Supporting Information

Iorio et al. 10.1073/pnas.1000138107

SI Text

Online Supporting Information (online SI figures, tables, and data) are available at the following URL: http://mantra.tigem.it (no login or password required).

Treatment with Test Drugs and Microarray Hybridizations. Drugs tested were chosen among well-known compounds, already present in the Connectivity Map (cMap) dataset, and new-generation compounds. They included HSP90 inhibitors *Tanespimycin* (1), *NVP-AUY922* (2), *NMS-E973* (3); Topoisomerase inhibitors *SN-38* (4) and *Doxorubicin* (5); cyclin-dependent kinases (CDKs) inhibitors *Flavopiridol* (6), *PHA-848125* (7), *PHA-690509*, and *PHA-793887* (8).

A2780 were treated with *Flavopiridol* (0.3 μM), *PHA-848125* (1 μM), *PHA-690509* (3 μM), and *PHA-793887* (1 μM), whereas MCF7 were treated with *PHA-848125* (8.5 μM), *PHA-793887* (6.0 μM), *Tanespimycin* 0.5 μM), *NVP-AUY922* (0.07 μM), *NMS-E973* (2 μM), *SN-38* (0.165 μM), and *Doxorubicin* (1.5 μM).

Additional data were collected by treating U251 and SF539 with *PHA-848125* (3 μM) (to assess the impact of the merging of data coming from different settings on the classification performances—see "Impact of Rank Merging on Performance").

Cell Lines, RNA Extraction, Treatments, and Data Preprocessing. A2780 (human ovary adenocarcinoma) and MCF7 (human mammary adenocarcinoma mammary) from European Collection of Cell Cultures were seeded in T-75 tissue culture flasks (Corning), 25.000 cells/cm^2 in RPMI medium 1640 (Gibco), pH 7.4, 10% FBS (EUROCLONE Australia-USDA approved), 2 mM L-Glutamine (Gibco), 1× penicillin–streptomycin (Gibco), and maintained in 5% CO_2 at 37 °C with 96% relative humidity. After 24 h, cells were treated with different compounds at a dose equal to 5× the IC50 for 6 h and collected using Qiagen RNeasy Lysis Buffer (Qiagen cat no. 79216). Total RNA was extracted using Qiagen Rneasy kit (Qiagen cat. no. 74104), starting from total cell lysates. The RNA was purified following manufacturer instructions. During the process, any genomic DNA contaminations were removed by DNAse treatment. Quantity and purity of the extracted RNA were assessed by spectrophotometric evaluation of light adsorbance at 260 and 280 nm; after extraction, RNA was stored at −80 °C. Biotin-labeled, fragmented cRNA probes were prepared starting from 1.5 μg of total RNA per replicate sample, using the "One-Cycle Target Labeling and Control Reagents" (Affymetrix) according to the protocols included in the Affymetrix GeneChip Expression Analysis Technical Manual (www.affymetrix.com). Samples were hybridized onto Affymetrix GeneChip® Human Genome U133 Plus 2.0 Arrays and processed as per manufacturer's instructions using "GeneChip® Hybridization, Wash, and Stain Kit" components (Affymetrix). Scanned images were first inspected for quality control (QC) using a variety of built-in QC tools from the Bioconductor package [www.bioconductor.org] of R, the open source environment for statistical analysis. Feature intensity values from scanned arrays were normalized and reduced to expression summaries using MAS5 implemented in the R statistical environment. A ranked list of genes was obtained for each compound treatment by sorting the microarray probe-set identifiers according to the differential expression values with respect to the untreated hybridization. These ranked lists were given as an input to the drug network (DN) tool.

Data are available at Gene Expression Omnibus (GEO) database, www.ncbi.nlm.nih.gov/geo (accession no. GSE18552).

Western Blot of Total MCF7 Cell Lysates. MCF7 cells were treated for 6 h with *PHA-793887*, *Doxorubicin*, or *SN-38* at a dose equal to 5× the IC50 for 6 h. Total protein lysates were resolved by SDS-PAGE, transferred onto nitrocellulose membrane (Hybond ECL GE Healthcare), and hybridized with specific antibodies: anti-p21 (BD pharmingen 556430), anti-pNPM Thr 199 (Cell signaling 3541), antitotal Rb (BD pharmingen 554136), anti-pRb Thr 821 (invitrogen 44-582G), anti-pRb Ser 807/811(Cell signaling 9308), antitotal RNA polI (millipore 05-623), anti-RNA pol II pSer5 (Santa Cruz sc-17794), anti-RNA pol II pSer2 (Abcam ab5131), and anti-MCL1 (Cell signaling 4572).

Evaluation of Autophagy. Synchronized wild-type human fibroblasts were treated with the following drugs: *Fasudil dihydrochloride* (Sigma) 10 μM, *Trifluoperazine* (Sigma) 1 μM, and *2DOG* (Sigma) 100 μM for 48 h. Following treatment, cells were lysed in cold lysis buffer (50 mM Tris-HCl, pH 7.4, 150 mM NaCl, 1% Triton X-100, 1 mM EDTA, and 0.1% SDS) in the presence of protease inhibitors (Sigma). Total proteins quantified by the Bradford method were resolved by SDS-PAGE and transferred onto PVDF membrane. Primary rabbit polyclonal anti-LC3 (Novus Biological), primary monoclonal anti-ß-actin (Sigma), and (HRP)-conjugated secondary antibodies were diluted in Tris-Buffered Saline Tween-20 1% BSA. Bands were visualized using a chemiluminescence detection system (Pierce). For immunofluorescence analysis, cells were permeabilized and incubated with primary anti-LC3 antibody and secondary fluorescent antibody (diluted in 1% BSA in PBS), mounted in glycerol/DAPI, and viewed on an epifluorescent microscope.

Drug Network Construction. At the heart of our approach is a notion of "distance" between two drugs. This is computed by combining differential gene expression profiles obtained with the same compound, but in different experimental settings, via an original rank-aggregation method (9), followed by a gene set enrichment analysis (10) whose results are combined into a distance value. The DN is then generated by considering each compound as a node and adding a weighted edge between two compounds if their similarity distance is below a given significance threshold.

Drug Distance Computation. To compute pairwise distances between drugs, we considered lists of genes ranked according to their differential expression following drug treatment, from the most up-regulated (at the top of the list) to the most down-regulated (at the bottom of the list). Observe that these lists include all of the genes that have been measured, even if not significantly differentially expressed. We merged all of ranked lists of genes obtained by treating cells with the same drug via a previously undescribed rank-aggregation algorithm (9), detailed in the next section. Once we had a single Prototype Ranked List (PRL) of genes for each drug in the dataset, we computed drug pairwise distances. The distance between two drugs A and B is computed as follows: for each of the two drugs, we composed an "optimal signature," i.e., a subset of the most differentially expressed genes in the corresponding PRL. To this aim, we extracted the top (respectively, bottom) 250 genes from the PRL of the drug. In order to assess how similar the PRLs of the two drugs are, we quantified the randomness in the distribution of the genes of the optimal signature of drug A along the PRL of drug B, and vice versa. To quantify this distribution, we used Gene Set Enrichment Analysis (GSEA) (10) using the optimal signature of drug A and the PRL of drug B, and vice versa. For example, if the

up-regulated genes (respectively, the down-regulated) in the optimal signature of drug A tend to be placed at the top (respectively, at the bottom) of the PRL of drug B, then the GSEA "enrichment score" will be high. Finally, the two "enrichment scores" (one for the optimal signature of drug A, and one for the optimal signature of drug B) are combined, ending up with the distance value between A and B.

Aggregation algorithm for Prototype Ranked List Generation. We built a PRL for each drug by aggregating all the ranked lists that have been obtained by treating cells with that drug (i.e., on different cell lines, with different concentrations, etc.). In our aggregation algorithm we made use of the following methods: a measure of the distance between two ranked lists (Spearman's Footrule), a method to merge two or more ranked lists (the Borda Merging Method), and an algorithm to obtain a single ranked list from a set of them in a hierarchical way (the Kruskal Algorithm) (11–13).

