bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Proteome-scale detection of drug-target interactions using correlations in transcriptomic perturbations

Nicolas A. Pabon[1,†], Yan Xia[2,†], Samuel K. 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 15260

[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

Corresponding Author:

Carlos Camacho

3077 Biomedical Science Tower 3

3501 Fifth Avenue

Pittsburgh, PA 15260

412-648-7776, Fax: 412-648-3163

E-mail: ccamacho@pitt.edu

http://structure.pitt.edu

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Keywords:

LINCS, polypharmacology, target prediction, drug discovery, gene expression, virtual screen, high-throughput

Availability:

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

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Summary

Systems biology seeks to understand how normal and disease protein networks respond when specific interactions are disrupted. A first step towards this goal is identifying the molecular target(s) of bioactive compounds. Here, we hypothesize that inhibitory drugs should produce network-level effects similar to silencing the inhibited gene and show that drug-protein interactions are encoded in mRNA expression profile correlations. We use machine learning to classify correlations between drug- and knockdown-induced expression signatures and enrich our predictions through a structure-based screen. Interactions manifest both as direct correlations between drug and target knockdowns, and as indirect correlations with up/downstream knockdowns. Cross-validation on 152 FDA-approved drugs and 3104 potential targets achieved top 10/100 prediction accuracies of 26/41%. We apply our method to 1680 bioactive compounds and experimentally validate five previously unknown interactions. Our pipeline can accelerate drug discovery by matching existing compounds to new therapeutic targets while informing on network and multi-target effects.

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Highlights

- Inhibitory drugs and gene knockdowns produce similar disruptions of cellular protein networks
- Drug targets can be identified from correlations in drug- and knockdown-induced mRNA expression
- Drug-target interactions manifest as both direct/indirect correlations with the target/pathway
- Five novel interactions are experimentally validated, including first-in-class CHIP inhibitors

eTOC Blurb

We delineate the role of small molecules in perturbing cellular protein networks by capturing direct correlations between drug- and target knockdown-induced mRNA expression profiles as well as indirect correlations with knockdowns of genes up/downstream of the actual target. Our findings could accelerate drug discovery by assessing the impact of promising bioactive chemistries in modulating gene expression for novel therapeutic targets.

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Introduction

Understanding how cellular pathways respond when specific interactions are disrupted is essential for developing therapies that might restore perturbed networks to their native states. Probing cells with bioactive small molecules can yield insight towards identifying and correcting aberrant pathways if the molecules' target(s) are known a priori, but this is often not the case (Drews, 2000; Gregori-Puigjane et al., 2012; Overington et al., 2006). The target identification problem has traditionally been approached from a target-centric paradigm, relying on high-throughput *in-vitro* screening of large libraries against a single protein (Swinney and Anthony, 2011). This approach has proven effective for kinases, GPCRs, and proteases, but yields for numerous other targets such as protein-protein interactions are poor (Bleicher et al., 2003; Pritchard et al., 2003). Moreover, these *in-vitro* biochemical screens often cannot provide any context regarding drug activity in the cell, multi-target effects, or toxicity (Mayr and Bojanic, 2009; Persidis, 1998).

As an alternative to target-centric high throughput screening, compound-centric computational approaches are now commonly applied to predict drug–target interactions by leveraging existing data. Many of these methods extrapolate from known chemistry, structural homology, and/or functionally related compounds and excel in target prediction when the query compound is chemically or functionally similar to known drugs (Gfeller et al., 2014; Keiser et al., 2009; Lo et al., 2015; Martinez-Jimenez

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

and Marti-Renom, 2015; Nickel et al., 2014; Yamanishi et al., 2010). Other structure-based methods such as molecular docking are able to evaluate novel chemistries, but are limited by the availability of protein structures (Li et al., 2006; Rognan, 2010; Wang et al., 2012), inadequate scoring functions and excessive computing times, which render structure-based methods ill-suited for genome-wide virtual screening (Meslamani et al., 2012). More recently, a new paradigm for predicting molecular interactions using cellular gene expression profiles has emerged (Faith et al., 2008; Lamb, 2007; Lamb et al., 2006). Studies have mapped drug-induced differential gene expression levels onto known protein interaction network topologies and prioritized potential targets by identifying highly perturbed subnetworks (Cosgrove et al., 2008; Isik et al., 2015; Laenen et al., 2013). These studies predict roughly 20% of known targets within the top 100 ranked genes (SI Appendix, SI Methods), but did not predict any new interactions.

The NIH's Library of Integrated Cellular Signatures (LINCS) project presents an opportunity to develop an unbiased approach to expression-based target prediction by integrating drug-induced expression signatures with signatures from other types of cellular perturbations. Specifically, the LINCS L1000 dataset contains cellular mRNA signatures from treatments with 20,000+ small molecules and 20,000+ gene over-expression (cDNA) or knockdown (sh-RNA) experiments. Based on our hypothesis that drugs which inhibit their target(s) should yield similar network-level effects to silencing the target gene(s) (Figure 1a), we calculated correlations between the expression signatures of thousands of small molecule treatments and gene knockdowns in the

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

same cells. We used the strength of these correlations to rank potential targets for a validation set of 29 FDA-approved drugs tested in the seven most abundant LINCS cell lines. We evaluate both direct signature correlations between drug treatments and knockdowns of their potential targets, as well as indirect signature correlations with knockdowns of proteins up/down-stream of potential targets. We combined these correlation features with additional gene annotation, protein interaction and cell-specific features using Random Forest (Andy Liaw, 2002; Qi et al., 2006) (RF) and achieved a top 100 target prediction accuracy of 55%, due primarily to our novel correlation features.

Finally, to filter out false positives and further enrich our predictions, we used molecular docking to evaluate the structural compatibility of the RF-predicted compound–target pairs. This orthogonal analysis significantly improved our prediction accuracy on an expanded validation set of 152 FDA-approved drugs, obtaining a top 10/100 accuracy of 26%/41%, more than double that of previous methods. Finally, we tested our method on 1680 small molecules profiled in LINCS and experimentally confirmed previously unknown and first-in-class targets for five different compounds. These compounds validated our hypothesis that drug- and target knockdown-induced gene expression perturbations correlate directly and/or indirectly via genes up/downstream of the actual target. More importantly, we provide enriched sets of likely active compounds to hundreds of human targets, providing a new avenue to identify suitable (multi-) targets

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

for novel chemistries and accelerate the discovery of chemical probes of protein function.

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Results

Preliminary prediction of drug targets using expression profile correlation features.

We constructed a validation set of 29 FDA-approved drugs that had been tested in at least seven LINCS cells lines, and whose known targets were among 2634 genes knocked down in the same cell lines. For these drugs, we ranked potential targets using the direct correlation between the drug-induced mRNA expression signature and the knockdown-induced signatures of potential targets (Figure 1b,c). For each cell line, the 2634 knockdown signatures were sorted by their Pearson correlation with the expression signature of the drug in that cell line. We used each gene's lowest rank across all cell lines to produce a final ranking of potential targets for the given drug. Using this approach, we predicted known targets in the top 100 potential targets for 8/29 validation compounds (Table S1). Indirect correlations were evaluated by the fraction of a potential target's known interaction partners (cf. BioGrid (Chatr-Aryamontri et al., 2015)) whose knockdown signatures correlated strongly with the drug-induced signature. Ranking by indirect correlations predicted the known target in the top 100 for 10 of our 29 validation compounds (Table S1). Interestingly, several of these compounds showed little correlation with the knockdown of their targets, with only 3/10 targets correctly predicted using the direct correlation feature alone (Figure 1d,e).

It is well known that expression profiles vary between cell types (Shen-Orr et al., 2010). Thus, we constructed a cell selection feature to determine the most "active" cell line,

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

defined as the cell line producing the lowest correlation between the drug-induced signature and the control signature. Ranking by direct correlations within the most active cell line for each drug predicted six known targets in the top 100 (Table S1). However, all six of these targets were already predicted by either direct or indirect correlations, strongly suggesting that scanning for the optimal correlation across all cell lines is a better strategy than trying to identify the most relevant cell type by apparent activity.

Finally, to incorporate the findings of previous studies that suggest that drug treatments often up/down regulate the expression of their target's interaction partners (Cosgrove et al., 2008; Isik et al., 2015; Laenen et al., 2013), we constructed two features to report directly on the drug-induced differential expression of potential targets' interaction partners. These features compute the maximum and the mean differential expression levels of potential targets' interaction partners in the drug-induced expression profile. The lowest rank of each potential target across all cell lines is used in a final ranking. Though neither expression feature produces top 100 accuracies better than those of our correlation features, maximum differential expression identifies three new targets that were not identified using any of the previous features (Table S1).

Combining individual features using random forest (RF). While each of the features in Table S1 performed better than random, combining them further improved results. Using Leave-One-Out Cross Validation (LOOCV) for each drug, logistic regression (Qi et

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

al., 2006) correctly identified known targets in the top 100 predictions for 11 out of 29 drugs and improved the average known target ranking of all drugs (Table S1). However, logistic regression assumes that features are independent, which is not the case for our dataset. Hence, we used RF, which is able to learn more sophisticated decision boundaries (Diaz-Uriarte and Alvarez de Andres, 2006). Following the same LOOCV procedure, the RF classifier led to much better results than the baseline logistic regression, correctly finding the target in the top 100 for 16 out of 29 drugs (55%) (Table S1). Without further training, we tested the RF approach on the remaining 123 FDA-approved drugs that had been profiled in 4, 5, and 6 different LINCS cell lines, and whose known targets were among 3104 genes knocked down in the same cells. We predicted known targets for 32 drugs (26%) in the top 100 (Table S2), an encouraging result given the relatively small size of the training set and the expected decline in accuracy as the number of cell lines decreases (see below, Table 1).

Re-training on the full set of 152 drugs and validating using LOOCV, we tested two alternative RF models: “on-the-fly”, which learns drug-specific classifiers trained on the set of drugs profiled in the same cell types, and “two-level”, which learns a single classifier trained on experiments from all training drugs (see SI Appendix, SI Methods). The performances of both methods as a function of the number of cell lines profiled are summarized in Table 1. On-the-fly RF correctly ranked the targets of 58 out of 152 drugs in the top 100 (38%), with 42 of them in top 50 (28%). Two-level RF produced better enrichment, correctly predicting targets for 63 drugs in the top 100 (41%), and for 54

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

drugs in the top 50 (36%). In sharp contrast, random rankings (based on 20000 permutations) leads to only 7% of drugs with targets in the 100, indicating that both our training/testing and LOOCV results are extremely significant (Figure S1). It is also noteworthy that the top-100 accuracy of the two-level RF analysis increases to 50% if we only consider drugs treated in 5 or more cell lines.

Structural enrichment of genomic predictions

After RF classification, we used structural data to further refine our predictions. For our 63 "hits", drugs in the validation set for which we correctly identified the known target in the top 100, we generated structural models of their potential targets by mining the Protein Data Bank (PDB) (Bernstein et al., 1978). We selected one or more representative crystal structures for each gene, optimizing for sequence coverage and structural resolution. We then docked hits to their top 100 potential targets and ranked using a prospectively validated pipeline (Baumgartner and Camacho, 2016; Koes et al., 2013; Koes et al., 2015; Ye et al., 2016).

On average, crystal structures were available for 69/100 potential targets for each compound, and structures of known targets were available for 53/63 hits. In order to avoid redocking into cocrystals of our hits, we made sure to exclude structures containing these 53 ligands. As shown in Figure 2, molecular docking scores improved the re-ranking of the known target for 40 of the 53 drugs, with a mean and median improvement of 13 and 9, respectively. Based on genomic data alone, the known target

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

was ranked in the top 10 for 40% of the 63 hits. After structural re-ranking, 65% had their known targets in the top 10 candidates, and this value improved to 75% in the subset of 53 drugs with known target structures. These results demonstrate the orthogonality of the genomic and structural screens, showing that molecular docking can efficiently screen false positives in our gene expression-based predictions.

Identifying new interactions in the LINCS dataset. The final output of our prediction pipeline (Figure 3) can be tailored to provide an enriched subset of roughly 10 predictions for experimental validation, whether these are potential targets (compound-centric) or potential inhibitors (target-centric). Compound-centric analyses proceed by performing molecular docking on the available structures of the input compound's top 100 RF-predicted targets. Target-centric analyses run the RF on all LINCS test compounds, identify compounds for which the input protein is ranked in the top 100 potential targets, and then dock these candidate inhibitors to the target. In both applications, we analyzed the final docking score distributions and applied a 50% cutoff threshold to identify highly enriched compound/target hits. Structural analysis further facilitates visual validation of the docking models of predicted hits, thereby minimizing false positives.

Compound-centric. We first demonstrated a compound-centric application of our pipeline by analyzing Wortmannin, a selective PI3K covalent inhibitor and commonly used cell biological tool. Drugbank (Wishart et al., 2006) lists four known human targets

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

of Wortmannin: PIK3CG, PLK1, PIK3R1, and PIK3CA. Of the 100 targets predicted for Wortmannin, the PDB contained structures for 75, which we used to re-rank these potential targets (Figure S2). Only one known kinase target of Wortmannin, PIK3CA, was detected, and ranked 5th. Our pipeline also ranked 4th the human kinase PDPK1 (PDK1). Although PDK1 is a downstream signaling partner of PI3Ks (Vanhaesebroeck and Alessi, 2000), there is no prior evidence of a direct Wortmannin-PDK1 interaction in the literature. Nevertheless, the drug induced expression signature of wortmannin showed strong *direct* correlation with the knockdown of PDK1 (Figure S3a), and the predicted binding affinity was comparatively high (Figure S2), both of which suggest a possible interaction.

We experimentally tested this interaction using an alphascreen PDK1 interaction-displacement assay. Since we predicted that Wortmannin binds to the PH domain of PDK1 (Figure S4), we measured the effect of increasing Wortmannin concentrations on the interaction of PDK1 with the second messenger PIP3. We found that Wortmannin specifically increased PDK1-PIP3 interaction, relative to control (Figure S5). Given that PIP3-mediated recruitment of PDK1 to the membrane is thought to play an important regulatory role in the activity of the enzyme (Gao and Harris, 2006; Masters et al., 2010), a disruptive increase in PDK1-PIP3 interaction following treatment with Wortmannin supports our prediction.

