Published in final edited form as:
Mol Biosyst. 2013 July ; 9(7): 1604–1619. doi:10.1039/c2mb25459j.

Predicting cancer drug mechanisms of action using molecular network signatures†

Justin R. Pritchard[a,b], **Peter M. Bruno**[a,b], **Michael T. Hemann**[a,b], and **Douglas A. Lauffenburger**[a,b,c]

Douglas A. Lauffenburger: lauffen@mit.edu

[a]Department of Biology M.I.T., Cambridge, MA, USA

[b]Koch Institute M.I.T., Cambridge, MA, USA

[c]Department of Biological Engineering M.I.T., 77 Massachusetts Ave., Cambridge, MA, USA

Abstract

Molecular signatures are a powerful approach to characterize novel small molecules and derivatized small molecule libraries. While new experimental techniques are being developed in diverse model systems, informatics approaches lag behind these exciting advances. We propose an analysis pipeline for signature based drug annotation. We develop an integrated strategy, utilizing supervised and unsupervised learning methodologies that are bridged by network based statistics. Using this approach we can: 1, predict new examples of drug mechanisms that we trained our model upon; 2, identify “New” mechanisms of action that do not belong to drug categories that our model was trained upon; and 3, update our training sets with these “New” mechanisms and accurately predict entirely distinct examples from these new categories. Thus, not only does our strategy provide statistical generalization but it also offers biological generalization. Additionally, we show that our approach is applicable to diverse types of data, and that distinct biological mechanisms characterize its resolution of categories across different data types. As particular examples, we find that our predictive resolution of drug mechanisms from mRNA expression studies relies upon the analog measurement of a cell stress-related transcriptional rheostat along with a transcriptional representation of cell cycle state; whereas, in contrast, drug mechanism resolution from functional RNAi studies rely upon more dichotomous (*e.g.*, either enhances or inhibits) association with cell death states. We believe that our approach can facilitate molecular signature-based drug mechanism understanding from different technology platforms and across diverse biological phenomena.

Introduction

While chemotherapy remains the standard of care for most disseminated malignancies, over the past decade-plus an increasingly wide range of prospective molecular interventions have become part of the drug discovery process for cancer treatment. Between the broadening repertoire of potential drug candidates, and the beckoned promise of stratified patient subpopulations likely to receive special benefit from particular therapeutics, a critical bottleneck lies in discerning and matching mechanism-of-action categories between prospective drugs and patient populations. This remains centrally true both for the array of chemotherapeutic candidates arising from modern high throughput screening technologies,

†Electronic supplementary information (ESI) available. See DOI: 10.1039/c2mb25459j

Correspondence to: Douglas A. Lauffenburger, lauffen@mit.edu.

as well as for rationally designed therapeutics where the molecular targets are specified but cell-level mechanism-of-action can vary across different tumor backgrounds.

To address this bottleneck, the data-driven classification of novel cancer therapeutics – first suggested in 1989[1] is an attractive approach. Diverse research groups, utilizing multiple model organisms and cultured human cells, now produce molecular and phenotypic signatures of drug action that can effectively annotate mechanisms of action at diverse stages of drug discovery.[2–7] However, the data analysis strategies employed in these studies are surprisingly limited. In many cases, these approaches generate lists of hypotheses that require post-hoc analyses, or use unsupervised learning to mine for unexplored connections in datasets. Though there is clear value in this unsupervised/list based approach, a more powerful integration of molecular signatures into drug discovery pipelines will benefit from the joint incorporation of explicitly supervised schemes that complement the unsupervised methodologies to enable quantification of predictive performance.

However, this aspiration presents serious technical challenges. Drug mechanism-of-action study datasets exhibit important features that distinguish them from other high throughput datasets in computational and systems biology.[8,9] They usually have as few as 2–4 structurally independent small molecules with well-annotated mechanisms of action. This paucity stands in stark contrast to microarray diagnostic training sets involving as many as hundreds of examples of a clinical class.[8–11] Furthermore, formal assessment of supervised machine learning algorithms typically are concerned solely with statistical generalization – *i.e.*, how well a test set drawn from the same class can be predicted from the training set.[12] While this is critical for drug mechanism of action prediction studies, we view it as merely a first step in incorporating supervised machine learning methodologies into a data driven pipeline for drug discovery.

We propose here that analysis strategies for data driven drug discovery should have three key properties:

1. *Statistical generalization in the classic machine learning sense*. Models for multiclass classification should accurately predict drugs whose mechanism of biologic action is a known mechanism that is already present in the training set.
2. *Ability to recognize a "New" drug mechanism-of-action*. Many multi-class machine learning methods yield predictions invariably falling within one the posed categories for all test observations. Test observations will either interpolate within their predicted class or else force an expansion of the predicted class, but regardless will be placed in the class. In order to recognize a "New" drug mechanism-of-action, we must instead quantitatively decide how big of an expansion is allowable, and then attempt to find robust groupings of drugs that fail this expansion test.
3. *Ability to gain predictive accuracy in the face of new biology (we term this biological generalization)*. We must be able to pick biologically relevant groupings, and predict new members of these groups using supervised approaches with quantifiable accuracy. This is a very difficult, but vital, test for molecular signatures.

Focusing our contribution here on studies most directly relevant to cancer as one important drug discovery realm, we developed a new strategy satisfying these three criteria and applied it to diverse datasets that incorporate molecular predictors of clinically adopted cancer therapeutics in mammalian cells.[3,4,7,13] While these studies derive from a variety of experimental methods (mRNA expression, drug–drug interactions, RNAi effects), we aim to focus the signature-based drug prediction field around centralized concepts that will facilitate adapting molecular signatures to a drug discovery workflow that involves

quantitative decision making, and to broaden our understanding of where predictive accuracy can originate.

We will take a moment in this Introduction section to explain our strategy conceptually. To begin, we define drug mechanism-of-action in terms of “subnetworks”. A drug mechanism subnetwork consists of nodes and edges, where the nodes are drugs and the edges represent weighted connections between these drugs. These weighted connections represent distances in molecular signature space. In our work, drug subnetworks are employed to create a set of statistical measures characterizing an integrated work-flow of unsupervised and supervised techniques for prediction of drug mechanism-of-action class whether from previously established training classes or de novo. Our goal is to satisfy our 3 criteria outlined above for data-driven drug mechanism-of-action.

Initial subnetwork membership can be defined by biochemical and genetic evidence from the literature (Fig. 1A). We start with well-validated mechanism-of-action classes. In this analysis, most of the common chemotherapeutic drug categories in clinical use and many drug categories in pre-clinical development are represented across the three datasets (mRNA, chemical interaction, and RNAi). These initially supervised category descriptions can be thought of as discrete subnetworks. All drugs within a class are completely connected by weighted edges in molecular signature space, but no edges between the subnetworks are allowed (Fig. 1A).

The different datasets that we have chosen to investigate have diverse data matrix structures (Fig. 1B), but all involve only a few chemically distinct small molecules possessing similar mechanisms-of-action with which to train category descriptions. The diverse data types necessitate distinct approaches to feature selection. As an overarching principle, our feature selection is done *via* a drug mechanism centric strategy. We implement this mechanism centric strategy by selecting features that are predictive across a mechanistic class (Fig. 1B).

Empirically, we have found that a *K*-nearest neighbor algorithm (*K*-NN) yields the best statistical generalization in functional RNAi datasets (Fig. 1C). Going beyond this, we evaluated each new supervised prediction with respect to how it might increase the size of a drug mechanism subnetwork relative to a negative control distribution (Fig. 1C). Since a given drug mechanism subnetwork is comprised of all possible edges for all nodes of a single drug mechanism of action, after each new prediction a prospective expansion of the average edge length is calculated within that subnetwork (Fig. 1C.1). A prediction can be interpolated and shrink the network size, yielding a linkage ratio (LR) less than 1. On the other hand, a prediction could be extrapolated and enlarge the network size leading to an LR > 1. Any change in subnetwork size is compared to the distribution of the same metric for all negative control (*i.e.*, out-of-network) compounds. To create this negative control distribution, all out-of-network compounds are, in turn, forced into a drug mechanism subnetwork to which they don’t belong (Fig. 1C.2). This creates a column vector of all negative control subnetwork expansions that is used to calculate a non-parametric estimate of the cumulative distribution function (cdf) for falsely including an out of category drug in a subnetwork. This cdf is then used to estimate a *p*-value and quantify whether a given subnetwork expansion is acceptable (Fig. 1C.3). If not, the drug is predicted to warrant a “New” mechanism-of-action category.

Prediction that a “New” category is warranted enables the datasets used for mechanism-of-action studies to be biologically generalized to more than merely further examples within the specified training set categories, permitting extension to drugs influencing entirely distinct biology. Signatures of the “New” drugs (*i.e.*, drugs that increase subnetwork size to such a degree that they likely belong to the negative control distribution) can be examined for

grouping in a manner that allows for the updating of the training set with new subnetworks – and ultimately the supervised prediction of yet additional untrained examples from these new drug subnetworks. This recognition of new groupings lends itself to an unsupervised methodology. For this extension purpose, we have developed a network-based consensus technique that attempts to find best agreement across a variety of alternative hierarchical clustering interpretations and topology thresholds by rewarding in cluster edges and penalizing out of cluster edges (Fig. 1D). Finally, to ascertain that quantitative supervised predictions can still be made, we refine the final network consensus cluster, keeping or adding only the groupings that are predictable in our supervised methodology.

In the body of this contribution we will explain in greater depth and detail our algorithmic choices and methodologies in the context of functional RNAi-based studies that our approach was initially developed for. We then show successful application of our strategy to two other literature studies based on different data types; in this context, we demonstrate that our approach can outperform the classical Connectivity Map analysis of mRNA expression datasets.

Materials and methods

Algorithm comparison

The algorithm comparison was performed to choose a predictive model, and then to assess the mechanism of the winning model's success. To perform the simulations for Fig. S1 (ESI†) that approximated RNAi drug datasets, we calculated the category averages and standard deviations using topoisomerase II poisons, alkylating agents, spindle poisons and nucleotide depletion agents (central tendencies and category standard deviations. For $N = 3$ drugs per mechanistic category and $N = 10$ drugs per mechanistic category, using a normal distribution (mean and standard deviation specified for each class), we generated 100 training sets and 4000 (1000 per mechanism) test signatures (Supplemental Data 1, ESI†). All supervised machine learning algorithms were implemented in Matlab 2011b. *K*-NN was implemented using knnclassify, with $K = 1$ and a "euclidean" or "correlation" metric.[14] Naive bayes used the NaïveBayes class in Matlab 2011b.[14] Ensemble learning with bootstrap aggregation was performed using the ADTM2 method of the fitensemble function.[15]

To examine how experimental error across individual operators and time affect RNAi signature *K*-NN predictions, we took signatures for which we have 7–31 experimental replicates over 3 years. These signatures included CBL, Dox, Vin, and 5-FU. We performed inverse transform sampling (1000 samplings) of the eCDF followed by nearest neighbor's predictions (Supplemental Data 2, ESI†). Predictions were scored as correct if the *K*-NN prediction predicted the right category.