In order to describe the algorithm, we introduce the following notation:

- D: the set of all the possible permutations of microarray probe-set identifiers (MPI);
- X: a set of ranked lists of probe-set identifiers computed by sorting, in decreasing order, the genome-wide differential expression profiles obtained by treating cell lines with the same drug;
- δ: $D^2 \to N$: the *Spearman's Footrule* distance associating to each pair of ranked lists in X, a natural number quantifying the similarity between them;
- B: $D^2 \to D$: the *Borda Merging Function* associating to each pair of ranked lists in X a new ranked list obtained by merging them with the *Borda Merging Method*;

A pseudocode description of the algorithm is the following:

1. $n = |X|$
2. while $n > 1$
3. find i,j: $\delta(x_i,x_j) = \min_{p,q=1,\ldots,n:\ p\neq q}\delta(x_p,x_q)$
4. $y = B(x_i,x_j)$
5. $X = (X/\{x_i,x_j\})\cup\{y\}$
6. $n = |X|$
7. end

The input of the algorithm is X. Following the *Kruskal Algorithm* strategy, the algorithm first searches for the two ranked lists of MPI in X with the smallest Spearman's Footrule distance [line 3]. Then it merges them using the *Borda Merging Method* [line 4], obtaining the new ranked list of MPI y. In the next step [line 5], the two merged lists are removed from X and the new one is added to it. The process restarts until only one list remains in X: the *Prototype Ranked List* of the drug.

Spearman's Footrule. Let r: $P \times D \to [1,\ldots,m]$ be a function defined on the set of MPI (P) and on all the possible ranked lists of MPI D, with values in the interval $[1,\ldots,m]$, assigning to each MPI $i \in P$ its position in the ranked list $d \in D$. For the micro-array platform used in our reference dataset (the cMap), $m = 22{,}283$.

We compute the Spearman's Footrule, neglecting normalization terms, as follows:

$$\delta(x,y) = \sum_{i=1}^{m} |r(i,x) - r(i,y)|,$$

where, in our case, $x,y \in X \subseteq D$.

Borda Merging Function. The Borda Merging Function, defined as $B(x,y) = z$, $(x,y,z \in D)$, implements a majority voting system. It computes the list of values

$$P = [p_1,\ldots,p_m],$$

as follows:

$$p_i = r(i,x) + r(i,y),$$

where r is the function previously defined. Finally a new ranked list of probes z is obtained by sorting them according to their values in P, in increasing order.

Distance Between Two Drugs. Once a PRL had been obtained for each drug in the dataset, we extracted a signature $\{p,q\}$ for each of them.

To this end, we selected the top-ranked 250 genes of each PRL and the bottom-ranked 250 ones (p and q, respectively). We considered this gene signature as a synthetic descriptor summarizing the general cellular response to the drug. In other words, we isolated sets of genes that seemed to consistently vary in response to the drug across different experimental conditions (e.g., different cell lines and different dosages).

We heuristically determined the size of p and q (i.e., 250) guided by the following considerations. We tested optimal signatures of different length k and for each value of k, we computed distances among drugs and derived a drug network, always using the same distance significance threshold. We observed that the network obtained with the smallest k always contained, as a subnetwork, the networks obtained with larger k values. This means that, as the signature length k increases, the overall structure of the network does not change substantially. We chose $k = 250$ as a good compromise, which takes into account the number of considered genes, the edge density of the obtained network, and the network prediction performances (i.e., number of literature-verified connections).

Given the optimal signature of drug d, with

$$p = \{p_1,\ldots,p_n\}$$

(up-regulated genes) and

$$q = \{q_1,\ldots,q_n\}$$

(down-regulated genes), we defined as the distance between drug d and drug x the *Inverse Total Enrichment Score* (TES) of the drug d signature $\{p,q\}$, with respect to the PRL of drug x, as follows:

$$\mathrm{TES}_{dx} = 1 - \frac{\mathrm{ES}_x^p - \mathrm{ES}_x^q}{2}.$$

Here, ES_x^y (with $r \in \{p,q\}$) is the Enrichment Score of the signature (the up-regulated part and the down-regulated one, respectively) with respect to the PRL of x.

ES_x^r ranges in $[-1,1]$, it is a measure based on the Kolmogorov–Smirnov statistics, and it quantifies how much a set of genes is at the top of a ranked list (10). The closer this measure is to 1, the closer the genes are to the top of the list. The closer to -1, the closer the genes are to the bottom of the list. TES_{dx} ranges in $[0,2]$, it takes as inputs a signature $\{p,q\}$ and a list x, and it quantifies how much the genes in the p set are placed at the top of the x PRL and how much the genes in the q set are placed at the bottom. The closer these two statements are to the truth, the closer to 0 is the value of TES_{dx}.

We defined two different distance measurements among drugs as follows: Given two drugs A and B:

Average Enrichment-Score Distance: $D = \frac{\mathrm{TES}_{A,B}+\mathrm{TES}_{B,A}}{2}$,

Maximum Enrichment-Score Distance: $D = \frac{\min(\mathrm{TES}_{A,B},\mathrm{TES}_{B,A})}{2}$.

We verified that the average distance is more stringent than the maximum distance (refer to online SI Table 2), whereas the maximum distance is more sensitive to weak similarities, providing a lower precision but a larger recall (see, the *NMS-doxorubicin* example in the section "Assessment of the Classification Performance and Comparison with the cMap Online Tool," online SI Table 4).

Estimation of a Significance Threshold for the Drug Distance. Because we have a large number of pairwise distance values equal to $\binom{1309}{2} = 856{,}086$, we decided to use the empirical probability distribution function (pdf) of these data to estimate a significance threshold for the distance. Specifically, we chose as distance significance threshold value the upper bound of the 5% quantile of this empirical pdf (online SI Fig. 1).

We obtained as threshold values 0.8065 and 0.8339 (respectively, for maximum and average distances, see online SI Fig. 1 and online SI Fig. 2).

Given a pairwise distance d, the corresponding empirical p values can now be computed by dividing the number of distances less than d in the whole set of all the possible ones by the cardinality of this set (i.e., 856,086). Obviously, the empirical p values of the computed threshold levels were equal to 0.05. Because a weighted edge was assigned to each pairwise distance below the threshold, all of the edges correspond to significant distances: The smaller the distance, the higher it is in significance.

We observed that the network structure, in terms of drug communities, does not change if we chose a different significance threshold value. This happens because the community-finding algorithm uses the weighted edges (i.e., distances) to generate communities and, therefore, is not very sensitive to the addition or removal of edges, due to different choices of the distance significance threshold.

Chemical Similarity and Drug Distance. In order to test whether drugs that are found to be similar according to our method could have also been identified simply by looking at their chemical similarities, we first collected the *canonical SMILES* (Simplified Molecular Input Line Entry Specification) (14) describing the chemical structure of the cMap drugs, and we then computed chemical similarities among them. We then checked if any correlation between chemical similarity and our definition of distance was present.

A SMILES is a specification for unambiguously describing the structure of chemical molecules using short text strings. SMILES were available on the *DrugBank* database (15, 16) for 579 cMap drugs (out of 1,309).

We focused on this subset of drugs by computing $\binom{579}{2} = 167{,}331$ pairwise chemical similarities with two different methods (both working on SMILES): The first one is based on a definition of distance between molecular *electrotopological states* (17, 18), whereas the second one is based on comparisons between *extended-connectivity fingerprints* and, making use of a software tool from SciTegic©, computes a *Property Distance* inversely proportional to chemical similarity (applications can be found in refs. 19–22).

In the online SI Fig. 3, each point represents a pair of drugs for which both the SMILES were available. The first coordinate of each point is equal to the distance between the two drugs (according to our definition). The second coordinate is equal to 1 minus the electrotopological states (ESF) similarity between the SMILES of the two drugs.

As apparent, there is no significant correlation between our distance and the ESF similarity (*Pearson Correlation Coefficient* between these two measurements is equal to 0.04).

In the same way, there is no significant correlation between our definition of distance and the *extended-connectivity fingerprints Property Distance*. In online SI Fig. 4, each point represents a pair of drugs for which both the SMILES were available. The first coordinate of each point is equal to the distance between the two drugs (according to our definition). The second coordinate is equal to the Property Distance between the SMILES of the two drugs.