Target-centric. To test a target-centric application of our pipeline, we chose a protein

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

that, to our knowledge, lacks specific inhibitors. STUB1, also known as CHIP (the carboxy-terminus of Hsc70 interacting protein), is an E3 ubiquitin ligase that manages the turnover of over 60 cellular substrates (Paul and Ghosh, 2015). CHIP interacts with the Hsp70 and Hsp90 molecular chaperones via its TPR motif, which recruits protein substrates and catalyzes their ubiquitination. Thus, treatment with small molecules that inhibit CHIP may prove valuable for pathologies where substrates are prematurely destroyed by the ubiquitin-proteasome system (Meacham et al., 2001).

The screening of the 1680 LINCS small molecules profiled in at least four cell lines predicted 104 compounds with CHIP among the top 100 targets. We docked these molecules to our representative structure of the TPR domain of CHIP (PDB ID: 2C2L (Zhang et al., 2005)), for which we had an available fluorescence polarization (FP) assay. The RF and docking score distributions were then plotted for all 104 compounds to select those highly enriched in one or both scoring metrics (Figure S6). We next visually examined the docking models of top ranking/scoring hits to select those that show suitable mechanisms of action, and purchased six compounds for testing (Table S3). In parallel, we performed a pharmacophore-based virtual screen of the ZINC database (Irwin and Shoichet, 2005) using the *ZincPharmer (Koes et al., 2015)* server, followed by the same structural optimization (Baumgartner and Camacho, 2016; Koes et al., 2013; Koes et al., 2015; Ye et al., 2016) performed on the LINCS compounds. We purchased seven of the resulting ZINC compounds for parallel testing.

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Our FP assay measured competition with a natural peptide substrate for the CHIP TPR domain. We found that four (out of six) of our LINCS compounds reliably reduced substrate binding (Figure 4a,b), while three (out of seven) ZINC compounds did so to a modest degree (Figure S7). The two strongest binders were LINCS compounds 2.1 and 2.2. A functional assay also verified that 2.1 and 2.2 prevented substrate ubiquitination and autoubiquitination (Figure 4c,d, S8, S9). Importantly, the predicted binding modes of these two compounds did not match the pharmacophore model of the TPR-HSP90 interaction (Zhang et al., 2005) that was used to screen the ZINC database (Figure S10). The latter emphasizes the power of our approach to identify novel compounds and mechanisms of action to targets without known inhibitors.

Comparison to existing target prediction methods. We next compared results for our 63 hits from the validation set to those produced by available structure and ligand-based methods. HTDocking (HTD) (Wang, 2012) is a structure-based target prediction method that docks and scores the input compound against a manually curated set of 607 human protein structures. For comparison, in our analysis we were able to extract high quality domain structures for 1245 (40%) of the 3104 potential gene targets. PharmMapper (PHM) (Liu et al., 2010) is a ligand-based approach that screens the input compound against pharmacophore models generated from publicly available bound drug-target cocrystal structures of 459 human proteins, and then ranks potential targets by the degree to which the input compound matches the binding mode of the cocrystalized ligands. The scope of HTD is limited by the availability of the target

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

structure, while PHM is limited by chemical and structural similarity of active ligands.

HTD and PHM rankings for known targets are shown in Table S4, and complete results are shown in Table S5. Our combined genomics-structure method outperforms the structure-based HTD server (average ranking of the known target is 13 for our method vs. 50 for the HTD server). This suggests that limiting the structural screening to our genomic hits allowed us to predict targets with higher accuracy than docking alone. Results when using the PHM server are on average similar to ours. However, PHM relies on the availability of ligand-bound crystal structures, which in practice makes this class of methods more suitable for drug repurposing than assessing new chemistries or targets.

With regards to our new validated interactions, a Wortmannin-PDK1 interaction at the catalytic site was ranked 540th by HTD and 56th by PHM. Although we cannot rule out a possible kinase domain interaction, a catalytic activity assay showed that Wortmannin had no measureable effect on the in vitro phosphorylation of the substrate T308tide by the isolated catalytic domain of PDK1 (Figure S11). On the other hand, CHIP inhibitors were not predicted by either of these methods, indicating their limitation assessing new chemistries or targets.

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Discussion

Delineating the role of small molecules in perturbing cellular interaction networks in normal and disease states in an important step towards identifying new therapeutic targets and chemistries for drug development. To advance on this goal, we developed a novel target prediction method based on the hypothesis that drugs that inhibit a given protein should have similar network-level effects to silencing the inhibited gene and/or its up/downstream partners. Using gene expression profiles from knockdown and drug treatment experiments in multiple cell types from the LINCS L1000 database, we developed several correlation-based features and combined them in a random forest (RF) model to predict drug-target interactions.

On a validation set of 152 FDA-approved drugs we achieve top-100 target prediction accuracy more than double that of previous approaches that use differential expression alone (Isik et al., 2015; Laenen et al., 2013). Consistent with our underlying hypothesis, the RF results highlight the importance of both direct expression signature correlations between drug treatment and knockdown of the gene target (Figure 1c) and indirect correlations between the drug and the target's interacting partners (Figure 1e, Figure 5). Contrary to earlier work (Cosgrove et al., 2008; Isik et al., 2015; Laenen et al., 2013), our methods and predictions are available for immediate download and testing (http://sb.cs.cmu.edu/Target2/), including predicted targets for 1680 LINCS small molecules from among 3000+ different human proteins.

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Unlike most available ligand-based prediction methods (Gfeller et al., 2014; Keiser et al., 2009; Lo et al., 2015; Martinez-Jimenez and Marti-Renom, 2015; Nickel et al., 2014; Yamanishi et al., 2010), the accuracy of our approach does not rely on chemical similarity between compounds in the training/test sets (Figure S12). The experimental validation of our predictions for CHIP and Wortmannin demonstrate the power of our combined genomic and structural pipeline in identifying novel targets and chemotypes. For instance, our screen against CHIP, a target with no known small molecule inhibitors, delivered four out six binding compounds, whereas a parallel analysis using solely a state-of-the-art structure-based virtual screening (Koes and Camacho, 2012; Ye et al., 2016) yielded only two weak-binding compounds. Moreover, the predicted mode of actions of the LINCS compounds suggest novel chemotypes that would have been difficult to prioritize in a ligand-based screen (Figure S10).

Indirect mRNA expression profile correlations with knockdowns of upstream/downstream interacting partners are an important determinant of drug-target interactions. However, they are also an important source of false positives in our analysis since profiles of knockdowns in the same pathway can correlate. Figure 5 shows that this is the case for compound 2.1, which based on indirect correlations with interacting partners UbcH5 and HSP90 predicts CHIP as a target, but also predicts UbcH5 and HSP90 as possible targets based on their corresponding direct correlations. Similarly, Figure 1 shows indirect correlations of vinblastine to its known target TUBA1,

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

based on direct correlations with interacting partners such as RUVBL1. Our pipeline eliminates some of these false positives using an orthogonal structure-based docking scheme that although limited to targets with known structure allows us to achieve a top-10 prediction accuracy of 26% for the compounds in our validation set.

Detailed analyses of our predictions suggest several new features to improve enrichment. We established a clear correlation between the number of cell-types screened and the target prediction accuracy. We identified that a significant source of false positives are indirect correlations that while important to detect the true target, also tend to predict interacting partners as potential targets. Incorporating compound- or target-specific features are also likely to improve our results. For instance, we noticed that our prediction results were less accurate for membrane proteins. Hence, we tested a cellular localization feature into our RF model, increasing the number of top-100 hits in our validation set from 63 to 66 (Si Appendix, SI Results).

In sum, our method represents a novel application of gene expression data for small molecule–protein interaction prediction, with structural analysis further enriching hits to an unprecedented level in our proteome-scale screens. Given the success of our proof-of-concept experiments, we are hopeful that our open source method and predictions will now be useful to other labs around the world for identifying new drugs for key proteins involved in various diseases and for better understanding the impact of drug modulation of gene expression. Moreover, our approach represents a new

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

framework for extracting robust correlations from intrinsically noisy gene expression data that reflect the underlying connectivity of the cellular interactome.

Author contributions

Y.X. and Z.B.J. performed genomic analysis. N.A.P., Z.Y. and C.J.C. performed structural analysis. S.E., J.B., J.G. and V.A. performed experiments involving CHIP. A.H., E.S. and R.B. performed experiments involving PDK1. N.A.P., Y.Z., Z.B.J. and C.J.C. prepared the manuscript.

Acknowledgements

This work was supported in part by the U.S. National Institute of Health (grants 1U54HL127624 to Z.B.J. and R01GM097082, to C.J.C.) and by the U.S. National Science Foundation (grants DBI-1356505 to Z.B.J. and 1247842, to N.A.P.). We thank the Connectivity Map team at the Broad Insititute for generation of the LINCS data set and query tools. The authors declare no competing financial interests.

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Materials and Methods

Data sources

A full description of the data used in our analysis can be found in the SI Appendix, SI Methods. Briefly, from the NIH LINCS library we extracted gene expression perturbations on 978 "landmark genes" from thousands of small molecule treatment and gene knockdown experiments in various cell lines. We then used ChEMBL (Gaulton et al., 2012), an open large-scale bioactivity database, to identify the LINCS compounds that were FDA approved and had known targets. To construct our validation set we selected the 152 FDA approved compounds that had been tested in at least four distinct LINCS cell lines, and whose known targets were knocked down in the same cell lines. Protein-protein interaction data used in feature construction was extracted from BioGRID (Chatr-Aryamontri et al., 2015) and HPRD (Keshava Prasad et al., 2009), both of which contain curated sets of physical and genetic interactions. Protein cellular localization data used in feature construction was obtained from the Gene Ontology database (Harris et al., 2004).

Extracting and integrating features from different data sources

The notation and symbols that we use in constructing and using the genomic features are described in Tables S6 and S7. Feature construction is summarized below and is explained in detail in the SI Appendix, SI Methods.

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Direct correlation: The first feature f_{cor} ,computes the correlation between the expression profiles resulting from a gene knockdown and treatment with the small molecule. Since we are considering multiple cells for each molecule/knockdown, the correlation feature for each molecule *d*, i.e. $f_{cor}(d,\cdot,\cdot)$, has a dimension of $|T_d|\times|C_d|$.

Indirect correlation: Information about protein interaction networks may be informative about additional knockdown experiments that we might expect to be correlated with the small molecule treatment profile. To construct a feature that can utilize this idea we did the following: for each molecule, protein, and cell line we computed $f_{PC}(d,g,c)$, which encodes the fraction of the known binding partners of *g* (i.e. the proteins interacting with *g*) in the top *X* knockdown experiments correlated with this molecule/cell compared to what is expected based on the degree of that protein (the number of interaction partners - this corrects for hub proteins). We used *X* = 100 here, though 50 and 200 gave similar results. See SI Appendix, SI Methods for complete details.

Cell selection: While the correlation feature is computed for all cells, it is likely that most drugs are only active in certain cell types and not others. Since the ability to consider the cellular context is one of the major advantages of our method we added a feature to denote the impact a drug has on a cell line. For each drug/molecule *d* we compute a cell specific feature, $f_{CS}(d,\cdot)$, which measures the correlation between the response expression profile and the control (WT) experiments for that cell. We expect a smaller

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

correlation if the drug/molecule is active in this cell, and a larger correlation if it is not.

Differential expression: In addition to determining the correlation-based rankings of interacting proteins, we also took their drug-induced differential expression into account. We constructed two features that summarize this information for each protein (see SI Appendix, SI Methods for details). These features either encode the average or the max (absolute value) expression level of the interaction partners of the potential target protein.

Generating structural models for docking

In order to use molecular docking to enrich of our random forest predictions, we needed to generate structural models for the genes profiled in LINCS. The union of our top 100 target predictions for the 1680 small molecules profiled in LINCS in at least four cell lines consisted of 3333 unique human genes. We used a python script (available on github[2]) to mine the PDB for structures of these genes and then select representative crystal structures for each. When multiple structures were available, a representative subset of structures were chosen so as to maximize sequence coverage, minimize strucutral resolution, and account for structural heterogeneity. Full details of this procedure can be found in the SI Appendix, SI Methods.

[2] https://github.com/npabon/generate_gene_models

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Docking procedure

Compounds were docked to representative structures of their predicted targets with smina (Koes et al., 2013), using default exhaustiveness and a 6 Å buffer to define the box around each potential binding site. Docked poses across predicted binding sites (Kozakov et al., 2015) on a given target were compared and the highest scoring pose of each compound was selected for further analyses (Baumgartner and Camacho, 2016; Koes et al., 2013; Koes et al., 2015; Ye et al., 2016) and comparison to other targets/compounds.

Experimental assays

Full details on all experimental assays involving CHIP and PDK1 can be found in the SI Appendix, SI Methods.

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figures

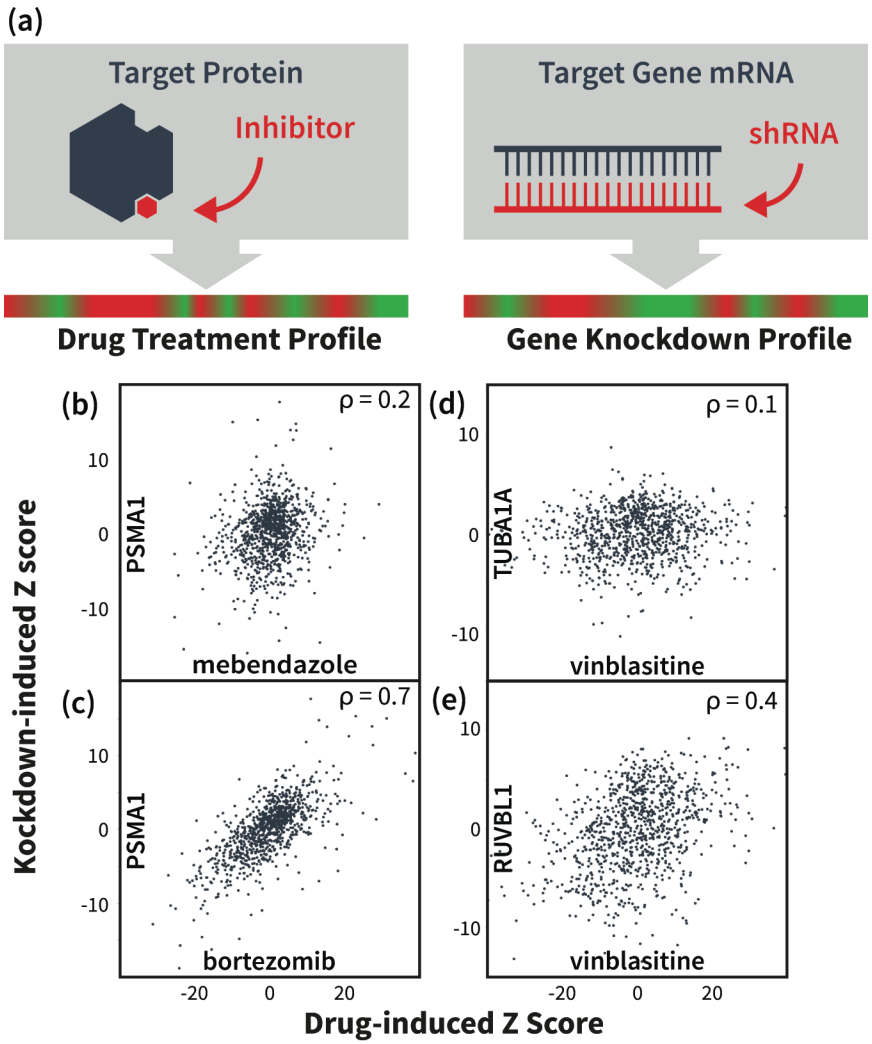


Figure 1. Drug- and gene knockdown-induced mRNA expression profile correlations reveal drug-target interactions. (a) Illustration of our main hypothesis: we expect a drug-induced mRNA signature to correlate with the knockdown signature of the drug's target gene and/or genes on the same pathway(s). (b,c) mRNA signature from knockdown of proteasome gene PSMA1 does not significantly correlate with signature

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

induced by tubulin-binding drug mebendazole, but shows strong correlation with signature from proteasome inhibitor bortezomib. Data points represent differential expression levels (Z-scores) the 978 landmark genes measured in the LINCS L1000 experiments. (d,e) Signature from tubulin-binding drug vinblastine shows little signature correlation with knockdown of its target TUBA1A, but instead correlates with the knockdown of functionally related gene RUVBL1.

