI60 (13) initiatives. By integrating these three data types, we have generated a similarity network composed of over 200 drugs (Supplementary Fig. 2). The drug similarity matrices computed from single data layers and fusion estimates are provided in Supplementary Table 1 for the CTRPv2 and NCI60 taxonomies.

To assess the relevance and benefits of our integrative drug taxonomies we first tested whether the different data layers were redundant or complementary. We then tested whether DNF enables prediction of drug MoA. While definitions may vary in the literature, for the context of this study, we refer to drug mechanisms of action specifically as determining drug targets for query drugs, as well as identification of the anatomical therapeutic classes (ATC) for different drugs. We further tested whether DNF enables clustering of drugs sharing common action mechanisms, and we demonstrated how identified drug communities from this clustering may be used towards novel discoveries for drug repositioning approaches and clinical applications.

Complementarity of drug structure, perturbation, and sensitivity profiles

We investigated the potential complementarity between drug structure, perturbation, and sensitivity profiles to assess their potential for drug taxonomies as part of DNF. We generated similarity networks (similarity matrices) representing each of the single-layer drug taxonomies of drug structure, drug perturbation, and drug sensitivity (Supplementary Table 1). In brief, we used the Tanimoto index to calculate the similarity between two drug structures. We used Pearson correlation to quantify the extent to which pairs of drugs affect similar genes at the expression level or whether drugs kill the same cancer cell lines for drug perturbation and drug sensitivity similarity matrices, respectively. We then computed the correlation between all pairs of similarity matrices (Supplementary Fig. 4) using the spearman correlation. The single-layers are only weakly correlated, with a maximum absolute spearman correlation coefficient of 0.091 between the perturbation and sensitivity layers using the CTRPv2 sensitivity dataset (Supplementary Fig. 4A), as well as a maximum of 0.085 between the structure and sensitivity layers using the NCI60 sensitivity dataset (Supplementary Fig. 4B). In contrast, the integrative drug taxonomy (DNF) is more highly correlated to each of the single layers, across both CTRPv2 (Supplementary Fig. 4A) and NCI60 (Supplementary Fig. 4B). These findings highlight that the structural, sensitivity, and transcriptional perturbation data are non-redundant, thereby presenting a diversity of drug-drug relationships across the set of drugs that commonly share all three data types. These findings also highlight that the DNF network yielded increased correlation between the three data types, such that it is not driven by only a single layer.

Performance of drug taxonomies against known drug targets

Determining novel drug-target interactions opens new avenues for drug repurposing efforts, and suggests mechanisms by which drugs can operate in cells. We explored the relevance of our DNF taxonomy by demonstrating its predictive value towards identification of drug targets. Of the 239 drugs represented in our DNF taxonomy generated using CTRPv2, 141 could be matched against the drug target benchmarks (Supplementary Fig. 2). Similarity, for the DNF taxonomy generated using the NCI60 dataset, 86 drugs out of 238 drugs could be matched to known drug target (Supplementary Fig. 2). We assessed the predictive value of our single-data layer and integrative drug taxonomies (DNF) against known drug targets (Supplementary Fig. 3). We performed a receiver operating characteristics (ROC) analysis to quantify how well our drug taxonomies align with established drug target, by statistically comparing the area under the ROC curve (AUROC) values for each drug taxonomy under evaluation. Similarly, we calculated the area under the precision-recall (PR) curve (AUPRC) (see Methods). Of the three single-layer taxonomies validated against annotated drug targets from CTRPv2, the drug sensitivity layer outperformed the structure and perturbation taxonomies (AUROC of 0.83, 0.71 and 0.64 for sensitivity, structural and perturbation data layers, respectively) (Fig. 2A). Importantly, DNF yielded the best predictive value (AUROC of 0.89), and was significantly higher than any single-layer taxonomy (one-sided Student t test p-value<1E-16, Supplementary Table 2). We further computed PR curves (Fig. 2B). These curves demonstrate that DNF supersedes the single layers taxonomies except for sensitivity, where it performs equivalently (AUPRC of 0.413 and 0.406 for DNF and drug sensitivity, respectively; Fig. 2B).

We further tested the predictive value of our DNF taxonomy using the set of drug sensitivity profiles obtained from the NCI60 dataset, where a much smaller panel of cell lines has been screened (60 vs 860 cell lines for NCI60 and CTRPv2, respectively; Supplementary Fig. 2). Our evaluation of single-layer taxonomies demonstrates that drug similarities based on structure and sensitivity profiles were the most efficient in predicting drug-target associations (AUROC of 0.8 for both layers; Supplementary Fig. 5A) compared to perturbation (AUROC of 0.62; Supplementary Fig. 5A). DNF was significantly more predictive of drug-target associations compared to single-layer taxonomies (AUROC of 0.88 and one-sided Student t test p-values<0.05, Supplementary Fig. 5A, Supplementary Table 2 and AUPRC of 0.552; Supplementary Fig. 5B).

Performance of drug taxonomies against anatomical classification (ATC)

Predicting the ATC of a drug provides insights about its pharmacological mechanism, and ultimately presents new potential indications for previously uncharacterized drugs. We demonstrated the relevance of our DNF network by testing its predictive value for ATC drug classifications (Fig. 2, Supplementary Fig. 5). We explored the value of the DNF taxonomies towards classifying drugs up to ATC level 4, which reports the chemical, therapeutic, and pharmacological subgroup of a given drug (23). A total of 51 and 72 drugs could be matched against the ATC benchmarks for the CTRPv2 and NCI60 taxonomies, respectively. We implemented a similar benchmarking approach to that previously conducted for drug target classification. We observed that drug sensitivity was no longer the most predictive layer for ATC classification, and instead exhibited comparable predictive power to drug perturbation (Fig. 2C, Supplementary Fig. 5C). Conversely, the structure-based taxonomy (Fig. 2C, Supplementary Fig. 5C) was the most predictive amongst single-layer taxonomies (AUROC of 0.72, 0.6 and 0.58 for structure, sensitivity, and perturbation layers, respectively, for the CTRPv2 taxonomy; see Supplementary Fig. 5 for the NCI60 taxonomy). DNF significantly outperformed single-layer taxonomies (Fig. 2C-D, Supplementary Fig. 5C-D) (AUROC of 0.8 and 0.85 for DNF based on CTRPv2 and NCI60, respectively, with one-sided Studennt t test p-value<0.05, Supplementary Table 2, and AUPRC of 0.558 and 0.492 vs. random classifiers' AUPRCs of 0.212 and 0.095, respectively).

Identification of drug communities

We assessed the biological relevance of DNF in discovering drugs with similar MoA by applying the affinity cluster propagation algorithm (30) to identify clusters of highly similar drugs, referred

Author Manuscript Published OnlineFirst on March 17, 2017; DOI: 10.1158/0008-5472.CAN-17-0096
Author manuscripts have been peer reviewed and accepted for publication but have not yet been edited.

to as *drug communities*. These communities can be represented by their most representative ('exemplar') drug, and similarities between communities are subsequently represented a network where each node is labeled by the exemplar drug. Community detection was implemented on the full set of 239 and 238 drugs of the DNF taxonomy (based on using CTRPv2 and NCI60 sensitivity data; Fig. 3 and Supplementary Fig. 6 respectively).

We identified 53 communities in the CTRPv2 DNF (Supplementary Table 3A), which resulted in a set of 39 drug communities with at least 2 drugs with known drug targets (permutation test p-value<1E-4; Supplementary Table 3B). Overall, DNF has produced a consistent classification of drugs for a variety of known functional classes (Supplementary Table 3C). Our classifications recapitulate most of the protein target-drug associations represented in CTRPv2 (Fig. 3). Receptor tyrosine kinases and non-receptor tyrosine kinases, including EGFR/ERBB2 (community C2), MEK1/2 (C41), TGFRB1 (C39), BRAF (C18), IGFR-1 (C6) KDR/FLT1 (C50) inhibitors. In addition to PI3K/mTOR inhibitors (C28), epigenetic regulators: HDACs (C45) and DNMT1 (C20) inhibitors, HMG CoA (C30) and proteasome inhibitors (C7) (Supplementary Tables 3A and 3D). Notably, all BRAF (V600E mutation) inhibitors were classified correctly, which include drugs already tested in metastatic melanoma (community C18: dabrafenib, GDC0879, PLX4720; Fig. 3) and mitogen-activated protein kinase/ERK kinase (MEK) inhibitors (C41: namely trametinib and selumetinib; Fig. 3). BRAF regulates the highly conserved MAPK/ERK signaling pathway, and BRAF mutational status has been proposed as a biomarker of sensitivity towards selumetinib and other MEK inhibitors (31). This explains the tight connection of these two communities (Fig. 3).