Support vector machines (SVM) ensemble

SVM models were trained in Matlab using the SVMtrain function with a linear basis function. They were trained individually, upon every drug mechanism subnetwork. These trained models were used to predict the entire drug test set in Fig. 2. Given N models for N drug mechanism subnetworks (A drug mechanism of action subnetwork is a set of D drugs that has the same subnetwork mechanism N_i): The predictions on D drugs form a $D \times N$ matrix S whose entries $\in[0,1]$. A row of S corresponding to a drug D_i for which

†Electronic supplementary information (ESI) available. See DOI: 10.1039/c2mb25459j

$$\sum_{i=1}^{N} N_i \neq 1 \quad (1)$$

indicated membership in two drug mechanism subnetworks, or a lack of prediction. Both of these possibilities were taken as evidence of the existence of a “New” drug mechanism.

Algorithm ensemble

Given *M* models for *M* algorithms (algorithms included *K*-NN, SVM, Naïve Bayes, LDA and ADTM2): The predictions on *D* drugs form a *D* × *M* matrix *A* whose entries ∈[1…*N*] where *N* is the number of drug mechanism subnetworks in the training set. A row of *A* corresponding to a drug D_i for which the value of an entry N_j corresponds to a prediction from the model M_j, was scored as follows, if any other model M_x across the *i*th row of *A* did not equal, the drug is predicted to have a “New” mechanism of action.

K-NN predictions

All *K*-NN predictions were performed in matlab using the knnclassify function. *K* was 1 unless otherwise specified. The distance metric was the “euclidean” metric or the “correlation” metric.

Negative control network expansion measurements

Given *N* drug subnetworks, for each subnetwork N_i there are $N - 1$ out of network drug mechanisms. Each drug mechanism subnetwork is of variable size (though generally 2–4 drugs exist per subnetwork). All drugs of all $N - 1$ out-of-network drug mechanisms of action are the set used to calculate the distribution of the negative control network expansion measurement. These measurements are linkage ratios for each of the negative control measurements. For specifics as to precise subnetwork membership and dataset provenance see Supplementary Data 3 (ESI†). For the top microarray features by category see Supplementary Data 5 (ESI†).

Linkage ratio

A drug mechanism of action subnetwork is a set of *D* drugs that has the same subnetwork mechanism N_i. The drugs form a completely interconnected subnetwork with *D* nodes and *E* edges where:

$$E = \frac{1}{2}(D-1)\,(D) \quad (2)$$

The edge lengths are described by the distance metric of the user’s choice and are in a vector *L*. The vector *L* of length *E* contains all of the pairwise distances in a drug mechanism of action subnetwork. To find an estimate for the size of the original subnetwork we take the average of these pairwise distances:

$$s = \frac{1}{E}\sum_{i=1}^{E} L_i \quad (3)$$

Next we take all $N - 1$ out-of-network drug mechanisms and add all drugs within them, one by one, to the subnetwork, creating a vector *L** of length *E* + *D*, for every drug, that describes a new falsely-expanded subnetwork with *D* + 1 nodes and *E* + *D* edges. An estimate of the size of this falsely-expanded subnetwork is:

$$s^* = \frac{1}{E+D} \sum_{i=1}^{E+D} L_i^* \quad (4)$$

Doing this for all out-of-network drug mechanisms creates a vector F, of length D_T, where D_T is the total amount of drugs from $N - 1$ subnetworks. F contains entries of s^* for each negative control false expansion.

The Linkage Ratio (LR) of a given false expansion is:

$$\mathrm{LR} = \frac{s^*}{s} \quad (5)$$

Where an LR > 1 indicates a category expansion and an LR < 1 indicates interpolation within a category. Calculating the LR across all entries of F gives the vector F_{LR} of size D_T containing all linkage ratios for all negative control network expansions.

Generation of *p*-values

Given F_{LR} we can plot an ecdf with D_T steps of step size $\frac{1}{D_T}$. We can compare this to a kernel density estimate of the cdf using the ksdenstiy function in matlab 2011b,[16] applied smoothing techniques for data analysis. We used a normal kernel on unweighted data. The density is unbounded and the bandwidth is 100 for all cdfs. All fits appeared to be good approximations of the ecdf.

Given a prediction p, to generate a p-value we add all pairwise distances between p and its predicted subnetwork to the baseline subnetwork vector L to create a new vector L^p of length $E + D$ edges. Given L^p we calculate the predicted expansion using (4) and the predicted LR using (5). We then use the kernel density estimate of the cdf to get an estimate of the p-value associated with a predicted expansion.

Classifier evaluation

Given this multiclass learning problem we adopted a modified version of a true positive rate (TPR) and false positive rate (FPR). A true positive (TP) is a drug whose mechanism of action is present in the training set, and is correctly predicted to belong to that mechanism of action by a classifier. A false positive (FP) is a drug whose mechanism of action is not in the training set and is predicted to significantly belong to any training set mechanism subnetwork, or a drug whose mechanism of action is in the training set, but is significantly predicted to belong to the wrong mechanism of action subnetwork. A false negative (FN) is a drug whose mechanism of action is in the training set, but is predicted to belong to a "New" drug mechanism. A true negative (TN) is a drug whose mechanism is not in the training set, and whose mechanism of action is predicted to be a "New" drug mechanism of action.

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \quad (6)$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}} \quad (7)$$

Hierarchical clustering

All hierarchical clustering was done using the clustergram function in Matlab 2011b. We used “euclidean” (RNAi) or “correlation” metrics (mRNA, chemical intervention data), in order to stay consistent with previous clusterings and to use the metric that was most appropriate for the dataset. Linkage was average.

Network consensus clustering

We assume that at least two clusters with >2 drugs exist in the dataset. Given this assumption, we find the maximum network interpretation for a given set of clustered solutions by using the equation:

$$\alpha\frac{c}{C}-\frac{o}{T} \quad (8)$$

Where *c* denotes the total number of observed edges (at a given threshold) between nodes that co-cluster, *C* is the total possible number of edges for a given clustering solution (all members of all clusters are completely connected by edges). *o* is the total number of edges between nodes that have different clustering assignments (out of cluster edges) and *T* is the total number of edges. α is a scaling parameter that was always 1 for our study but could be tuned for future applications to give a greater reward for in cluster edges.

Network consensus clustering returns clusters that maximize the network–cluster agreement for a given set of clusters. It returns clusters of size 2 or more whose members share edges in the consensus network solution.

Network rendering

All networks were rendered in Cytoscape[17] using a weighted force-directed layout. The pairwise distance for each edge was used as the weighting attribute.

Criterion for updates

Given a network consensus cluster, final updates were refined using our supervised approach. If “ties” existed in the consensus clustering we selected the most parsimonius solution. If two clusters were composed of experimental replicates of the same compound, those clusters were combined.

Given a consensus solution, we update the original training set (if 2 drugs make up a cluster we update only 1, if 3 or 4 drugs make up a cluster we update 2 and 3). To pass our supervised refinement criterion, a drug from a 2 drug consensus cluster (for example Cantharidin and AA2 in the RNAi dataset) must predict the other drug from the cluster using *K*-NN alone because a linkage ratio cannot be calculated (*e.g.* AA2 must predict that Cantharidin is in the same mechanism subnetwork). In the case of AA2 and cantharidin, AA2 cannot predict cantharidin. In fact cantharidin is predicted to be a “New” mechanism, so this mechanistic subnetwork is not added to the training set.

In larger classes the *N*th most distant member of a new consensus subnetwork is predicted using the $N-1$ least distant subnetwork members.

1. If the *N*th member is significantly predicted to belong to the consensus subnetwork using our supervised methodology, the update stands, and updates are allowed until this is no longer possible.

2. If the prediction fails, the *N*th most distant member is iteratively eliminated from the network consensus cluster until a correct prediction can be made.

Connectivity map analysis

Connectivity map searches were done using the instance query web tool. A list of all searches, instance id's, search thresholds and search results are given in Supplementary Data 4 (ESI†). Only MCF7 arrays were chosen to minimize across cell line variation. All array instance ids can be found in Supplementary Data 3 (ESI†). The best searches for HDAC inhibitors and topoisomerase II poisons were considered the best because of the search rankings of true positives. Best search probe sets were used for the error analysis in Fig. 6. All data was from the processed, rank normalized dataset for Connectivity Map build2, and is available on the Broad Institute website [http://www.broadinstitute.org/]. We made a concerted effort to present representative and best case results.

The best predicting features were built across the five drug mechanism training set in Fig. 4. *T*-tests and the Wilcoxon rank sum test was used to rank all features for mechanism specific predictive capability. *T*-test and Wilcoxon ranked features had high degrees of overlap. We chose the top 1, 10, or 100 *T*-test features to perform leave-one-out cross validation. Top features and gene names are listed in the Supplementary Data 5 (ESI†). The final predictor used 10 features per mechanism because 1 feature could not accurately cross validate and 10 features was the smallest number of features that cross validated accurately.

Functional annotation

Functional annotation was performed using DAVID.[18]

RNAi signatures

New RNAi signatures for Fig. S2 (ESI†) were generated and normalized as in[7] and are available in Supplementary Data 2 (ESI†).

Results

Our first goal, to achieve statistical generalization (*i.e.*, how well a test set drawn from the same class can be predicted from a training set), is best explained in the context of an RNAi study: the original dataset for which we developed our first generation analysis method.[7] In this initial study, 29 shRNAs targeting diverse aspects of the DNA damage response and programmed cell death biology were validated for knockdown and examined for their ability to affect drug response in lymphoma cells *in vitro.* Drug response was assessed by a mixed population of shRNA-infected cells in which 10–20% of the population harbors the shRNA (along with GFP as a marker). An RNAi signature of a drug was derived as the pattern across all shRNAs of how each shRNA affects cell response to the drug relative to uninfected cells. We found that we could use as few as 8 shRNAs to accurately predict new examples of drugs for topoisomerase I and II poisons, spindle poisons, alkylating agents, nucleotide depletion agents, and transcription/translation inhibitors, as well as histone-deacetylase inhibitors (HDACs) that the model had not been trained on.