Also in this case, both the correlation plot and the *Pearson Correlation Coefficient* (0.05) show that there is no significant correlation between these two distances.

This is a first evidence that chemical commonalities between two drugs have no significant influences on their distance. As a matter of fact, in very few cases (i.e., points on the figure) with DN distance less that 0.5 (which is a value lower than the selected significance threshold of 0.8065), there is a tendency for chemical distance and DN distance to both be small, but for the majority of the cases (i.e., those with a DN below the 0.8065 threshold) the chemical similarity does not correlate at all with the DN distance.

In addition, also the opposite effect happens; that is, drugs with very small chemical distance have very high DN distance. Therefore, the two measures are not correlated, although there are a few cases where very small chemical distance corresponds with small DN distance.

Moreover, we measured the tendency of our *community-identification* algorithm to group together drugs that are similar by the chemical point of view. To this end we considered the empirical pdf of the pairwise ESF similarity, computed on the whole set of drugs with a SMILES. Then we considered the pairwise ESF similarity computed only between drugs in the same community. Finally, we tested the null hypotesis that this set (similarities in the same community) was sampled from the first distribution. The obtained p value was equal to 1, meaning that the composition of our communities is not significantly influenced by chemical similarities.

In the online SI Fig. 5, we can observe that the empirical pdf of the pairwise ESF similarity computed between drugs in the same community (red line) almost perfectly overlaps the pdf of the pairwise ESF computed on the whole set of drugs with a SMILES (blue line). Very similar results were obtained by considering the Property Distance measures, reported in the online SI Fig. 6.

Finally, we computed the average ESF similarity for all the communities that are enriched for a given mode of action (MoA) (those contained in Table S3) and containing at least two drugs with an available chemical descriptor. Results are contained in the online SI Table 7 and show that just for few communities the average ESF is significantly greater than the average value (1.7).

In this table the first column contains the community identifiers, the second one contains the community enrichment (Literature evidence/ATC-code/Direct Target Gene), the third one contains the fraction of drugs in the community for which chemical descriptors were available, and the last column contains the average ESF similarity for the community.

Assessment of the Classification Performance and Comparison with the cMap Online Tool. In order to compare our classification results with those provided by the cMap online tool (23, 24), we computed a traditional signature of differentially expressed genes [i.e., list of significant genes according to t test corrected with false discovery rate (FDR)] for each microarray experiment, as described in section "Canonical Construction of the Signatures." The experiments were relative to four groups of related drugs (online SI Table 1).

We used these signatures to query the cMap online tool. We then compared the results obtained with our approach with those provided by the cMap online tool by means of Receiver Operating Characteristic (ROC) analysis. The cMap tool provided in

output a list of drugs connected to each of the input signatures. In these lists, we filtered out the drugs that were predicted to be negatively connected to the input signature, and we considered each of the remaining drugs as true positives if they belonged to at least one of four different reference "golden standard" sets. The reference sets included both the counterpart of the tested drugs (already present in the cMap) and also well-known related drugs (if available); the drugs included in these sets were all those known to have the same MoA as the tested drugs (respectively, HSP90 inhibitors, TopoI inhibitors, TopoII inhibitors, and CDK2 inhibitors) according to either Drugbank (15, 16) or ChemBank (25, 26). All of the signatures obtained with the traditional approach, which have been used to query the cMap online tool, are available at http://mantra.tigem.it (in a unique compressed folder, containing each signature in a .grp file) together with the corresponding results (in xls format).

The result assessment shows that the DN approach performed comparably and, in many cases, better than the cMap classic online tool. The percentage of cases in which the first neighbor of a tested compound in the DN is a true positive is equal to 89% for the average distance and 77% for the maximum distance (see the section "Distance Between Two Drugs"). This value raises to 100% if we consider the case in which there is at least a true positive among the first two neighbors of each tested compound, for both the distances (as depicted in the online SI Table 2).

All the results of the ROC analysis and the comparison with the performances of the cMap classic online tool are provided in the online SI Table 3.

The particular case of the tested compound Nerviano Medical Sciences (NMS)-Doxorubicin shows that our DN approach is able to correctly classify drugs with high precision and sensitivity where the cMap classic online tool clearly fails (online SI Table 4).

Moreover, the usefulness of our DN approach and its output format is demonstrated in the following example: When the *NMS-Tanespimycin* signature, including the 142 maximally up-regulated and the 61 maximally down-regulated probe sets (available at the previously provided URL), was used to interrogate the cMap in the classic way, *Geldanamycin*, *Tanespimycin*, *Alvespimycin*, and *Monorden* ranked among the top six hits, that also included the protein synthesis inhibitor *Emetine*. However, the next top hits up to position 29 were a miscellaneous of chemicals most of which cannot clearly be related to the HSP90 and/or ubiquitin protein degradation inhibition. Known proteasome inhibitors ranked position 29 and 30. Similar results were obtained by querying the cMap classic online tool with the gene signatures of the other two HSP90 tested inhibitors. On the contrary, the subnetwork containing the tested compounds (Fig. 3*A* of the main text) and all their significant neighbors provides a modular and meaningful view of the DN approach output. This allows users to easily interpret the obtained output and to make a hypothesis on the MoA of a new drug in a clearer way.

Canonical Construction of the Signatures. Scanned microarray images were first inspected for QC using a variety of built-in QC tools from the Bioconductor (27) package of R, the open source environment for statistical analysis.

Feature intensity values from scanned arrays were normalized and reduced to expression summaries using the Robust Multiarray Algorithm and normalized by the quantiles method (28, 29).

To assess differential expression, we used a moderated t test together with a FDR correction of the p value (30, 31).

Thus, the list of differentially expressed genes was generated using a FDR ≤ 0.05 together with an absolute fold-change threshold of 2 (i.e., $|\log_2(\text{fold change})| \geq 1$).

Impact of Rank Merging on Performance. Some recent approaches attempted to use cMap data to build a drug similarity network by selectively comparing pairs of individual genome-wide expression profiles (GEPs) (32) rather than pairs of drug PRLs, as done in our approach, which are obtained by merging individual GEPs for the same drug, prior to the comparison

The use of individual GEPs will tend to group together profiles coming from the same cMap batch experiment, or the same cell line, rather then grouping drugs with similar MoA. To avoid this problem, it is necessary to merge together all of the differential expression profiles obtained with the same drug, on different cell lines and at different dosages, prior to computing distances.

To show the effect of using individual GEPs, for each GEP we considered the K closest GEPs in the cMap dataset, according to the distance. We then computed the percentage (PPV in the online SI Fig. 7) of these closest GEPs that were obtained by treating cells with the same drug (green line in the online SI Fig. 7) as the GEP under consideration, or in the same cell lines, regardless of the drug (blue line in the online SI Fig. 7), or in the same batch experiment, regardless of the drug (red line in the online SI Fig. 7).

We therefore conclude from online SI Fig. 7 that using individual GEPs to compute the similarity distance between drugs is not able to catch similarities in MoAs because of its inability to discriminate treatments obtained with different drugs in the same experimental setting.

In order to additionally assess the impact of the PRL merging procedure on the classification performance of our tool, we specifically produced additional microarray data by treating U251 human glioblastoma cell line (NCI) with *PHA-848125* at 3 μM, a dose equal to 5× the IC50 for 6 h.

We then merged the set of gene expression profiles by using different combinations of them, and we evaluated the ability of our tool to classify the resulting different PRLs.

Results of this assessment are summarized in the online SI Table 5; the list of neighbors is available online at http://mantra.tigem.it.

Performances are measured by means of ROC analysis, assuming the neighborhoods as sets of predictions and drugs in the online SI Table 1 are considered as correct predictions.

As expected the best performance is obtained when the *PHA-848125* PRL derives from treatments of all three cell lines.