Using the NCI60 DNF, we identified 51 communities (Supplementary Fig. 6, Supplementary Table 3E) with 20 of those containing at least two known drug targets (permutation test p-value<1E-4; Supplementary Table 3F). In this case, we also identified a number of well-characterized drug communities. These include the community composed of EGFR inhibitors (C20; Supplementary Fig. 6). The community C14, including cardiac glycosides also concur with the study of Khan et al. (5), showing that these compounds inhibit Na+/K+ pumps in cells. Bisacodyl, a laxative drug, is part of the C14 community, which concurs with Khan et al. who demonstrated a sharing a MoA similar to cardiac glycosides, despite its structural dissimilarity to that class of compounds (5).

Enrichment of DNF drug communities for drug targets and ATC classifications

We conducted a quantitative community enrichment analysis to test whether DNF succeeds in identifying drug communities that are enriched for drug targets and ATC classifications. Fisher's exact test was conducted between all the drugs in each community versus all drugs attributed to a specific drug target or ATC class. This approach allowed us to test which specific drug targets or ATC classifications are significantly enriched in the computed communities (Fig. 4, Supplementary Fig. 7). The analysis was conducted using all of the communities of DNF based on CTRPv2 (n=53 communities, Fig. 4, Supplementary Table 4A,B) as well as DNF based on NCI60 sensitivity data (n=51 communities, Supplementary Fig. 7, Supplementary Table 4C,D).

Clustering of the DNF drug taxonomies identified a wide range of community sizes (Supplementary Fig. 8), with a median community size of 4 drugs both CTRPv2 and NCI60-based DNF taxonomies. Many of these communities were significantly enriched for drug targets and ATC classes (Fig. 4, Supplementary Fig. 7). Among these, for example, community C2 in CTRPv2 is statistically enriched against the ERBB2 and EGFR targets and contains well known inhibitors for these targets, such as afatinib, erlotinib and lapatinib (Fig. 3, Fig. 4A). Similarly, C30 hosts almost all of the members of the statin family, which are known to affect the mevalonate pathway and HMGCR (Fig. 3, Fig. 4A). We identified enrichment of DNF communities (based on NCI60) against several representative ATC categories (Supplementary Fig. 7B). These include communities enriched for known anticancer and other therapeutic classes, including antimalarial drugs (C4, ATC P01BE [Artemisinin and derivatives], anthracyclines (C12, L01DB [anthracyclines and related substances]) , antimetabolites (C9, ATC L01BB [purine analogues]), cholesterol lowering agents (C33, ATC C10AA [HMG CoA reductase inhibitors]), corticosteroids (C18, overrepresented in many ATC categories since they are indicated for a large number of medical conditions), vinca and taxanes alcaloids (C38 and C49; L01CD [taxanes] and L01CA [Vinca alkaloids and analogues], respectively) and protein kinase inhibitors (C20, L01XE). As expected, communities containing very few annotated targets or ATC classes do not demonstrate significant enrichment (Supplementary Table 4 for CTRPv2 and NCI60).

Identification of novel drug-drug relationships and drug action mechanisms

We conducted an explorative analysis to identify potential mechanisms for existing drugs and for poorly characterized drugs in the set of drugs constituting the DNF network. We identified a community of HMG Co-A reductase inhibitors (statins) composed of fluvastatin, lovastatin, and simvastatin (C30; Fig. 3). These are a class of cholesterol-lowering drugs, and which have been found to reduce cardiovascular disease. Interestingly, parthenolide clusters with this community, and has been experimentally observed to inhibit the NF-Kb inflammatory pathway in atherosclerosis and in colon cancer (32,33), thereby suggesting similar behavior to statin compounds. We also classified correctly drugs with unannotated mechanisms/targets in CTRPv2 such as ifosfamide, cyclophosphamide and procarbazine (C17; Fig. 3) which are known alkylating agents (ATC code: L01A; Fig. 4). Furthermore, this was also true for docetaxel and paclitaxel (C21; Fig. 3), two taxanes drugs with unannotated target in CTRPv2 although known as sharing similar MoA (ATC code: L01CD; Fig. 4).

Our integrative drug taxonomy was also able to identify targets for drugs with poorly understood mechanisms and to infer new mechanism for other drugs. Community C15, for example, contains tigecycline and Col-3 (Fig. 3); both are derivatives of the antibiotic tetracycline (34). Tigecycline is an approved drug, however its target is not characterized in humans. Col-3 showed antitumor activities by inhibiting matrix metalloproteinase (34).

Interestingly, tosedostat (CHR-2797), a metalloenzyme inhibitor with antiproliferative potential (35), is also a member of this community. Another drug in this community, phloretin, is a natural compound with uncharacterized targets and has been recently shown to deregulate matrix metalloproteinases at both gene and protein levels (36). Our results suggest that matrix metalloproteinases would be the preferred target for drugs in this community. DNF also consolidated previous findings for drugs that may serve as tubulin polymerization disruptors, and which have not been previously classified as such. We identified a community of three drugs (C49; Fig. 3) in which LY2183240, and YK-4-279 have been recently identified to decrease alpha-tubulin levels (14)._Tivantinib, a c-MET tyrosine kinase inhibitor, also blocked microtubule polymerization (37). Interestingly, this community is tightly connected to known microtubule perturbagens (C21; Fig. 3).

The identification of community C33 including the BCL-2 inhibitors ABT-737 and navitoclax (Fig. 3) concurs with the study of Rees et al. (38) where a high expression of BCL-2 was reported to confer sensitivity to these two drugs. This was not the case for another BCL-2 inhibitor, obatoclax, for which they proposed that its metabolic modification in cells impacts its interaction with BCL-2 proteins, therefore reducing its potency. We showed indeed that obatoclax did not cluster with the other two BCL-2 inhibitors (ABT-737 and navitoclax). Such an example demonstrates how the structural and sensitivity profiles of these two BCL-2 inhibitors are largely coherent, as opposed to obatoclax, which previously showed off-target effects compared to ABT-737 (39). This provides a good evidence to consider sensitivity profiles when developing new potent and specific BCL-2 inhibitors.

Our results also suggest the existence of “super communities”, that are a grouping of several communities contributing to a larger, systems-based MoA. An example is provided by the tightly connected communities C3, C21, C23, C43 (Fig. 3). One of these communities (C3: Alvocidib, PHA-793887 and staurosporine) includes well-characterized inhibitors of cyclin dependant kinases (CDKs) that are known to be major regulators of the cell cycle. BMS-345541 for example, which also clusters with drugs in C3, is an ATP non-competitive allosteric inhibitor of CDK (40). Those compounds are positioned close in the community network to topoisomerase I and II inhibitors (C43: SN-38, topotecan, etoposide, teniposide), microtubule dynamics perturbators (C21: paclitaxel, docetaxel, vincristine, parbendazole) and polo-like kinase inhibitors (C23: GSK461364, GW843682X). Iorio *et al.* reported that the similarity between CDK inhibitors and the other DNA-damaging agents is mediated through a p21 induction, which explains the interconnection and rationale of similar transcriptional and sensitivity effects of these regulators of cell cycle progression (7).

