Supplementary Information:

Proteome-scale detection of small-molecule protein interactions using data from gene expression perturbations

Nicolas A. Pabon[1,†], Yan Xia[2,†], Sam Estabrooks[3], Zhaofeng Ye[4], Amanda K. Herbrand[5], Evelyn Süß[5], Ricardo M. Biondi[5], Victoria A. Assimon[6], Jason E. Gestwicki[6], Jeffrey L. Brodsky[3], Carlos J. Camacho[1,*], and Ziv Bar-Joseph[2,*]

[1] Department of Computational and Systems Biology, University of Pittsburgh , Pittsburgh, PA 15213
[2] Machine Learning Department, School of Computer Science, Carnegie Mellon University, 15213
[3] Department of Biological Sciences, University of Pittsburgh, Pittsburgh, PA 15213
[4] School of Medicine, Tsinghua University, Beijing, China 100084
[5] Department of Internal Medicine I, Universitätsklinikum Frankfurt, 60590 Frankfurt, Germany
[6] Department of Pharmaceutical Chemistry, University of California, San Francisco, San Francisco, CA 94158

[†] These two authors contributed equally
* To whom correspondence should be addressed

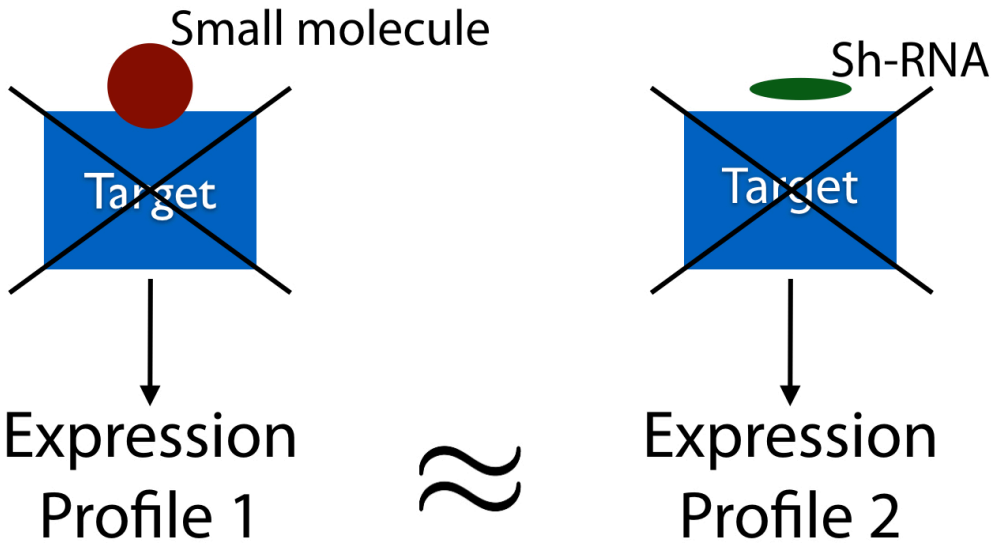
Availability

Supplementary Methods, Results, Data and Matlab code are available at the supporting website http://sb.cs.cmu.edu/Target2/.