Specifically, by using the profiles individually the best classified was the one obtained by treating the MCF7 cell line. This is quite obvious, first of all because MCF7 is the most recurrent cell line among those treated in the cMap dataset. Moreover for A2780 and U251 there are no treatments at all in the cMap. However, once combining the profiles from MCF7 with that from A2780 or U251, classification performances are still good, although the combination with U251 gives a less efficient classification. U251 cell line is the more diverse cell line among the three, since glioblastoma is a very heterogeneous disease where different pathways are known to be disrupted, which might explain the observed signal dilution.

Despite this, when combining the profile coming from all three cell lines together (MCF7, A2780, and U251), we obtain the best performances in classification, supporting the hypothesis that a sufficiently large combination of treated cell lines provides a sufficiently general summary of the drug activity, which is well classified in the majority of the cases.

We further explored the robustness of our method in classifying drugs by pooling together profiles coming from treatments on cell lines with a very different genetic background, potentially causing a significant signal dilution. To this aim we collected additional gene expression data by treating SF539 human glioma cell line with *PHA-848125* for 6 h.

The SF539 cell line is genotipically characterized by a mutation in the *Rb* gene (encoding for the retinoblastoma tumor suppressor protein), whereas the other three treated cell lines (A2780, MCF7, and U251) are *Rb* wild type.

DNA replication and the regulation of the G_1/S transition is under the control of the *Rb*/*E2F* pathway. In wild-type cells *Rb* binds the *E2F-1* transcription factor, thus inhibiting its regulatory activity. When *Rb* is phosphorylated by CDK2, it releases *E2F-1* that mediates the cell cycle progression (33).

In the SF539 cell line, *Rb* is no longer able to block E2F-1, which is constitutively active in this cell line as a result. As a consequence, inhibiting CDK2 with *PHA-848125* on SF539 will not have the same effect on the *E2F* mediated transcription that is elicited in the *Rb* wild-type cell lines.

Following the strategy previously described, we merged the set of gene expression profiles obtained by treating A2780, MCF7, U251, and SF539 with *PHA-848125*, and we evaluated the ability of our tool to classify the resulting different PRLs.

Results of this assessment are summarized in the online SI Table 5; the complete list of neighboring drugs is available online at http://mantra.tigem.it.

Performances were measured by means of ROC analysis, assuming the drug neighborhood as predictions, and the set of drugs in the online SI Table 1 as the golden standard.

Results, in the online SI Table 5, show that by using expression profiles individually from a single cell line, the best classification is obtained with the MCF7 cell line. This is to be expected, because MCF7 is the most recurrent cell line among those treated in the cMap dataset. The worst classification was instead obtained with the SF539 cell line, which is coherent with the *Rb* inactivation that mediates the MoA of the *PHA-848125* compound.

Nevertheless, once we combined the profile coming from all four cell lines together (MCF7, A2780, U251, and SF539), or even of two cell lines (MCF7 and SF539) only, we improved the classification performance considerably compared to using just the SF539 cell line.

These results support the hypothesis that a sufficiently large combination of treated cell lines provides a sufficiently general summary of the drug activity, which is well classified by the DN.

Community Identification. We used the affinity propagation algorithm (34) for identifying communities in our DN. This algorithm takes in the drug distance matrix and outputs a set of clusters. The algorithm also indicates, for each cluster, an element called the cluster exemplar: the element whose features best interpolate the features of all the other points in the cluster. The algorithm requirement consists in a pairwise distance matrix and a set of probabilities, one for each node to be elected as exemplar. We assumed this probability uniform.

In the first step of our procedure, by applying the affinity propagation algorithm, the whole set of drugs was partitioned in a finite number of clusters. In each of these clusters, a drug was indicated as the cluster exemplar.

By adding significant edges to nodes corresponding to drugs in a cluster, we obtained communities. With "significant edges" we meant edges whose weight was below the significance distance threshold we selected to generate the drug network.

We then, recursively, clustered again the exemplars, in order to obtain second-level communities ("rich clubs"). The procedure was recursively applied over cluster exemplars until convergence (no exemplars were clustered together).

Impact of Community Identification on the Performances. With our community-identification algorithm we basically perform a pruning of the edges of the network in order to make it "modular." Most of the identified communities are enriched for a given mode of action, as shown in the main text. This means that, once we integrate a previously undescribed compound into the network, we can make a hypothesis on its MoA by looking at communities to which the previously undescribed compound is connected. This represents a strong improvement compared to existing methods (9, 23, 24).

Here, we additionally show that our community-identification-based pruning is able to keep the "right" connections and to eliminate the "wrong" ones. In other words, we measure how the tendency of drugs with a similar MoA of being linked together changes after removing edges following the application of the community-identification algorithm.

We first labeled the compounds in the cMap according to their Anatomical Therapeutic Chemical (ATC) classification code (35), which classifies drugs according to their therapeutic and chemical characteristics. Because only 768 out of the 1,309 compounds have an ATC code, we reduced our analysis to this subset of compounds.

We then ranked the edges of both the pruned network, following the community-identification algorithm, and the original network, in ascending order (according to the associated distance value), and we computed the percentage of edges that connect drugs sharing the same ATC-code prefix of length 3, as shown in the online SI Fig. 8.

As we can see in online SI Fig. 8, the pruned network shows a better performance of the unpruned network.

Drug-to-Community Distance. We defined the Drug-to-Community distance as follows: Let x be the testing drug and C a network community containing a subset C_x of, at least, two drugs that are connected to x through significant edges (i.e., through edges whose weights are below the significance threshold). Then we define the distance between x and C as

$$\sqrt[|C_x|]{\prod_{d\in C_x} D(d,x)}/|C_x|,$$

where $D(d,x)$ is the distance between drug d and drug x (max distance), as defined in the section "Distance Between Two Drugs." So, the distance between the testing drug x and the network community C is given by the ratio between the geometric mean of the significant distances between drugs in C and x and the cardinality of this set of distances. If $|C_x| < 2$, then we assume the distance between C and x is equal to ∞.

Observations on Data Quality. We observed that 78% of the compounds contained in the cMap dataset have been tested on, at least, three different cell lines (out of five) and just 6% have been tested on a single cell line. Therefore, for the majority of the compounds in the cMap dataset, we have multiple treatments suitable for the extraction of a general cellular response (the PRL). Only for a minority of drugs (6%) that have been tested on a single cell line at a single concentration, we had a single ranked list of genes, and therefore we used this single list as the PRL of the drug.

Community Gene Ontology (GO) Fuzzy-Enrichment Analysis. Let us consider the community

$$C = \{d_1,\ldots,d_n\},$$

composed by n drugs. For each drug d_i in this community, we select the top-ranked 2,000 genes from its PRL (the set Up_i) and the bottom-ranked 2,000 ones (the set Down_i). We then compute the following unions: $\mathrm{U_{UP}} = \bigcup_{i=1}^{n} \mathrm{Up}_i$, and $\mathrm{U_{DOWN}} = \bigcup_{i=1}^{n} \mathrm{Down}_i$.

Then, for each gene j in $\mathrm{U_{UP}}$ (respectively, $\mathrm{U_{DOWN}}$) we define a membership score, as follows: $m_j^{\mathrm{UP}} = \frac{|\{i|j\in \mathrm{Up}_i\}|}{n}$ (respectively, $m_j^{\mathrm{DOWN}} = \frac{|\{i|j\in \mathrm{Down}_i\}|}{n}$).