DNF based on NCI60 sensitivity information enabled identification of three tightly connected drug communities: C2, C5, C32 (Supplementary Fig. 6). These communities contain a number of compounds which showed antitumor activity by generating reactive oxygen species (ROS) (communities C2: elesclomol, fenretinide; C5: ethacrynic acid, curcumin; C32: bortezomib, menadione). Interestingly, ethacrynic acid, an FDA approved drug indicated for hypertension, clustered with curcumin, a component of turmeric. Ethacrynic acid inhibits glutathione S-transferase (GSTP1) and induced mitochondrial dependant apoptosis through generation of ROS and induction of caspases (41). Curcumin showed antitumor activity by production of ROS and promotion of apoptotic signaling. Thus, our results suggest that GSTP1 could be a potential target of the widely-used natural compound curcumin. Interestingly, some of the identified communities using NCI60, such as the tight connection between BRAF/MEK inhibitor drugs (C42; Supplementary Fig. 6), had also been identified in our original analysis using CTRPv2 sensitivity profiles. This demonstrates a high degree of specificity of drug-target associations across cell lines and experimental platforms, which is crucial for precision medicine.

DISCUSSION

Identification of drug mechanisms for uncharacterized compounds is instrumental for determination of on-targets responsible of pharmacological effects, and off-targets associated with unexpected physiological effects. Shortcomings of current approaches include a degree of reliance on pharmacological, biochemical, and functional annotations that pertain to existing, well-characterized drugs, and which may not be applicable towards prediction of a new small compounds (12,42). To address this issue, we developed DNF, a high-throughput, systematic and unbiased approach that relies on basic and complementary drug characteristics, and harnesses this integrative classification to provide a global picture of drug relationships.

In this analysis, we have conducted to our knowledge the first large-scale integration of high-throughput drug-related data that encompasses drug structure, sensitivity and perturbation signatures towards elucidating drug-drug relationships. We have removed any reliance on existing annotations that pertain to existing drugs, such as drug-target classifications or knowledge of the anatomical and organ system targeted by the drug compounds. As a consequence, we developed a scalable approach that relies only on basic drug information, making DNF both flexible for comprehensive drug classification, but also adaptable for testing new experimental compounds with minimal information (Fig. 5).

Our high-throughput drug similarity network (DNF) capitalized upon our integrative Similarity Network Fusion method (15) to construct a global drug taxonomy based on the fusion of drug structure, sensitivity, and perturbation data. The construction of drug-similarity networks and their subsequent fusion allows us to harness the complementary nature of several drug datasets to infer an integrative drug taxonomy. Testing how well different drug taxonomies can correctly predict drug targets (Fig. 2A-B) and anatomical (ATC) drug classifications (Fig. 2C-D) indicates that DNF constitutes a significant improvement towards drug classification, compared to single data type analyses using either drug sensitivity, structure, or perturbation information alone. The marked improvement of drug classification using the similarity network-based method is sustained even with the use of a different source of *in vitro* sensitivity data (NCI60;

Supplementary Fig. 5) to generate the DNF similarity matrix. Indeed, testing DNF using the NCI60 sensitivity information reveals that our integrative taxonomy consistently supersedes single-layer drug taxonomies across the target and ATC benchmarks despite the reduced number of cell lines used for sensitivity screening (Supplementary Fig. 5). While DNF was not intended as a supervised approach to predict drug targets and ATC classifications *per se*, the ability to efficiently predict different drug classes provide credence to using our novel similarity network-based method to discover drug relationships. DNF is the only method that is consistently the top performer, while each single layer taxonomy performs well in only a few cases (Fig. 2 and Supplementary Fig. 5). Overall, these observations indicate that our integrative approach succeeds in combining several drug data types into a single comprehensive network that efficiently leverages the spectrum of the underlying data.

Our explorative analysis stresses the importance of drug sensitivity information as an important asset for prediction of drug-target associations (Fig. 2A,B and Supplementary Fig. 5A,B). Such findings support the relevance of pharmacological assays to predict drug targets, and underscore the comprehensive nature of the CTRPv2 dataset (860 cell lines screened with 16 drug concentrations, tested in duplicate) (14). Similarity, we have observed a priority for drug structure information towards prediction of ATC drug classification (Fig. 2C,D, Supplementary Fig. 5C,D). We have also conducted a quantitative comparison of the predictive performance of DNF against four published drug classification algorithms (3,7,8,43) that could be directly compared to the DNF approach (Supplementary Methods, Supplementary Fig. 9, Supplementary Table 5). DNF outperforms the published methods in all cases (Supplementary Fig. 9).

We performed a community detection analysis on our integrated drug networks and highlighted many cases of drug clusters (drug communities) with known MoA, thereby capturing context-specific features associated to drug sensitivity and transcriptomic profiles in cancer cells (Fig. 3, Supplementary Fig. 6). These cases serve both as ‘positive controls’ as well as validation of our methods. We demonstrated that DNF correctly identified communities of BRAF inhibitors, and MAPK/MEK inhibitors, among others in the CTRPv2 taxonomy (Fig. 3). We also highlight several well-characterized drug communities using the DNF taxonomy based on the NCI60 dataset (Supplementary Fig. 6). Our quantitative assessment of the clusters identified from the DNF network revealed several communities that were significantly enriched for drug targets as well as ATC classes (Fig. 4, Supplementary Fig. 7), which underscores the biological relevance of the drug clusters that we had identified. Collectively, these findings support the relevance of DNF for the classification of drug relationships across several classes of drugs. Importantly, we compared the DNF drug communities to previously published data, and found an overlap between our results and clinical observations. For example, ibrutinib, which is a Bruton tyrosine kinase inhibitor (BTK) approved for the treatment of Mantle cell lymphoma and chronic lymphocytic leukemia, clustered with the known EGFR inhibitors (C2: erlotinib, gefitinib, afatinib and others). The effect of ibrutinib in EGFR Mutant Non-Small Cell Lung Cancer has been reported in a recent clinical trial (ClinicalTrials.gov Identifier: NCT02321540). This was also the case for MGCD265, a Met inhibitor, which clustered with most of the VEGFR (vascular endothelial growth factor receptor) inhibitors (C50: pazopanib, cediranib and others). In this community, pazopanib is the only FDA approved drug for the treatment of renal cell carcinoma. Recent evidence shows that the clinical drug candidate MGCD265 can be used to treat renal malignancies (ClinicalTrials.gov Identifier: NCT00697632).

Since polypharmacology provides great opportunities for drug repurposing through simultaneous blockade of multiple targets or pathways (44), we investigated whether our integrative drug taxonomy can be used to identify compounds with multiple targets. For example, we showed that the two specific ALK inhibitors in CTRPv2 (crizotinib and TAE684) cluster closely to the IGF-1R inhibitors (Linsitinib and BMS536924). This finding concurs with the results of Seashore-Ludlow et al. showing that TAE684 and crizotinib contribute to IGF-1R inhibition in neuroblastoma cancer cell lines (45). Our approach is able to capture other relevant off-target effects by showing that a known c-MET kinase inhibitor (tivantinib) clusters with microtubule polymerization inhibitors, which was confirmed by Katayama et al. (37).

The current availability of sensitivity data and its overlap with drug perturbation and drug structure information remains a limiting factor. LINCS L1000 (46) was favored over CMAP (9) for drug perturbation datasets as it contains a larger set of drug compounds (L1000=20,326 v.s. CMAP=1,309). Similarly we used the pharmacological profiles from the CTRPv2 (14) and NCI60 datasets because they contain a large number of drugs compared to GDSC (47) or CCLE (48) or other smaller datasets (CTRPv2=481 and NCI60=49,938 v.s. GDSC=251 and CCLE=24). Our taxonomy is currently composed of nearly 240 drugs with drug structure, perturbationa nd sensitivity data available. The number of drugs with multiple data layer is likely to increase, as the LINCS L1000 is expanding at fast pace and new sensitivity data are frequently released (47,49). Therefore, we developed DNF with scalability in mind as we expect the number of drugs in our network to increase continuously over time. Recognizing that the exploration of large-scale drug similarity network is challenging, we developed the DNF web-application to interactively display the CTRPv2 and NCI60 drug similarity networks (dnf.pmgenomics.ca). We leveraged the cytoscape.js library (28) to allow users to easily navigate drug communities and investigate the drugs within each community and their similarities. In particular, users can use the DNF web-application to determine, for a given drug-drug similarity, which data layer contributed most to the similarity score, therefore providing detailed information about drug communities detected using our integrative taxonomy.