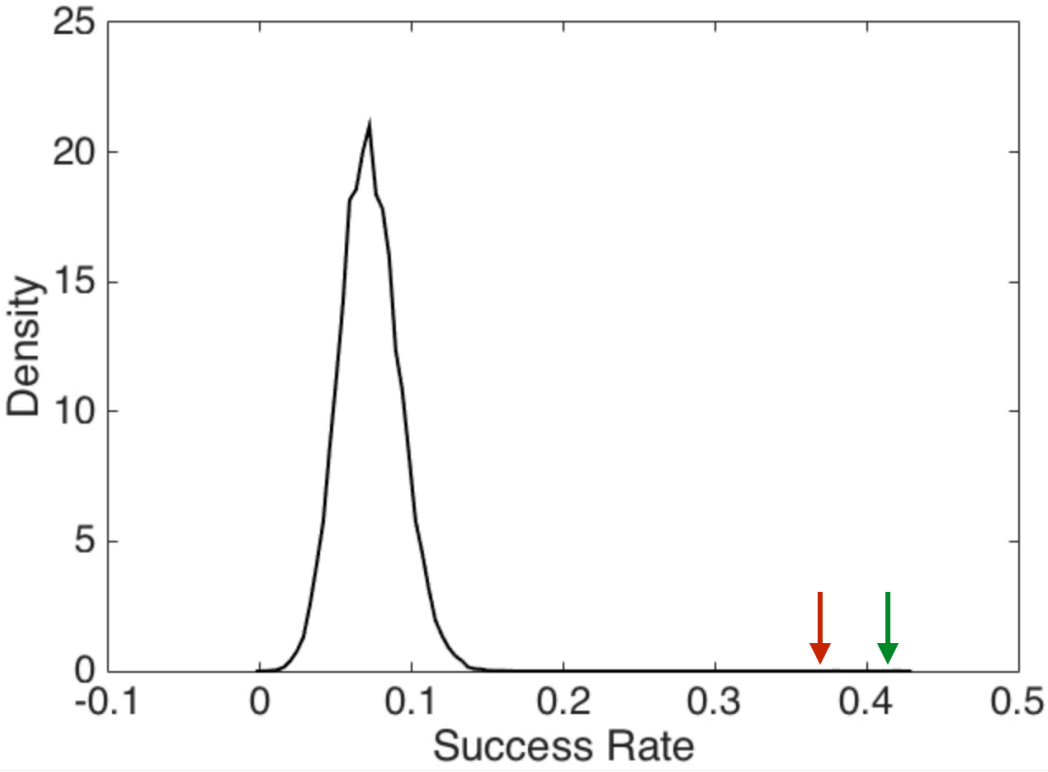
Contact

Ziv Bar-Joseph: zivbj@cs.cmu.edu
Carlos J. Camacho: ccamacho@pitt.edu

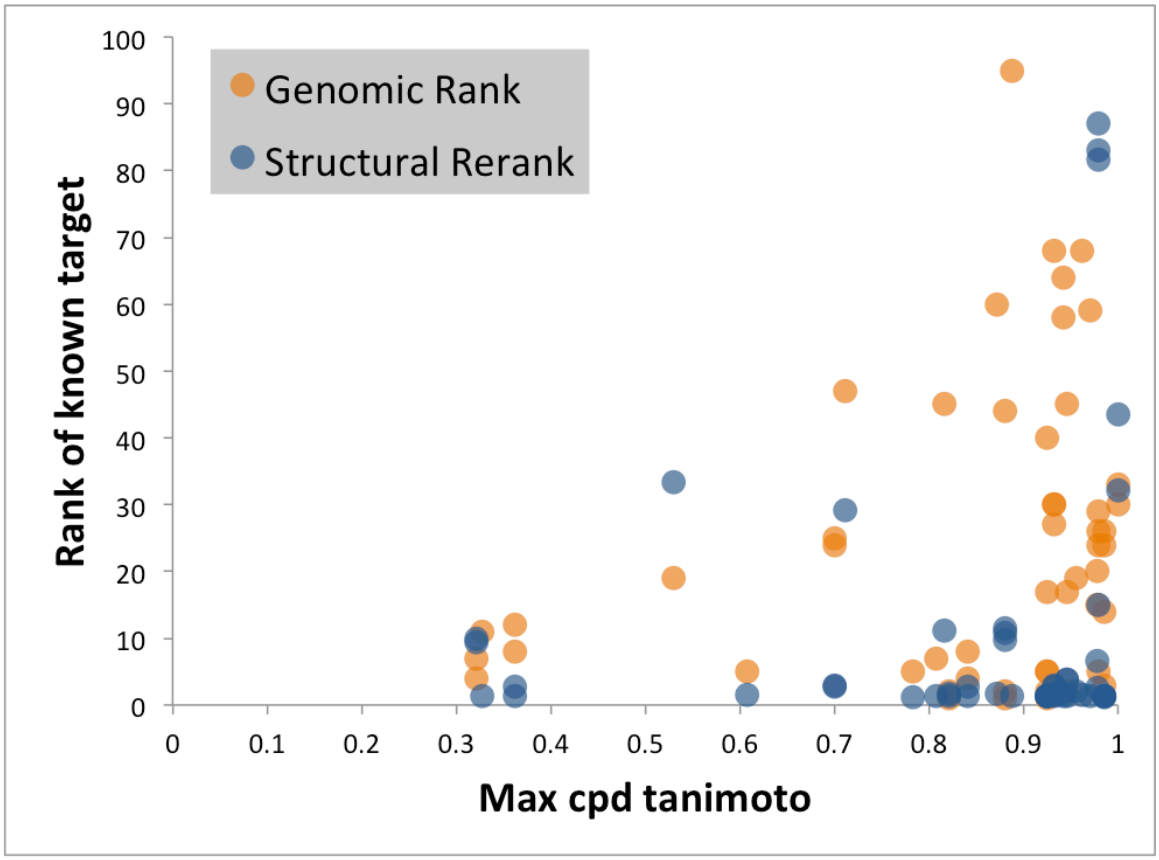
Supplementary Figure 1. Hypothesis of our genomic target prediction analysis: the effect of inhibiting a target by a small molecule is similar to that of knocking down the same target with sh-RNA. These effects can be manifested by the resulting expression profiles as well as the activation/inhibition of specific pathways. As we discuss in detail in the Online Methods, this assumption leads to the construction of several different types of features that we use to evaluate the similarity of the effects of the two treatments within the LINCS dataset.



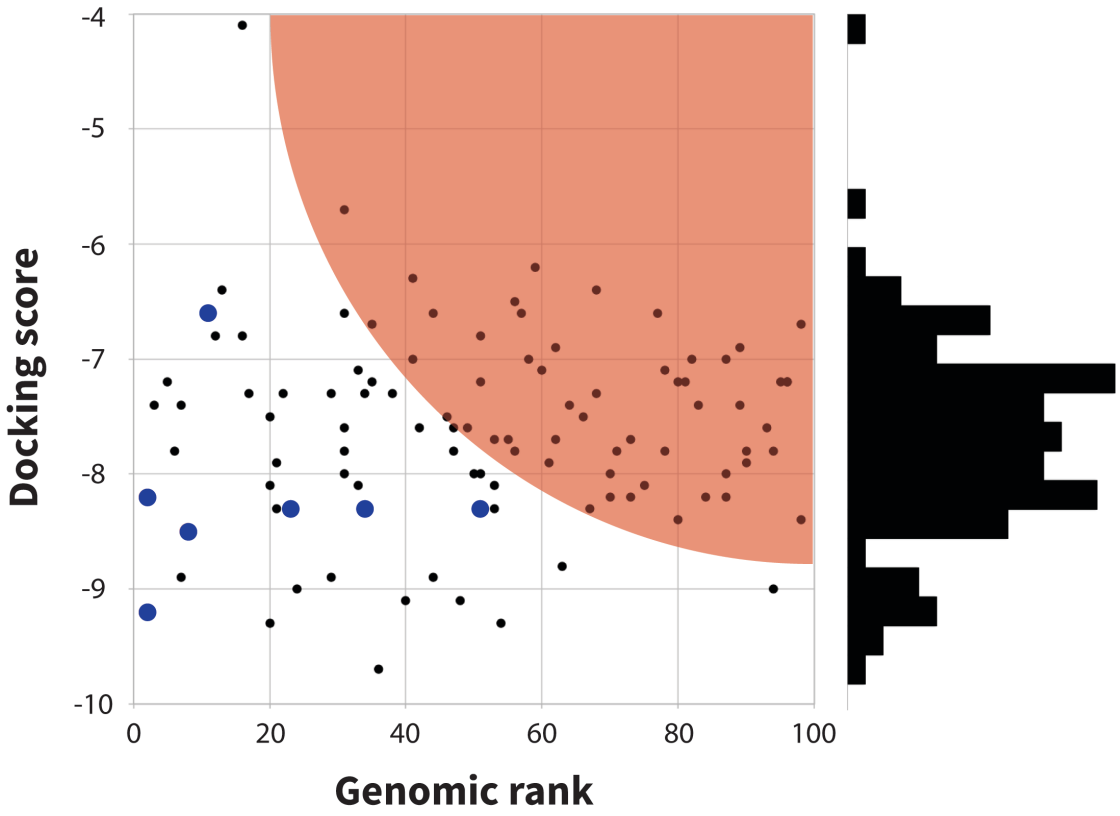
Supplementary Figure 2. **Comparing the random forest approaches with a random classifier** for predicting known targets of the 152 drugs in the validation set. The red arrow indicates the success rate of on-the-fly random forest and the green arrow represents the two-level random forest.