Clearly the following relations are verified: $\frac{1}{n} \le m_j^{\mathrm{UP}}$, $m_j^{\mathrm{DOWN}} \le 1$. Without loss of generality, in what follows, we limit our discussion to the case of the up-regulated genes. When $m_j^{\mathrm{UP}} = 1$, gene j is in Up_i for each i (i.e., the gene is up-regulated

when treating with each drug in the community). On the other hand, $m_j^{\mathrm{UP}} = \frac{1}{n}$ when gene j is in Up_i for only one (i.e., the gene is up-regulated when treating with only one drug in the community). Now, fixing a chosen membership threshold level k, such that $\frac{1}{n} \leq k \leq 1$, we define the fuzzy intersection of the up-regulated genes, in the drug community C, with membership k, as follows: $F_{\mathrm{UP}}(C,k) = \{j | m_j^{\mathrm{UP}} \geq k\}$. Note that $F_{\mathrm{UP}}(C,1) = \cap_{i=1}^{n} \mathrm{Up}_i$ In the same way, the fuzzy intersection of the down-regulated genes, in the drug community C, with membership k, is computed as $F_{\mathrm{DOWN}}(C,k) = \{j | m_j^{\mathrm{DOWN}} \geq k\}$. Once we have these two fuzzy intersections, we perform a classical GO term enrichment analysis (36, 37) on them. We do this by assessing how much the occurrence of each GO term, among those associated to the genes in $F_{\mathrm{UP}}(C,k)$ (respectively, $F_{\mathrm{DOWN}}(C,k)$), is surprising and far from the expected values when genes are randomly grouped. We indicate with $\mathrm{GO}_{\mathrm{UP}}(k)$ the set of GO terms overrepresented in $F_{\mathrm{UP}}(C,k)$ and with $\mathrm{GO}_{\mathrm{DOWN}}(k)$ the set of GO terms overrepresented in $F_{\mathrm{DOWN}}(C,k)$. The following pseudocode describes the heuristic approach we used to fix an appropriate value of k, in order to maximize both k and the number of fuzzy-enriched GO terms. The input to the algorithm is the drug community C. The output consists of the fuzzy-enriched GO terms and the determined k.

1. $k = 1$
2. $n\mathrm{Up} = n\mathrm{Down} = 0$
3. totalGO $= \{\}$
4. while $n\mathrm{Up} < 2{,}000$ and $n\mathrm{Down} < 2{,}000$
5. compute $F_{\mathrm{UP}}(C,k)$ and $F_{\mathrm{DOWN}}(C,k)$
6. compute $\mathrm{GO}_{\mathrm{UP}}(k)$ and $\mathrm{GO}_{\mathrm{DOWN}}(k)$
7. if$|\mathrm{GO}_{\mathrm{UP}}(k)| + |\mathrm{GO}_{\mathrm{DOWN}}(k)| < |\mathrm{totalGO}|$
8. then return $\{\mathrm{totalGO}, k + 1/n\}$
9. else
10. totalGO $= \mathrm{GO}_{\mathrm{UP}}(k) \cup \mathrm{GO}_{\mathrm{DOWN}}(k)$
11. $n\mathrm{Up} = |\mathrm{GO}_{\mathrm{UP}}(k)|$, $n\mathrm{Down} = |\mathrm{GO}_{\mathrm{DOWN}}(k)|$
12. $k = k - 1/n$
13. endif
14. endwhile
15. return $\{\mathrm{totalGO}, k + 1/n\}$
16. end

When the computation begins, k is set to 1 [line 1]. The cardinality of the two fuzzy intersections is set to zero [line 2] and the set of fuzzy-enriched GO terms is set to empty [line 3].

Then a cycle iterates until one of the two fuzzy-sets contains more than 2,000 genes [line 4]. In each of the iterations, the fuzzy intersections and the sets of fuzzy-enriched GO terms are recomputed [lines 5 and 6], according to the actual value of k.

Then if the total number of fuzzy-enriched GO terms does decrease [line 7], then the total set of fuzzy-enriched GO terms and the value of k, which have been computed in the previous iteration, are given in output and the procedure ends [line 8]. If the total number of fuzzy-enriched GO terms does not decrease [line 9], then the variables are updated [lines 10–11] and the membership threshold value is decreased [line 12]. The remaining code [lines 13–15] is executed if the total number of fuzzy-enriched GO terms never decreases and the total number of genes in the two fuzzy intersections is greater than 2,000.

Examples of Other GO Fuzzy-Enriched Communities. The whole list of GO fuzzy-enriched communities is available at http://mantra.tigem.it (in a unique xls file).

Among the 57 GO fuzzy-enriched communities, most of the enriched GO terms are strictly linked to the mode of action of the drugs in the community. One of the most representative GO fuzzy-enriched communities is n. 28, which is enriched for a well-defined mode of action (HSP90 inhibition). For this community, our algorithm gave in output an optimal threshold level for the membership functions equal to 80% (meaning that the computed fuzzy intersections were composed by genes that were significantly differentially expressed when treating with four among five drugs in this cluster). The fuzzy intersection of up-regulated genes contained 209 genes, whereas the down-regulated one contained 236 (see *online SI Data* at http://mantra.tigem.it).

HSP90 is a chaperone protein responsible for the correct folding, stabilization, and function of multiple proteins (38). Inhibition of HSP90 increases the amount of unfolded client proteins in the cellular environment. This leads to a stress condition for the cell, resulting in the activation of a proper response via the activation of several pathways, as those involved in the *ubiquitin-proteasome* degradation system. Looking at the GO terms enriched in the up-regulated fuzzy intersection for this community, we can infer the response induced in the cell by these compounds [i.e., unfolded protein response (see *online SI Data*)].

The genes contained in the fuzzy intersections of this community were differentially expressed across drugs in this community according to the following proportions: 98% were differentially expressed following *Alvespimycin* treatment, 95% following *Geldanamycin*, 89% following *Monorden*, 84% following *Tanespimycin*, and 45% following *Fulvestrant*.

Interestingly this percentage of differential expression is approximately proportional to the relation occurring between these drugs and the MoA characterizing this community (i.e., HSP90 inhibition). This is because *Alvespimycin* and *Geldanamycin* directly bind the HSP90 protein inhibiting its cytosolic chaperone function, and they are very similar by the chemical point of view; *Monorden* is a less specific HSP90 inhibitor with effects also on Topoisomerases I and II; *Fulvestrant* binds the estrogen receptor, dissociates HSP90, and triggers its intracellular degradation; therefore it indirectly inhibits this chaperone functionality in the cell.

Another interesting GO fuzzy-enriched community is number 63, which is enriched for the sodium/potassium membrane pump blocking activity (100% of the drug in the community). The fuzzy intersection of up-regulated genes contained 40 genes (and the fuzzy-enriched GO terms reported in Table 3), whereas the down-regulated one contained 39 genes and no enriched GO terms. The GO terms that are fuzzy enriched in the up-regulated genes of this community are reported in the *online SI Data*. These GO terms could be linked to a specific effect of *cardiac glycosides* (the majority of the drugs in this comunity), i.e., the enhancement of heart phosphatides (i.e., *ethanolamine* and *phopshatidylethanolamine*) activity (39). The majority of the genes contained in the computed fuzzy intersections were differentially expressed in most of the PRLs of the cardiac glycosides in this community (>90%).

Community 43 provides another interesting example. This community contains estrogens and estrogen inhibitors. The fuzzy-intersection of up-regulated genes contained 425 genes, whereas the down-regulated one contained 335 genes. The fuzzy-enriched GO terms (see the *online SI Data*) are related to interactions between estrogens and the Golgi apparatus (40, 41) and in metabolic processes of organic compounds interacting with estrogens (*cobalamin*, *porphyrin*, and others).

Statistical Tests. We validated each community by checking if ATC codes or target genes were surprisingly overrepresented among those associated to its composing drugs (or vice versa checking if drugs with similar MoA, i.e., same ATC codes or target gene, were found in the same communities). In a similar way, we searched for enriched GO terms when we analyzed sets of genes that were differentially expressed after treatments with all drugs in a community.

In both cases we had to analyze frequencies of terms (ATC codes/target genes and GO terms, respectively) within given sets (drug communities and set of genes, respectively). Therefore, we performed the same statistical test in both analyses.

In order to test the enrichment significance of each ATC code/target gene in a drug community, and to quantify it through a *p*-value assignment, we had to compute the probability of counting, by chance, at least k occurrences of a given ATC code/target gene among those associated to the n drugs within community. If we know that, in the total drug set of N drugs, m of them are associated to the given ATC code/target gene, then the probability follows the hypergeometric distribution and is given by

$$\Pr\{X \geq k\} = \sum_{x=k}^{\infty} \binom{m}{x} \binom{N-m}{n-x} / \binom{N}{n}.$$

In the same way, p values were computed for assessing the significance of a given GO-term enrichment in those associated to genes in a given set. Finally, correction for multiple hypothesis testing was applied to the obtained p values.