Given that *K*-NN worked well in this original study, we wished to test here whether it outperforms other machine learning algorithms in RNAi datasets. We initially chose to simulate drug categories instead of the drugs themselves. To do this we chose four drug categories: topoisomerase I poisons, alkylating/alkylating like agents, spindle poisons and nucleotide depletion drugs. These categories were defined by multiple examples of different drugs possessing distinct chemistries within each individual category. A category-based average and standard deviation were generated, and these parameters were then used to

simulate “theoretical drugs” from these categories to generate 100 different training sets and 1000 tests per drug category. In these simulated RNAi training datasets, *K*-NN dramatically outperformed all other techniques when 3 training drug examples were present (Fig. S1A, ESI†). However, as more training drug examples were added ($N = 10$) the performance gap between *K*-NN and other algorithms decreased (Fig. S1B, ESI†). This indicated that for RNAi studies based on small numbers of validated drug training examples *K*-NN is a good choice.

Next, to probe utility for statistical generalization, we examined how *K*-NN predictions would perform over the progression of time and with actual experimental error. Following the original study, we have generated many signatures across a variety of drug categories. Focusing on the deepest coverage, we chose four of those categories: topoisomerase II poisons, alkylating/alkylating like agents, spindle poisons and nucleotide depletion drugs for which 7–31 replicate drug signatures were tested over three years. We now randomly sampled these much larger datasets from the expected cumulative distribution function of individual shRNAs for various specific drugs using inverse transform sampling. Using 1000 simulated Chlorambucil (an alkylating agent), Doxorubicin (topoisomerase II poison), 5-FU (a nucleotide depletion agent), and vincristine (a spindle poison) signatures, we estimated the robustness of the *K*-NN predictions for the more recent drug additions to the dataset. We observed that *K*-NN had an accuracy of 98–100% (Fig. S2, ESI†). These results offered confidence that a *K*-NN approach achieves comfortably strong statistical generalization in RNAi datasets.

Having shown statistical generalization, we next probed categories of drug mechanisms-of-action that the original shRNA signature had not been trained upon. Given drugs whose mechanisms-of-action were represented in the training set plus those whose mechanisms-of-action were not (kinase inhibitors, phosphatase inhibitors, HSP90 inhibitors, and others) we sought to examine the capability for discerning between these two cohorts of drugs – representing true positives and false positives, where a false positive is a “New” drug example that is not in the original training set.

We implemented a variety of machine learning techniques with the goal of identifying “New” drugs that did not belong to the specified training set categories, and found that many classic techniques failed in doing so (Fig. 2A). Most algorithms produce a false positive prediction when a drug’s mechanism is not represented in the training set. The best of the unmodified approaches, the *K*-NN approach, could achieve a true positive ratio (TPR) of 100% but its false positive ratio (FPR) was 100% concomitantly.

We thus considered whether an ensemble algorithm approach might improve performance, so we tested two such methods. The first method employed an ensemble of binary support vector machines (SVMs) each trained separately on each of the distinct drug mechanisms of action in the training set. SVMs seek to define decision boundaries on the basis of multi-variate “support vectors”. By making a binary (single class, “in” *versus* “out”) model for each drug category, we could impose rules on the family of predictions made by the ensemble of models. If a drug was predicted to belong to two groups, or was never predicted to belong to a training group, it was classified as having a new mechanism-of-action. The second method was an ensemble of the individual algorithms previously tested (Fig. 2A), as we had noticed that correlation between algorithms for different test sets is low (Fig. S1, ESI†). This raised the notion that the different algorithms were intrinsically relying on different features in the data. With this in mind, we created a scoring criterion whereby drugs with prediction disagreements among diverse algorithms are counted as “New” mechanisms of action. Both kinds of ensemble approaches improved prediction performance by reducing false positives modestly, though with high losses of true positives. The SVM

and algorithm ensembles had TPRs of 88% and 42% and FPRs of 67% and 0% respectively (Fig. 2A).

Toward a yet more effective approach, we recognized that the signature-based drug mechanism-of-action subnetworks exhibit diverse sizes (*i.e.*, as characterized by mean edge length) (Fig. 2B). The range of weighted pairwise distances between nodes in a given subnetwork produces complex shapes along with heterogeneous sizes in molecular signature space. Moreover, different subnetworks should be able to tolerate the inclusion of novel drugs to a different extent based upon their size and relative spacing to each other in subnetworks, but a sound theoretical estimate of the intra-subnetwork distribution for a given drug network is elusive because the "true" shape of the mechanistic subnetwork across all possible efficacious chemistries on a given biochemical target is unknown. Furthermore, the generally small number of distinct chemistries per class for model training precludes firm empirical estimation of a distribution within a subnetwork. Accordingly, we proceeded to take advantage of the "negative control" drugs (*i.e.*, those not falling into a given category) to build distributions of false positive predictions; the negative controls offer a reasonable approximation of a background distribution because they are sampled broadly across many mechanisms of action with diverse biological functions. The empirical cumulative distribution function of these false positive network expansions is plotted for each category (Fig. 2B bottom), and a kernel density estimate of the cdf permits estimate of the probability that a predicted linkage ratio could be generated in the negative control dataset (Fig. 2B bottom). Implementing this methodology across all drug categories clearly separates false positives into a "New" drug category. Calculated results shown in Fig. 2C show this methodology to be a major improvement over the algorithms tested in Fig. 2A, including the ensemble algorithms (Fig. 2C). Using a *p*-value cutoff of 0.05, this advanced method yielded a TPR of 100% and an FPR of 8.3%; a more stringent cutoff of 0.01 yielded a TPR = 94% and an FPR of 0%.

In our original RNAi study we had found that there was predictive resolution among HSP90 inhibitors as well as some classes of kinase inhibitors, which represented entirely new biological classes of drug mechanisms-of-action that the shRNA signature had not been trained on. This was a surprising result, but we had performed that study in an ad hoc manner by simply adding two new drug examples to the trained signature and predicting a third. Nevertheless, it suggested a potential for signatures of molecular predictors to be sufficiently broad in their biological comprehension to garner predictive resolution for untrained classes. We thus sought to apply our new network-based signature strategy to other datasets as a systematic way to test this potential more broadly.

While our network based statistic utilizes the negative control distribution for out-of-network expansion and is clearly able to define drugs not belonging to a training set subnetwork, the main challenge becomes identifying high confidence functional groupings between these "New" compounds such that the groupings have similar mechanisms-of-action. These new functional groupings must then be incorporated back into the training set and integrated with our supervised methodology so that during an iterative process new examples of these unknown subnetwork mechanisms might be identified and compiled. Because "New" groupings must be found, unsupervised learning is a natural choice. However, one such method, hierarchical clustering, will always produce clusters, and it requires the user to decide how to interpret the hierarchy. Furthermore, the structure between groups in a cluster tree is often significantly altered as neighboring branches are combined in the next level of hierarchy (often using average or nearest linkage measurements). Using the "New" mechanisms of action set of drugs from Fig. 2C, we, see that hierarchical clustering (Fig. 3A) can raise multiple options for interpreting natural groupings in the data.

In an alternative approach, networks of pairwise distances can also be used to examine relationships between nodes (here drugs), especially if the topology is thresholded at distance cutoffs.[19] This network based representation offers the advantage of displaying all edges of a certain distance, and grouped distances are not recalculated as a hierarchy changes. In our data set, as the topology threshold is increased (Fig. 3B), the network loses nodes and coalesces around the more densely interconnected subnetworks. Hence, the network representation may more faithfully represent some aspects of the natural data connections within and across cluster groupings even if it fails to provide an unbiased or non-trivial stopping point. Given these strengths and weaknesses, we developed a consensus method that combines the natural representation of thresholded network topologies and the discrete interpretations of hierarchical clustering. Under the assumption that two or more clusters exist in the data, we vary cluster size and the number of edges included in the network topology. In this procedure, we reward consensus solutions that maintain within cluster edges and eliminate between cluster edges; the best scoring solution is selected. Finally, we refine the groupings in an iterative fashion using our supervised approach, and test statistical and biological generalization.

Given that we had previously examined inhibitors of HSP90 and EGFR in our earlier study, we wanted to see if our proposed methodology could identify them *a priori*. Here we are not looking for “New” mechanisms in the singular, but “New” subnetworks in a more unbiased and systematic manner. We performed the search for “New” subnetworks in the presence of control out of category compounds (sunitinib, gliotoxin, and cantharidin and AA2). Here, network consensus clustering identified an optimal solution comprising 4 clusters (Fig. 3A–C). This solution separated gliotoxin into its own 1-member category; the other three clusters were: a 2-member cluster containing cantharidin and AA2; a 3-member cluster with the 3 EGFR inhibitors; and, a 6-member cluster containing all of the HSP90 inhibitors plus sunitinib. Using this solution we ran our iterative supervised refinement. Upon refinement, AG1478 and erlotinib could accurately predict gefitinib and no other drugs, settling on a 3-member EGFR inhibitor cluster. The closest four HSP90 inhibitors could accurately predict 17AAG, but all five failed to accurately predict sunitinib. This created a subnetwork update that only included the 5 HSP90 inhibitors. Finally, AA2 could not identify cantharidin as its nearest neighbor, so this 2-member cluster –one a direct activator of the apoptosis downstream of the mitochondria, and the other a phosphatase inhibitor – was not updated as a new subnetwork. Thus, using network consensus clustering and supervised prediction refinement we could identify the “New” drug subnetworks for HSP90 and EGFR inhibitors that had appeared in our previous study (Fig. 3D).

With this success we wanted to test our approach in two other datasets, the modulatory profiling approach of Wolpaw[13] and the Connectivity Map of Lamb.[3] We aimed to test the generalization of the analysis strategies in different data types, and to examine a more natural workflow in datasets where more powerful analysis strategies were not fully developed.

In the Wolpaw dataset, small molecule tool compounds and some RNAi perturbations that affect broad areas of biology were used to create drug action signatures. To measure how an effect modulates small molecule action, Wolpaw quantified shifts across an entire dose effect curve. This spectrum of the dosing effects across different compounds and cell lines formed a dose response modulation signature able to hierarchically cluster drugs by putative mechanisms-of-action. Interestingly, while their clustering did not resolve topoisomerase II and I poisons, or a coherent DNA damage cluster (defined by the presence of topoisomerase poisons and akylating agents), it could identify clusters that possessed similar chemical structure (hydrophobic amines), acted in a Bax/Bak-independent manner, and seemed to kill cells by non-specific reactive mechanisms. This suggested to us that their dose effect

modulation approach could provide effective resolution, albeit with a disparate biological focus. We decided to examine their data with our new analysis strategy to compare capabilities.