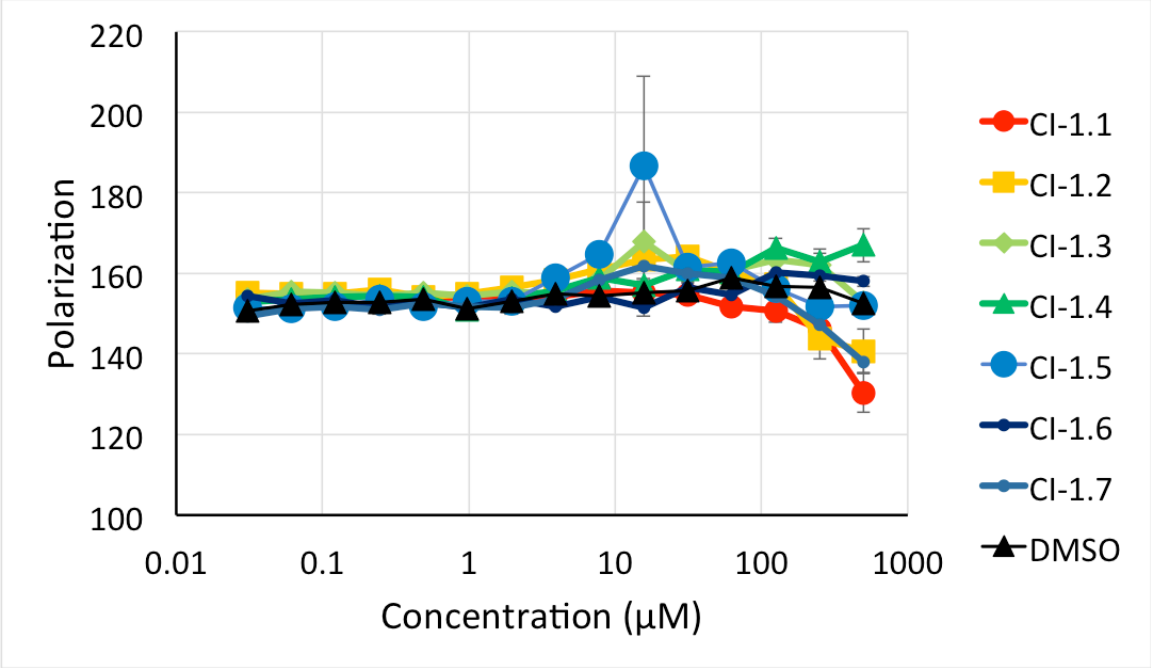
Supplementary Figure 3. Correlation of target prediction accuracy and "structural uniqueness" of the query compound with respect to the training compounds. Each point in the plot represents one of the 53 compounds in our enrichment analysis. The structural uniqueness of a compound (x-axis) is defined as its maximum Tanimoto distance to any of the training compounds. The predicted ranking of the known target for each compound is shown on the y-axis. Orange and blue points represent the ranking pre- and post- structural filtering.



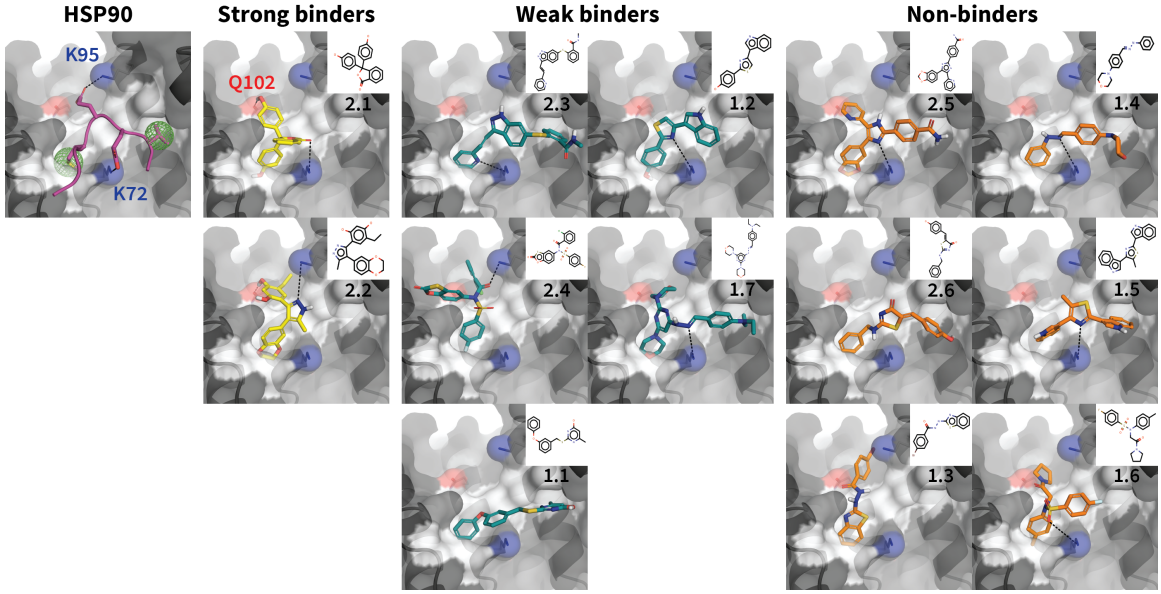
Supplementary Figure 4. Result of the target-centric screen against CHIP. The plot on the left shows the 104 compounds predicted by random forest to bind CHIP, plotted according to the rank of CHIP in their predicted targets list (x - axis), vs. their CHIP docking score (y – axis). The shaded red area of the plot represents compounds that were filtered out of analysis due to low rank/score. The blue dots represent the compounds that were purchased for experimental validation. The histogram on the right shows the distribution of compounds by docking score.