The recent release of large-scale pharmacogenomic datasets, such as those generated within the CTRPv2 and LINCS L1000 projects, provides a unique opportunity to further investigate the effects of approved and experimental drugs on cancer models and their potential mode of action. Methods leveraging DNF in combination with the wealth of molecular profiles from cancer cell lines and patient tumors hold the potential to identify robust biomarkers and improve drug matching for individual patients, which would constitute a major step towards precision medicine and drug repurposing in cancer.

This study has several potential limitations. First, the number of drugs with all data layers available is limited by the small overlap between drug sensitivity and perturbation datasets. As these datasets grow, the DNF taxonomy will expand proportionally. Second, the L1000 is a relatively new dataset and there is no consensus yet on how to best normalize the data and compute the transcriptomic perturbation signature for each drug. In this study, we used the gene expression data as normalized by the Broad Institute (QNORM, level 3 data) and the signature model implemented in PharmacoGx (21); however we recognize that other pipelines could be used to mine the L1000 dataset and potentially improve DNF, such as the multilevel mixed-effect model recently published by Vis et al. (50). While it remains unclear that the use of more complex models will improve the drug perturbation signatures, our study provides a benchmark to assess the benefits of such approaches. Third, the low number of cell lines in specific tissue types in L1000 prevented us from creating tissue-specific integrative taxonomies to better explore the molecular context of drug MoA. This limitation could be overcome with the availability of perturbation profiles for more cell lines in the future. Fourth, given that SuperPred and DrugE-Rank websites only report the top predictions but not the full list of drugs with similar targets or ATC codes, we computed the distance between these partial rankings to compute similarities between drugs (Supplementary Methods). Similarly, a direct comparison between DNF and the taxonomy inferred using Iorio *et al.* and Iskar *et al.* methods is not possible due their reliance on CMAP. The adaptation of these methods for the L1000 dataset was challenging due to the reduced set of 1000 genes. Consequently, the results of the comparison between DNF and published methods should be interpreted cautiously. Finally, while Pearson's correlation coefficient have been shown to be suitable to detect similarities between drug sensitivity profiles for drugs with broad or narrow inhibitory effects (51,52), there is currently no consensus on the best similarity metric to compare pharmacological profiles and new methods are under active development (53). We also had to use different drug sensitivity measures for CTRPv2 and NCI60 as both projects released different types of pharmacological profiles: CTRPv2 reported percentage of cell viability while NCI60 reported percentage of growth inhibition controlled for population doubling time. Despite these heterogeneous drugs sensitivity data, we observed similar communities for the drugs in common between the two sensitivity datasets, supporting the robustness of DNF.

In conclusion, we developed an integrative taxonomy inference approach, DNF, leveraging the largest quantitative compendiums of structural information, pharmacological phenotypes and transcriptional perturbation profiles to date. We used DNF to conduct a comparative assessment between our integrative taxonomy, and single-layer drug taxonomies, as well as published methods used to predict drug targets or functional annotations. Our results support the superiority of DNF towards drug classification, and also highlights singular data types that are pivotal towards prediction of drug categories in terms of anatomical classification as well as drug-target relationships. Overall, the DNF taxonomy has produced the best performance, most consistent classification of drugs for multiple functional classes. The comprehensive picture of drug-drug relationships produced by DNF has also succeeded in predicting new drug MoA. The integrative DNF taxonomy has the potential to serve as a solid framework for future studies involving inference of MoA of new, uncharacterized compounds, which represents a major challenge in drug development for precision medicine.

ACKNOWLEDGEMENTS

The authors would like to thank Jacques Archambault for sharing his expertise in pharmacology and supporting Nehme El-Hachem financially, and Dr. Fei-Fei Liu for supporting Laleh Soltan Ghoraie financially. We thank John Chen and Ali Nawed for their contribution to the DNF web-application. They also thank Dr. Shanfeng Zhu for running the DrugE-Rank prediction algorithm on the set of drugs analyzed in this study.

REFERENCES

1. Chen B, Butte AJ. Leveraging big data to transform target selection and drug discovery. Clin Pharmacol Ther. 2016;99:285–97.

2. Sheridan RP, Kearsley SK. Why do we need so many chemical similarity search methods? Drug Discov Today. 2002;7:903–11.

3. Dunkel M, Günther S, Ahmed J, Wittig B, Preissner R. SuperPred: drug classification and target prediction. Nucleic Acids Res. 2008;36:W55–9.

4. Martin YC, Kofron JL, Traphagen LM. Do structurally similar molecules have similar biological activity? J Med Chem. 2002;45:4350–8.

5. Khan SA, Virtanen S, Kallioniemi OP, Wennerberg K, Poso A, Kaski S. Identification of structural features in chemicals associated with cancer drug response: a systematic data-driven analysis. Bioinformatics. 2014;30:i497–504.

6. Chen B, Greenside P, Paik H, Sirota M, Hadley D, Butte AJ. Relating Chemical Structure to Cellular Response: An Integrative Analysis of Gene Expression, Bioactivity, and Structural Data Across 11,000 Compounds. CPT Pharmacometrics Syst Pharmacol. 2015;4:576–84.

7. Iorio F, Bosotti R, Scacheri E, Belcastro V, Mithbaokar P, Ferriero R, et al. Discovery of drug mode of action and drug repositioning from transcriptional responses. Proc Natl Acad Sci U S A. 2010;107:14621–6.

8. Iskar M, Campillos M, Kuhn M, Jensen LJ, van Noort V, Bork P. Drug-induced regulation of target expression. PLoS Comput Biol [Internet]. 2010;6. Available from: http://dx.doi.org/10.1371/journal.pcbi.1000925

9. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science. 2006;313:1929–35.

10. Campillos M, Kuhn M, Gavin A-C, Jensen LJ, Bork P. Drug target identification using side-effect similarity. Science. 2008;321:263–6.

11. Kuhn M, Al Banchaabouchi M, Campillos M, Jensen LJ, Gross C, Gavin A-C, et al. Systematic identification of proteins that elicit drug side effects. Mol Syst Biol. 2013;9:663.

12. Napolitano F, Zhao Y, Moreira VM, Tagliaferri R, Kere J, D'Amato M, et al. Drug repositioning: a machine-learning approach through data integration. J Cheminform. 2013;5:30.

13. Shoemaker RH. The NCI60 human tumour cell line anticancer drug screen. Nat Rev Cancer. 2006;6:813–23.

14. Seashore-Ludlow B, Rees MG, Cheah JH, Cokol M, Price EV, Coletti ME, et al. Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset. Cancer Discov. 2015;5:1210–23.

15. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, et al. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods. 2014;11:333–7.

16. Shankavaram UT, Varma S, Kane D, Sunshine M, Chary KK, Reinhold WC, et al. CellMiner: a relational database and query tool for the NCI-60 cancer cell lines. BMC

Genomics. 2009;10:277–277.

17. NIH, Broad Institute. The LINCS Connectivity Map Project [Internet]. The LINCS Connectivity Map Project. 2015 [cited 2016]. Available from: https://clue.io/

18. Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, et al. PubChem Substance and Compound databases. Nucleic Acids Res. 2016;44:D1202–13.

19. Tanimoto TT. An Elementary Mathematical Theory of Classification and Prediction. International Business Machines Corporation; 1958.

20. Guha R, Others. Chemical informatics functionality in R. J Stat Softw. 2007;18:1–16.

21. Smirnov P, Safikhani Z, El-Hachem N, Wang D, She A, Olsen C, et al. PharmacoGx: An R package for analysis of large pharmacogenomic datasets. Bioinformatics [Internet]. 2015; Available from: http://dx.doi.org/10.1093/bioinformatics/btv723

22. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, et al. DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res. 2008;36:D901–6.

23. Nahler G. Anatomical therapeutic chemical classification system (ATC). In: Nahler G, editor. Dictionary of Pharmaceutical Medicine. Springer Vienna; 2009. page 8–8.

24. Bento AP, Gaulton A, Hersey A, Bellis LJ, Chambers J, Davies M, et al. The ChEMBL bioactivity database: an update. Nucleic Acids Res. 2014;42:D1083–90.