Following our RNAi-based network signature approach, we found that the Wolpaw potency modulation data alone was sufficient for accurate classification. We constructed a training set comprising topoisomerase II poisons, topoisomerase I poisons, spindle poisons, proteasome inhibitors, and alkylating agents that was able to accurately predict "New" mechanism-of-action classes, including erastin (a novel Bax/Bak-independent death-inducing compound), compounds sharing structures with hydrophobic amines, and mitochondrial disruptors azide and valinomycin (Fig. S3A, ESI†). Given these "New" compounds we employed network consensus clustering and received a three-way tie in the network–cluster solution. Given a tie, the most parsimonious solution included only 2 clusters (Fig. S3, ESI†). Supervised refinements on the most parsimonius solution reinforced this two cluster solution with 2 nodes and 1 edge (Fig. S3C, ESI†). These 2 clusters contained the erastin profiles, and the hydrophobic amines (Fig. S3B–D, ESI†). With these new subnetworks defined, an updated dataset incorporating these 2 subnetworks was generated to examine the statistical and biological generalization of our approach in the Wolpaw dataset. With respect to statistical generalization, our method was able to predict Mitoxantrone as a topoisomerase II poison, as well as numerous novel spindle poisons that were also accurately predicted to be spindle poisons. To examine biological generalization, we tested our extended model's predictive capability on a Bax/Bak-independent erastin-like compound along with diverse new hydrophobic amines that were not part of the subnetwork update (Fig. S3E, ESI†). At *p*-values of 0.1 and 0.05, our algorithm produced a TPR of 88% and 75% respectively, both with an FPR of 0%. Thus, our analysis methodology can be gainfully applied to different data types.

A current standard in the molecular signature field is the Connectivity Map (CMap).[3] In this approach, human cancer cell lines are treated with drugs, then at 6 hours harvested for genome-wide microarray mRNA expression measurement. The CMap employs a rank-based pattern matching methodology to connect a query signature of mRNAs to other microarrays in the database. While in theory the lists generated by querying a signature can provide many potential hypotheses, in practice the lists require post hoc analysis to gain confidence in a proposed prediction.

As a first step toward testing our approach on the CMap data set, we examine the associated CMap analysis search tool by exploring lists thereby generated; we evaluated a variety of inclusion and exclusion criteria and present here a post hoc analysis of the best-case search results for two categories (see ESI† and Supplemental Data). To score the searches, a drug was counted in the top 5 or 25 even if it was the drug used to search the database (to aid in producing best case results). HDAC inhibitors are one of the best categories with respect to results, and are featured in the Lamb publication.[3] Nonetheless, even within good prediction categories such as the HDAC inhibitors, the HSP90 inhibitor 17-AAG and PI3K inhibitor LY-294002 present substantial false positive issues. 17AAG and LY-294002 are always present in the top 25 results, and they are present more often than Scriptaid (a true HDACi) in the top 5 results. Accordingly, if one were to examine a TPR and FPR given the quantification of the top 5 or 25 results, the top 25 tally for HDACs would yield a best case TPR of 100% with an FPR of 66%. This threshold would allow one to accurately group Scriptaid with the other HDAC inhibitors, but at the expense of also grouping LY-294002 and 17AAG with the HDAC inhibitors (Fig. 4A). Using the top 5 as a quantification test eliminates these false positives, but concomitantly also eliminates Scriptaid. This results in a best case TPR of 75% and an FPR of 0%. Examining another category, topoisomerase II poison searches appear much poorer than HDAC signatures. The HDACi, trichostatin A,

and the topoisomerase I poison camptothecin occur more often than all of the topoisomerase II poisons in the top 25 search results, and all but doxorubicin in the top 5 search results (Fig. 4A). For topoisomerase II poisons this generates a top 25 TPR of 0% and FPR of 100%, or a top 5 TPR of 20% and FPR of 0%. These results display a clear resolution gap between the Connectivity Map approach and the previously examined functional RNAi approaches.

Given this performance for the Connectivity Map, we wondered whether our previous successes in similar drug categories could be attributed to the disparate biology associated for functional cell death measurement *vs* global mRNA measurement, or instead is a direct attribute of our analysis strategy. Thus, we aimed to improve the Connectivity Map performance for mRNA expression signatures using our drug network analysis method. Since the Connectivity Map data clusters by batch rather than by drug mechanism-of-action[3] we took a "bottom-up" approach to feature selection (a deviation from our original technique in ref. 7) (Fig. 4B). We seek to select n features from each drug subnetwork based upon the entire drug mechanism of action subnetwork, and not single agents. To do this we add 1, 10, or 100 highly predictive mRNA subnetwork features at a time. To train the model, we added features for topoisomerase II poisons, topoisomerase I poisons, alkylating agents, nucleotide depletion agents, and spindle poisons. As we iteratively added features we assessed the leave-one-out cross-validation (LOOCV) across only these training classes. We found that 10 mRNAs per training category could adequately train a *K*-NN classifier across all five training categories (Fig. 4C).

Based on this training of the CMap dataset using our approach, we asked whether "New" mechanisms-of-action might now be identified, and with statistical and biological generalization as we earlier proved for RNAi and potency modulation datasets. To pose a stringent test set, we selected four drug categories: PI3K inhibitors (wortmanin), HDAC inhibitors (trichostatin A, valproic acid), MTOR inhibitors (sirolimus) and HSP90 inhibitors (geldanamycin, alvespimycin), as well as negative control microarrays chosen randomly across CMap batches that could be assumed to be random hits with no common mechanism. The latter should represent "New" mechanisms-of-action for which no partner compound exists in the limited set of 5 mechanisms of action that we trained our mRNA classifier on. All of the test set compounds were predicted to have "New" mechanisms-of-action (Fig. S5, ESI[†]), so we went on to perform network consensus clustering with the goal of finding new subnetworks. Network consensus clustering was able to identify four clusters corresponding to PI3K, HDAC, and HSP90 inhibitors. Supervised refinement reinforced this cluster selection and did not alter the subnetwork assignments (Fig. 5A–C). As two of these clusters were uniformly comprised of experimental replicates of wortmanin, they were merged. Importantly, PI3K and HSP90 inhibitors – which had represented false positives for HDAC inhibitors when using the Connectivity Map's analysis approach – were properly separated in our network consensus solution.

To probe statistical and biological generalization, we updated the training set with wortmanin, alvespimycin, geldanamycin, trichostatin A, and valproic acid, and found that we could produce a statistically generalized result. We accurately predicted mitoxantrone and ellipticine as topoisomerase II poisons and podophylotoxin (along with others), as spindle poisons. Further, we found that our analysis produces a Connectivity Map mRNA signature that is biologically generalizable. We could use the same updated model that was trained on only 5 drug categories to accurately predict the identity of LY-294002, SAHA, Scriptaid, and 17AAG (examples of PI3k, HDAC, and HSP90 inhibitors respectively) while still recognizing the randomly chosen mRNA signatures from diverse batches as "New" drugs. Using our analysis approach with the Connectivity Map mRNA data, we can achieve a TPR of 90% and an FPR of 4% at a *p* value of 0.1 (Fig. 5D). Furthermore, LY-294002,

scriptaid, and 17AAG, which presented difficult tradeoffs when using the Connectivity Map methodology, were correctly predicted using our analysis.

Because we were able to enhance analysis of the Connectivity Map dataset for anti-cancer drug discovery, we sought an explanation for our methodological advance. It has been previously shown that the distribution of noise across a microarray compendium dataset can be highly correlated with the distribution of noise across experimental replicates.[6] To examine the noise distribution in the Connectivity Map dataset, we plotted the error within a drug category against the error across the entire subset of the data we examined. Reminiscent of the Hughes compendium paper, the Connectivity Map yields a within-category error distribution that maps predominantly linearly with the error distribution across the data set (Fig. 6A and D). This suggests that the majority of the error is due to experimental issues typically common to microarray measurements. However, some points lie off this distribution; they have a large distribution of error across the dataset, but low error within a drug category. The best Connectivity Map searches (*i.e.*, those for HDAC inhibitors) have a larger proportion of search features within this second distribution. Compared to even the best topoisomerase II Connectivity Map search feature sets, there were many more low category noise features in the HDAC search set (feature sets shown are from the Fig. S4, ESI[†] searches) (Fig. 6B and E). Features that were selected using our bottom up approach within the topoisomerase II poisons are dramatically enriched in this low noise section of the data set. Surprisingly, this enrichment extends beyond topoisomerase II poisons to HDAC inhibitors, which they were not trained upon (Fig. 6C and F). We suggest that sampling from this low within category error distribution could be part of the statistical rationale for the better performance of our approach. Furthermore, the high error distribution was enriched for phosphorylation or acetylation keywords, neither of which appeared as functional annotations in our highly predictive feature sets (Fig. 6G).

Since we were able to produce surprisingly accurate predictions of drug mechanism of action in the Connectivity Map, we aimed to interpret that predictive ability in a biological manner. Starting with DAVID, an online functional annotation tool, we were able to develop descriptions of the signature set, with respect to gene ontology/keyword designations, encompassing approximately 50% of the probes in the mRNA set. Next we plotted some of the average values across an entire drug category in the mRNA data set for mRNAs that were clearly associated with a consistent biology or GO term (Fig. 7). It is important to note that these interpretations are rationalizations, and only represent a snapshot of the mRNA cell state at the time point that they were taken (6 hours). Surprisingly, features trained on 5 of the 8 total drug mechanism networks generated biological interpretations in all drug categories.

For the alkylating agents, functional annotation revealed many terms related to protein stress and heat shock. A heat shock protein (HSP) response representing three Affymetrix probes that measure heat shock protein mRNA levels (HSP70A/B and HSP105) appeared to be an analog gauge, seemingly measuring the drug induced stress response state in a continuous fashion across the various mechanism of action categories in the dataset (Fig. 7A). A similar analog gauge was observed for the transcriptional regulation GO term genes identified in the topoisomerase II trained feature set (Fig. 7B) with drug categories displaying differences in the relative ordering of the amount of this transcriptional regulation signature in a continuous fashion.