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

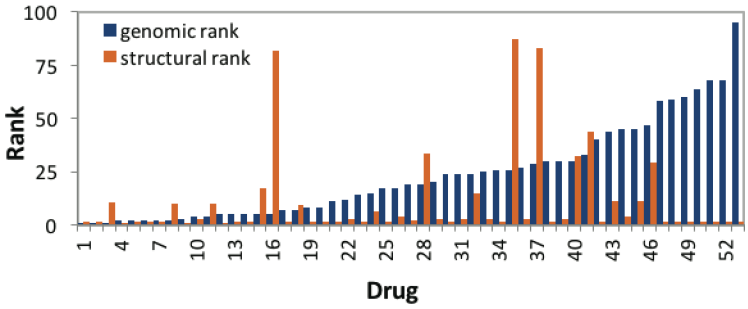


Figure 2. Structural enrichment of genomic target predictions. Predicted ranking (lower is better) of the highest-ranking known target for the 53 hits in our validation set with known target structures. Percentile rankings are shown following RF analysis (blue), and following structural re-ranking (orange). Drug names/IDs are listed in Table S8.

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

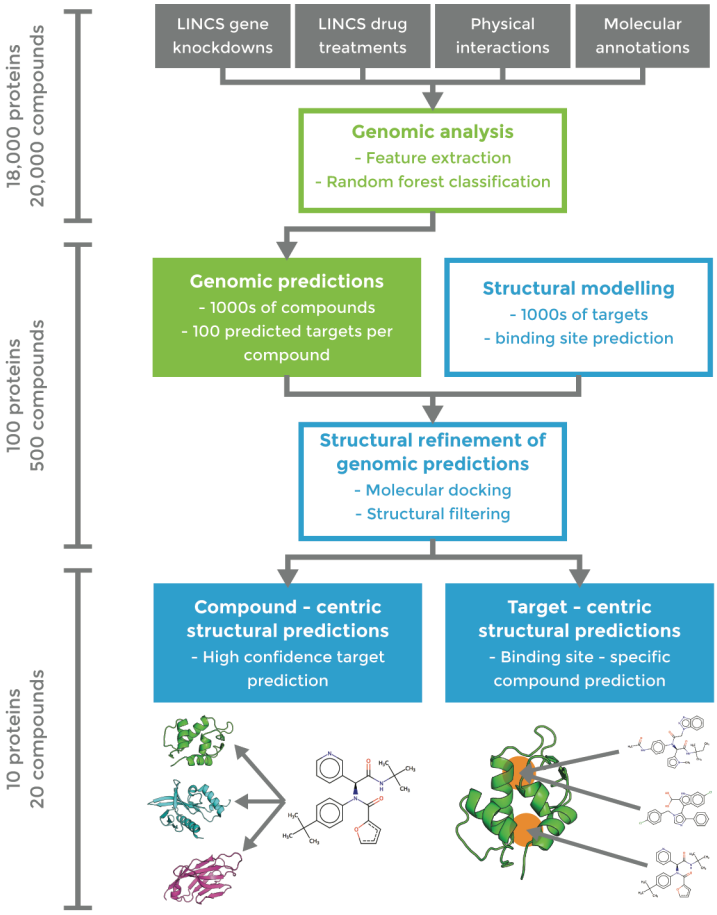


Figure 3. Flowchart of combined genomic (green) and structural (blue) pipeline for drug-target interaction prediction. The approximate number of proteins/compounds in each phase is indicated on the left in grey.

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

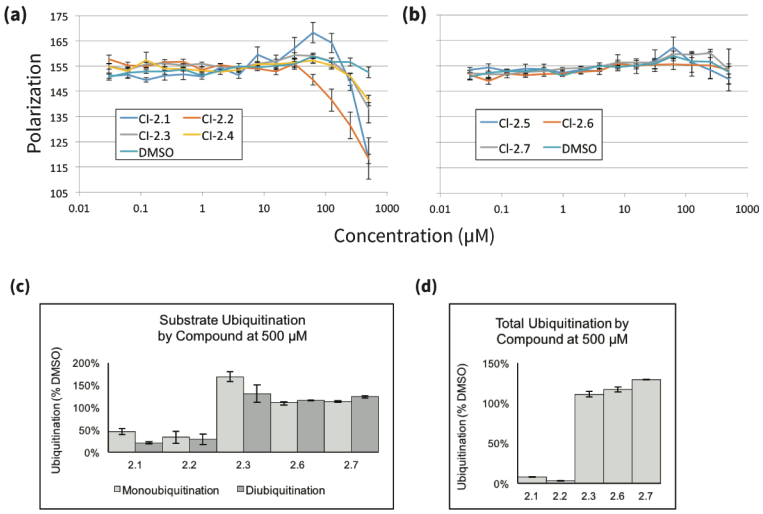


Figure 4. (a,b) Predicted CHIP inhibitors disrupt binding to chaperone peptide by fluorescence polarization. High ranked (a) and low ranked (b) compounds were tested for the ability to compete with a known TPR ligand (5-FAM-GSGPTIEEVD, 0.1 μM) for binding to CHIP (0.5 μM). Results are the average and standard error of the mean of two experiments each performed in triplicate. **(c,d) CHIP inhibitors prevent ubiquitination by CHIP in vitro. (**c) Quantification of substrate ubiquitination by CHIP from Anti-GST western blot experiments with tested compounds at 500μM, blotted as in Figure S11a and normalized to DMSO treated control (2.1, 2.2: N=4; all other compounds: N=2). **(**d) Quantification of total ubiquitination by CHIP from Anti-GST western blot experiments with tested compounds at 500μM, blotted as in Figure S11b and normalized to ubiquitination by a DMSO treated control (all compounds: N=2).

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

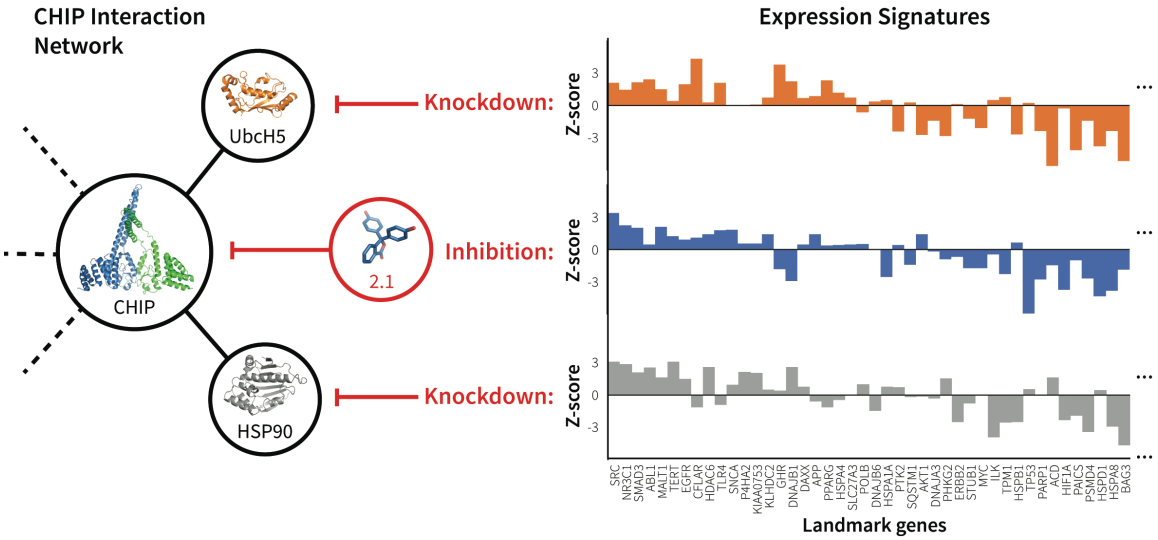


Figure 5. mRNA expression signature of CHIP inhibitor 2.1 correlates with knockdown of CHIP interacting partners. The figure illustrates the correlation between the mRNA expression profile signatures produced by treating cells with 2.1 and by knocking down CHIP interaction partners UbcH5 and HSP90. These three perturbations have similar network effects (left), as illustrated by their resulting differential expression signatures (right). For clarity, expression signatures show only the subset of LINCS landmark genes that are functionally related to CHIP according to BioGRID (Chatr-Aryamontri et al., 2015).

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Tables

# of Cells	All	7	6	5	4
# of Drugs	152	29	30	42	51
On-the-fly					
Top 100	58	13	15	16	14
Top 50	42	10	10	12	10
Top 100%	38%	45%	50%	38%	27%
Top 50%	28%	34%	33%	29%	20%
Two-level					
Top 100	63	14	15	22	13
Top 50	54	12	14	20	8
Top 100%	41%	48%	50%	52%	25%
Top 50%	36%	41%	47%	48%	16%

Table 1. Performance of two random forest models on validation set of 152 FDA-approved drugs. The number of drugs with targets ranked in top 100/50 are shown for the "on-the-fly" and "two-level" RF classification models. Results are divided into subsets of drugs profiled in different numbers of cell lines. Note that the success rate for RF is significant with $p < 10^{-6}$ based on randomization tests (Figure S1).

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

SI Appendix: Proteome-scale detection of drug-target interactions using correlations in transcriptomic 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

Corresponding Author:

Carlos Camacho

3077 Biomedical Science Tower 3

3501 Fifth Avenue

Pittsburgh, PA 15260

412-648-7776, Fax: 412-648-3163

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

E-mail: ccamacho@pitt.edu

http://structure.pitt.edu

Availability:

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

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

1. SI 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 (SI Appendix, SI 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 (Table S9).

Based on this result we further incorporated cellular component as a feature in our two-

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

level random forest. We encode this feature by assigning 1 to the intracellular genes and -1 to the extracellular ones. 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.

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

2. SI Methods

Data sources

LINCS: LINCS is an NIH program that generates and curates gene expression profiles across multiple cell lines and perturbation types at a massive scale. To date, LINCS has generated millions of gene expression profiles (over 150 gigabytes of data) containing small-molecules and genetic gain- (cDNA) and loss-of-function (sh-RNA) constructs across multiple cell types. Specifically, the LINCS dataset contains experiments profiling the effects of 20,143 small-molecule compounds (including known drugs) and 22,119 genetic constructs for over-expressing or knocking-down genes performed in 18 different cell types selected from diverse lineages which span established cancer cell lines, immortalized (but not transformed) primary cells, and both cycling and quiescent cells.

The gene expression profiles were measured using a bead-based assay termed the L1000 assay[2]. To increase throughput and save costs, this assay only profiles a set of 978 so-called "landmark genes" and the expression values of other genes can be computationally imputed from this set. Note however, that in our analysis we do not rely on such imputation and our methods only need to use the values for the measured genes. In our analysis we used level-4 signature values (containing z-scores for each gene in each experiment based on repeats relative to population control). Data

[2] http://support.lincscloud.org/hc/en-us/sections/200437157-L1000-Assay

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

processing of LINCS was done using the l1ktool.[3]

ChEMBL: To obtain a list of known targets for the drugs in our validation set we used ChEMBL, an open large-scale bioactivity database (Gaulton et al., 2012). We retrieved the records of all FDA-approved drugs using the ChEMBL web service API[4]. These records contain the designed targets for the drugs along with their synonyms (alternate names) and unique chemical IDs. We used this information to cross-reference these drugs with those in LINCS.

Protein-protein interaction and gene ontology: We obtained PPI information for our feature sets from BioGRID (Chatr-Aryamontri et al., 2015) and HPRD (Keshava Prasad et al., 2009), both of which contain curated sets of physical and genetic interactions. We retrieved all the records corresponding to protein-protein interactions (PPI) from these data sources and converted them to an adjacency list representation. We obtained the cellular localization of proteins from the Gene Ontology database (Harris et al., 2004). We relied on prior analysis (Navlakha et al., 2014) to assign the location of for each protein as either "intracellular" (inside of cell) or "extracellular" (outside of cell). See Supplementary Methods for details on how various compartments are assigned.

Extracting experiments from LINCS

[3] http://code.lincscloud.org/
[4] https://www.ebi.ac.uk/chembl/

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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)[5]. 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: $\Delta_{d..}$ ($978 \times |C_d|$)

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

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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

Control experiments

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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: $\Psi_{.c}$ (978 × 1)

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

Building a validation dataset from LINCS

We used ChEMBL to retrieve the reported targets and other meta-information of all FDA-approved drugs, and then cross referenced these drugs with the small molecules profiled in LINCS using their primary product names, synonyms, canonical SMILES strings and standard InChIKey. Based on this analysis we identified 1031 out of approximately 1300 FDA-approved drugs reported in LINCS. However, most of these drugs were profiled in only one or very few cell lines, which meant that relatively little response data was available for them. We thus further reduced this set to 152 drugs profiled in at least 4 cell lines (Table 1) and used these drugs and their known targets as the positive training set. Table S10 lists the number of drugs and knockdown experiments available for the seven most abundant cell lines in terms of known targets profiled that we used in our analysis.