25. Cheng J, Xie Q, Kumar V, Hurle M, Freudenberg JM, Yang L, et al. Evaluation of analytical methods for connectivity map data. Pac Symp Biocomput. 2013;5–16.

26. Schröder MS, Culhane AC, Quackenbush J, Haibe-Kains B. survcomp: an R/Bioconductor

package for performance assessment and comparison of survival models. Bioinformatics. 2011;27:3206–8.

27. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13:2498–504.

28. Franz M, Lopes CT, Huck G, Dong Y, Sumer O, Bader GD. Cytoscape.js: a graph theory library for visualisation and analysis. Bioinformatics. 2016;32:309–11.

29. Sandve GK, Nekrutenko A, Taylor J, Hovig E. Ten simple rules for reproducible computational research. PLoS Comput Biol. 2013;9:e1003285.

30. Frey BJ, Dueck D. Clustering by passing messages between data points. Science. 2007;315:972–6.

31. Friday BB, Yu C, Dy GK, Smith PD, Wang L, Thibodeau SN, et al. BRAF V600E disrupts AZD6244-induced abrogation of negative feedback pathways between extracellular signal-regulated kinase and Raf proteins. Cancer Res. 2008;68:6145–53.

32. Riganti C, Doublier S, Costamagna C, Aldieri E, Pescarmona G, Ghigo D, et al. Activation of nuclear factor-kappa B pathway by simvastatin and RhoA silencing increases doxorubicin cytotoxicity in human colon cancer HT29 cells. Mol Pharmacol. 2008;74:476–84.

33. López-Franco O, Hernández-Vargas P, Ortiz-Muñoz G, Sanjuán G, Suzuki Y, Ortega L, et al. Parthenolide modulates the NF-kappaB-mediated inflammatory responses in experimental atherosclerosis. Arterioscler Thromb Vasc Biol. 2006;26:1864–70.

34. Syed S, Takimoto C, Hidalgo M, Rizzo J, Kuhn JG, Hammond LA, et al. A phase I and

pharmacokinetic study of Col-3 (Metastat), an oral tetracycline derivative with potent matrix metalloproteinase and antitumor properties. Clin Cancer Res. 2004;10:6512–21.

35. Krige D, Needham LA, Bawden LJ, Flores N, Farmer H, Miles LEC, et al. CHR-2797: an antiproliferative aminopeptidase inhibitor that leads to amino acid deprivation in human leukemic cells. Cancer Res. 2008;68:6669–79.

36. Ma L, Wang R, Nan Y, Li W, Wang Q, Jin F. Phloretin exhibits an anticancer effect and enhances the anticancer ability of cisplatin on non-small cell lung cancer cell lines by regulating expression of apoptotic pathways and matrix metalloproteinases. Int J Oncol. 2016;48:843–53.

37. Katayama R, Aoyama A, Yamori T, Qi J, Oh-hara T, Song Y, et al. Cytotoxic activity of tivantinib (ARQ 197) is not due solely to c-MET inhibition. Cancer Res. 2013;73:3087–96.

38. Rees MG, Seashore-Ludlow B, Cheah JH, Adams DJ, Price EV, Gill S, et al. Correlating chemical sensitivity and basal gene expression reveals mechanism of action. Nat Chem Biol. 2016;12:109–16.

39. Vogler M, Weber K, Dinsdale D, Schmitz I, Schulze-Osthoff K, Dyer MJS, et al. Different forms of cell death induced by putative BCL2 inhibitors. Cell Death Differ. 2009;16:1030–9.

40. Bogoyevitch MA, Fairlie DP. A new paradigm for protein kinase inhibition: blocking phosphorylation without directly targeting ATP binding. Drug Discov Today. 2007;12:622–33.

41. Wang R, Liu C, Xia L, Zhao G, Gabrilove J, Waxman S, et al. Ethacrynic Acid and a Derivative Enhance Apoptosis in Arsenic Trioxide--Treated Myeloid Leukemia and Lymphoma Cells: The Role of Glutathione S-Transferase P1-1. Clin Cancer Res. AACR; 2012;18:6690–701.

42. Ma'ayan A, Rouillard AD, Clark NR, Wang Z, Duan Q, Kou Y. Lean Big Data integration in systems biology and systems pharmacology. Trends Pharmacol Sci. 2014;35:450–60.

43. Yuan Q, Gao J, Wu D, Zhang S, Mamitsuka H, Zhu S. DrugE-Rank: improving drug–target interaction prediction of new candidate drugs or targets by ensemble learning to rank. Bioinformatics. 2016;32:i18–27.

44. Antolin AA, Workman P, Mestres J, Al-Lazikani B. Polypharmacology in Precision Oncology: Current Applications and Future Prospects. Curr Pharm Des. 2016;22:6935–45.

45. Seashore-Ludlow B, Rees MG, Cheah JH, Cokol M, Price EV, Coletti ME, et al. Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset. Cancer Discov. 2015;5:1210–23.

46. Duan Q, Flynn C, Niepel M, Hafner M, Muhlich JL, Fernandez NF, et al. LINCS Canvas Browser: interactive web app to query, browse and interrogate LINCS L1000 gene expression signatures. Nucleic Acids Res. 2014;42:W449–60.

47. Iorio F, Knijnenburg TA, Vis DJ, Bignell GR, Menden MP, Schubert M, et al. A Landscape of Pharmacogenomic Interactions in Cancer. Cell [Internet]. 2016; Available from: http://dx.doi.org/10.1016/j.cell.2016.06.017

48. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012;483:603–7.

49. Haverty PM, Lin E, Tan J, Yu Y, Lam B, Lianoglou S, et al. Reproducible pharmacogenomic profiling of cancer cell line panels. Nature. 2016;533:333–7.

50. Vis DJ, Bombardelli L, Lightfoot H, Iorio F, Garnett MJ, Wessels LF. Multilevel models

improve precision and speed of IC50 estimates. Pharmacogenomics. 2016;17:691–700.

51. Safikhani Z, Smirnov P, Freeman M, El-Hachem N, She A, Rene Q, et al. Revisiting inconsistency in large pharmacogenomic studies. F1000Res. 2016;5:2333.

52. Cancer Cell Line Encyclopedia Consortium, Genomics of Drug Sensitivity in Cancer Consortium. Pharmacogenomic agreement between two cancer cell line data sets. Nature. 2015;528:84–7.

53. Safikhani Z, El-Hachem N, Smirnov P, Freeman M, Goldenberg A, Birkbak NJ, et al. Consistency in drug response profiling: reply. Nature. 2016;540:E6–8.

FIGURE LEGENDS

Figure 1: Schematic representation of the SNF method and its use in integrating of different types of drug information. Datasets representing drug similarity, drug sensitivity, and drug perturbation profiles are first converted into drug-drug similarity matrices. Similarity matrices are fully integrated within the SNF method to generate a large-scale, multi-tier, Drug Fusion Network (DNF) taxonomy of drug-drug relationships.

Figure 2: Validation of the DNF taxonomy using CTRPv2 sensitivity data and single dataset taxonomies against the ATC and Drug-target benchmarks. ROC and PR curves are shown for each of the taxonomies generated with the CTRPv2 sensitivity dataset, tested against ATC annotations and drug-target information from CHEMBL or internal benchmarks. Lines representing random (“rand”) classifications are drawn in grey for both ROC and PR curves. **(A)** ROC curve against drug-targets **(B)** PR curve against drug-targets **(C)** ROC curve against ATC drug classifications **(D)** PR curve against ATC drug classifications

Figure 3: Network representation of 53 exemplar drugs that are representative of the drug communities identified by the DNF taxonomy using CTRPv2 sensitivity data. Each node represents the exemplar drugs, and node sizes reflect the size of the drug community represented by the exemplar node. Nodes are colored to reflect shared MoA as determined using known drug targets. Communities sharing similar MoA and proximity in the network are highlighted, with the community number indicated next to each community. Drug communities pertaining to the super-community are labelled in red.