Further inspection of the data revealed that in order to provide a broader interpretation of two of the training categories, we had to combine them. The nucleotide depletion and spindle poison features appear to form a molecular signature of cell cycle state. 9 of 10 Spindle poison features were associated with a mitotic cell cycle state, and included

canonical mitotic kinases, as well as regulators of the mitotic checkpoint and cytokinesis. Spindle poisons exhibited clear evidence of mitotic arrest, with all of these 9 probes greatly upregulated. Nucleotide depletion agents (GO term: DNA replication) exhibited clear evidence of an S-phase arrest (Primase and CPT1 are involved in lagging strand synthesis and pre-replication complex assembly respectively and are highly upregulated). Furthermore, adding more resolution, the G1/S cyclin Cyclin E2 appeared most strongly in the S phase arrested nucleotide depletion category (consistent with a peak in expression at S-phase). The presence of Cyclin E2 in the absence of the core replication components (as well as the absence of the mitotic signature) suggested a G1 cell cycle phase for Alkylating agents and HDACs. Absence of Cyclin E2, and mitotic genes would suggest topoisomerase I and topoisomerase II poisons are arresting the MCF7 cells in G2 by 6 hours. Interestingly the heterogeneous nature of the PI3K data may suggest asynchronous cells at the time the mRNA measurements are taken (Fig. 7C). Importantly these cell cycle rationalizations are supported by the literature since trichostatin A has been shown to induce Cyclin D1 degradation and a G1 arrest in MCF-7 cells.[20]

In contrast, the most obvious feature in the RNAi dataset is the presence or absence of a functional signature of DNA damage (p53 + CHK2). Within the DNA damage category, the character of the response (*i.e.*, whether non homologous end joining (DNAPK) is necessary for repair or apoptosis) seems to add predictive value. Furthermore, the lack of p53 and the presence of BCL-2 family members helps predictive resolution outside of DNA damage category in RNAi datasets. Thus in the RNAi dataset, the type of cell death appears to be interpretable as a set of sensitivities that is either present or absent, whereas the Connectivity Map contains an mrna impression of cell cycle state and a cell stress/transcription gauge (Fig. 7D).

Discussion

Molecular and phenotypic signatures are becoming an increasingly common tool in drug discovery. From yeast and zebrafish, to mouse and human cell lines[2–7,21] diverse model systems and phenotypes are being used to produce large multidimensional datasets describing drug action. Historically, the analysis of these datasets has been predominantly exploratory and unsupervised, or it has focused on identifying novel connections between drugs.[22,23] Here we suggest an integrated approach to data analysis that could be used to solve many of the problems associated with adopting molecular signatures into routine drug development practices, and jointly allows for quantitative and exploratory data analysis.

Towards this goal we have attempted to develop and test the performance of specific criteria that are important to molecular signature based drug prediction and discovery. The first, statistical generalization is an obvious but important one. The finding that *K*-NN performs better in small training sets is likely due to the fact that highly non-parametric and non-linear decision boundaries can be accommodated by the algorithm.[24] This may also explain why the gap in prediction capability decreases as sample size increases.

An important goal of signature based drug discovery is the ability to recognize a drug that has a novel mechanism of action. Beyond this recognition of a new drug mechanism, a signature based approach must also be capable of predicting other drugs that share this mechanism. It is often assumed that because they are unbiased, genome wide approaches are intrinsically sufficient for this task.[3] However, beyond the intrinsic bias of a single type of genome-wide measurement, we show that prediction using a large database of genome wide measurements can be improved by reducing the number of mRNAs that are utilized. This reduction appears to eliminate noisy features that inhibit the signature based recognition of similar drugs or even replicates of the same compound across batches. In our original study,

we investigated a very limited set of biology that was directly related to cell death and DNA damage.[7] However, this limited set of cell death genes proved to be sufficient to describe entirely distinct drug mechanisms of action that the set was never derived upon. Given this success, we wondered whether this biological generalization was a function of the biology that we were focusing upon (cell death/DNA damage) or our analysis strategies (such as the drug mechanism centric feature selection). Given the data in Fig. 5–7 we suggest that a portion of the generalization is due to our analysis strategies. This would suggest that the analysis pipeline presented here can be implemented with molecular signatures that are implemented on diverse platforms for different purposes.

A large strength of our strategy lies in its implementation of quantitative decision making, and its potential for increasing automation. We suggest that the ability to rapidly recognize new drugs, update entirely new drug mechanisms that the models are not trained upon, and produce accurate supervised predictions provides a simple, intuitive, and iterative informatics approach in an inherently iterative process. Furthermore, the grouping of “New” drugs into high confidence functional groups should accelerate the biochemical investigation of completely novel compounds by providing a family of similar compounds to investigate. However, while we have attempted to estimate the sensitivity and specificity of our approach across diverse drug categories and data sources, a large scale implementation will be necessary to estimate the true sensitivity and specificity of the approach in a particular screen during the applied practice of drug discovery. Furthermore, a thorough understanding of the exact predictive performance of the analysis strategy in a particular dataset will need to be informed by the chemical library that is utilized in a given screen. Thus the relative prior probability of obtaining “new” *versus* established drug mechanisms of action in chemical libraries of diverse sizes will ultimately dictate the usability of even the most sensitive and specific approaches.

Finally, because our signatures can achieve biological generalization across diverse drug categories, and the signatures are biologically interpretable (Fig. 7), they hold the potential for novel biological insight into diverse processes. For instance, 9 out of 10 mRNAs selected to predict spindle poisons were enriched for a mitotic signature by Gene Ontology. The 10th mRNA was a poorly characterized protein, ZNF165, that is thought to be expressed specifically in the testis.[25] We propose that its similarity in expression to mitotic genes may suggest cell cycle regulation and potentially a mitotic function in MCF7 cells. Thus, novel biological hypotheses can be generated from these signature based approaches as well.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was partially supported by the NCI Integrative Cancer Biology Program grant U54-CA112967. We would like to thank Boyang Zhao for a critical and insightful reading of this manuscript.

References

1. Paull KD, Shoemaker RH, Hodes L, Monks A, Scudiero DA, Rubinstein L, Plowman J, Boyd MR. J Natl Cancer Inst. 1989; 81:1088–1092. [PubMed: 2738938]
2. Rihel J, Prober DA, Arvanites A, Lam K, Zimmerman S, Jang S, Haggarty SJ, Kokel D, Rubin LL, Peterson RT, Schier AF. Science. 2010; 327:348–351. [PubMed: 20075256]
3. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet JP, Subramanian A, Ross KN, Reich M, Hieronymus H, Wei G, Armstrong SA, Haggarty SJ, Clemons PA, Wei R, Carr SA, Lander ES, Golub TR. Science. 2006; 313:1929–1935. [PubMed: 17008526]

4. Wolpaw AJ, Shimada K, Skouta R, Welsch ME, Akavia UD, Pe'er D, Shaik F, Bulinski JC, Stockwell BR. Proc Natl Acad Sci U S A. 2011; 108:E771–E780. [PubMed: 21896738]
5. Krutzik PO, Crane JM, Clutter MR, Nolan GP. Nat Chem Biol. 2008; 4:132–142. [PubMed: 18157122]
6. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD, Kidd MJ, King AM, Meyer MR, Slade D, Lum PY, Stepaniants SB, Shoemaker DD, Gachotte D, Chakraburtty K, Simon J, Bard M, Friend SH. Cell. 2000; 102:109–126. [PubMed: 10929718]
7. Jiang H, Pritchard JR, Williams RT, Lauffenburger DA, Hemann MT. Nat Chem Biol. 2011; 7:92–100. [PubMed: 21186347]
8. van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH. Nature. 2002; 415:530–536. [PubMed: 11823860]
9. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Science. 1999; 286:531–537. [PubMed: 10521349]
10. Lenz G, Wright G, Dave SS, Xiao W, Powell J, Zhao H, Xu W, Tan B, Goldschmidt N, Iqbal J, Vose J, Bast M, Fu K, Weisenburger DD, Greiner TC, Armitage JO, Kyle A, May L, Gascoyne RD, Connors JM, Troen G, Holte H, Kvaloy S, Dierickx D, Verhoef G, Delabie J, Smeland EB, Jares P, Martinez A, Lopez-Guillermo A, Montserrat E, Campo E, Braziel RM, Miller TP, Rimsza LM, Cook JR, Pohlman B, Sweetenham J, Tubbs RR, Fisher RI, Hartmann E, Rosenwald A, Ott G, Muller-Hermelink HK, Wrench D, Lister TA, Jaffe ES, Wilson WH, Chan WC, Staudt LM. For the Lymphoma/Leukemia Molecular Profiling Project. N Engl J Med. 2008; 359:2313–2323. [PubMed: 19038878]
11. Chuang HY, Lee E, Liu YT, Lee D, Ideker T. Mol Syst Biol. 2007; 3:140. [PubMed: 17940530]
12. Bousquet, O.; Boucheron, S.; Lugosi, G. Introduction to Statistical Learning Theory. Springer; Heidelberg, Germany: 2004.
13. Wolpaw AJ, Shimada K, Skouta R, Welsch ME, Akavia UD, Pe'er D, Shaik F, Bulinski JC, Stockwell BR. Proc Natl Acad Sci U S A. 2011; 108:E771–E780. [PubMed: 21896738]
14. Mitchell, TM. Machine Learning. McGraw-Hill; New York: 1997.
15. Freund Y, Schapire RE. J Comput Syst Sci. 1997; 55:119–139.
16. Bowman, AW.; Azzalini, A. Applied smoothing techniques for data analysis: the kernel approach with S-Plus illustrations. Clarendon Press; Oxford University Press; Oxford New York: 1997.
17. Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T. Bioinformatics. 2011; 27:431–432. [PubMed: 21149340]
18. Huang da W, Sherman BT, Lempicki RA. Nat Protoc. 2009; 4:44–57. [PubMed: 19131956]
19. Atkinson HJ, Morris JH, Ferrin TE, Babbitt PC. PLoS One. 2009; 4:e4345. [PubMed: 19190775]
20. Alao JP, Lam EW, Ali S, Buluwela L, Bordogna W, Lockey P, Varshochi R, Stavropoulou AV, Coombes RC, Vigushin DM. Clin Cancer Res. 2004; 10:8094–8104. [PubMed: 15585645]
21. Parsons AB, Lopez A, Givoni IE, Williams DE, Gray CA, Porter J, Chua G, Sopko R, Brost RL, Ho CH, Wang J, Ketela T, Brenner C, Brill JA, Fernandez GE, Lorenz TC, Payne GS, Ishihara S, Ohya Y, Andrews B, Hughes TR, Frey BJ, Graham TR, Andersen RJ, Boone C. Cell. 2006; 126:611–625. [PubMed: 16901791]
22. Campillos M, Kuhn M, Gavin AC, Jensen LJ, Bork P. Science. 2008; 321:263–266. [PubMed: 18621671]
23. Iorio F, Bosotti R, Scacheri E, Belcastro V, Mithbaokar P, Ferriero R, Murino L, Tagliaferri R, Brunetti-Pierri N, Isacchi A, di Bernardo D. Proc Natl Acad Sci U S A. 2010; 107:14621–14626. [PubMed: 20679242]
24. Jain AK, Raudys SJ. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1991; 13:252–265.
25. Tirosvoutis KN, Divane A, Jones M, Affara NA. Genomics. 1995; 28:485–490. [PubMed: 7490084]