The odds ratio (number of observed terms divided by the expected value) was computed as follows:

$$\frac{k}{E(X)} = k\frac{N}{nm}.$$

Biochemical Assay to Test Inhibition of CDKs by *SN-38* and *Doxorubicin*. Inhibition of CDK activity by two Topoisomerase inhibitors (*Doxorubicin* and *SN-38*) and two NMS CDK inhibitors (*PHA-00848125* and *PHA-00793887*), used as controls, was tested in a biochemical assay.

CKD2/CycA activity was measured using 1.08 nM of recombinant protein in 50 mM Tris pH 7.5, 10 mM $MgCl_2$, 1 mM DTT, 3 mM Na_3VO_4, 0.2 mg/mL BSA, 4 mM Histone H1 (Sigma H5505), and 10 uM ATP (Aldrich A2620-9) with 1 nM $^{33}\gamma$ *P* ATP (Redivue [$^{33}\gamma$ *P*] ATP, Amersham 2,500 Ci/mmole, 10 mCi/mL, 3.33 μmoles/L). Compounds were 3-fold serially diluted from 10 to 0.00051 mM for IC_{50} determination. The final concentration of DMSO was 1%.

After 60 min of incubation, an amount of 60 mL of Dowex resin (SIGMA-Supelco cat.13858-U)/formate, pH 3.0 was added to stop the reaction and capture unreacted $^{33}\gamma$ *P* ATP, separating it from the phosphorylated substrate in the solution according to the literature (42).

A volume of 27 mL of supernatant was transferred into 384 white flat-bottom plates, with 50 mL of Microscint 40 (Perkin-Elmer), and the radioactivity was counted in the Top Count (Perkin-Elmer). Experimental data were analyzed by an internally customized version of the SW package "Assay Explorer" (MDL Information System) using the four parameter logistic equation:

$$y = \text{bottom} + (\text{top} - \text{bottom})/(1 + 10^{(\log \text{IC50} - x) * \text{slope}}),$$

where x is the logarithm of the inhibitor concentration, y is the response; y starts at bottom and goes to top with a sigmoid shape.

As shown in Fig. S2 and in the online SI Table 6, no biochemical inhibition of CDKs by *SN-38* and *Doxorubicin* could be observed.

1. Schulte TW, Neckers LM (1998) The benzoquinone ansamycin 17-allylamino-17-demethoxygeldanamycin binds to HSP90 and shares important biologic activities with geldanamycin. *Cancer Chemother Pharmacol* 42:273–279.
2. Eccles SA, et al. (2008) NVP-AUY922: A novel heat shock protein 90 inhibitor active against xenograft tumor growth, angiogenesis, and metastasis. *Cancer Res* 68:2850–2860.
3. Fogliatto G, et al. (2009) Identification of a potent and specific inhibitor of Hsp90 showing in vivo efficacy. *American Association for Cancer Reasearch—Annual Meeting* Poster 37 (Abstract #4685).
4. Kawato Y, Aonuma M, Hirota Y, Kuga H, Sato K (1991) Intracellular roles of SN-38, a metabolite of the camptothecin derivative CPT-11, in the antitumor effect of CPT-11. *Cancer Res* 51:4187–4191.
5. Arcamone F, et al. (2000) Adriamycin, 14-hydroxydaunomycin, a new antitumor antibiotic from S. peucetius var. caesius. Reprinted from Biotechnology and Bioengineering, Vol. XI, Issue 6, Pages 1101–1110 (1969). *Biotechnol Bioeng* 67:704–713.
6. Senderowicz AM (1999) Flavopiridol: the first cyclin-dependent kinase inhibitor in human clinical trials. *Invest New Drugs* 17:313–320.
7. Brasca MG, et al. (2009) Identification of N,1,4,4-Tetramethyl-8-{[4-(4-methylpiperazin-1-yl)phenyl]amino}-4,5-dihydr o-1H-pyrazolo[4,3-h]quinazoline-3-carboxamide (PHA-848125), a Potent, Orally Available Cyclin Dependent Kinase Inhibitor. *J Med Chem* 52:5152–5163.
8. Brasca MG, et al. (2010) Optimization of 6,6-dimethyl pyrrolo[3,4-c]pyrazoles: Identification of PHA-793887, a potent CDK inhibitor suitable for intravenous dosing. *Bioorg Med Chem* 18:1844–1853.
9. Iorio F, Tagliaferri R, di Bernardo D (2009) Identifying network of drug mode of action by gene expression profiling. *J Comput Biol* 16:241–251.
10. Subramanian A, et al. (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 102:15545–15550.
11. Cormen TH, Leiserson CE, Rivest RL (1990) Minimum spanning trees. *Introduction to Algorithms* (Cambridge, MA, MIT Press).
12. Diaconis P, Graham R (1977) Spearman's footrule as a measure of disarray. *J R Stat Soc* 39:262–268.
13. Lin S (2010) Space oriented rank-based data integration. *Stat Appl Genet Mol Biol* 9: Article20.
14. Weininger D (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comp Sci* 28:31–36.
15. Wishart DS (2008) DrugBank and its relevance to pharmacogenomics. *Pharmacogenomics* 9:1155–1162.
16. Wishart DS, et al. (2006) DrugBank: A comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res* 34(Database issue):D668–672.
17. Hall LH, Mohney B, Kier LB (1991) The electrotopological state: structure information at the atomic level for molecular graphs. *J Chem Inf Comp Sci* 31:76–82
18. Hall LH, Kier LB (2000) The E-state as the basis for molecular structure space definition and structure similarity. *J Chem Inf Comp Sci* 40:784–791.
19. McIntyre TA, Han C, Davis CB (2009) Prediction of animal clearance using naive Bayesian classification and extended connectivity fingerprints. *Xenobiotica* 39:487–494.
20. Hu Y, Lounkine E, Bajorath J (2009) Improving the search performance of extended connectivity fingerprints through activity-oriented feature filtering and application of a bit-density-dependent similarity function. *ChemMedChem* 4:540–548.
21. Jensen BF, Vind C, Padkjaer SB, Brockhoff PB, Refsgaard HH (2007) In silico prediction of cytochrome P450 2D6 and 3A4 inhibition using Gaussian kernel weighted k-nearest neighbor and extended connectivity fingerprints, including structural fragment analysis of inhibitors versus noninhibitors. *J Med Chem* 50:501–511.
22. Rogers D, Brown RD, Hahn M (2005) Using extended-connectivity fingerprints with Laplacian-modified Bayesian analysis in high-throughput screening follow-up. *J Biomol Screen* 10:682–686.
23. Lamb J (2007) The Connectivity Map: A new tool for biomedical research. *Nat Rev Cancer* 7:54–60.
24. Lamb J, et al. (2006) The Connectivity Map: Using gene-expression signatures to connect small molecules, genes, and disease. *Science* 313:1929–1935.
25. Petri Seiler K, Kuehn H, Pat Happ M, Decaprio D, Clemons PA (2008) Using ChemBank to probe chemical biology. *Current Protocols in Bioinformatics* (Wiley, New York), Chap 14, Units 14, 15.
26. Seiler KP, et al. (2008) ChemBank: A small-molecule screening and cheminformatics resource database. *Nucleic Acids Res* 36(Database issue):D351–359.
27. Gentleman RC, et al. (2004) Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biol* 5:R80.
28. Wu Z, Irizarry RA, Gentlemant R, Martinez Murillo F, Spencer F (2004) A model based background adjustment for oligonucleotide epression arrays. Johns Hopkins University, Dept. of Biostatistics (Working Paper 1).
29. Irizarry RA, et al. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4:249–264.
30. Smyth GK (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* 3:Article3.
31. Westfall PH, Young SS (1993) *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment* (New York, Wiley).
32. Hu G, Agarwal P (2009) Human disease-drug network based on genomic expression profiles. *PloS One* 4:e6536.
33. Nevins JR (2001) The Rb/E2F pathway and cancer. *Hum Mol Genet* 10:699–703.
34. Frey BJ, Dueck D (2007) Clustering by passing messages between data points. *Science* 315:972–976.
35. Schwabe U (1995) *ATC-Code* (Wissenschaftliches Institut der AOK, Bonn, Germany).
36. Ashburner M, et al. (2000) Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25:25–29.
37. Rivals I, Personnaz L, Taing L, Potier MC (2007) Enrichment or depletion of a GO category within a class of genes: Which test? *Bioinformatics* 23:401–407.
38. Taldone T, Sun W, Chiosis G (2009) Discovery and development of heat shock protein 90 inhibitors. *Bioorg Med Chem* 17:2225–2235.
39. Marinetti GV, Temple K, Stotz E (1961) The in vivo effect of digitoxin on rat heart phosphatides. *J Lipid Res* 2:188–190.
40. Greenfield JP, et al. (2002) Estrogen lowers Alzheimer beta-amyloid generation by stimulating trans-Golgi network vesicle biogenesis. *J Biol Chem* 277(14):12128–12136.