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 ($\Delta_{d..}$):

  - 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 $\Delta_{d..}$ ($978 \times |C_d|$) 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

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

response to the gene knockdown. 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 $\Psi_{\cdot c}$

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

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

$f_{CS}(d,\cdot)$ 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.

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

- 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:

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

- 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: $f_{CS}(d,\cdot)$ ($|C_d| \times 1$)

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

There are a total of D = 152 drugs in Ω, and the number of drugs with different values of $|C_d|$ are summarized in Table 1.

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[6]. We further assign the locations as either "intracellular" (inside of cell) or "extracellular" (outside of cell). The detailed assignments are shown in Table S9.

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

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

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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:

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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

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 $(m + 100)$.

Extending Random forests to Drugs with Missing Features

Since our goal here is to predict targets for as many small molecule as possible, we did not want to restrict our analysis to molecules that were only profiled in a large number of cell lines. As noted above, requiring at least seven cell lines reduces the number of known drugs that can be evaluated from 152 to 29 and leads to a similar reduction in the number of novel small molecules that can be evaluated. Thus, it is highly desirable that our classifiers can handle missing data (i.e., cells for which experiments were not performed). To this end, we developed two distinct methods to deal with different compound-specific cell line combinations and extended the random forest (Andy Liaw, 2002; Qi et al., 2006) model so that can handle molecules profiled in less than seven (but more than four) cell types.

In the first method we simply build the random forest "on-the-fly". For a given drug *i*,

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

we iterate through all other drugs in Ω and test if a drug *d* was profiled in at least all cells which drug *i* was profiled in. In other words, we test if $C_i \subseteq C_d$ and if so we extract the features of corresponding cell lines in C_i from $\Omega^{(d)}$ and include them in the training data. After we include data for all compatible drugs we can use the training data to train and apply a random forest for the given drug *i*. We note that for any drug in Ω, there are at least 28 compatible drugs because 29 drugs have been applied to all seven cell lines. However, the main disadvantage of this method is that we need to train separate random forest for every test drug.

In the second method we perform a "two-level" random forest construction process. Here, in addition to the standard step of selecting a (random) subset of the features for each of the trees in the forest we included a step that selected a (random) subset of cells for each of the trees. Specifically, in the first step, we randomly sample four cell lines from the seven total cell lines (denoted as C_i). In the second step, we find all drugs $d \in \Omega$ such that $C_i \subseteq C_d$, extract their features, and use them to train that tree. We repeat this process 3500 times, such that each combination of four cell lines is expected to have roughly 100 trees ($\binom{7}{4} = 35$). To apply this two-level random forest to a test drug *t* with cell line profile C_t, we select from the forest those decision trees *i* for which $C_i \subseteq C_t$ and use them to predict the targets for *t*. Note that unlike the on-the-fly method above, here we only need to train one forest for the entire prediction task.

Generating structural models for docking

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Using a python script[7] We queried the PDB via its RESTful Web Service interface[8] using Uniprot primary gene name as the search criteria and found crystal structures for 1245 of the 3333 human genes in our analysis. The mean and median numbers of structures per gene were 11 and 3, respectively. We then analyzed the structures for each gene and selected representative structures that would be used for docking. Representative structure selection was performed automatically using a procedure (explained below) that attempts to optimize for sequence coverage, structural resolution, and structural diversity.

To select representative structures, we first divided each gene's structures into "high" and "low" resolution categories using a 2.0 Å threshold. Small structures with less than 20 amino acids were discarded. We then used a greedy algorithm to assess sequence coverage for the remaining structures and select (as representative) the fewest and highest resolution structures that would cover the most of the protein sequence. Redundant structures, defined as structures that did not contain at least 10 residues that were not contained in any of the larger or higher resolution structures, were discarded unless they represented a unique conformation of the protein. Protein conformation was evaluated using ProDy (Bakan et al., 2011) and was considered "unique" if the redundant structure had an all atom RMSD to each of the other representative structures that was above a cutoff threshold that could range between 4.0 Å and 10.0 Å. The specific value of the threshold used for each gene was chosen to

[7] https://github.com/npabon/generate_gene_models
[8] https://www.rcsb.org/pdb/software/rest.do

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

try to minimize the number of redundant structures that would be docked against, and higher cutoffs were used for genes that had many redundant structures representing different conformations. After selection, the mean and median numbers of representative structures per gene were 2 and 1, respectively. Each representative structure consisted of exactly one amino acid chain and coordinated ions but without cocrystal ligands or crystallographic waters. We note that this automated procedure is not necessarily tailored to produce representative structures for functional oligomers, since only one chain is considered at a time.

Comparison to previous expression perturbation target prediction methods

Unlike our method which uses both drug-induced and knockdown-induced mRNA expression perturbations, previous target prediction methods analyzed only the drug data within the context of protein interaction networks (Isik et al., 2015; Laenen et al., 2013). As their primary measurement of prediction accuracy, these works generally report the aggregate Area Under the Curve (AUC) of their gene rankings across all validation compounds. The studies mentioned above achieve AUC values of 0.9 and higher in ranking between 11,000 and 18,000 potential gene targets for each compound. To compare these results against our method, we examined the reported AUC curves and calculated the percentage of compounds for which the correct target was ranked within the top 100 potential targets. Both studies achieved top-100 accuracy of 20-21%.

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Experimental assays involving CHIP

Materials: Rabbit anti-GST polyclonal antibody conjugated to HRP was purchased from Abcam (ab3416), mouse anti-ubiquitin monoclonal antibody was purchased from Santa Cruz Biotechnology (sc-8017), and horse anti-mouse polyclonal antibody conjugated to HRP was purchased from Cell Signaling Technology (7076S). E2 enzyme UbcH5b and recombinant human ubiquitin were obtained from Boston Biochem (E2-662 and U-100H, respectively).

Protein purifications: His-Ube1, His-CHIP, GST-Hsc70$_{395-646}$, and GST-AT-3 JD were expressed in and purified from *E. coli* BL21(DE3) competent cells (New England Biolabs). Ube1/PET21d was a gift from Dr. Cynthia Wolberger (Addgene plasmid #34965) (Berndsen and Wolberger, 2011), pET151/D-TOPO CHIP and pGST‖2 Hsc70$_{395-646}$ were gifts from Dr. Saurav Misra (Sheffield et al., 1999; Zhang et al., 2015), and pGEX6p1 AT-3 JD was a gift from Dr. Matthew Scaglione (Faggiano et al., 2013; Todi et al., 2010). Transformed cultures were incubated in Luria broth with 100 µg/mL ampicillin at 37°C and shaken at 225 rpm until an OD_{600} of 0.3 was attained. Protein expression was then induced with 500µM isopropyl β-D-1-thiogalactopyranoside (IPTG) and cultures were incubated for 24 hrs. at 18°C (15°C for cells expressing GST-Hsc70$_{395-646}$ or GST-AT-3 JD) before the cells were harvested at 5000 rpm for 10 min at 4°C using an F7S-4x1000y rotor for the Sorvall RC-5B Plus Superspeed centrifuge. Cell pellets were stored at -80°C.

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Cells harboring His-Ube1 or His-CHIP were thawed and lysed by incubation in lysis buffer (10 mM imidazole, 50 mM $NaPO_4$ pH 8, 300 mM NaCl, 5 mM 2-mercaptoethanol, 0.25% Triton-100X, 2 mg/mL lysozyme) for 30 min on ice followed by sonication. Purification of Ube1 required addition of protease inhibitors (1% PMSF, 0.2% leupeptin, 0.1% pepstatin A) during lysis and throughout purification. After centrifugation, lysates were applied to Ni-NTA agarose resin (Qiagen), the column was washed with 30 mM imidazole, and proteins were eluted with 200 mM imidazole. Peak fractions containing His-Ube1 were pooled, dialyzed into 20 mM HEPES pH 7.4, 20 mM NaCl, and further purified by anion exchange chromatography over DEAE-Sepharose (GE Healthcare). Bound protein was eluted with a 50-300 mM NaCl gradient. Purified His-Ube1 and His-CHIP were dialyzed into 50 mM HEPES pH 7, 50 mM NaCl, and His-CHIP was further concentrated by centrifugal filtration (Millipore).

Cells harboring GST-$Hsc70_{395-646}$ or GST-AT-3 JD were similarly thawed and lysed by incubation in lysis buffer (50 mM Tris pH 7.5, 150 mM NaCl, 5 mM 2-mercaptoethanol, 0.25% Triton-100X, 2 mg/mL lysozyme, with protease inhibitors) followed by sonication. After centrifugation, lysates were applied to glutathione agarose (Sigma), the column was washed, and proteins were eluted in 6.8 mg/mL reduced glutathione. Peak fractions for each substrate were pooled and dialyzed into 50 mM HEPES pH 7, 50 mM NaCl.

After isolation, the purity of all proteins was verified by SDS-PAGE followed by Coomassie Brilliant Blue staining. Protein concentration was determined by either

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Bradford (Bio-Rad) or BCA (Thermo Scientific) protein concentration assays. Purified proteins were flash frozen in liquid nitrogen and stored at -80°C.

Fluorescence polarization assay: Fluorescence polarization (FP) studies were carried out as previously described (Assimon et al., 2015). Briefly, the FP tracer was composed of a peptide derived from Hsp72/HSPA1A (GSGPTIEEVD) that was coupled at the N-terminus to 5-carboxyfluorescein (5-FAM) via an aminohexanoic acid spacer. This tracer (K_D ~ 0.51 ± 0.03 µM) was used in a competition FP format to estimate binding to CHIP. Tracer concentration was 1 µM, and the CHIP concentration was 0.5 µM in a total volume of 20 µL in 50 mM HEPES, 10 mM NaCl, 0.01% Triton X-100, pH 7.4. The final DMSO concentration was approximately 1%. After mixing the components, each black 384 well plate (Corning) was covered from light and incubated at room temperature for 30 min. Polarization values were measured at Excitation 485 nm and Emission 530 nm using a Molecular Devices Spectramax M5 plate reader (Sunnyvale, CA). Data were analyzed using GraphPad Prism 6 software.

CHIP in vitro ubiquitination assay: Reactions were initiated by pre-incubating 125 nM Ube1, 1 µM UbcH5b, and 200 µM ubiquitin for 30 min at 37°C in 50 mM HEPES pH 7.0, 50 mM NaCl, 2 mM ATP, and 4 mM $MgCl_2$. In a separate reaction tube, 10 µM purified CHIP and up to 500 µM compound dissolved in DMSO were combined and incubated for 15 min on ice, followed by the addition of 3 µM of either GST-$Hsc70_{395-646}$ or GST-AT-3 JD, which served as substrates for CHIP-dependent ubiquitination. DMSO in these

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

reactions was <5%. After pre-incubation, the ubiquitin-charged E1/E2 mixture was dispensed after which all reactions proceeded for 15 min at 37°C. Reactions were quenched by addition of SDS sample buffer supplemented with 50 mM EDTA, 20 mM DTT. Quenched reactions were resolved by 10% SDS-PAGE, transferred to nitrocellulose membranes and western blotted with either anti-GST HRP-conjugated antibody to visualize substrate ubiquitination, or anti-ubiquitin primary antibody, followed by an HRP-conjugated secondary antibody to visualize the amount of total ubiquitination. Products were visualized using a Bio-Rad ChemiDoc XRS+ imaging system and quantified using ImageJ software.

Experimental assays involving PDK1

Materials: Soluble biotin-phosphatidylinositol3,4,5-triphosphate, biotin-PIP3, labeled with biotin at sn1-position, was from Echelon Biosciences Inc. Bio-GST, used as a control in the alphascreen system, corresponds to biotinylated GST, (Perkin-Elmer). The peptide substrate T308tide (KTFCGTPEYLAPEVRR; > 75% purity) were synthesized using Pepscan.

PDK1 constructs: PDK1 CD (1-359) and PDK1 PH (360-556) were cloned in pEBG2T vector in frame with GST, expressed in HEK293 by transient transfection and purified using glutathione-sepharose, as described previously for different GST-fusion constructs (Dettori et al., 2009).

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Alphascreen interaction assay: The interaction between GST-PDK1 PH (10 nM) and biotin-PIP3 (20 nM) was measured using alphascreen technology (Perkin-Elmer), a bead-based proximity assay. The displacement of the interaction by Wortmannin was performed as previously described for the catalytic domain of PDK1 (Schulze et al., 2016; Zhang et al., 2014). Briefly, the assays were performed in a final volume of 25 µL in white 384-well microtiter plates (Greiner Bio-One), including the interacting partners in a buffer containing 50 mM Tris-HCl pH 7.4, 100 mM NaCl, 2 mM DTT, 0.01% (v/v) Tween-20, 0.1% (w/v) BSA, and the corresponding concentration of the compound (1% final DMSO concentration). 5 µL of beads (anti-GST conjugated acceptor beads and streptavidin-coated donor beads) at a 20 µg/ml (microg/ml) were then added to the mixture and after an incubation of 60 minutes, alphascreen counts were measured in an EnVision Multiplate reader. To set-up the assays, cross-titration experiments were performed, where the concentration of both interacting partners were varied. The concentration of binding partners in the assays were chosen so that both inhibitors and enhancers of the interaction could be identified. Controls using Bio-GST were performed to rule out unspecific effects on the biotin-GST alphascreen interaction assay system.