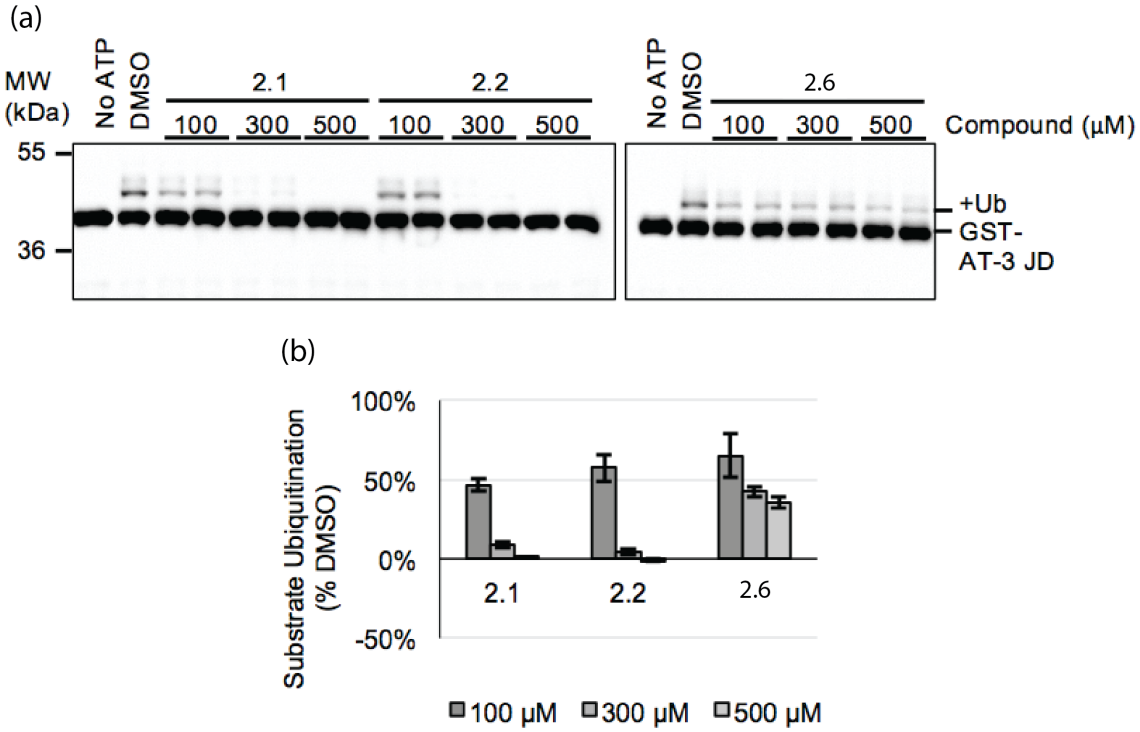
Supplementary Figure 5. Disruption of CHIP binding to chaperone peptide measured by fluorescence polarization. Results are the average and standard error of the mean of two experiments each performed in triplicate.



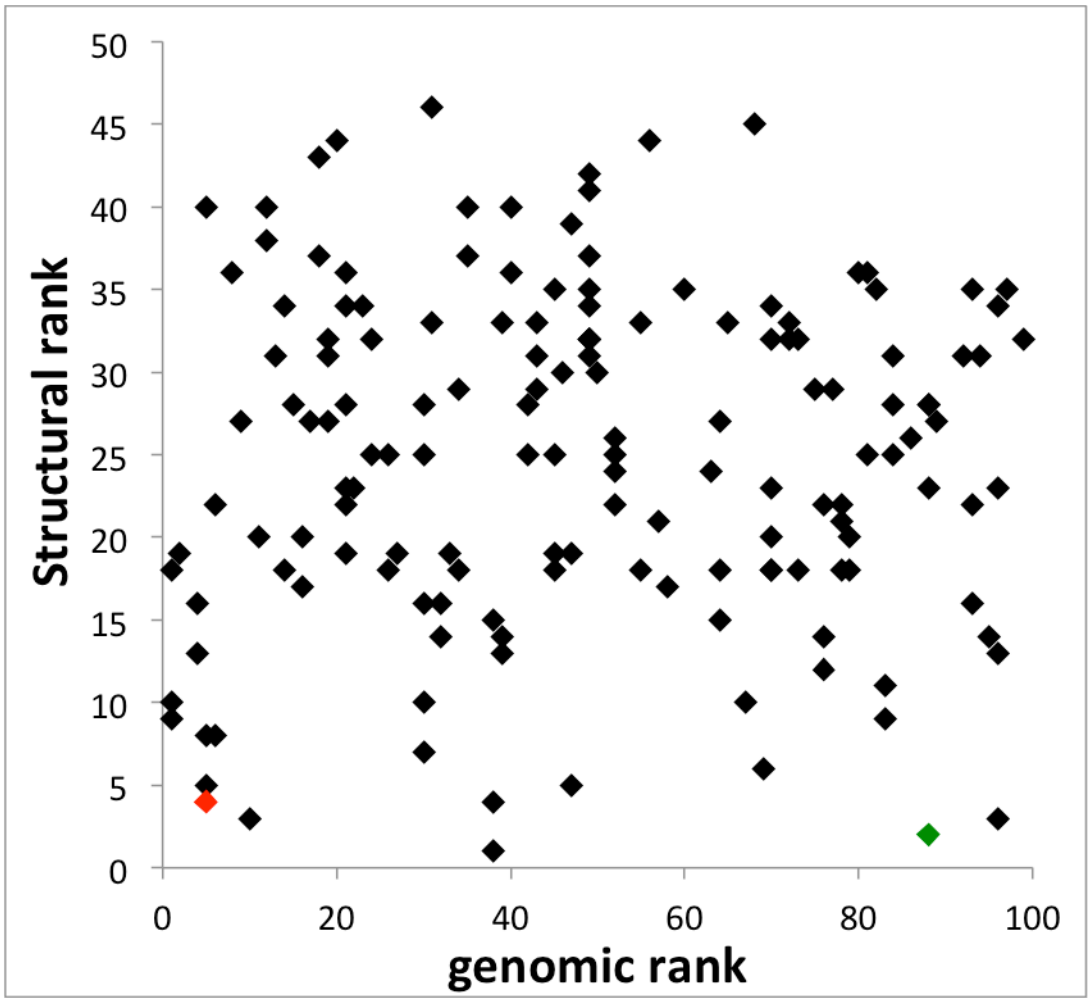
Supplementary Figure 6. Comparison of virtual screens against CHIP. **HSP90** shows structure of the CHIP (grey) - HSP90 (magenta) interface (PDB ID: 2C2L[1]), indicating the hydrophobic (green spheres) and polar contact (blue surface / dashed lines) pharmacophores used to screen the ZINC database. **Strong binders** show predicted binding modes for compounds 2.1 and 2.2 from the LINCS screen, which showed the strongest FP signal and robust inhibition of CHIP ligases activity. Interestingly, 2.1 and 2.2 are the only predicted hits to make a novel hydrogen bond to CHIP residue Q102, a contact whose importance is not obvious from the cocrystal structure show predicted binding modes. **Weak binders** show predicted binding modes for compounds 2.3 and 2.4 from the LINCS screen, and compounds 1.1, 1.2, and 1.7 from the ZINC screen, which showed modest FP signal. **Non-binders** show predicted binding modes for non-binding LINCS compounds 2.5 and 2.6, and non-binding ZINC compounds 1.3 – 1.6.