41. Poole MC, Easley CS, Hodson CA (1991) Alteration of the mammotroph Golgi complex by the dopamine agonist 2 Br-alpha-ergocryptine (CB-154) in ovariectomized estrogen primed rats. *Anat Rec* 231:339–346.

42. Kasten TP, Currie MG, Moore WM (2002) Ion-exchange resin/enzyme activity assay. US Patent Appl 2002/072082,2002).

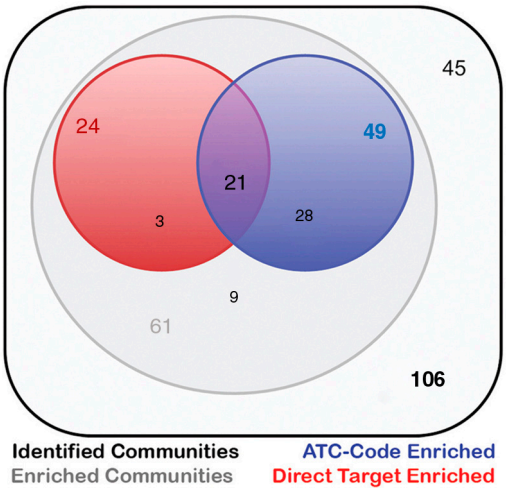

Fig. S1. Drug Community Enrichments. Of 106 identified communities (black thick line), 61 are enriched for at least one common feature (ATC code, direct target gene, and literature derived evidence). Forty-nine of these enriched communities are enriched for at least one ATC code (blue area) and 24 (red area) are enriched for, at least, one direct target gene. The intersection of these two sets contains 21 communities (purple area) enriched both for ATC codes and direct target genes. The enriched MoA of the remaining nine communities (gray area) has been verified by searching the literature.

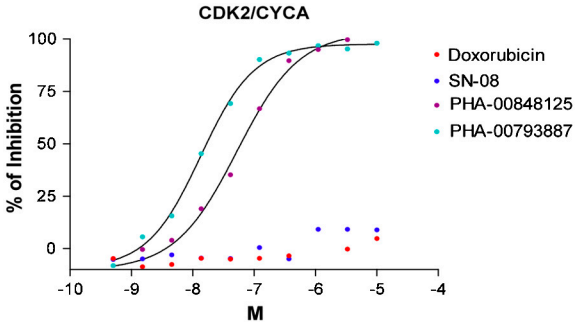

Fig. S2. Inhibition of CDKs by *Doxorubicin* and *SN38*: Biochemical Essay. Inhibition of CDK2/cyclinA complex (CDK2/CYCA) activity by two Topoisomerase inhibitors (*Doxorubicin* and *SN-38*) and two CDK inhibitors (*PHA-00848125* and *PHA-00793887*) developed at Nerviano Medical Sciences, used as controls, tested in a biochemical assay. Compound concentration is on the x axes, expressed in Moles (M), whereas percentage of inhibition is on the y axes. Different colors represent different compounds. No biochemical inhibition of CDKs by SN-38 and Doxorubicin could be observed.

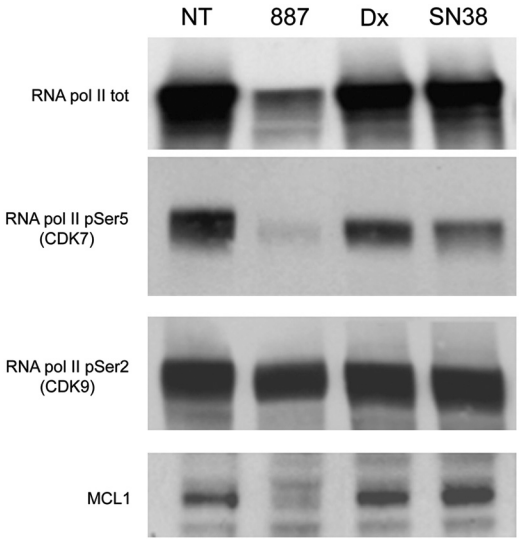

Fig. S3. Western blots following treatments with CDK2 and Topo inhibitors. Decreased phosphorylation of RNA polymerase II (RNA Pol II) on Serine 5 (RNA Pol II pSer 5, [CDK7 substrate]) and to a lesser extent on Serine 2 (RNA Pol II pSer2, [CDK9 substrate]) by *PHA-00793887* (887), a CDK inhibitor developed at Nerviano Medical Science, coupled to loss of MCL1 protein. Minor effects on Serine 5 are observed also with *Doxorubicin* (*Dx*) and *SN-38*, but they do not affect MCL1 levels.

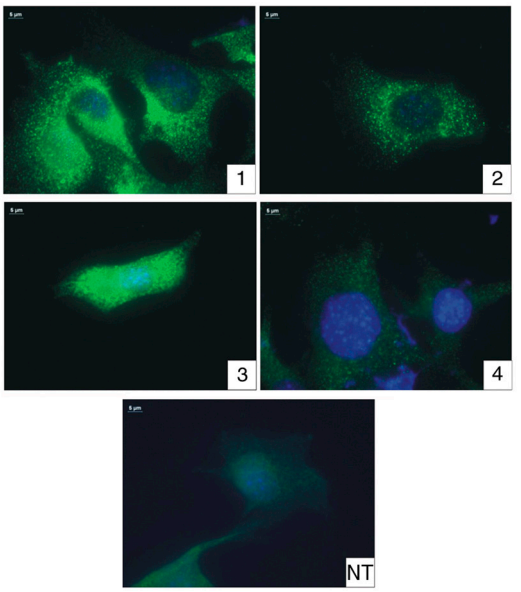

Fig. S4. Effects of *Fasudil* on autophagy. Immuno-fluorescence with anti-LC3 antibody in fibroblasts treated with drugs promoting autophagy. Evaluation of LC3 levels in human fibroblasts after treatment with drugs: Immunofluorescence (1, Rapamycin; 2, Fasudil; 3, Trifluoperazine; 4, 2DOG; NT, untreated). The experiments were performed in triplicate and representative results are shown.

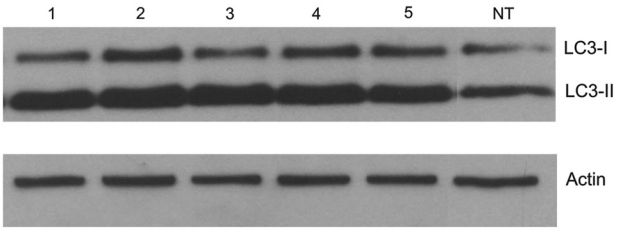

Fig. S5. Effects of *Fasudil* on HeLa Cells. Western blot with anti-LC3 antibody in drug treated Hela cells treated: 1, Rapamycin; 2, Fasudil 10 μm; 3, Fasudil 30 μM; 4, Trifluoperazine 1 μM; 5, 2DOG 100 μM; NT, untreated. LC3II levels are increased following treatment with *Fasudil*, *Trifluoperazine*, and *2DOG* as compared to the untreated control. The experiments were performed in duplicate and representative results are shown.