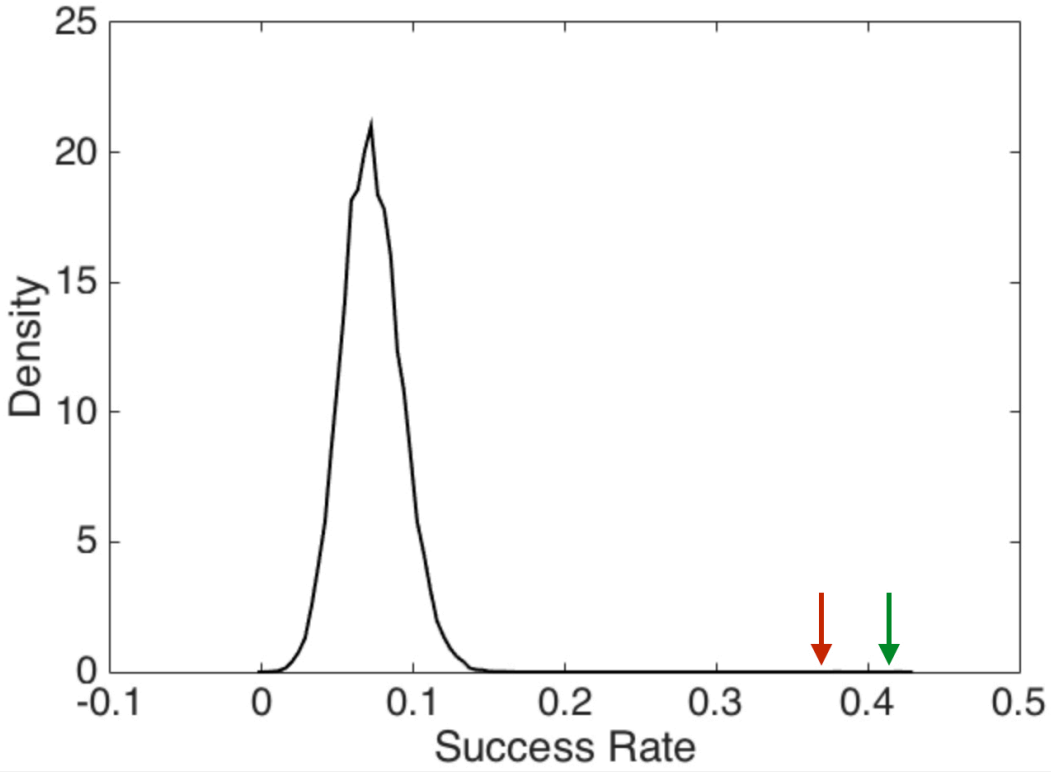
PDK1 protein kinase activity assay: The in vitro activity of PDK1 was tested using 100-300 ng purified protein, following the transfer of ^{32}P from radiolabelled [g^{32}P]ATP to the polypeptide substrate T308tide at room temperature (22 °C) in a mix containing 50 mM Tris pH 7.5, 0.05 mg/ml BSA, 0.1% mercaptoethanol, 10 mM MgCl2, 100 µM [g^{32}P]ATP (5-50 cpm/pmol) and 0.003% Brij, as previously performed. (Schulze et al., 2016)

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

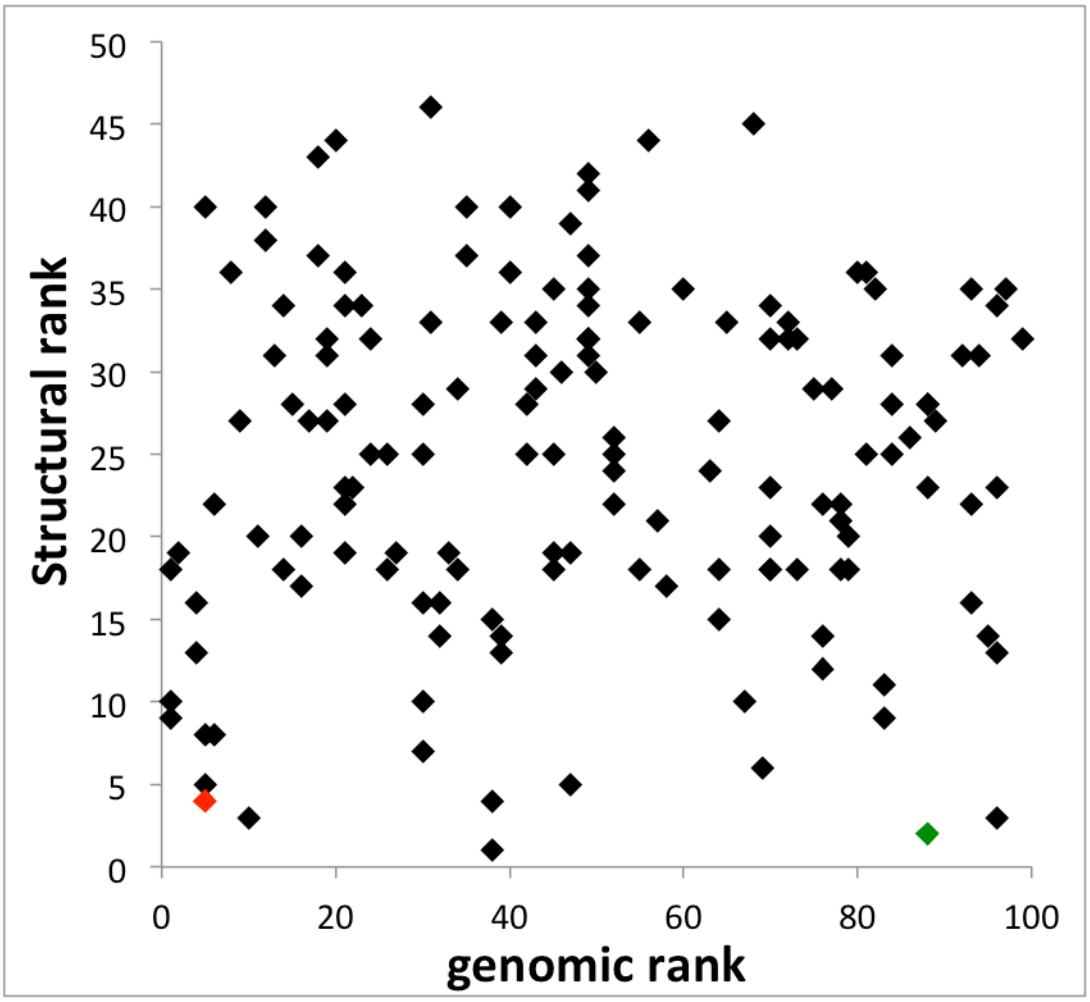
3. SI Figures

Figure S1. **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.



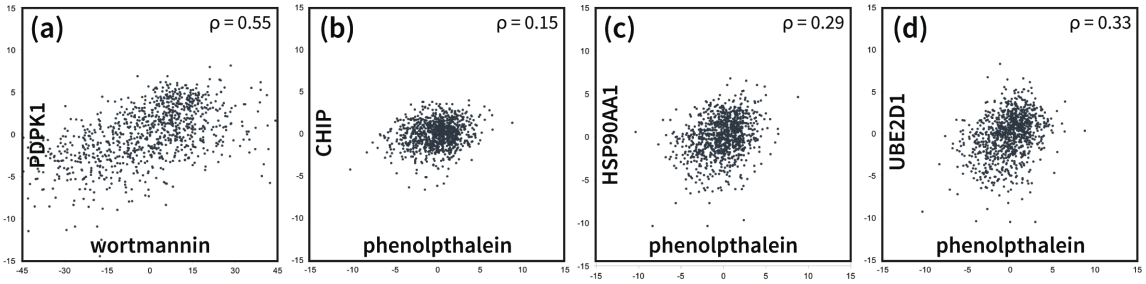
bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure S2. Predicted targets for the drug wortmannin. Points represent structural models of the top-100 RF-predicted potential targets for wortmannin. The RF 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.



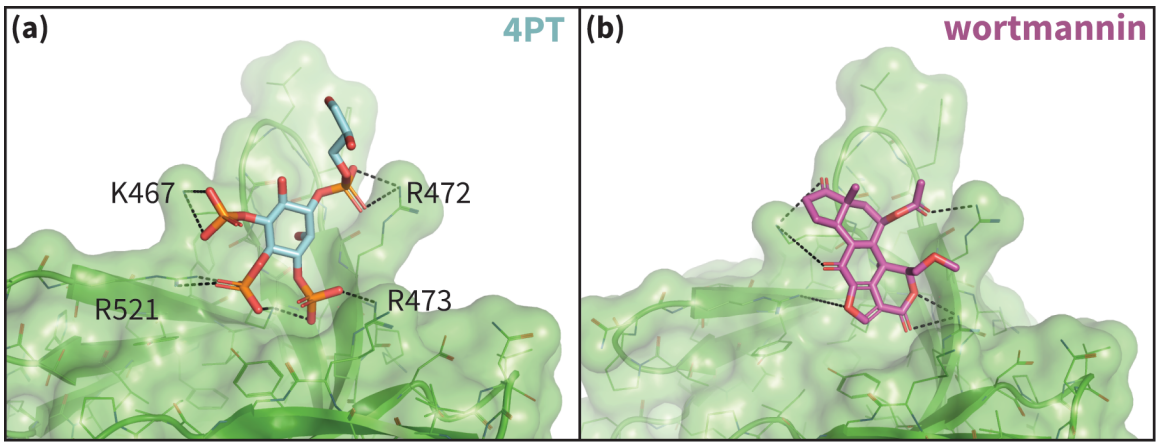
bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure S3. Expression profile correlations of newly discovered interactions. (a) Drug-induced expression signature of wortmannin shows strong direct correlation with the knockdown of newly validated target PDPK1. (b,c,d) Drug induced expression signature of phenolphthalein shows little direct correlation with newly validated target CHIP, but shows comparatively stronger indirect correlation with CHIP interaction partners HSP90AA1 and UBE2D1.



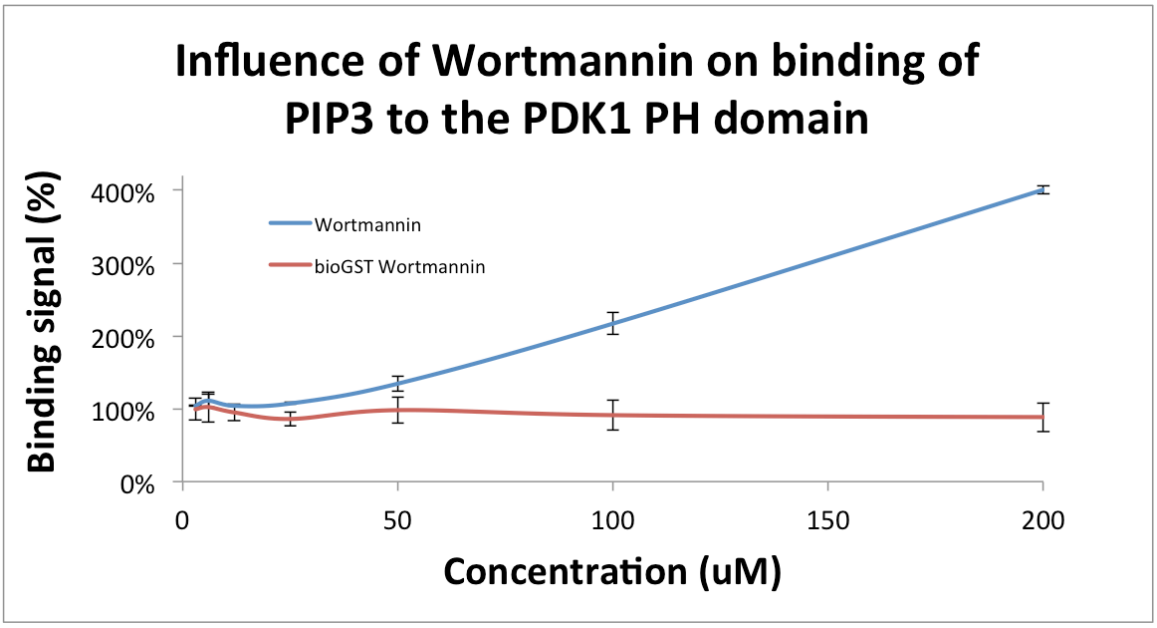
bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure S4. Docking model of wortmannin bound to the PH domain of PDK1. (a) Cocrystal structure (PDB ID: 1W1G (Komander et al., 2004)) 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.



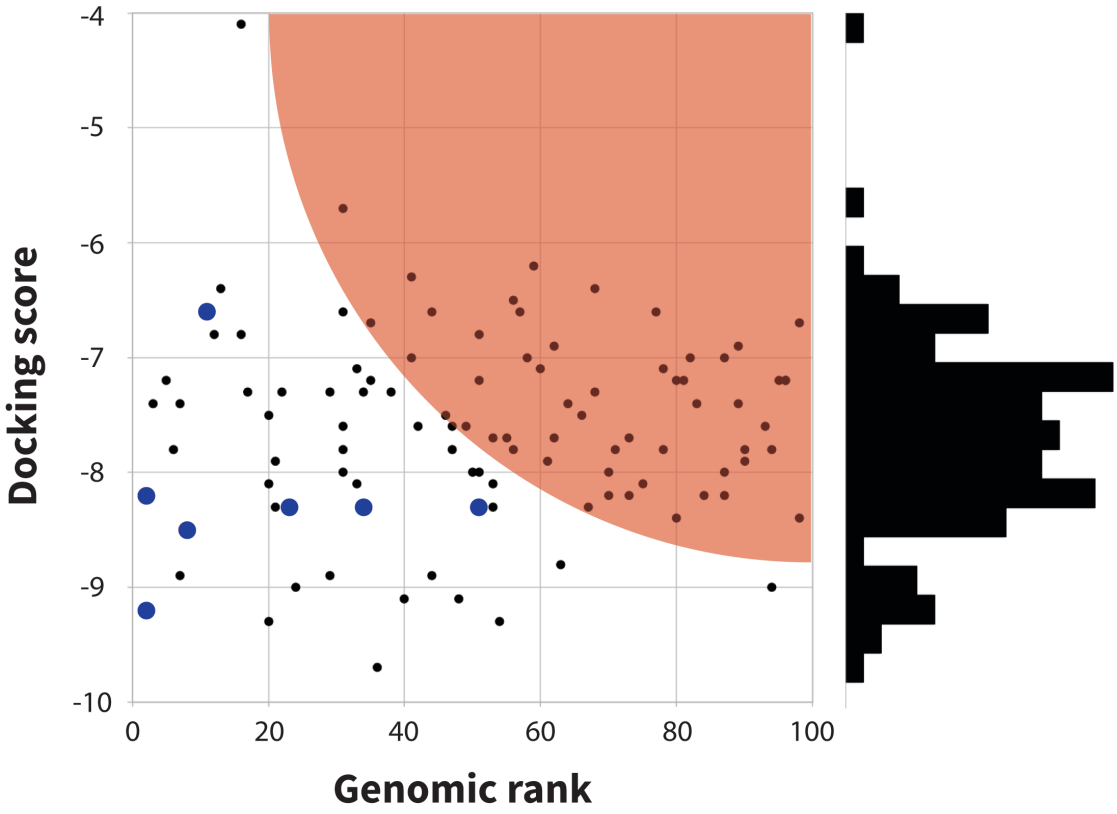
bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure S5. Alphascreen PDK1-PIP3 interaction-displacement assay results for increasing concentrations of wortmannin. Error bars represent the standard error on the mean from two parallel runs.