Figure 4: Enrichment of Drug Communities of the DNF taxonomy (using CTRPv2

sensitivity data). A total of 53 communities were tested for enrichment against drug target annotations from the CTRPv2 data and ATC annotations from ChEMBL (see methods). Fisher's exact test was performed between all the drugs in each community versus all drugs attributed to a specific drug target or ATC class, and corrected for multiple testing (FDR correction). **(A)** Enrichment of communities for Drug target annotations, with -log10 FDR values indicated in the heat map, which has been reduced to show significantly enriched communities. Communities are labelled by community number as determined by the affinity propagation clustering algorithm. **(B)** Enrichment of communities for ATC classes, with -log10 FDR values indicated in the heat map, which has been reduced to show significantly enriched communities. Communities are labelled by community number as determined by the APC algorithm.

Figure 5: Schematic of the adaptability of DNF towards prediction of new experimental compounds. A user can cluster his own drug (e.g., bioactive molecule with uncharacterized MoA) in DNF to characterize its mechanism of action by providing the following data: (*i*) the chemical structure as a canonical SMILES string; (*ii*) list of genes perturbed by the chemical compound, in at least 3 cell lines (genome-wide transcriptional studies are preferred to get a maximum overlap with the set of 1000 landmark genes); (*iii*) sensitivity profiles (e.g., area under the drug-dose response curve), including measurements of cell growth in culture medium in 9 concentrations for at least 50 cell lines from multiple tissue types. By adding these new data to our current datasets, a user will be able to assess in which community their drug of interest clusters and subsequently infer its potential MoA. No information regarding the compound’s target or anatomical therapeutic classification is required for DNF taxonomy.

FIGURE 1

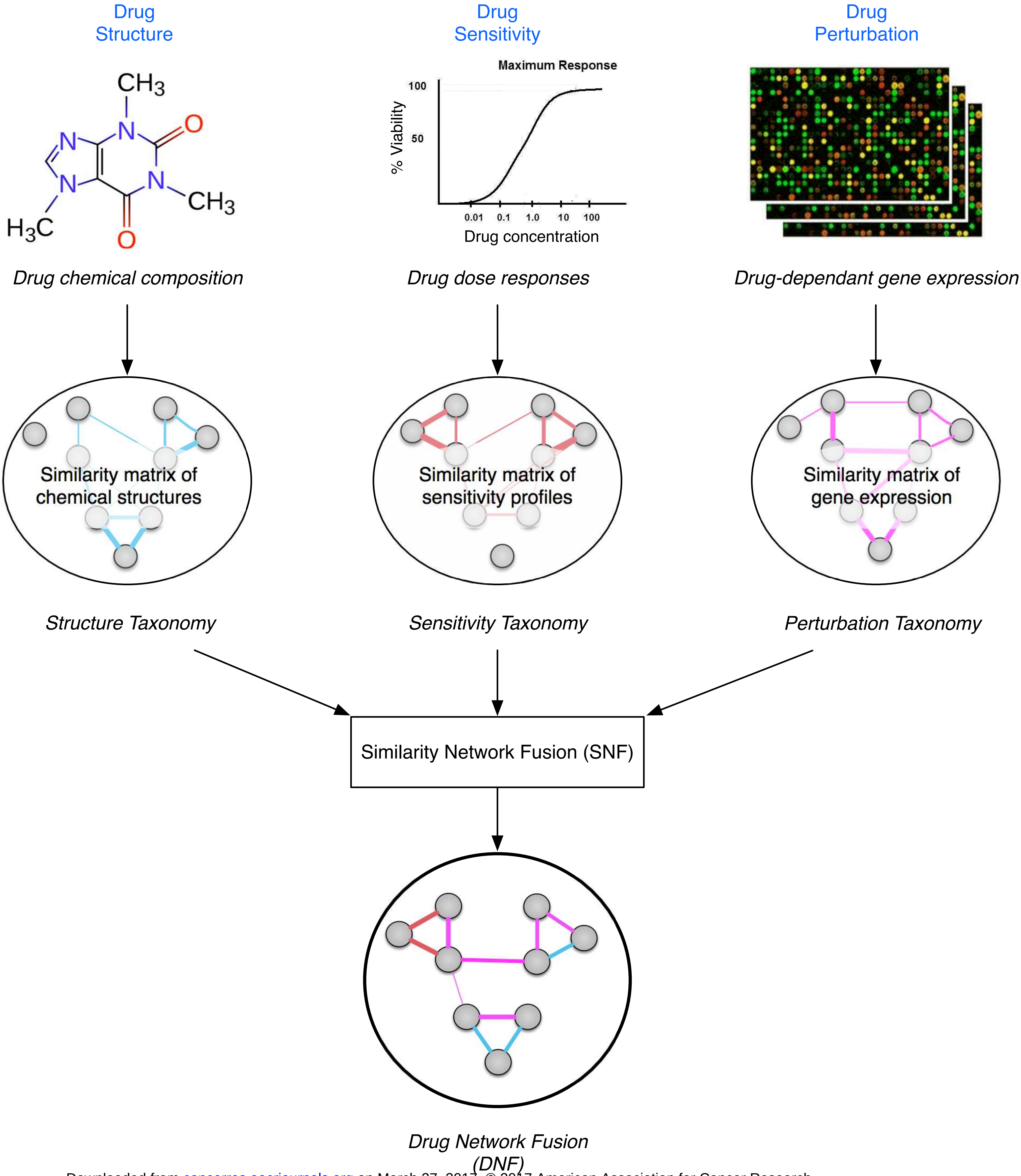
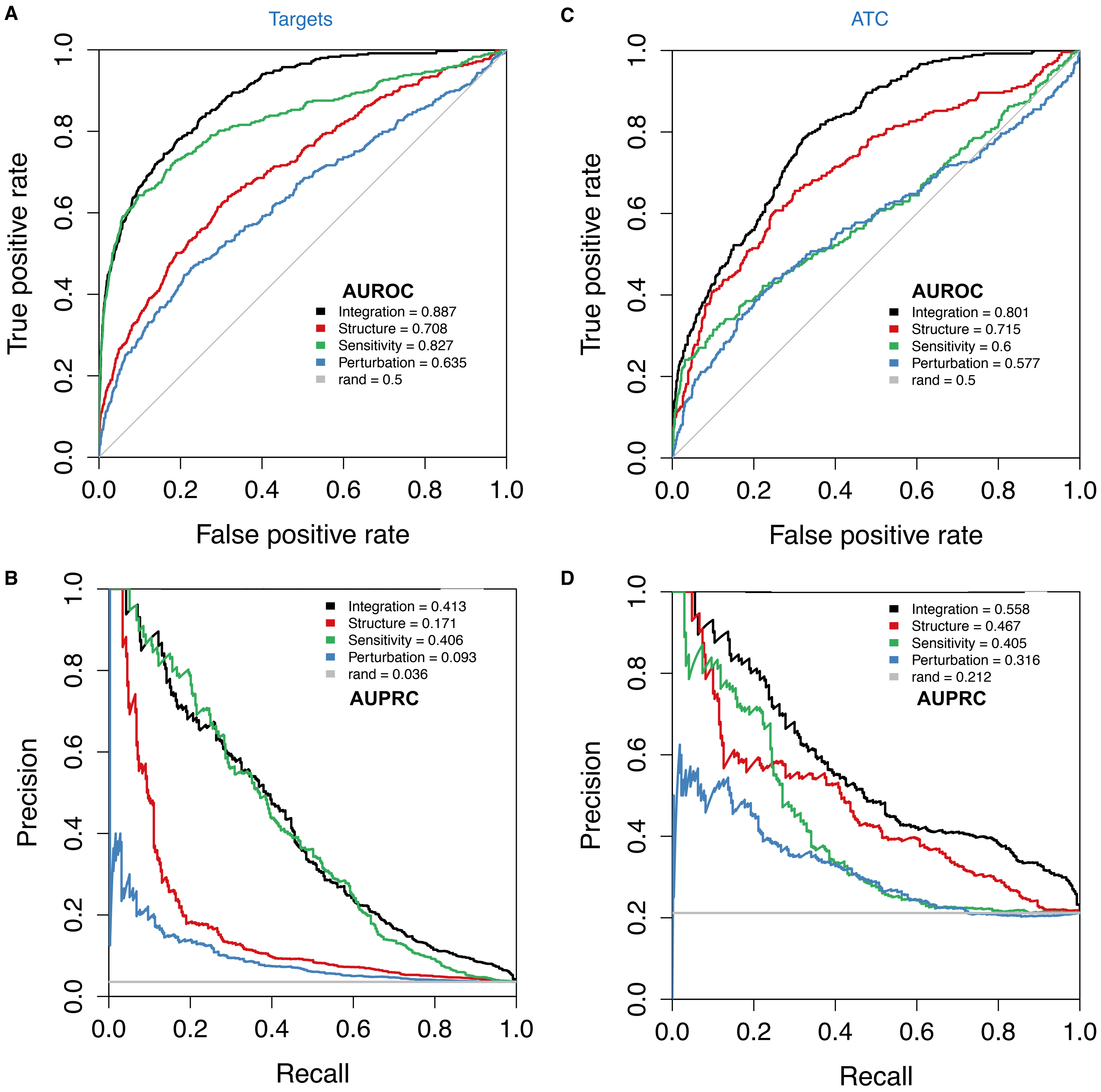