Table S1. First 10 neighbors and 10 closest communities of the tested compounds

Closest 10 neighbors

NMS-Tanespimycin		NMS-E973		NVP-AUY922	
Distance	*Compound*	*Distance*	*Compound*	*Distance*	*Compound*
0.436	Alvespimycin*	0.4436	Alvespimycin*	0.6084	Alvespimycin*
0.4913	Geldanamycin*	0.4891	Geldanamycin*	0.6391	Monorden*
0.5176	Monorden*	0.5294	Monorden*	0.7123	Geldanamycin*
0.6315	Tanespimycin*	0.6568	Tanespimycin*	0.7506	Puromycin
0.6533	Puromycin	0.6723	Puromycin	0.7608	Tanespimycin*
0.7178	Trifluoperazine	0.7308	Trifluoperazine	0.7756	Gefitinib
0.7542	Parthenolide	0.7638	Disulfiram		
0.7561	Thiostrepton	0.7842	Methylbenzethonium_chloride		
0.7608	Withaferin_A	0.785	Parthenolide		
0.7724	Disulfiram	0.7903	Lanatoside_C		
NMS-Doxorubicin		**NMS-SN38**		**Flavopiridol**	
Distance	*Compound*	*Distance*	*Compound*	*Distance*	*Compound*
0.5587	Daunorubicin*	0.3215	Irinotecan*	0.454	Alsterpaullone*
0.6495	GW-8510	0.5641	Camptothecin*	0.4857	GW-8510*
0.6536	Hycanthone	0.6158	Apigenin*	0.5374	Apigenin*
0.6555	Ellipticine*	0.6251	Phenoxybenzamine	0.5534	0175029-0000
0.6689	Irinotecan	0.6363	Etoposide	0.5789	Daunorubicin
0.69	Camptothecin	0.6596	Luteolin*	0.5966	Doxorubicin
0.6921	Etoposide*	0.6675	Tyrphostin_AG-825	0.5976	Camptothecin
0.6926	Mycophenolic_acid	0.6877	Daunorubicin	0.6196	Ellipticine
0.6996	Phenoxybenzamine	0.6882	Thioguanosine	0.627	H-7*
0.7175	Doxorubicin*	0.6903	Hycanthone	0.6301	Tyrphostin_AG-825
PHA-690509		**PHA-793887**		**PHA-848125**	
Distance	*Compound*	*Distance*	*Compound*	*Distance*	*Compound*
0.3838	GW-8510*	0.4715	0175029-0000	0.6212	0175029-0000
0.4613	Doxorubicin	0.4846	GW-8510*	0.6352	Apigenin*
0.4794	Alsterpaullone*	0.5145	Alsterpaullone*	0.6504	Harmine*
0.5001	H-7*	0.537	Apigenin*	0.6672	Thioguanosine
0.5593	Daunorubicin	0.5694	Daunorubicin	0.6711	GW-8510*
0.5873	Camptothecin	0.5976	Doxorubicin	0.6746	Luteolin*
0.5956	Ellipticine	0.6014	Ellipticine	0.6795	Daunorubicin
0.6048	Mitoxantrone	0.6353	Tyrphostin_AG-825	0.6828	Irinotecan
0.6144	Tyrphostin_AG-825	0.6582	Luteolin*	0.6877	Camptothecin
0.6274	Fisetin*	0.6607	Camptothecin	0.6886	Piperlongumine

Closest 10 communities

NMS-Tanespimycin		NMS-E973		NVP-AUY922	
Distance	*Community*	*Distance*	*Community*	*Distance*	*Community*
0.1285	28†	0.1310	28†	0.1285	28†
0.1296	104	0.1996	63	0.1296	40
0.1329	63	0.2481	40		
0.1863	40	0.2566	100		
0.2567	100	0.2640	104		
NMS-Doxorubicin		**NMS-SN38**		**Flavopiridol**	
Distance	*Community*	*Distance*	*Community*	*Distance*	*Community*
0.0978	14†	0.0888	32†	0.048	14†
0.1458	3	0.1174	14	0.0603	90
0.19	16	0.1434	3	0.0625	32†
0.2374	32	0.2581	89	0.0954	89
0.3955	40	0.3798	75	0.1929	52
				0.1995	85
				0.2527	40
				0.2564	63
				0.3781	104
				0.3874	61

Closest 10 communities

PHA-690509		PHA-793887		PHA-848125	
Distance	*Community*	*Distance*	*Community*	*Distance*	*Community*
0.03	90	0.0527	14†	0.0721	14†
0.0464	14†	0.0916	32†	0.0845	32†
0.0585	32†	0.0947	63	0.0927	63
0.0639	89	0.1927	3	0.255	89
0.1283	85	0.383	104	0.2590	104
0.1299	52			0.3762	69
0.1931	74			0.3763	100
0.1933	61			0.3847	3
0.2561	13				
0.3837	40				

*True positives, drugs sharing the mode of action with the testing one.
†True positives, communities enriched for the mode of action of the testing drug.

Table S2. Selectivity profile of the tested CDK inhibitors

	PHA-793887	PHA-848125	PHA-690509	Flavopiridol
Enzyme	Average IC50 (uM)	Average IC50 (uM)	Average IC50 (uM)	Average IC50 (uM)
CDK1	0,060	0,398	0,160	0,034
CDK2	0,008	0,045	0,031	0,040
CDK4	0,062	0,160	>10	0,090
CDK5	0,005	0,265	0,090	0,102
CDK7	0,010	0,150	nt	0,754
CDK9	0,138	1,112	0,141	0,025
GSK3	0,079	>10	1,900	0,971
TRKA	>10	0,053	nt	nt

Table S3. 2-deoxy-D-glucose (2DOG) Analysis

Whole neighborhood		
1	Fasudil	0.5162
2	Thapsigargin	0.5644
3	Trifluoperazine	0.577
4	Gossypol	0.633
5	Niclosamide	0.6539
6	Tyrphostin_AG-1478	0.6682
7	Valinomycin	0.678
8	Ivermectin	0.6792
9	Sodium_phenylbutyrate	0.6833
10	BW-B70C	0.6905
11	Calmidazolium	0.6912
12	5224221	0.6968
13	MG-132	0.6971
14	Desipramine	0.7007
15	Rottlerin	0.7013
16	Clotrimazole	0.7054
17	Mefloquine	0.7066
18	Ionomycin	0.7087
19	Tamoxifen	0.7143
20	Cytochalasin_B	0.7164
21	Ciclosporin	0.7201
22	Puromycin	0.7268
23	Pyrvinium	0.7283
24	Astemizole	0.729
25	Alexidine	0.7305
26	Disulfiram	0.7311
27	Fendiline	0.7329
28	Prochlorperazine	0.7387
29	Anisomycin	0.7397
30	Pararosaniline	0.7417
31	Chlorprothixene	0.742
32	Loperamide	0.7422
33	Mometasone	0.7439
34	Iloprost	0.7475
35	0297417-0002B	0.748
36	Thioridazine	0.7488
37	MG-262	0.75

Whole neighborhood		
38	Spiperone	0.7556
39	Arachidonyltrifluoromethane	0.7599
40	Methylbenzethonium_chloride	0.7615
41	5707885	0.763
42	Oligomycin	0.7701
43	Podophyllotoxin	0.7725
44	Homochlorcyclizine	0.7736
45	Perphenazine	0.7742
46	Celastrol	0.7752
47	Vanoxerine	0.776
48	Idoxuridine	0.776
49	5666823	0.7765
50	Hydroxyzine	0.7766
51	Nordihydroguaiaretic_acid	0.7776
52	Geldanamycin	0.7776
53	Metergoline	0.7777
54	Novobiocin	0.7779
55	Terfenadine	0.7781
56	Butoconazole	0.7787
57	Piroxicam	0.7808

2DOG community
2-deoxy-D-glucose
fasudil
sodium_phenylbutyrate
tamoxifen
arachidonyltrifluoromethane
novobiocin

2DOG Rich-club

Member	Community
Trifluoperazine*	100
Ciclosporin*	43
Astemizole*	34
Oligomycin*	78
Gefitinib	60
5114445	4
Esculetin	54
Dimethyloxalylglycine	51
Demecolcine	48
Zardaverine	106
CP-319743	10
Terconazole	92
3-aminobenzamide	2
Mycophenolic_acid	75
HC_toxin	16

*Rich-club members significantly connected to 2DOG.

Other Supporting Information Files

Dataset S1 (XLS)