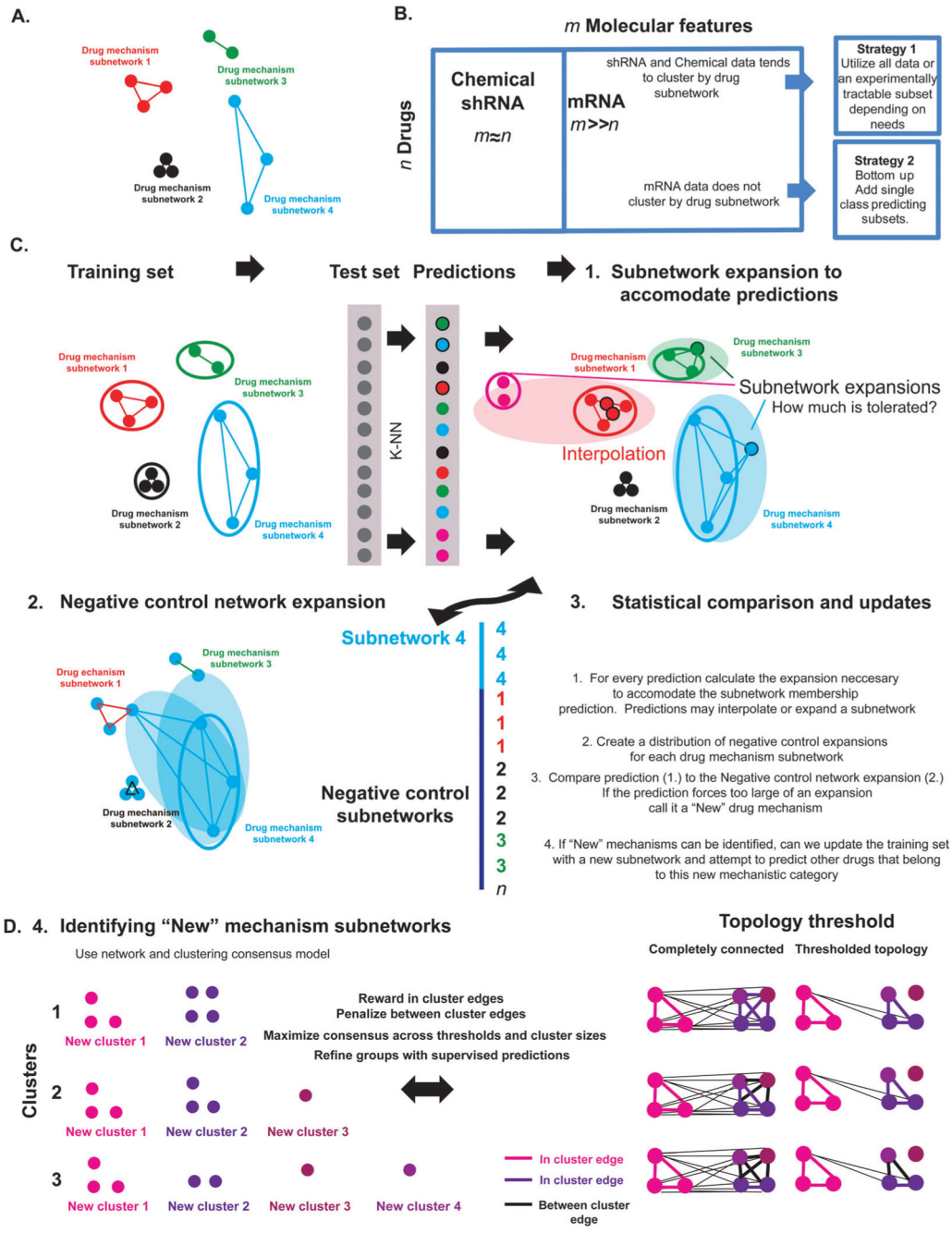


Fig. 1.

An overview of the proposed methodology. (A) Drugs are represented as nodes in a network. The network is supervised, such that each drug belongs to a single subnetwork (colored differently). Every node is connected to every other node by an edge with a pairwise distance attribute. These drug mechanism subnetworks are the starting point for all analyses. (B) In this paper we use three distinct datasets, whose matrices are represented by the blue boxes. The matrices are n drugs by m features. Different datasets have different matrix sizes and error/performance considerations, and as such require distinct methods of feature selection. (C) Given a training set as denoted in A, a drug mechanism subnetwork has a distinct size (denoted by the colored circles). Given a test set with uncharacterized

drugs (grey rectangle + circles) we perform *K*-NN predictions to get a putative mechanism of action (colored circles in the grey rectangle). Given this prediction (outlined circles), a prediction may interpolate within a subnetwork (red circles), or be extrapolated and form a category expansion (blue, green and pink circles where opaque colored ovals represent subnetwork expansions sizes). Given these expansions in (1), how much expansion of a subnetwork should be allowed to accommodate a prediction? To approach this question we perform negative control network expansion measurements. To estimate an allowable expansion for the blue subnetwork 4, we force all other subnetworks to become subnetwork 4, and create a vector of how much each of these negative controls expands the blue subnetwork 4 (blue opaque disks). The vector of negative control distributions is then used to generate a *p*-value (3) for how likely a given prediction's subnetwork expansion would be if it were from the negative control distribution. (D) To identify "New" mechanism subnetworks, we implement a consensus approach, that given clustering solutions (denoted 1,2,3) with different clusters, we use the topology of a thresholded drug–drug network to systematically choose the best network–cluster combination. We reward within cluster edges (colored lines) and penalize between cluster edges (black lines). Finally, we iteratively refine this solution utilizing our supervised methodology from C.

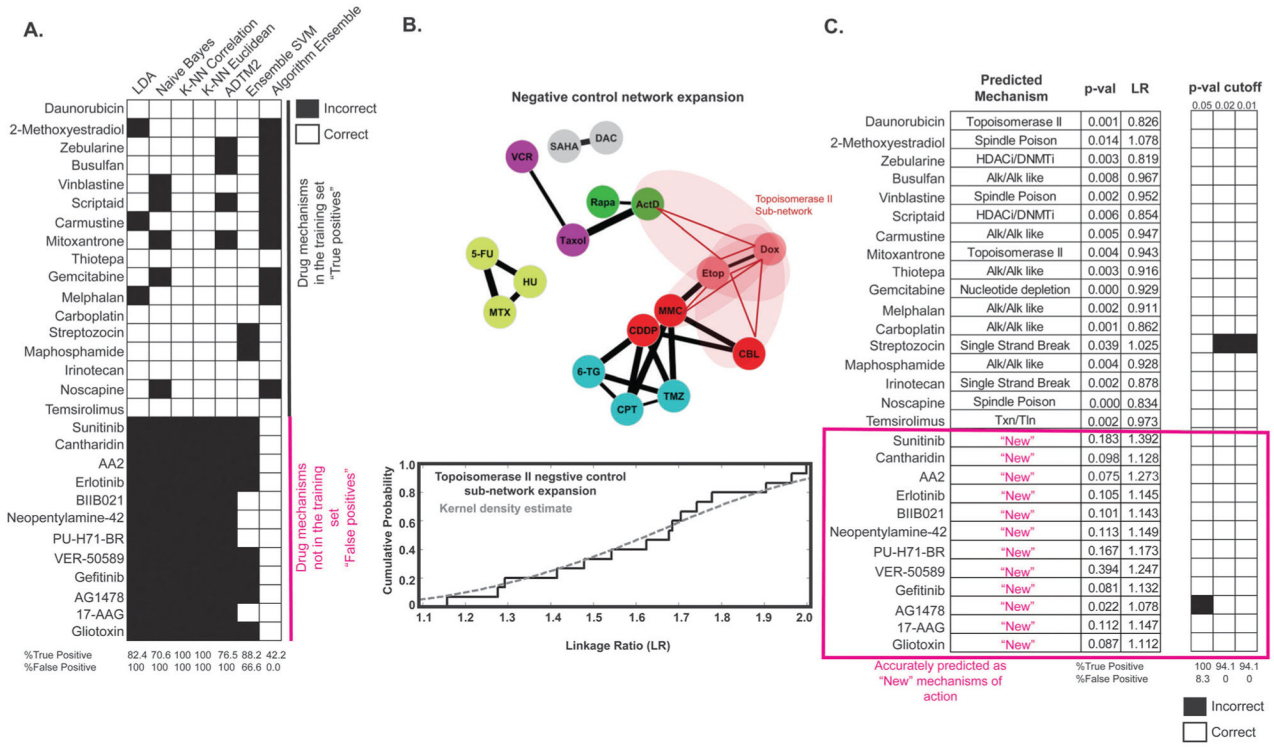


Fig. 2.

A strategy to predict "New" mechanisms of action. (A) A test set of drugs, some of which have mechanisms within the training set, and some which don't (pink) were tested with a variety of algorithmic approaches. Linear discriminate analysis (LDA), Naïve Bayes, *K*-nearest neighbors (*k*-NN) with different metrics, multiclass alternating decision trees (ADTM2), an ensemble of binary support vector machines (SVM), and a multi-algorithm ensemble were used. Their performance across these drugs is denoted in black and white. (B) Top: a graphic illustrating negative control network expansion is superimposed upon a thresholded topology of the training set, where line thickness is weighted by the pairwise euclidean distance (small = close). Bottom: a graph of the negative control network expansion ecdf and the kernel density estimate of the cdf for topoisomerase II poisons. The fit is representative of all categories. (C) Using the same drugs as in A, the subnetwork expansion statistic is applied to the *K*-NN neighbors predictions. The results are shown. *p*-Values that are greater than a threshold are considered "New" drugs.