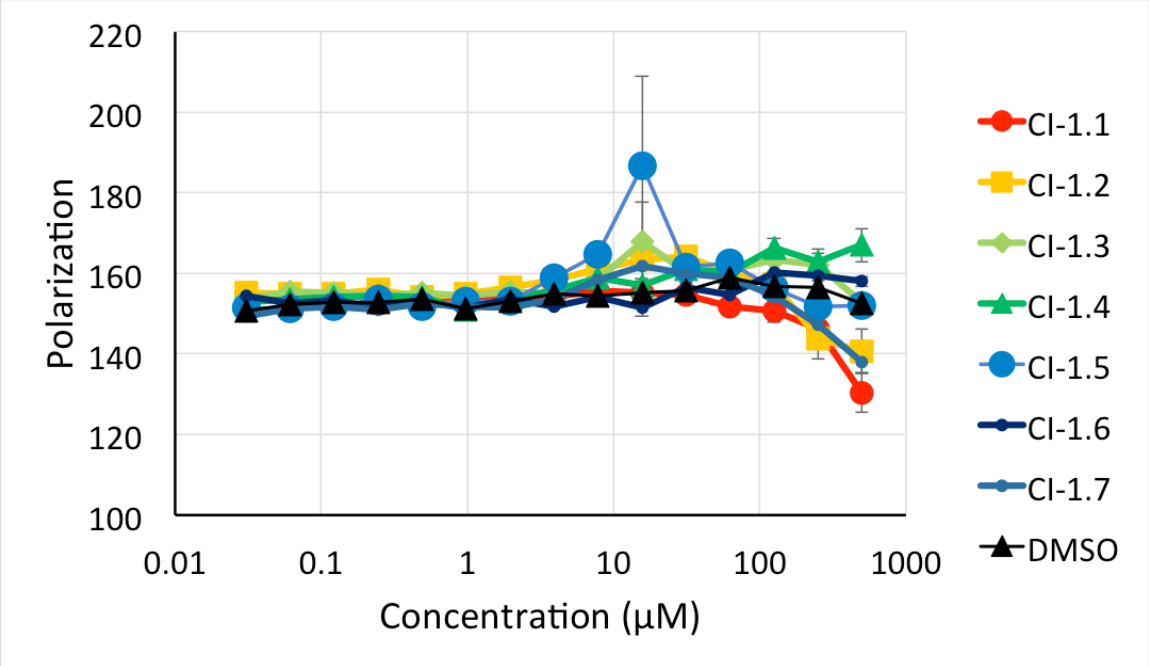
bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure S6. 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.



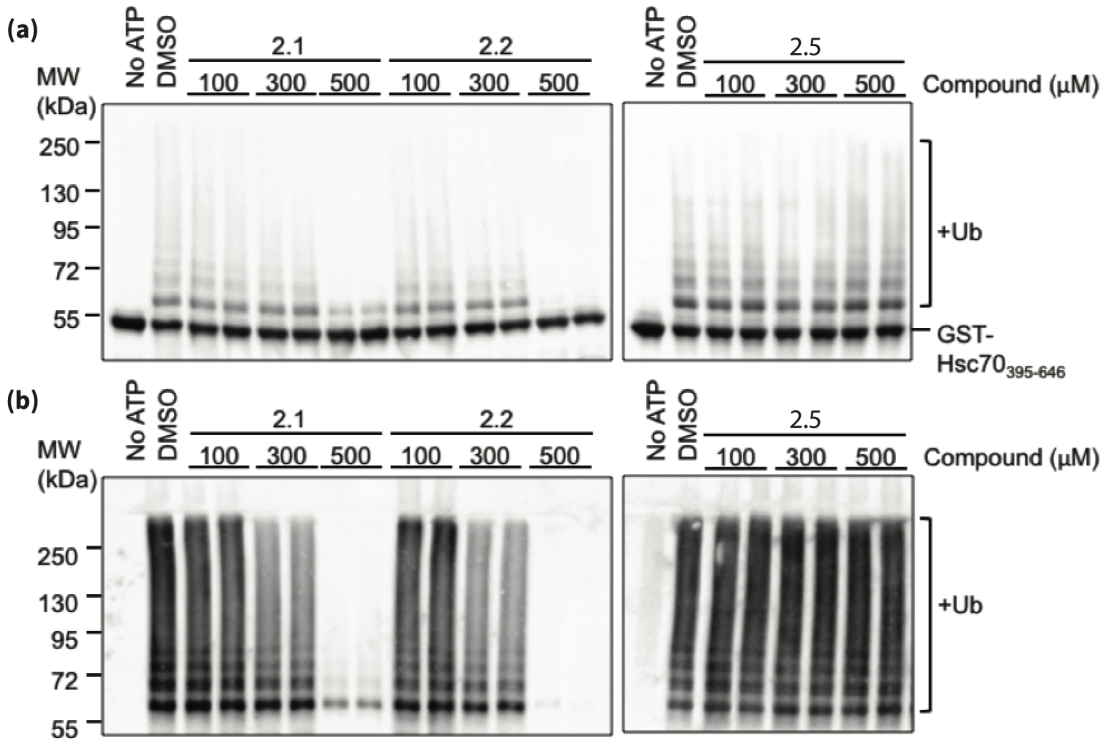
bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure S7. 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.



bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

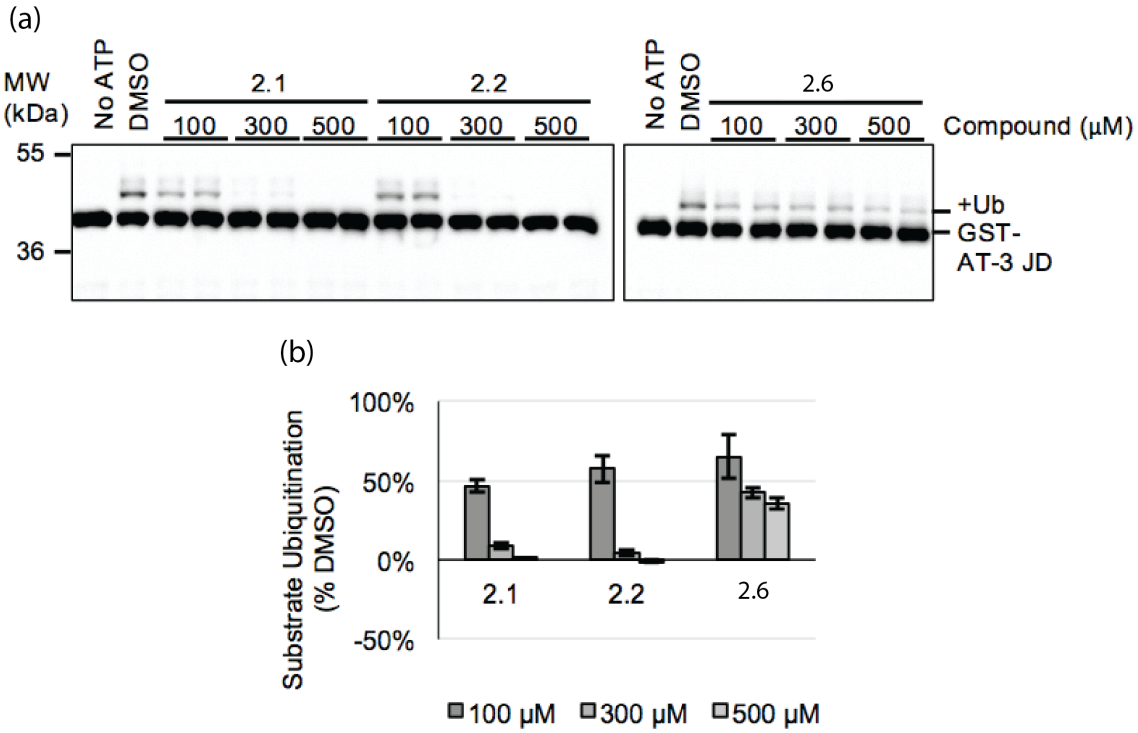
Figure S8. CHIP inhibitors prevent ubiquitination by CHIP in vitro. (a) Anti-GST western blot showing substrate ubiquitination by CHIP in reactions treated with high ranked (2.1, 2.2) and low ranked (2.5) compounds. **(**b) Anti-ubiquitin western blot showing total ubiquitination by CHIP in reactions treated with high ranked (2.1, 2.2) and low ranked (2.5) compounds.



bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

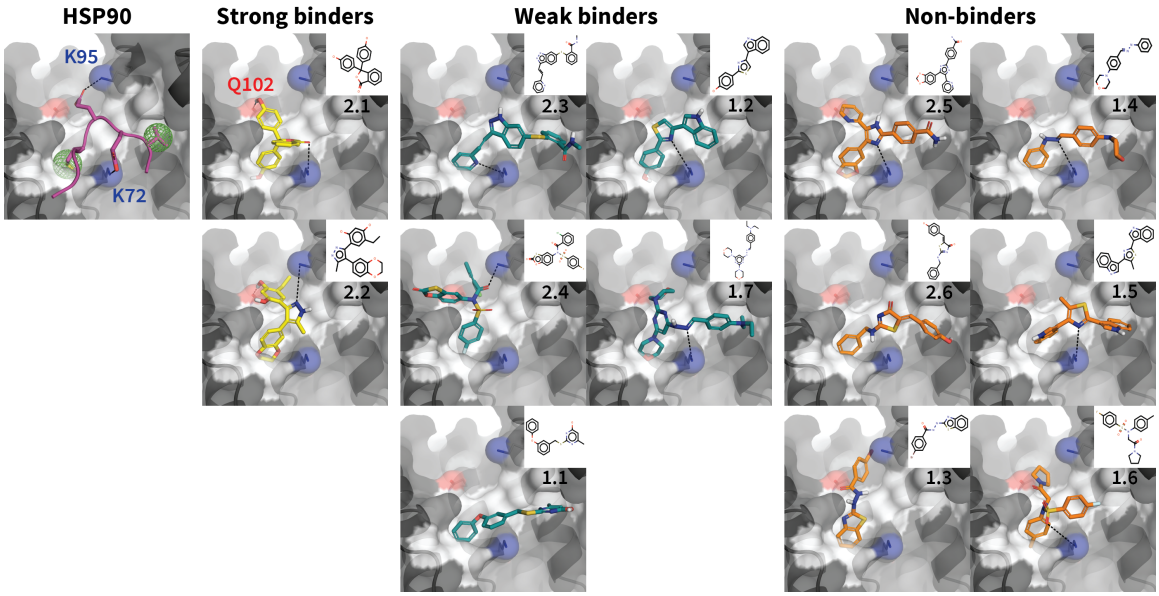
bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure S9. **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).



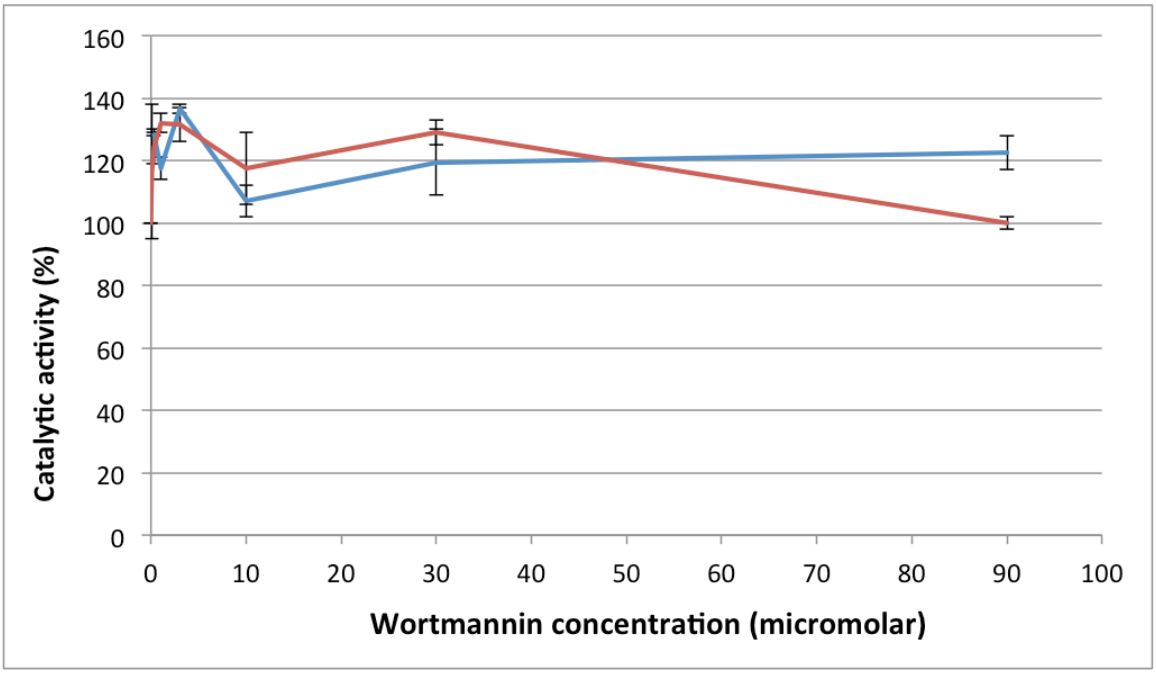
bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure S10. Comparison of virtual screens against CHIP. **HSP90** shows structure of the CHIP (grey) - HSP90 (magenta) interface (PDB ID: 2C2L(Zhang et al., 2005)), 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. **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.