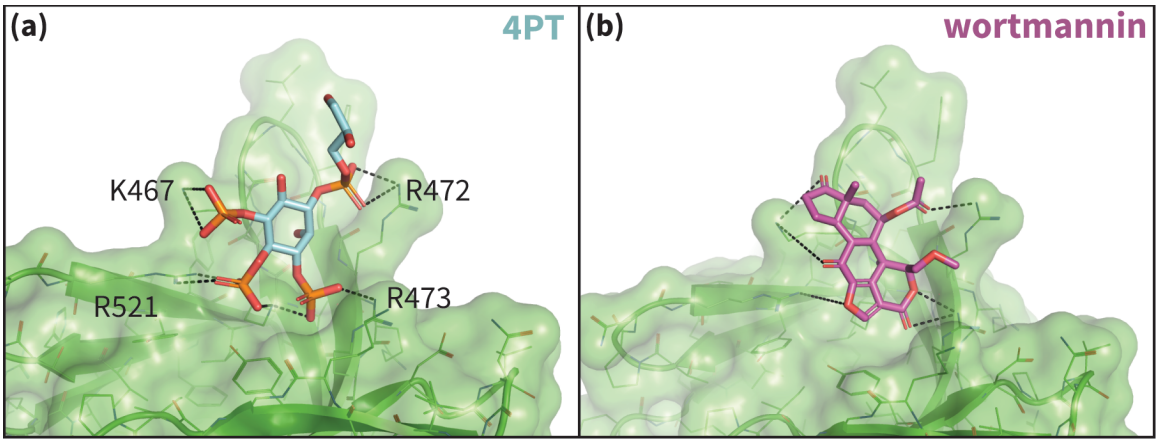
Supplementary Figure 7. **Predicted CHIP Inhibitors Prevent Ubiquitination of an Alternate Substrate.** (A) Anti-GST western blot showing AT-3 JD substrate ubiquitination by CHIP in reactions treated with compounds. (B) Quantification of all reactions as in A treated with up to 500 μM compound 2.1, 2.2, or 2.6, normalized to ubiquitination by a DMSO treated control (all compounds: N=4).



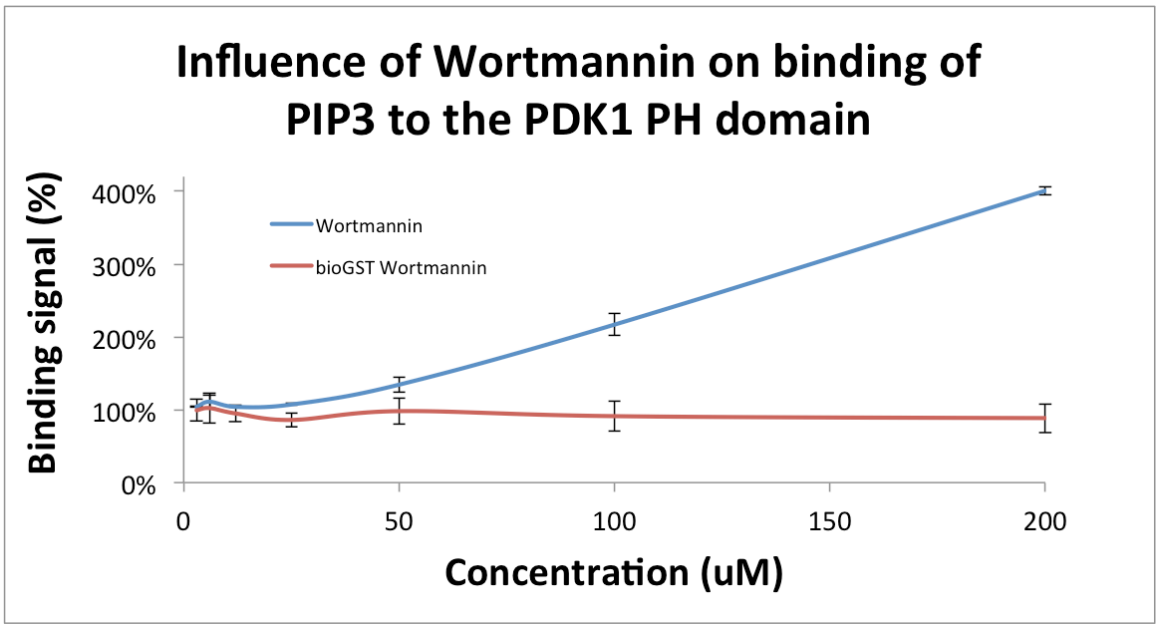
Supplementary Figure 8. Pipeline-predicted targets for the drug wortmannin. Each point on the plot represents one structural model of one potential target predicted in the top 100 for wortmannin by the random forest regressor. The random forest ranking for each target (x axis) is plotted against the docking score ranking (y axis). The red dot indicates the ranking of the known target PIK3CA. The green dot indicates the ranking for the previously unknown target, PDPK1.



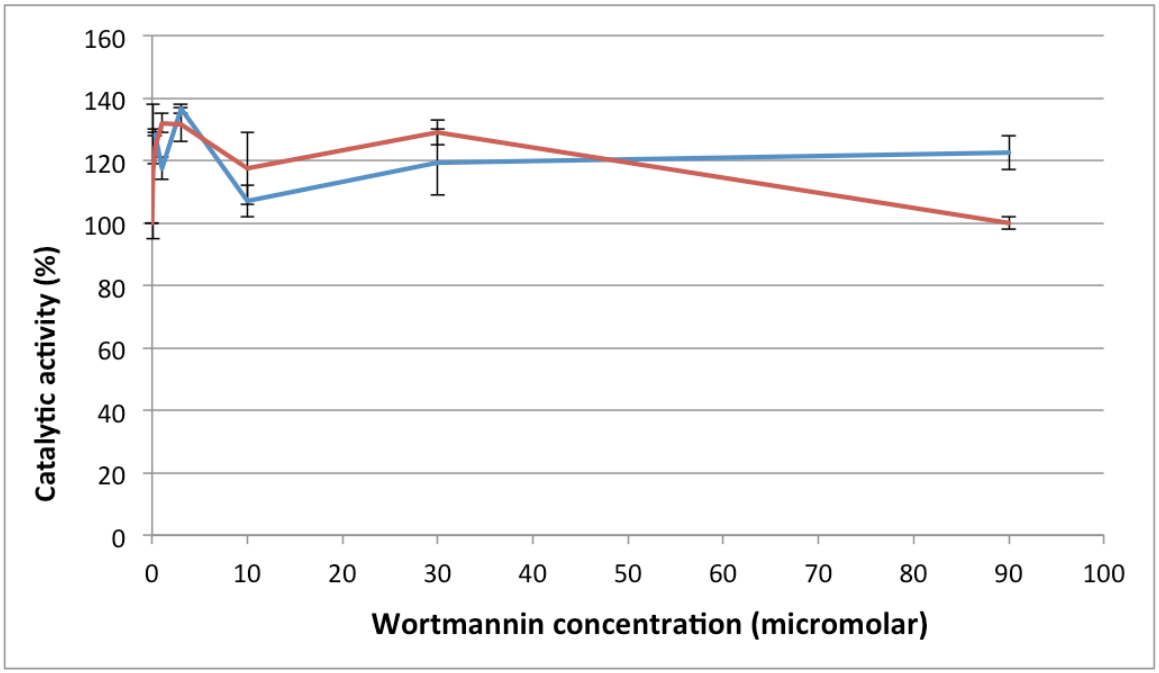
Supplementary Figure 9. Docking model of wortmannin bound to the PH domain of PDK1. (a) Cocrystal structure (PDB ID: 1W1G[2]) of the PH domain of PDK1 bound to the 4PT ligand which mimics the head group of it's natural ligand PIP3. Dashed lines indicate key polar interactions. (b) Docking model of wortmannin bound to the PDPK1 PH domain, which captures many of the same polar interactions seen in the cocrystal.