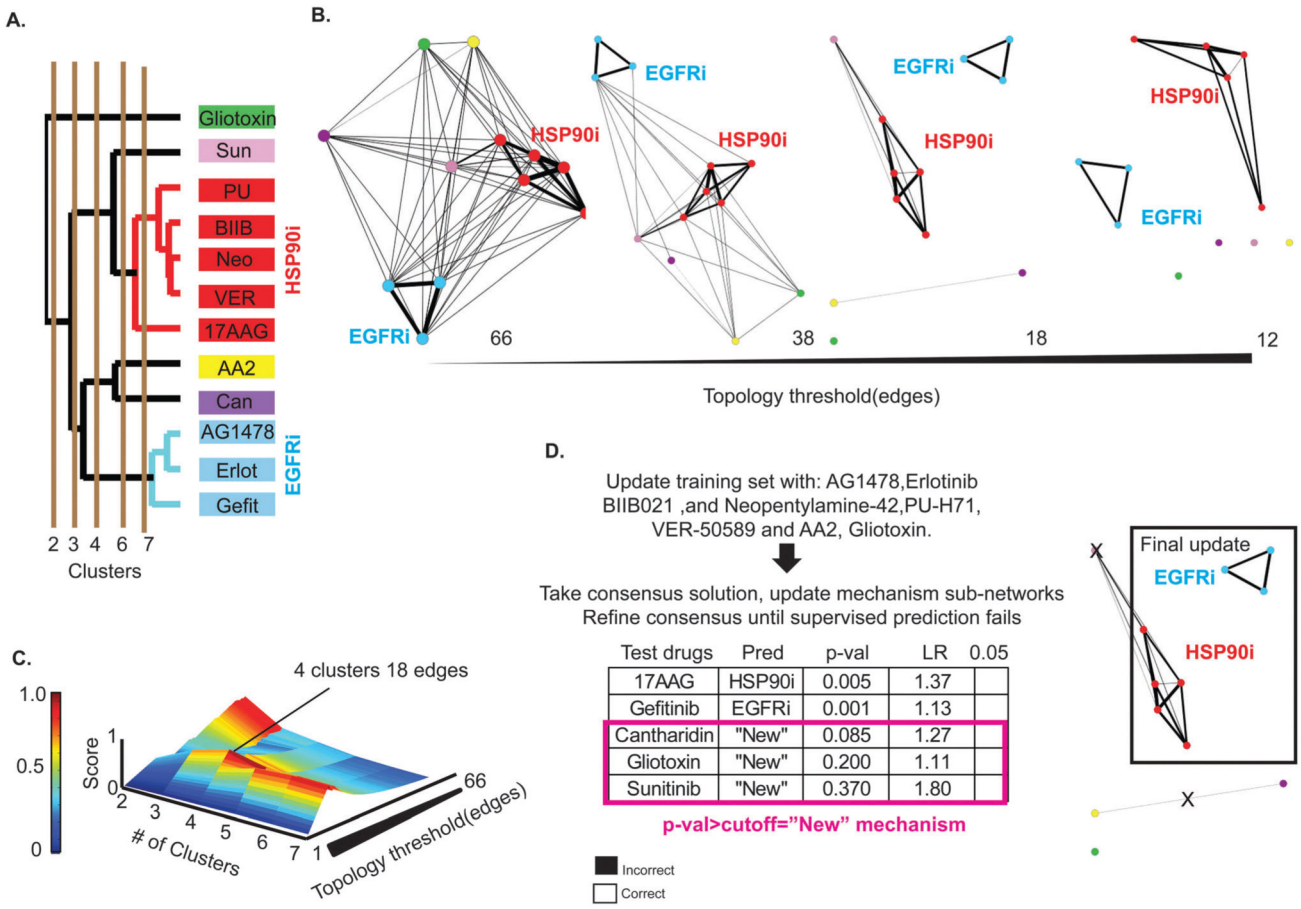


Fig. 3.

A strategy to update new drug mechanism subnetworks. (A) A hierarchical clustergram with average linkage. This clustergram contains all of the "New" drugs in an attempt to test an approach to updating new subnetworks to the training set. Brown lines represent potential clustering interpretations. (B) As network edges are thresholded by their pairwise distances, the number of the edges decreases. Node colors in B match colors in A and are consistent with the final update interpretation. Line thickness is proportional to distance, with thicker lines indicating closer distances. Layouts are weighted force directed layouts (weightings are based upon the pairwise distance attribute of the edges). (C) A heatmap and surface plot that gives the values of the scoring function across multiple cluster interpretations, and all possible thresholded topologies. The color and height represent the value of the score. (D) Final updates are refined using the supervised methodology. Predictions and *p*-values are shown in the box. The final refinement is indicated in the box adjacent to the prediction table. Xs indicate refinements made in this final step. The final solution is the biologically generalized solution that was previously shown in Jiang *et al.* 2011.

A.

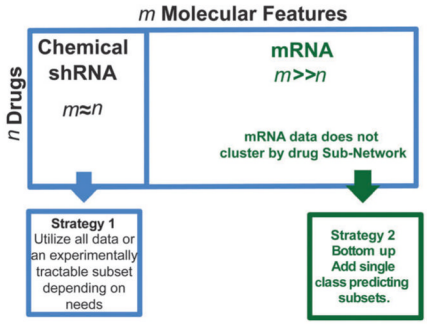
Compiled HDAC searches

Drug	%in Top 25	%in Top 5
SAHA(Vorinostat)	100	100
Trichostatin A	100	100
Valproic Acid	100	56.3
Scriptaid	100	0
17AAG(Tanespimycin)	100	25
Geldanamycin	87.5	31.3
LY-294002	100	18.8

Compiled Topoisomerase II searches

Drug	%in Top 25	%in Top 5
Etoposide	9.7	3.2
Doxorubicin	41.9	35.5
Daunorubicin	45.2	19.4
Ellipticine	41.9	6.5
Mitoxantrone	29.0	19.4
Trichostatin A	74.2	38.7
Camptothecin	71.0	25.8

B.



C.

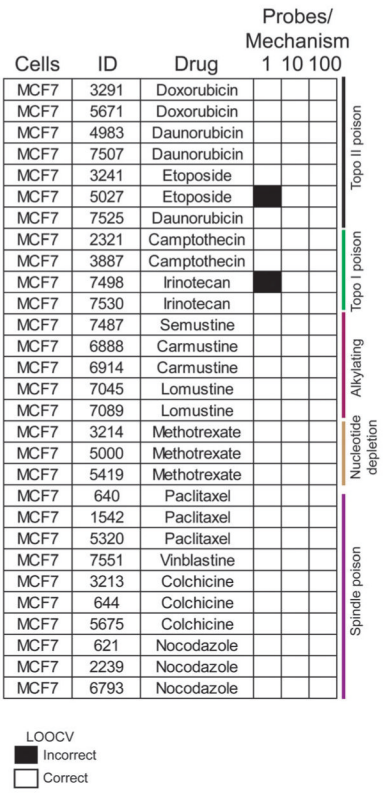


Fig. 4.

An examination of the Connectivity Map data. (A) HDAC inhibitor searches (top) and topoisomerase II poison searches (bottom) are compiled from Connectivity Map searches. Red indicates a drug that would be considered a false positive. % in top refers to the percentage of the time that a given drug appeared in the top n searches. (B) A schematic of the manner in which the Connectivity Map feature subset that we use was built. The Connectivity Map has dramatically different matrix dimensions. These are denoted by the box size. The number of molecular features is very high, but the data in this high dimensional feature space does not cluster drugs by drug mechanism or even drug replicates outside of a common batch.[3,13] Failure to cluster across batches suggested correlated batch noise, so after selecting drug arrays from across batches, we took a bottom up approach to feature selection.(C) The results of leave-one-out cross validation (LOOCV) for the five mechanisms of action that the model was trained upon (denoted on the right of the grid), as features were added from the bottom up.

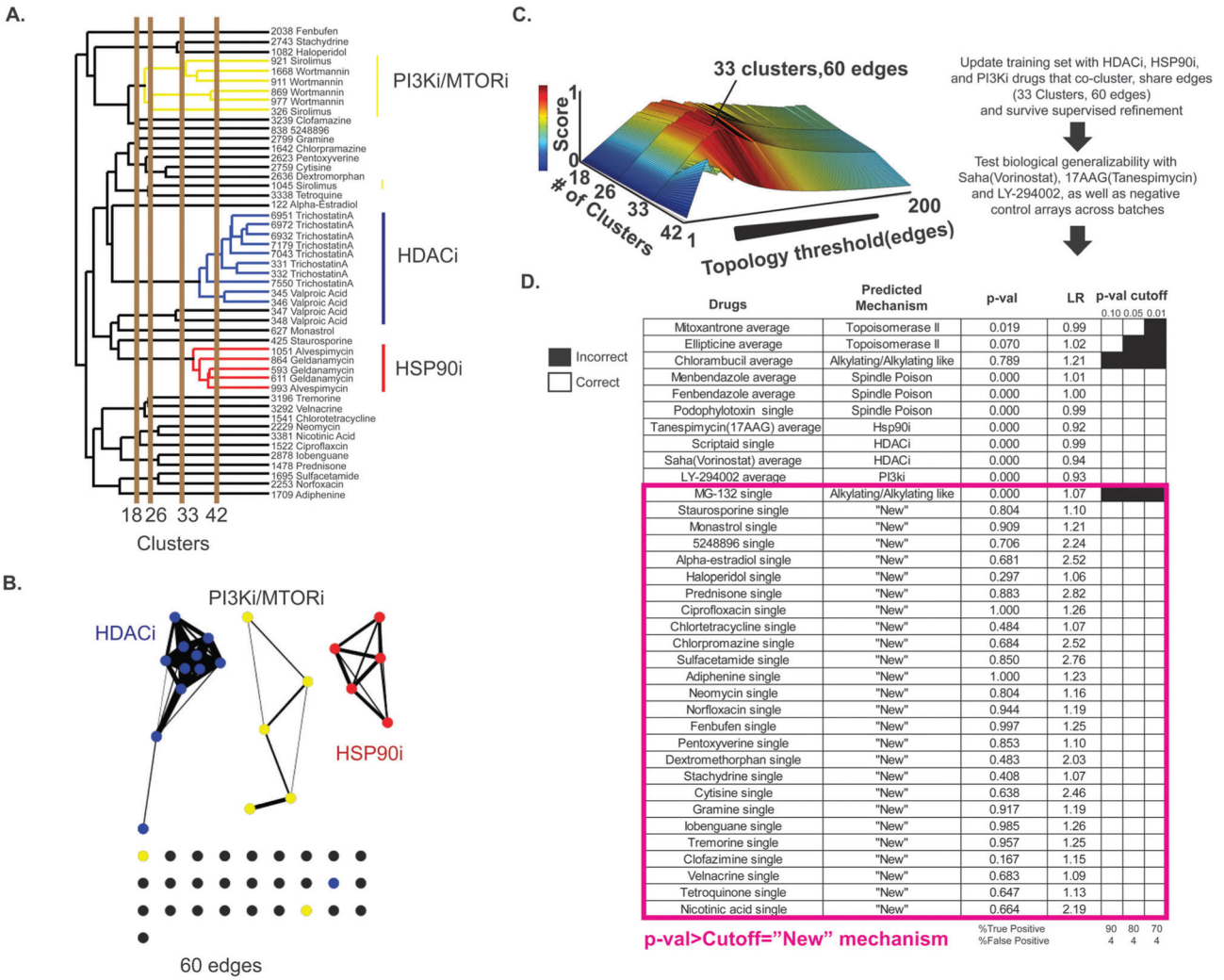
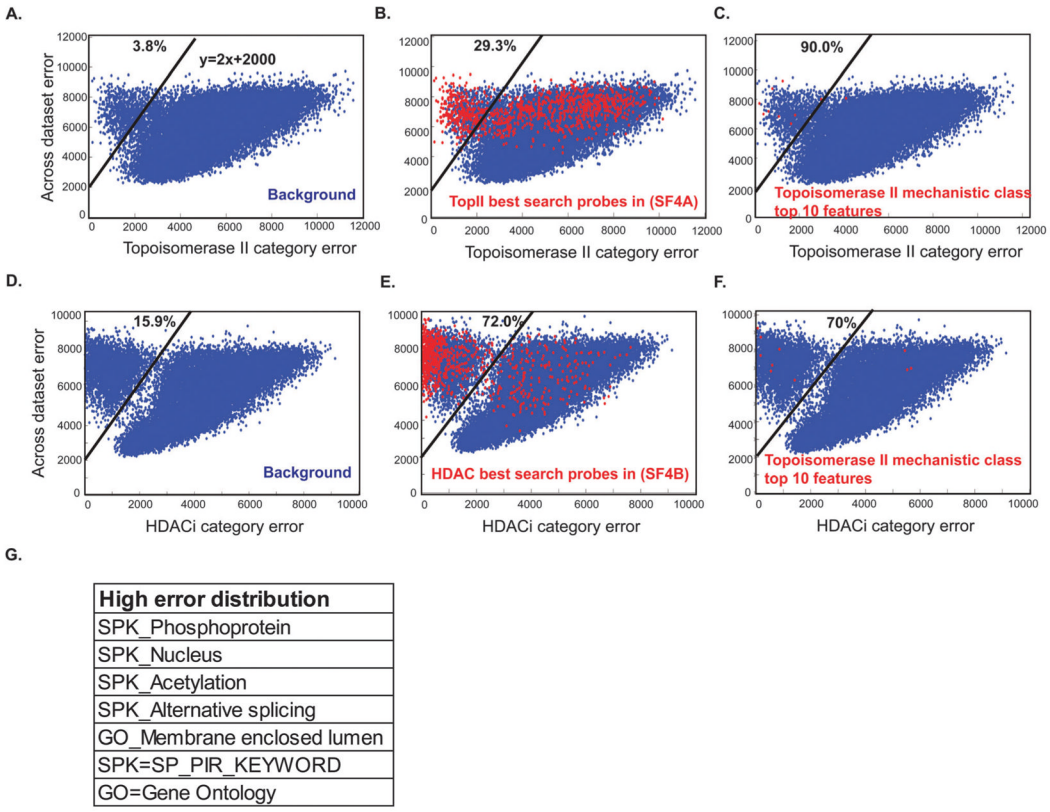