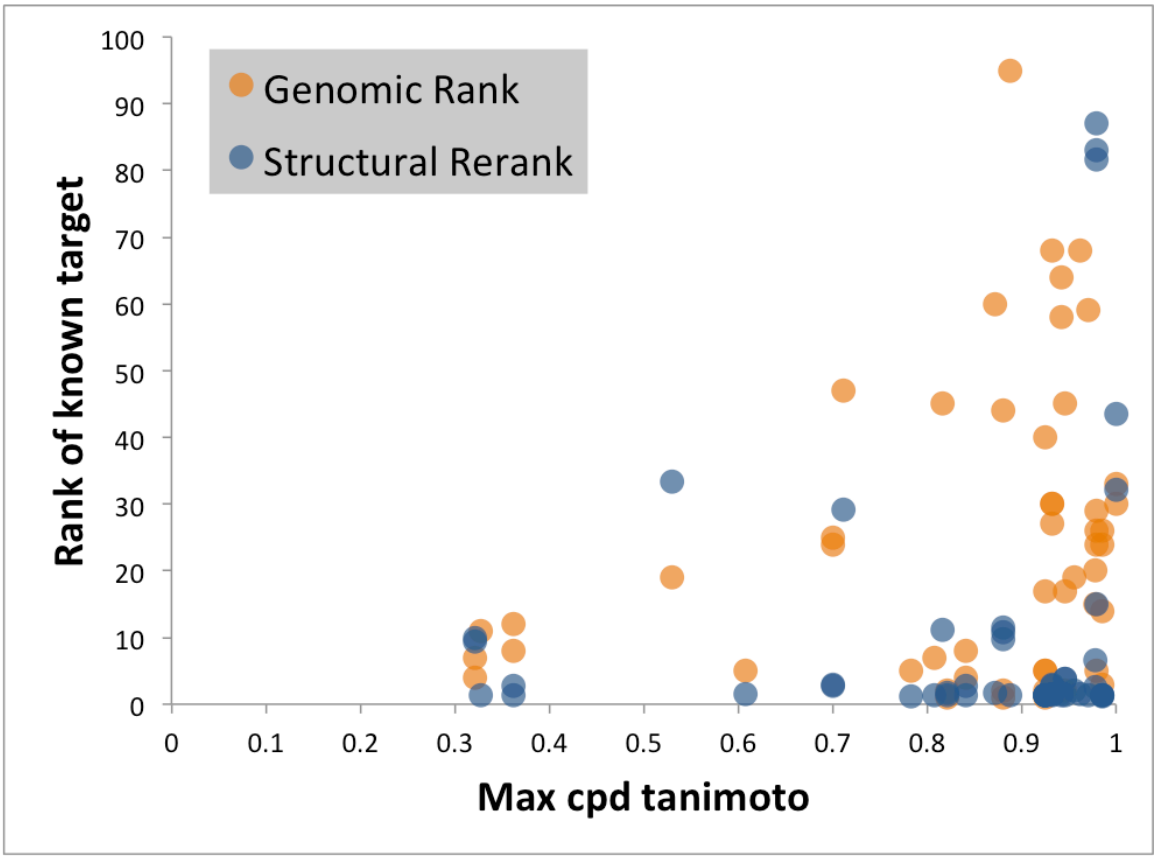
bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure S11. 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.



bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure S12. 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.



bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

4. SI Tables

Table S1. Performance of target prediction using different features and methods on 29 FDA-approved drugs. DIR: direct correlation feature; IND: indirect correlation feature; CS: cell selection feature; MAX: maximum differential expression feature; MEAN: mean differential expression feature; LR: logistic regression; RF: random forest. Values are for the ranking of the top known target for each drug.

Drug	Random	DIR	IND	CS	MAX	MEAN	LR	RF
vinorelbine	310	126	128	1318	1690	425	28	88
dexamethasone	1498	1891	284	943	315	1143	757	157
dasatinib	2325	1009	94	222	290	2621	182	532
vincristine	1979	473	439	386	2231	2196	456	37
mycophenolate-mofetil	564	1100	1263	2986	100	301	3064	3086
amlodipine	995	1338	2439	1801	1875	974	3037	650
lovastatin	1712	72	811	2078	1124	1068	1334	55
clobetasol	2194	820	21	157	74	15	38	65
calcitriol	2514	1059	2938	221	125	1814	1299	252
flutamide	919	2604	69	2806	463	298	702	647
prednisolone	2382	1439	206	787	402	1068	257	23
nifedipine	940	1225	1465	1285	88	322	3037	2249
vemurafenib	1042	1	82	1	1149	1403	22	2
glibenclamide	29	1415	2028	409	1059	740	1300	366
digoxin	2376	73	1470	118	828	567	732	44
bortezomib	1882	1	1	2	2546	2513	24	5
vinblastine	1612	515	56	100	224	377	38	2

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

digitoxin	573	89	430	216	521	653	79	50
losartan	645	489	988	770	636	31	735	1931
pitavastatin	1855	1976	1036	1117	90	527	1632	373
digoxin	69	521	776	194	127	559	208	64
hydrocortisone	303	312	72	58	93	122	29	17
paclitaxel	2299	74	121	47	371	1862	79	19
lovastatin	988	1	735	1587	1698	1484	128	100
irinotecan	1742	1023	20	236	128	1886	46	160
vincristine	1394	96	74	17	1272	69	28	9
vinblastine	1359	490	75	1383	373	1735	35	2
raloxifene	2080	2883	1818	1172	1064	479	1114	2520
digoxin	1005	102	1066	112	2096	2027	252	167
Mean Ranking	1365	800.6	724.3	776.9	794.9	1009.6	712.8	471.4
Top 100	2	8	10	6	5	3	11	16

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table S2. **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.

LINCS ID	*Rank of known target*
BRD_0x2D_K38775274	*29*
BRD_0x2D_A82238138	*23*
BRD_0x2D_A22032524	*66*
BRD_0x2D_A92177080	*86*
BRD_0x2D_K46137903	*1245*
BRD_0x2D_A82371568	*1472*
BRD_0x2D_A97437073	*18*
BRD_0x2D_K28307902	*1206*
BRD_0x2D_A35108200	*1402*
BRD_0x2D_A23770159	*605*
BRD_0x2D_A81233518	*850*
BRD_0x2D_A01643550	*155*
BRD_0x2D_K56343971	*611*
BRD_0x2D_K36927236	*1163*
BRD_0x2D_K76723084	*394*
BRD_0x2D_K56851771	*564*
BRD_0x2D_A76528577	*1260*

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

BRD_0x2D_K47635719	*1441*
BRD_0x2D_K76205745	*2887*
BRD_0x2D_A16998493	*496*
BRD_0x2D_K18194590	*1046*
BRD_0x2D_A75144621	*79*
BRD_0x2D_A78391468	*2697*
BRD_0x2D_A26711594	*1942*
BRD_0x2D_K55395145	*1196*
BRD_0x2D_K73999723	*606*
BRD_0x2D_K77554836	*565*
BRD_0x2D_A23637604	*2015*
BRD_0x2D_A65449987	*135*
BRD_0x2D_K28936863	*2*
BRD_0x2D_K15108141	*131*
BRD_0x2D_K77175907	*2294*
BRD_0x2D_A20126139	*2356*
BRD_0x2D_K18135438	*23*
BRD_0x2D_A69512159	*1123*
BRD_0x2D_A23723433	*547*
BRD_0x2D_K31627533	*69*
BRD_0x2D_K82109576	*2607*
BRD_0x2D_A79672927	*3188*
BRD_0x2D_K72238567	*266*
BRD_0x2D_A70155556	*909*
BRD_0x2D_K84937637	*2088*

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

BRD_0x2D_K84036904	*225*
BRD_0x2D_A90131694	*2*
BRD_0x2D_A02180903	*2596*
BRD_0x2D_A68723818	*1784*
BRD_0x2D_A27887842	*316*
BRD_0x2D_K56429665	*70*
BRD_0x2D_A79768653	*77*
BRD_0x2D_K32821942	*7*
BRD_0x2D_K60511616	*499*
BRD_0x2D_K38003476	*15*
BRD_0x2D_A63894585	*377*
BRD_0x2D_K88510285	*589*
BRD_0x2D_K28143534	*2390*
BRD_0x2D_K02637541	*2822*
BRD_0x2D_K00824317	*1164*
BRD_0x2D_A48720949	*59*
BRD_0x2D_K97810537	*367*
BRD_0x2D_A83237092	*2513*
BRD_0x2D_K32744045	*2041*
BRD_0x2D_K67174588	*2716*
BRD_0x2D_A46186775	*137*
BRD_0x2D_K49328571	*190*
BRD_0x2D_A77824596	*382*
BRD_0x2D_A23359898	*1193*
BRD_0x2D_K27721098	*80*

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

BRD_0x2D_A69636825	*2513*
BRD_0x2D_K60640630	*279*
BRD_0x2D_K33106058	*67*
BRD_0x2D_A30815329	*13*
BRD_0x2D_A29426959	*11*
BRD_0x2D_A49225603	*151*
BRD_0x2D_K10916986	*1191*
BRD_0x2D_K35483542	*1508*
BRD_0x2D_K43736954	*261*
BRD_0x2D_K23478508	*1547*
BRD_0x2D_K66296774	*357*
BRD_0x2D_A37780065	*3041*
BRD_0x2D_A81772229	*127*
BRD_0x2D_K65146499	*176*
BRD_0x2D_A15297126	*2623*
BRD_0x2D_K96354014	*238*
BRD_0x2D_A55393291	*40*
BRD_0x2D_K17674993	*270*
BRD_0x2D_A49765801	*2669*
BRD_0x2D_A62025033	*26*
BRD_0x2D_K23566484	*81*
BRD_0x2D_A94756469	*4*
BRD_0x2D_A07765530	*2555*
BRD_0x2D_A34299591	*719*
BRD_0x2D_K11129031	*45*

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

BRD_0x2D_A36010170	*67*
BRD_0x2D_K27316855	*705*
BRD_0x2D_K53790871	*82*
BRD_0x2D_K08547377	*1069*
BRD_0x2D_K47832606	*1678*
BRD_0x2D_A28746609	*2083*
BRD_0x2D_K09416995	*2646*
BRD_0x2D_K97514127	*1485*
BRD_0x2D_A60414806	*295*
BRD_0x2D_A22783572	*15*
BRD_0x2D_A46335897	*531*
BRD_0x2D_K81169441	*2764*
BRD_0x2D_K81709173	*680*
BRD_0x2D_A07440155	*643*
BRD_0x2D_K34776109	*203*
BRD_0x2D_K89626439	*2384*
BRD_0x2D_K49577446	*1516*
BRD_0x2D_A55594068	*35*
BRD_0x2D_A60571864	*16*
BRD_0x2D_A69951442	*4*
BRD_0x2D_K41260949	*1313*
BRD_0x2D_A07000685	*22*
BRD_0x2D_A13133631	*282*
BRD_0x2D_K90553655	*1655*
BRD_0x2D_A92439610	*2294*

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

BRD_0x2D_M30523314	*242*
BRD_0x2D_K92428153	*807*
BRD_0x2D_A26095496	*122*
BRD_0x2D_K12994359	*2171*
BRD_0x2D_A96107863	*1110*
BRD_0x2D_K99369265	*7*

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table S3. 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

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table S4. Comparison of our pipeline to existing drug-target prediction methods. The average ranking of the highest ranked known target is listed for all 63 validation 'hits' and for the subset of 53 hits with known structures. Rankings are compared between the initial random-forest ranking (GEN), the structural re-ranking of the top 100 RF predicted targets (STR), the HTDocking server (HTD), and the PharmMapper server (PHM).

	Avg # top-100 structures available	GEN	STR	HTD	PHM
All hits (n=63)	69	22	24	56	23
Known structures (n=53)	71	23	13	50	12

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table S5. 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.

COMPOUND	LINCS ID	Num Structs	GEN	STR	HTD	PHM
alclometasone	BRD-A90131694	76	15%	7%	54%	3%
beclometasone	BRD-K97810537	77	20%	3%	49%	1%
betamethasone	BRD-A02180903	81	3%	1%	9%	0%
betamethasone	BRD-A92177080	74	24%	1%	9%	0%
betamethasone	BRD-K39188321	72	26%	1%	9%	0%
bortezomib	BRD-K88510285	66	19%	33%	100%	100%
budesonide	BRD-A60571864	72	2%	10%	60%	5%
budesonide	BRD-A82238138	74	1%	11%	60%	5%
budesonide	BRD-A34299591	70	44%	11%	60%	5%
clobetasol	BRD-A26095496	73	2%	1%	77%	0%
clobetasol	BRD-A63894585	58	1%	2%	77%	0%
clocortolone	BRD-K38003476	72	7%	1%	79%	0%

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

cortisone	BRD-K43736954	62	60%	2%	2%	10%
desoximetasone	BRD-A49447682	72	59%	1%	39%	0%
dexamethasone	BRD-A35108200	79	2%	1%	80%	0%
dexamethasone	BRD-A10188456	75	5%	1%	80%	0%
dexamethasone	BRD-K38775274	75	5%	1%	80%	0%
dexamethasone	BRD-A93424738	74	1%	1%	80%	0%
dexamethasone	BRD-K47635719	71	40%	1%	80%	0%
diflorasone	BRD-K17674993	78	17%	1%	85%	1%
fludroxycortide	BRD-K00824317	75	8%	1%	72%	8%
fludroxycortide	BRD-A49765801	75	4%	3%	72%	8%
flunisolide	BRD-A65449987	74	58%	1%	42%	10%
flunisolide	BRD-K49577446	61	64%	2%	42%	10%
fluocinonide	BRD-A15297126	48	19%	2%	86%	2%
fluorometholone	BRD-A13133631	77	14%	1%	84%	0%
flutamide	BRD-K28307902	79	47%	29%	49%	1%
fulvestrant	BRD-A90490067	75	7%	9%	0%	26%
fulvestrant	BRD-A83237092	60	4%	10%	0%	26%
halcinonide	BRD-K81709173	83	5%	1%	75%	1%
hydrocortisone	BRD-A46186775	74	25%	3%	60%	3%
hydrocortisone	BRD-A65767837	68	24%	3%	60%	3%
irinotecan	BRD-K08547377	76	95%	1%	0%	100%
lovastatin	BRD-A70155556	81	30%	32%	15%	10%
medrysone	BRD-A20126139	75	68%	1%	70%	4%
medrysone	BRD-K56515112	63	2%	2%	70%	4%
mometasone	BRD-K60640630	63	5%	2%	30%	0%

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

paclitaxel	BRD-A23723433	72	41%	100%	100%	100%
paclitaxel	BRD-A28746609	77	5%	100%	100%	100%
prednicarbate	BRD-K46137903	54	45%	11%	65%	4%
prednisolone	BRD-A27887842	78	45%	4%	8%	0%
prednisolone	BRD-A01643550	78	17%	4%	8%	0%
rimexolone	BRD-K31627533	77	2%	1%	84%	0%
simvastatin	BRD-A81772229	62	33%	44%	0%	3%
sirolimus	BRD-A79768653	60	5%	82%	45%	15%
sirolimus	BRD-K89626439	53	29%	83%	45%	15%
sirolimus	BRD-K84937637	62	26%	87%	45%	15%
temsirolimus	BRD-A62025033	67	24%	15%	12%	3%
testosterone	BRD-K90553655	68	30%	1%	1%	5%
testosterone	BRD-A48720949	74	27%	3%	1%	5%
testosterone	BRD-A55393291	70	30%	3%	1%	5%
toremifene	BRD-K67174588	73	8%	1%	54%	2%
toremifene	BRD-K51350053	72	12%	3%	54%	2%
triamcinolone	BRD-A37780065	67	68%	1%	51%	21%
vemurafenib	BRD-K56343971	70	11%	1%	84%	0%
vinblastine	BRD-A22783572	60	1%	100%	100%	100%
vinblastine	BRD-A55594068	60	1%	100%	100%	100%
vincristine	BRD-K82109576	53	46%	100%	100%	100%
vincristine	BRD-A60414806	64	4%	100%	100%	100%
vincristine	BRD-A76528577	59	1%	100%	100%	100%
vinorelbine	BRD-M30523314	51	25%	100%	100%	100%
vinorelbine	BRD-K97514127	54	24%	100%	100%	100%

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

vinorelbine	BRD-K10916986	62	17%	100%	100%	100%

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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

Table S6. Symbols and notations

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table S7. 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**
Direct Correlation	f_{cor}	Correlation between a drug treatment experiment and a gene knockdown experiment
Indirect Correlation	f_{PC}	Fraction of the known binding partners of a gene in the top *X* correlated knockdown experiments
Cell Selection	f_{CS}	Correlation between a drug treatment experiment and the control experiment for the cell line
PPI Expression	f_{PE}	The average or the max (absolute value) expression for

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

		the known binding partners of a gene

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table S8. Enrichment compound names. The names and LINCS IDs of the validation compounds shown in Figure 3.