Supplementary Figure 10. Alphascreen PDK1-PIP3 interaction-displacement assay results for increasing concentrations of wortmannin.



Supplementary Figure 11. Effect of wortmannin on the in-vitro phosphorylation of the substrate T308tide by the isolated catalytic domain of PDK1. The two lines are from two replicates of the activity assay, with error bars representing the standard error on the mean from two parallel runs for each replicate.



Supplementary Table 1. **Results of testing our random forest classifier on the 123 FDA approved drugs profiled in 4-6 LINCS cell lines**, after having trained our model on the 29 FDA approved drugs profiled in all 7 LINCS cell lines. The rank of the highest-ranking known target for each compound is listed next to their LINCS ID. We achiece top-100 predictions for 32 drugs, a 26% success rate.

**** Table provided as a separate tabular data file *supp_table_1.txt***

Supplementary Table 2. Structural enrichment of random forest predictions for validation hits and comparison with existing methods. Our 63 `hits' are listed with their LINCS ID and the number of top-100 predicted targets that had structures available in the PDB. The ranking of the known targets are shown after our genomic random forest target prediction (GEN), and after our structural re-ranking (STR), along with the percentile rankings produced by alternative target prediction methods HTDocking (HTD) and PharmMapper (PHM). STR, HTD, and PHM values of 100 indicate that the structure of the known target either is not known or was not included in the set of potential targets used by the method.

** **Table provided as a separate tabular data file *supp_table_2.txt***

Supplementary Table 3. Predicted CHIP-targeting compounds out of 104 candidate molecules. 'CHIP RANK' indicates the ranking of CHIP in the random-forest predicted list of potential targets for each compound. 'CPD RANK' indicates the structure-based ranking of the compound after docking of all 104 candidate compounds to the HSP90 binding site on the CHIP-TPR domain.

Cpd #	NAME	ID	CHIP RANK	CPD RANK
2.1	phenolphthalein	BRD_K19227686	2	22
2.2	HSP90_inhibitor	BRD_K65503129	2	4
2.3	axitinib	BRD_K29905972	8	13
2.4	BRD_K59556282	BRD_K59556282	11	92
2.5	SB_431542	BRD_K67298865	34	17
2.6	MW_STK33_2B	BRD_K78930611	51	16

Supplementary Table 4. Symbols and notations

Symbol	Meaning
d	Index for a drug
c	Index for a cell line
g	Index for a gene
N_D	Total number of genes
N_C	Total number of cell lines
C_d	The set of cell line indeces for drug d
P_d	The set of protein target indeces for drug d
G_c	The set of knockdown gene indeces for cell line c
T_d	The intersection of knockdown gene indeces G_c for all cell lines in C_d
N_{dc}	Number of experiments for applying drug d to cell line c
N_{gc}	Number of experiments for knocking down gene g in cell line c
N_g	Neighbors, or protein-protein interaction partners, of gene g
Δ	Drug-response data
Γ	Gene-knockdown data
Ψ	Control data
Ω	Full feature data
X_d	Training data derived from drug d
y_d	Training label derived from drug d
v_d	Negative (non-target) genes for drug d

Supplementary Table 5. Summary of constructed feature sets. Note that different feature sets can have different dimensions (some contain values for each of the cell lines, etc…). The exact dimension and content of each feature set is discussed in the text.

Feature Name	Symbol	Meaning
Correlation	f_{cor}	Correlation between a drug treatment experiment and a gene knockdown experiment
Cell Selection	f_{CS}	Correlation between a drug treatment experiment and the control experiment for the cell line
PPI Correlation	f_{PC}	Fraction of the known binding partners of a gene in the top X correlated knockdown experiments
PPI Expression	f_{PE}	The average or the max (absolute value) expression for the known binding partners of a gene

Supplementary Table 6. The number of drugs profiled for different number of cell lines included in the validation dataset. While several drugs were profiled in at least four of these cell lines (152), only 29 were profiled in all seven cell lines.

# Cells	7	6	5	4	Total
# Drugs	29	30	42	51	152

Supplementary Table 7. Seven cell lines were included in the validation dataset. The number of drugs, knockdown genes, and control experiment are shown. For a given cell line, we only include drugs that have their target knockdown experiments available in that cell line.

Cell Line	Drugs	Knockdowns	Controls
A549	188	11947	52
MCF7	180	12031	54
VCAP	175	13225	56
HA1E	172	11968	53
A375	143	11696	58
HCC515	129	7828	52
HT19	96	10185	52

Supplementary Table 8. The cellular localization of successful and unsuccessful drug targets enriched by gene ontology.

	Cellular Component	p-value
Successful Targets	proteasome core complex	7.81E-37
	proteasome core	1.10E-28
	proteasome alpha-subunit	5.68E-18
	cytosol	7.53E-12
	protein complex	1.88E-11
Failed Targets	transmembrane transporter complex	7.77E-15
	sodium-exchanging ATPase complex	4.42E-14
	cation-transporting ATPase complex	8.74E-13
	plasma membrane part	2.19E-11
	chloride channel complex	2.33E-09

Supplementary Results

Gene ontology analysis of protein targets

While the success rate of our Random Forest genomic analysis is promising, there are still several drugs for which we fail to correctly identify the target. We attempted to determine if the genomic data we used is more appropriate to specific drug / protein characteristics. By characterizing the set of drugs and / or proteins for which we expect the method to be more accurate we improve the ability of experimentalists to use our methods when studying one of these molecules.

We divided the 152 drugs in our training data into "successful" predictions (the 63 drugs for which the correct target was ranked in the top 100), and "unsuccessful" predictions. We also divided the known targets into those that were correctly predicted and those that were not. We considered several different ways to characterize small molecules including molecular weight, solubility, and hydrophobicity, but none of these seemed to significantly correlate with our "successful" and "unsuccessful" classifications. Next, we used gene ontology (Online Methods) to test for enrichment of "successful" and "unsuccessful" targets. Interestingly, we found that "successful" targets were significantly associated with intracellular categories, while the "unsuccessful" targets were mostly associated with transmembrane and extracellular categories (Supplementary Table 8).