Fig. 5.

Examining statistical and biological generalization in the Connectivity Map data. (A) The hierarchical clustering of “New” drugs from the Connectivity Map data. Connectivity Map instance ids are appended to the front of the drug name. Brown lines denote representative clustering solutions. Colored bars denote some of the mechanisms being tested for biological generalization across a background of randomly chosen arrays. (B) A 60 edge thresholded network, organized using a weighted force directed layout. Edge width and layout was weighted based upon pairwise distances. Colors are coordinated to A. Black circles are for the randomly chosen negative control arrays. (C) A surface/heatmap of the network consensus clustering scores across the cluster interpretations in A. All edge thresholds less than 200 are shown. The maximum solution is indicated above the surface. (D) A test of the statistical and biological generalization of the Connectivity Map. The top 10 drugs that are true positives from 6 different drug mechanism subnetworks. True positives with multiple arrays were averaged. Arrays with only one instance were predicted on the basis of that single array data. These are denoted “average” or “single”.



G.

High error distribution
SPK_Phosphoprotein
SPK_Nucleus
SPK_Acetylation
SPK_Alternative splicing
GO_Membrane enclosed lumen
SPK=SP_PIR_KEYWORD
GO=Gene Ontology

Fig. 6.

An error analysis in the Connectivity Map. (A) The within category (topoisomerase II poison error in standard deviation) plotted against the standard deviation across the dataset for all probe sets. The line $y = 2x + 2000$ is the filter used to calculate the percentages in the upper left hand corner. 3.8% is the percentage of all probe sets in the background distribution that are across this line. (B) Superimposed on the background distribution (in blue) is the error for the probe sets that were used in the "Best" topoisomerase II poison search that we identified in Fig. S4 (in red) (ESI†). 29.3% is the percentage of these probes across the filter. (C) Superimposed on the background distribution (in blue) is the error for the 10 probes that were used in the creation of our predictor. 90% are across the filter. (D) The within category (HDAC inhibitor error in standard deviation) plotted against the standard deviation across the dataset for all probe sets. The line $y = 2x + 2000$ is the filter used to calculate the percentages in the upper left hand corner. 15.9% is the percentage of all probe sets in the background distribution that are across this line. (E) Superimposed on the background distribution (in blue) is the error for the probe sets that were used in the "Best" HDAC inhibitor search that we identified in Fig. S4 (in red) (ESI†). 72.0% is the percentage of these probes across the filter. (F) Superimposed on the background distribution (in blue) is the error for the 10 topoisomerase II poison probes that were used in the creation of our predictor (in red). 70% of topoisomerase II poison probes are across the filter in the HDAC inhibitor data. (G) Gene ontology (GO) terms, Swiss-Prot (SP) and Protein information resource(PIR) keywords for the background distribution (1.25XMedian) in the HDAC category.

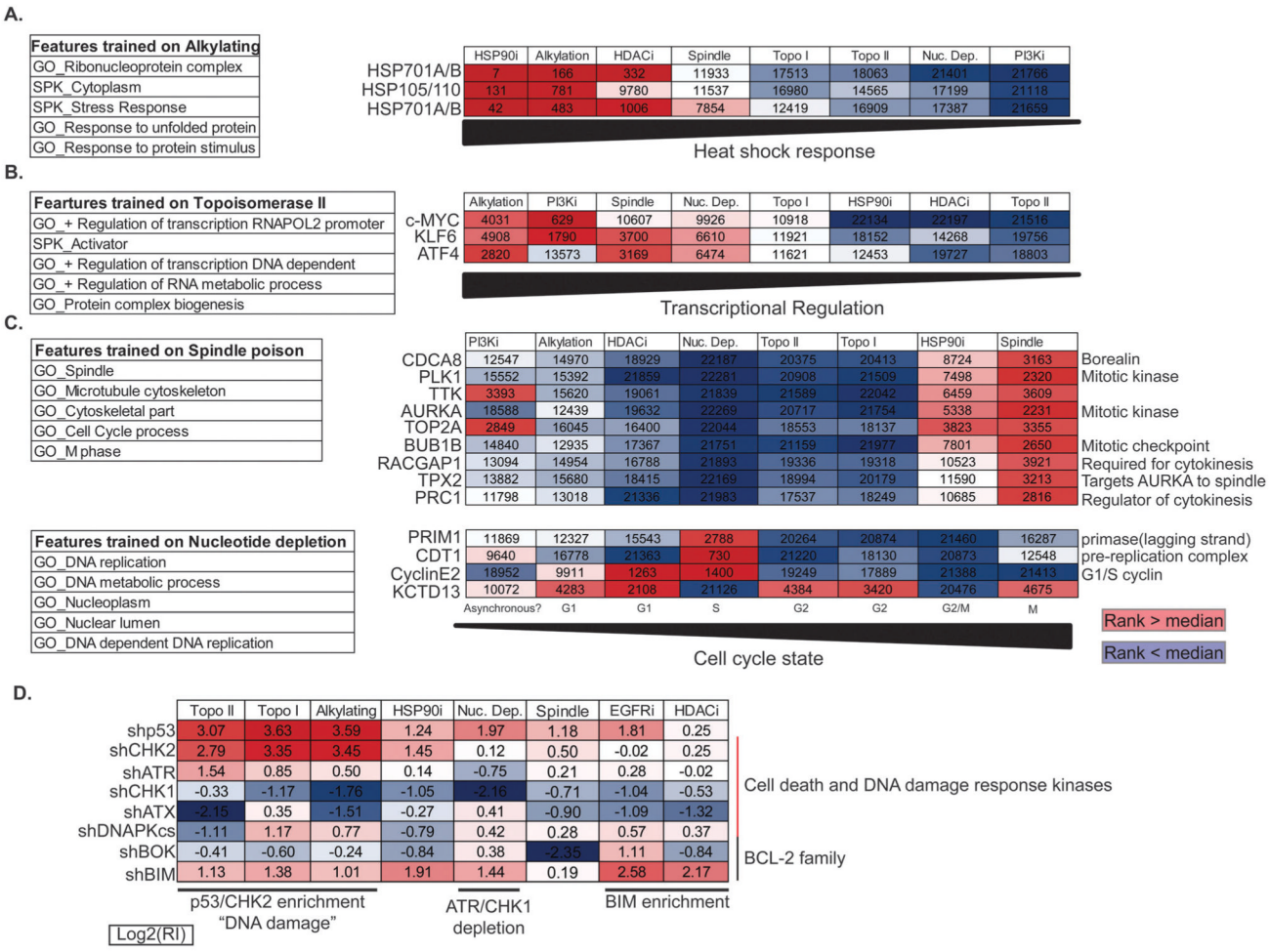


Fig. 7.

An examination of the mechanisms of biological generalization. (A) Left: Keywords for the features that were trained solely upon the alkylating agents. Right: A heatshock protein signature heatmap where entries are the average rank for a gene across all drugs in a category. Blue indicates an mRNA measurement below the median. Red represents an mRNA measurement above the median. (B) Left: keywords for the features that were trained solely on the topoisomerase II poisons. Right: a transcriptional regulation signature heatmap, where entries are the average rank for a gene across all drugs in a category. Blue indicates an mRNA measurement below the median. Red represents an mRNA measurement above the median. (C) Left: keywords for the nucleotide depletion and spindle poison categories are displayed together. Right: heatmaps for genes associated with cell cycle state. Interpretations of the heatmaps are described below. (D) RNAi signature data is displayed in heatmap form and annotated. Numbers are the average Log2(RI) value for an shRNA across all drugs in a category. Red is enrichment and blue is depletion.