Drug Structure
Drug Sensitivity
Drug Perturbation
CH3
N
O
N
N
N
CH3
H3C
O
Maximum Response
100
50
% Viability
0.01
0.1
1.0
10
100
Drug concentration
Drug chemical composition
Drug dose responses
Drug-dependant gene expression
Similarity matrix of chemical structures
Similarity matrix of sensitivity profiles
Similarity matrix of gene expression
Structure Taxonomy
Sensitivity Taxonomy
Perturbation Taxonomy
Similarity Network Fusion (SNF)
Drug Network Fusion (DNF)

FIGURE 2



A
Targets
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
AUROC
Integration = 0.887
Structure = 0.708
Sensitivity = 0.827
Perturbation = 0.635
rand = 0.5
C
ATC
True positive rate
False positive rate
AUROC
Integration = 0.801
Structure = 0.715
Sensitivity = 0.6
Perturbation = 0.577
rand = 0.5
B
Precision
Recall
Integration = 0.413
Structure = 0.171
Sensitivity = 0.406
Perturbation = 0.093
rand = 0.036
AUPRC
D
Precision
Recall
Integration = 0.558
Structure = 0.467
Sensitivity = 0.405
Perturbation = 0.316
rand = 0.212
AUPRC

FIGURE 3

Author Manuscript Published OnlineFirst on March 17, 2017; DOI: 10.1158/0008-5472.CAN-17-0096
Author manuscripts have been peer reviewed and accepted for publication but have not yet been edited.