Based on this result we further incorporated cellular component as a feature in our two-level random forest. We encode this feature by assigning 1 to the intracellular genes and -1 to the extracellular ones (see Online Methods for detail). We ran the two-level random forest with this additional feature included and demonstrated that the cellular component increases the number of top 100 genes to 66 and top 50 genes to 55.

Supplementary Methods

Extracting experiments from LINCS

After determining the subsets of small molecules and cell lines, we obtained the associated experiment identifiers known as "distil IDs" from LINCS meta-information. We included only the reproducible distil IDs known as "Gold" IDs. We then extracted the corresponding signature values from LINCS using the L1000 Analysis Tools (l1ktools)[1]. We only extracted the signature values of the 978 "landmark" genes because their expression was directly measured, whereas the values of other genes were imputed from the data of these landmark genes.

Drug response experiments

There exist multiple experiments (distil IDs) corresponding to a combination of drug *d* and cell line *c* (applying drug *d* to cell line *c*). Denote the N_{dc} as the number of experiments for the combination *d,c*. We extracted a matrix of signature values of size $978 \times N_{dc}$ (number of landmark genes × number of experiments) per combination. We next took the median of signature values across different experiments, and obtained a 987×1 signature vector per combination. The overall drug-response data **Δ**, therefore, is implemented as a MATLAB structure with D = 152 entries, each containing the following fields.

name: $PertID_d$ (string)
cells: $Cells_{C_d}$ ($|C_d| \times 1$ string array)
signature: $\Delta_{d..}$ ($978 \times |C_d|$)

where $PertID_d$ is the unique internal identifier of a small molecule *d* in LINCS. $\Delta_{d..}$ contains the expression values of drug *d* across C_d different cell lines. The $Cells_{C_d}$ field contains cell line names corresponding to the column of $\Delta_{d..}$.

Gene knockdown (KD) experiments

We follow a similar protocol to extract the signature values of gene knockdown experiments. Denote N_{gc} as the number of experiments for the combination of gene *g* and cell line *c* (knocking down gene *g* in cell line *c*). Then, for each combination of *g* and *c* we extracted signature values of size $978 \times N_{gc}$. After taking the medians across different experiments, we obtain a 978×1 vector per combination. The overall gene knockdown data Γ has $C = 7$ entries and each entry contains the following fields:

name: $Cells_c$ (string)
genes: $Symbols_{G_c}$ ($|G_c| \times 1$ string array)
signature: $\Gamma_{c..}$ ($978 \times |G_c|$)

where $Cells_c$ is the name of the cell line indexed by *c*. $\Gamma_{c..}$ contains the signature values of the knockdown of genes in cell line *c*. The $Symbols_{G_c}$ field is a subset of

[1] https://github.com/cmap/l1ktools

gene symbols corresponding to the column identifiers of $\Gamma_{c\cdot\cdot}$ under the HGNC naming scheme.

Control experiments

We also extracted the signatures of control experiments. The signature values for each cell line were extracted and we obtained a 978 × 1 vector after taking the medians. We denote the overall control experiment data as Ψ. Ψ is of size 978 × *C* and implemented with the following format:

name: $Cells_c$ (string)
control: $\Psi_{\cdot c}$ (978 × 1)

where $\Psi_{\cdot c}$ is the signature column vector for a cell line c.

Extracting and integrating features from different data sources

Correlation feature

The correlation feature, denoted as f_{cor}, is constructed as follows:

- For each drug *d* in Δ ($\Delta_{d\cdot\cdot}$):

  - Denote T_d as the intersection of gene symbol indices for cells in C_d:

$$T_d = \bigcup_{c \in C_d} G_c$$

  - Obtain the knockdown signature values of T_d from Γ. Denote this data matrix as $\Gamma_{C_d \cdot T_d}$, which is of size $|C_d| \times 978 \times |T_d|$, where for each cell line in C_d there is a signature matrix of size $978 \times |T_d|$.

  - Compute the Pearson's correlation between $\Delta_{d\cdot\cdot}$ ($978 \times |C_d|$) and $\Gamma_{C_d \cdot T_d}$ ($|C_d| \times 978 \times |T_d|$). Specifically, for each cell line $c \in C_d$, we compute the correlation between $\Delta_{d\cdot c}$ and $\Gamma_{c \cdot T_d}$, and obtain a correlation vector of size $|T_d|$. This is the correlation between the responses of the cells to the drug treatment and their response to the gene KD. Each entry in this vector is the correlation of 978 landmark genes of the drug *d* in one cell line ($\Delta_{d\cdot c}$) and a knockdown of gene *g* in the same cell line ($\Gamma_{c\cdot g}$). In other words, if we collect these correlation vectors for all cell lines in C_d and denote the overall correlation feature as f_{cor}:

$$f_{cor}(d, g, c) = corr(\Delta_{d\cdot c}, \Gamma_{c\cdot g}) \quad \forall g \in T_d$$

The correlation feature for one drug *d*, $f_{cor}(d,\cdot,\cdot)$, has a dimension of $|T_d| \times |C_d|$.

Cell selection feature

The cell selection feature, denoted as f_{CS}, is computed as follows:

- For each drug *d* in Δ ($\Delta_{d\cdot\cdot}$):
    - For each cell line *c* in C_d:
        - Compute the correlation between $\Delta_{d\cdot c}$ and $\Psi_{\cdot c}$

$$f_{CS}(d,c) = corr(\Delta_{d\cdot c}, \Psi_{\cdot c})$$

$f_{CS}(d,\cdot)$ produces a $|C_d| \times 1$ vector, and each entry corresponds to the correlation between the drug-response and control experiments for one cell line in C_d. This feature is used to determine the relevance of the drug to the cell type being studied.

PPI correlation score:

The PPI correlation Score, denoted as f_{PC} is constructed as follows:

- For each drug *d* in Δ ($\Delta_{d\cdot\cdot}$):
    - Obtain T_d, as defined above.
    - For each cell line *c* in C_d:
        - Sort T_d in descending order using the correlation values $f_{cor}(d,\cdot,c)$
        - Denote the sorted gene symbol indices for cell line *c* as $\sigma_c(T_d)$
        - For each knockdown gene *g* in T_d:
            - Obtain the set of neighbor gene symbol indices from the PPI adjacency list, and denote it as N_g.
            - Compute f_{PC} as:

$$f_{PC}(d,g,c) = \frac{\left|N_g \cap \sigma_c(T_d)_{1:100}\right|}{\left|N_g \cap \sigma_c(T_d)\right| + 50}$$

$f_{PC}(d,g,c)$ has the same dimension as f_{cor} ($|T_d| \times |C_d|$). It reflects the fraction of gene *g*'s binding partners that are more correlated with drug *d* in the context of cell line *c*. We use 50 as the pseudo-count to penalize hub proteins, which have substantially more neighbors than others.