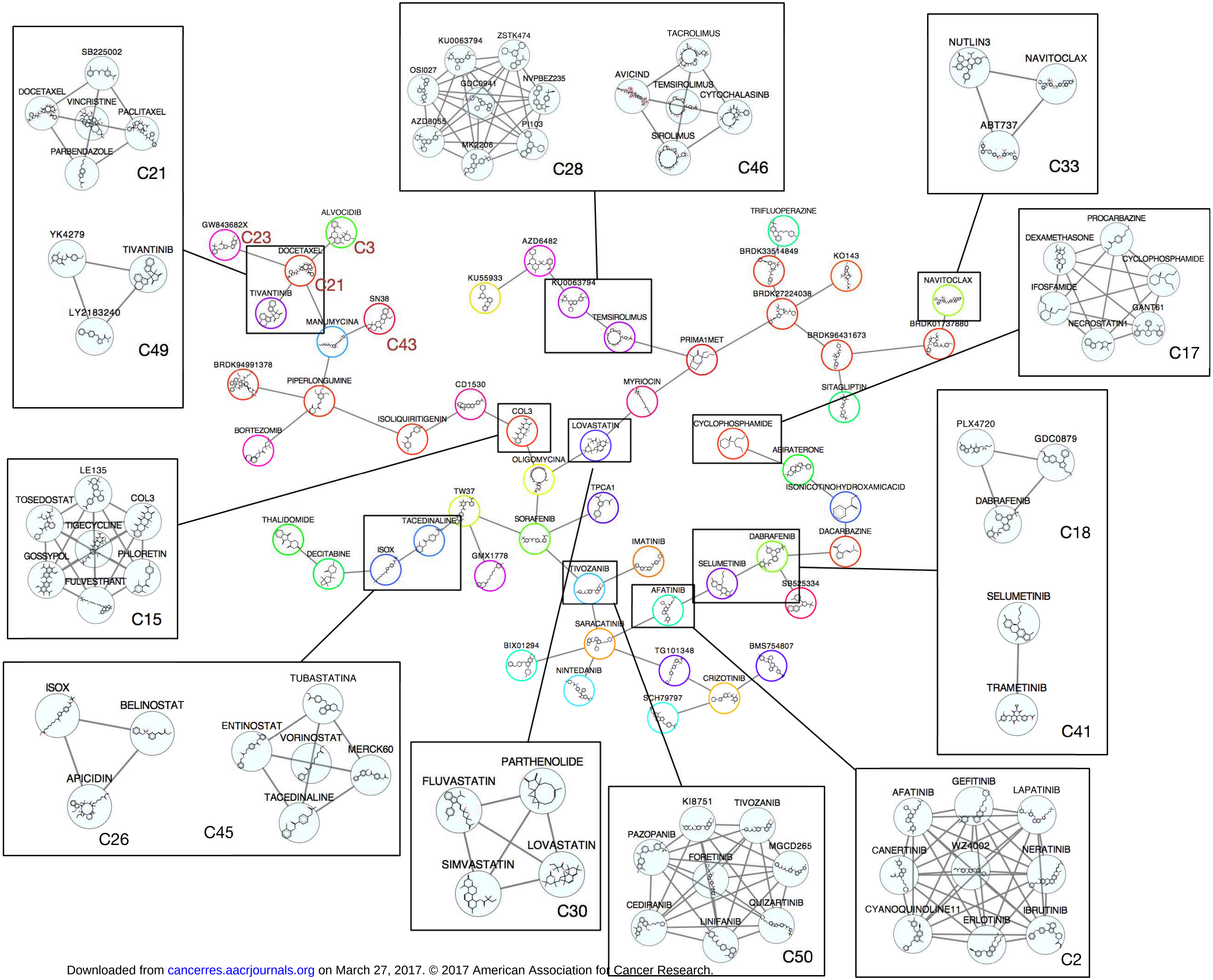


FIGURE 4

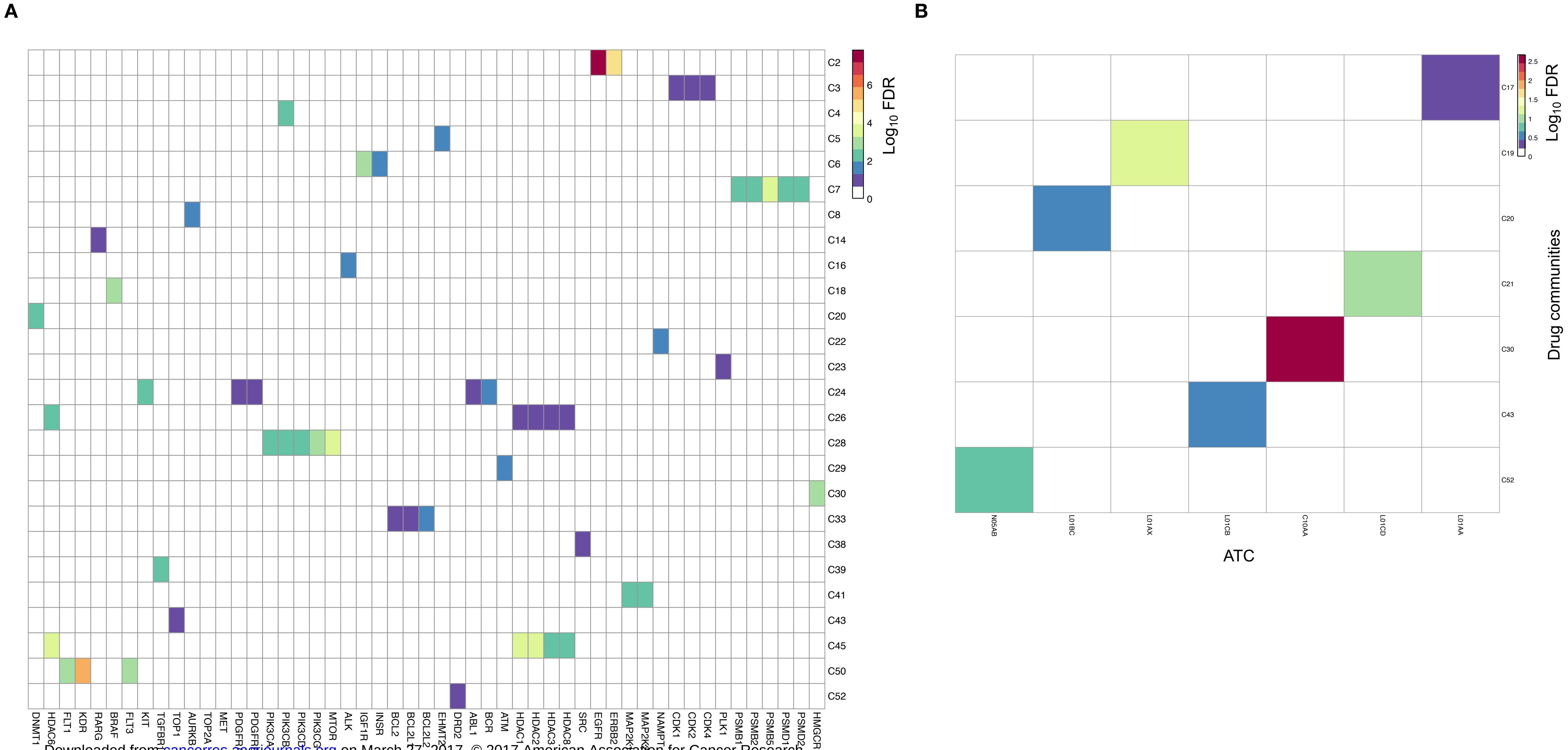
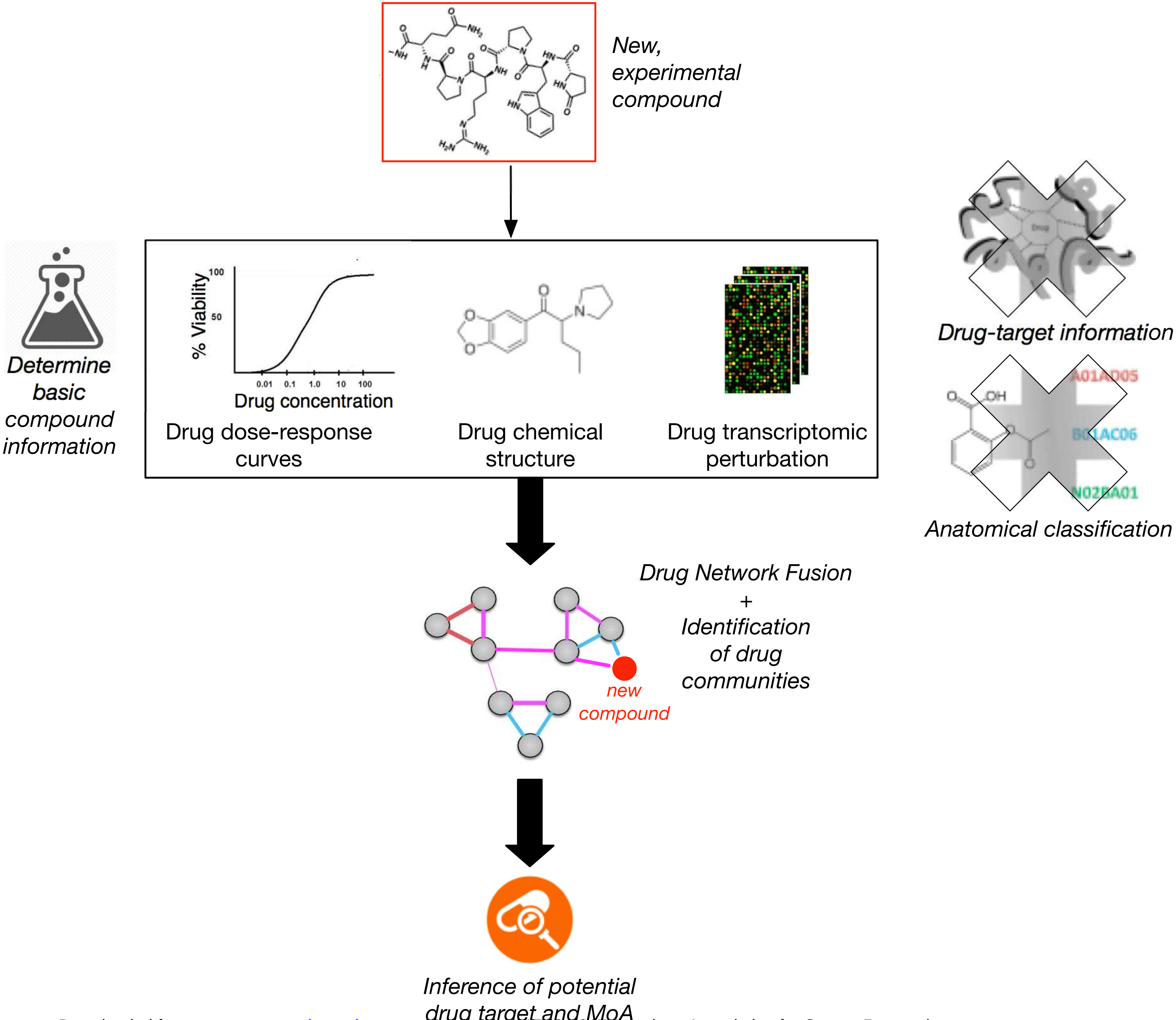

A
C2
C3
C4
C5
C6
C7
C8
C14
C16
C18
C20
C22
C23
C24
C26
C28
C29
C30
C33
C38
C39
C41
C43
C45
C50
C52
DNMT1
HDAC6
FLT1
KDR
RARG
BRAF
FLT3
KIT
TGFBR1
TOP1
AURKB
TOP2A
MET
PDGFRA
PDGFRB
PIK3CA
PIK3CB
PIK3CD
PIK3CG
MTOR
ALK
IGF1R
INSR
BCL2
BCL2L1
BCL2L2
EHMT2
DRD2
ABL1
BCR
ATM
HDAC1
HDAC2
HDAC3
HDAC8
SRC
EGFR
ERBB2
MAP2K1
MAP2K2
NAMPT
CDK1
CDK2
CDK4
PLK1
PSMB1
PSMB2
PSMB5
PSMD1
PSMD2
HMGCR
Log$_{10}$ FDR
6
4
2
0
B
C17
C19
C20
C21
C30
C43
C52
N05AB
L01BC
L01AX
L01CB
C10AA
L01CD
L01AA
ATC
Drug communities
Log$_{10}$ FDR
2.5
2
1.5
1
0.5
0

FIGURE 5


New, experimental compound
Determine basic compound information
% Viability
100
50
0.01
0.1
1.0
10
100
Drug concentration
Drug dose-response curves
Drug chemical structure
Drug transcriptomic perturbation
Drug-target information
A01AD05
B01AC06
N02BA01
Anatomical classification
Drug Network Fusion
+
Identification of drug communities
new compound
Inference of potential drug target and MoA

Author Manuscript Published OnlineFirst on March 17, 2017; DOI: 10.1158/0008-5472.CAN-17-0096
Author manuscripts have been peer reviewed and accepted for publication but have not yet been edited.

Cancer Research
The Journal of Cancer Research (1916–1930) | The American Journal of Cancer (1931–1940)

AACR American Association for Cancer Research

Integrative cancer pharmacogenomics to infer large-scale drug taxonomy

Nehme El-Hachem, Deena M.A. Gendoo, Laleh Soltan Ghoraie, et al.

Cancer Res Published OnlineFirst March 17, 2017.