PPI expression score

We compute two types of PPI expression scores, denoted as $f_{PE_{max}}$ and $f_{PE_{avg}}$, as follows:

- For each drug *d* in Δ ($\Delta_{d..}$):

    - For each knockdown gene *g* in T_d:

        - Obtain N_g, as above (the list of neighbors, or interaction partners, of *g*)

        - For each cell line *c* in C_d:

            - Find the set of signature values for the neighbors of *g*, $\Delta_{d,N_g,c}$ (size $|N_g| \times 1$)

            - Compute the two PPI expression scores as:

$$f_{PE_{max}}(d,g,c) = \max\left(\Delta_{d,N_g,c}\right)$$

$$f_{PE_{avg}}(d,g,c) = \text{avg}\left(\Delta_{d,N_g,c}\right)$$

Feature data structure

We combined the features for all drugs in a MATLAB structure Ω. Ω has *D* entries, and each entry $\Omega^{(d)}$ has the following fields:

name:	$PertID_d$ (string)
targets:	P_d (protein targets for *d)*
cells:	$Cells_{C_d}$ ($\|C_d\| \times 1$ string array)
genes:	T_d (common genes across G_c)
correlation:	$f_{cor}(d,\cdot,\cdot)$ ($\|T_d\| \times \|C_d\|$)
PPI correlation:	$f_{PC}(d,\cdot,\cdot)$ ($\|T_d\| \times \|C_d\|$)
max PPI expression:	$f_{PE_{max}}(d,\cdot,\cdot)$ ($\|T_d\| \times \|C_d\|$)
avg PPI expression:	$f_{PE_{avg}}(d,\cdot,\cdot)$ ($\|T_d\| \times \|C_d\|$)
cell selection:	$f_{CS}(d,\cdot)$ ($\|C_d\| \times 1$)

There are a total of *D* = 152 drugs in Ω, and the number of drugs with different values of $|C_d|$ are summarized in Supplementary Table 6.

Subcellular Localization Assignment

We obtained the cellular localization of genes from the Gene Ontology Consortium. The GO database provides web services to query genes in terms of their associated biological processes, cellular components and molecular

functions in a species-independent manner[2]. We further assign the locations as either "intracellular" (inside of cell) or "extracellular" (outside of cell). The detailed assignments are shown in Supplementary Table 8.

Classification procedure

Criterion of successful classification

Due to the intrinsic noise from the data, we define a successful classification for a drug if any of its correct targets is enriched into the top *K* ranked genes, where *K* can be either 50 or 100.

Analysis of feature importance

The evaluation of single features was performed using the drugs that have been applied on all seven cell lines. There are 29 of these drugs from Ω. We sort (descendingly) the common genes T_d for a drug *d* and cell line *c* using an individual feature $f(d,\cdot,c)$, where f is either f_{cor} or f_{PC}. Denote $\sigma_d(g,c)$ as the ranking of a gene $g \in T_d$ in the context of cell line *c*. Then, we define the overall ranking of a gene, $\sigma_d(g)$, to be the best ranking across all seven cell lines: $\sigma_d(g) = \min(\sigma_d(g,c))$ for $c \in C_d$.

Constructing training dataset

Next, we wish to learn and evaluate classifiers that predict drug targets using all features from the feature dataset Ω. We first construct a training data set (design matrix *X* and its associated labels *y*) from the feature dataset Ω.

For each drug *d* in Ω, we select the rows corresponding to the targets in P_d from the other feature matrices and concatenate them into a row vector. The same cell selection vector is appended to every row of targets. These rows are assigned with a positive label 1. We then randomly sampled 100 non-target genes (denoted as ν_d) and construct the row vectors the same way as the target genes, and these rows are assigned with a negative label 0. In other words, the training matrix and label vector constructed from a drug *d* are of the following format:

$$X_d = \begin{bmatrix} f_{cor}(d,P_{d1},\cdot) & f_{PC}(d,P_{d1},\cdot) & f_{PE_{max}}(d,P_{d1},\cdot) & f_{PE_{avg}}(d,P_{d1},\cdot) & f_{CS}(d,\cdot) \\ f_{cor}(d,P_{d2},\cdot) & f_{PC}(d,P_{d2},\cdot) & f_{PE_{max}}(d,P_{d2},\cdot) & f_{PE_{avg}}(d,P_{d2},\cdot) & f_{CS}(d,\cdot) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ f_{cor}(d,P_{dm},\cdot) & f_{PC}(d,P_{dm},\cdot) & f_{PE_{max}}(d,P_{dm},\cdot) & f_{PE_{avg}}(d,P_{dm},\cdot) & f_{CS}(d,\cdot) \\ f_{cor}(d,\nu_{d1},\cdot) & f_{PC}(d,\nu_{d1},\cdot) & f_{PE_{max}}(d,\nu_{d1},\cdot) & f_{PE_{avg}}(d,\nu_{d1},\cdot) & f_{CS}(d,\cdot) \\ f_{cor}(d,\nu_{d2},\cdot) & f_{PC}(d,\nu_{d2},\cdot) & f_{PE_{max}}(d,\nu_{d2},\cdot) & f_{PE_{avg}}(d,\nu_{d2},\cdot) & f_{CS}(d,\cdot) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ f_{cor}(d,\nu_{d100},\cdot) & f_{PC}(d,\nu_{d100},\cdot) & f_{PE_{max}}(d,\nu_{d100},\cdot) & f_{PE_{avg}}(d,\nu_{d100},\cdot) & f_{CS}(d,\cdot) \end{bmatrix}; y_d = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

[2] http://geneontology.org/page/go-enrichment-analysis

where $m = |P_d|$, the total number of targets for drug *d*. Therefore, the training matrix X_d for drug d is of size $(m + 100) \times 5|C_d|$, and label vector y_d has length $(m + 100)$.

References

1 Zhang, M. *et al.* Chaperoned ubiquitylation--crystal structures of the CHIP U box E3 ubiquitin ligase and a CHIP-Ubc13-Uev1a complex. *Mol. Cell* **20**, 525-538, doi:10.1016/j.molcel.2005.09.023 (2005).

2 Komander, D. *et al.* Structural insights into the regulation of PDK1 by phosphoinositides and inositol phosphates. *EMBO J.* **23**, 3918-3928, doi:10.1038/sj.emboj.7600379 (2004).