Num	Compound Name	LINCS ID
1	dexamethasone	BRD-A93424738
2	clobetasol	BRD-A63894585
3	budesonide	BRD-A82238138
4	dexamethasone	BRD-A35108200
5	rimexolone	BRD-K31627533
6	clobetasol	BRD-A26095496
7	medrysone	BRD-K56515112
8	budesonide	BRD-A60571864
9	betamethasone	BRD-A02180903
10	fludroxycortide	BRD-A49765801
11	fulvestrant	BRD-A83237092
12	halcinonide	BRD-K81709173
13	dexamethasone	BRD-A10188456
14	dexamethasone	BRD-K38775274
15	mometasone	BRD-K60640631
16	sirolimus	BRD-A79768653
17	clocortolone	BRD-K38003476
18	fulvestrant	BRD-A90490067
19	fludroxycortide	BRD-K00824317
20	toremifene	BRD-K67174588
21	vemurafenib	BRD-K56343971

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

22	toremifene	BRD-K51350053
23	fluorometholone	BRD-A13133631
24	alclometasone	BRD-A90131694
25	diflorasone	BRD-K17674993
26	prednisolone	BRD-A01643550
27	fluocinonide	BRD-A15297126
28	bortezomib	BRD-K88510285
29	beclometasone	BRD-K97810537
30	betamethasone	BRD-A92177080
31	hydrocortisone	BRD-A65767837
32	temsirolimus	BRD-A62025033
33	hydrocortisone	BRD-A46186775
34	betamethasone	BRD-K39188321
35	sirolimus	BRD-K84937637
36	testosterone	BRD-A48720949
37	sirolimus	BRD-K89626439
38	testosterone	BRD-K90553655
39	testosterone	BRD-A55393291
40	lovastatin	BRD-A70155556
41	simvastatin	BRD-A81772229
42	dexamethasone	BRD-K47635719
43	budesonide	BRD-A34299591
44	prednisolone	BRD-A27887842
45	prednicarbate	BRD-K46137903
46	flutamide	BRD-K28307902

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

47	flunisolide	BRD-A65449987
48	desoximetasone	BRD-A49447682
49	cortisone	BRD-K43736954
50	flunisolide	BRD-K49577446
51	medrysone	BRD-A20126139
52	triamcinolone	BRD-A37780065
53	irinotecan	BRD-K08547377

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table S9. 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

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table S10. 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

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References

Andy Liaw, M.W. (2002). Classification and regression by randomforest. R news *2*, 18-22.

Assimon, V.A., Southworth, D.R., and Gestwicki, J.E. (2015). Specific Binding of Tetratricopeptide Repeat Proteins to Heat Shock Protein 70 (Hsp70) and Heat Shock Protein 90 (Hsp90) Is Regulated by Affinity and Phosphorylation. Biochemistry *54*, 7120-7131.

Bakan, A., Meireles, L.M., and Bahar, I. (2011). ProDy: protein dynamics inferred from theory and experiments. Bioinformatics *27*, 1575-1577.

Baumgartner, M.P., and Camacho, C.J. (2016). Choosing the Optimal Rigid Receptor for Docking and Scoring in the CSAR 2013/2014 Experiment. J Chem Inf Model *56*, 1004-1012.

Berndsen, C.E., and Wolberger, C. (2011). A spectrophotometric assay for conjugation of ubiquitin and ubiquitin-like proteins. Anal Biochem *418*, 102-110.

Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer, E.F., Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., and Tasumi, M. (1978). The Protein Data Bank: a computer-based archival file for macromolecular structures. Arch Biochem Biophys *185*, 584-591.

Bleicher, K.H., Bohm, H.J., Muller, K., and Alanine, A.I. (2003). Hit and lead generation: beyond high-throughput screening. Nat Rev Drug Discov *2*, 369-378.

Chatr-Aryamontri, A., Breitkreutz, B.J., Oughtred, R., Boucher, L., Heinicke, S., Chen, D., Stark, C., Breitkreutz, A., Kolas, N., O'Donnell, L., *et al.* (2015). The BioGRID interaction database: 2015 update. Nucleic Acids Res *43*, D470-478.

Cosgrove, E.J., Zhou, Y., Gardner, T.S., and Kolaczyk, E.D. (2008). Predicting gene targets of perturbations via network-based filtering of mRNA expression compendia. Bioinformatics *24*, 2482-2490.

Dettori, R., Sonzogni, S., Meyer, L., Lopez-Garcia, L.A., Morrice, N.A., Zeuzem, S., Engel, M., Piiper, A., Neimanis, S., Frodin, M., *et al.* (2009). Regulation of the interaction between protein kinase C-related protein kinase 2 (PRK2) and its upstream kinase, 3-phosphoinositide-dependent protein kinase 1 (PDK1). J Biol Chem *284*, 30318-30327.

Diaz-Uriarte, R., and Alvarez de Andres, S. (2006). Gene selection and classification of microarray data using random forest. BMC Bioinformatics *7*, 3.

Drews, J. (2000). Drug discovery: a historical perspective. Science *287*, 1960-1964.

Faggiano, S., Menon, R.P., Kelly, G.P., McCormick, J., Todi, S.V., Scaglione, K.M., Paulson, H.L., and Pastore, A. (2013). Enzymatic production of mono-ubiquitinated proteins for structural studies: The example of the Josephin domain of ataxin-3. FEBS Open Bio *3*, 453-458.

Faith, J.J., Driscoll, M.E., Fusaro, V.A., Cosgrove, E.J., Hayete, B., Juhn, F.S., Schneider, S.J., and Gardner, T.S. (2008). Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata. Nucleic Acids Res *36*, D866-870.

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Gao, X., and Harris, T.K. (2006). Role of the PH domain in regulating in vitro autophosphorylation events required for reconstitution of PDK1 catalytic activity. Bioorg Chem *34*, 200-223.

Gaulton, A., Bellis, L.J., Bento, A.P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B.*, et al.* (2012). ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res *40*, D1100-1107.

Gfeller, D., Grosdidier, A., Wirth, M., Daina, A., Michielin, O., and Zoete, V. (2014). SwissTargetPrediction: a web server for target prediction of bioactive small molecules. Nucleic Acids Res *42*, W32-38.

Gregori-Puigjane, E., Setola, V., Hert, J., Crews, B.A., Irwin, J.J., Lounkine, E., Marnett, L., Roth, B.L., and Shoichet, B.K. (2012). Identifying mechanism-of-action targets for drugs and probes. Proc Natl Acad Sci U S A *109*, 11178-11183.

Harris, M.A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C.*, et al.* (2004). The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res *32*, D258-261.

Irwin, J.J., and Shoichet, B.K. (2005). ZINC--a free database of commercially available compounds for virtual screening. J Chem Inf Model *45*, 177-182.

Isik, Z., Baldow, C., Cannistraci, C.V., and Schroeder, M. (2015). Drug target prioritization by perturbed gene expression and network information. Sci Rep *5*, 17417.

Keiser, M.J., Setola, V., Irwin, J.J., Laggner, C., Abbas, A.I., Hufeisen, S.J., Jensen, N.H., Kuijer, M.B., Matos, R.C., Tran, T.B.*, et al.* (2009). Predicting new molecular targets for known drugs. Nature *462*, 175-181.

Keshava Prasad, T.S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A.*, et al.* (2009). Human Protein Reference Database--2009 update. Nucleic Acids Res *37*, D767-772.

Koes, D.R., Baumgartner, M.P., and Camacho, C.J. (2013). Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise. J Chem Inf Model *53*, 1893-1904.

Koes, D.R., and Camacho, C.J. (2012). ZINCPharmer: pharmacophore search of the ZINC database. Nucleic Acids Res *40*, W409-414.

Koes, D.R., Pabon, N.A., Deng, X., Phillips, M.A., and Camacho, C.J. (2015). A Teach-Discover-Treat Application of ZincPharmer: An Online Interactive Pharmacophore Modeling and Virtual Screening Tool. PLoS One *10*, e0134697.

Komander, D., Fairservice, A., Deak, M., Kular, G.S., Prescott, A.R., Peter Downes, C., Safrany, S.T., Alessi, D.R., and van Aalten, D.M. (2004). Structural insights into the regulation of PDK1 by phosphoinositides and inositol phosphates. EMBO J *23*, 3918-3928.

Kozakov, D., Grove, L.E., Hall, D.R., Bohnuud, T., Mottarella, S.E., Luo, L., Xia, B., Beglov, D., and Vajda, S. (2015). The FTMap family of web servers for determining and characterizing ligand-binding hot spots of proteins. Nat Protoc *10*, 733-755.

Laenen, G., Thorrez, L., Bornigen, D., and Moreau, Y. (2013). Finding the targets of a drug by integration of gene expression data with a protein interaction network. Mol Biosyst *9*, 1676-1685.

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Lamb, J. (2007). The Connectivity Map: a new tool for biomedical research. Nat Rev Cancer *7*, 54-60.

Lamb, J., Crawford, E.D., Peck, D., Modell, J.W., Blat, I.C., Wrobel, M.J., Lerner, J., Brunet, J.P., Subramanian, A., Ross, K.N., *et al.* (2006). The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science *313*, 1929-1935.

Li, H., Gao, Z., Kang, L., Zhang, H., Yang, K., Yu, K., Luo, X., Zhu, W., Chen, K., Shen, J., *et al.* (2006). TarFisDock: a web server for identifying drug targets with docking approach. Nucleic Acids Res *34*, W219-224.

Liu, X., Ouyang, S., Yu, B., Liu, Y., Huang, K., Gong, J., Zheng, S., Li, Z., Li, H., and Jiang, H. (2010). PharmMapper server: a web server for potential drug target identification using pharmacophore mapping approach. Nucleic Acids Res *38*, W609-614.

Lo, Y.C., Senese, S., Li, C.M., Hu, Q., Huang, Y., Damoiseaux, R., and Torres, J.Z. (2015). Large-scale chemical similarity networks for target profiling of compounds identified in cell-based chemical screens. PLoS Comput Biol *11*, e1004153.

Martinez-Jimenez, F., and Marti-Renom, M.A. (2015). Ligand-target prediction by structural network biology using nAnnoLyze. PLoS Comput Biol *11*, e1004157.

Masters, T.A., Calleja, V., Armoogum, D.A., Marsh, R.J., Applebee, C.J., Laguerre, M., Bain, A.J., and Larijani, B. (2010). Regulation of 3-phosphoinositide-dependent protein kinase 1 activity by homodimerization in live cells. Sci Signal *3*, ra78.

Mayr, L.M., and Bojanic, D. (2009). Novel trends in high-throughput screening. Curr Opin Pharmacol *9*, 580-588.

Meacham, G.C., Patterson, C., Zhang, W., Younger, J.M., and Cyr, D.M. (2001). The Hsc70 co-chaperone CHIP targets immature CFTR for proteasomal degradation. Nat Cell Biol *3*, 100-105.

Meslamani, J., Li, J., Sutter, J., Stevens, A., Bertrand, H.O., and Rognan, D. (2012). Protein-ligand-based pharmacophores: generation and utility assessment in computational ligand profiling. J Chem Inf Model *52*, 943-955.

Navlakha, S., He, X., Faloutsos, C., and Bar-Joseph, Z. (2014). Topological properties of robust biological and computational networks. J R Soc Interface *11*, 20140283.

Nickel, J., Gohlke, B.O., Erehman, J., Banerjee, P., Rong, W.W., Goede, A., Dunkel, M., and Preissner, R. (2014). SuperPred: update on drug classification and target prediction. Nucleic Acids Res *42*, W26-31.

Overington, J.P., Al-Lazikani, B., and Hopkins, A.L. (2006). How many drug targets are there? Nat Rev Drug Discov *5*, 993-996.

Paul, I., and Ghosh, M.K. (2015). A CHIPotle in physiology and disease. Int J Biochem Cell Biol *58*, 37-52.

Persidis, A. (1998). High-throughput screening. Advances in robotics and miniturization continue to accelerate drug lead identification. Nat Biotechnol *16*, 488-489.

Pritchard, J.F., Jurima-Romet, M., Reimer, M.L., Mortimer, E., Rolfe, B., and Cayen, M.N. (2003). Making better drugs: Decision gates in non-clinical drug development. Nat Rev Drug Discov *2*, 542-553.

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Qi, Y., Bar-Joseph, Z., and Klein-Seetharaman, J. (2006). Evaluation of different biological data and computational classification methods for use in protein interaction prediction. Proteins *63*, 490-500.

Rognan, D. (2010). Structure-Based Approaches to Target Fishing and Ligand Profiling. Mol Inform *29*, 176-187.

Schulze, J.O., Saladino, G., Busschots, K., Neimanis, S., Suss, E., Odadzic, D., Zeuzem, S., Hindie, V., Herbrand, A.K., Lisa, M.N.*, et al.* (2016). Bidirectional Allosteric Communication between the ATP-Binding Site and the Regulatory PIF Pocket in PDK1 Protein Kinase. Cell Chem Biol *23*, 1193-1205.

Sheffield, P., Garrard, S., and Derewenda, Z. (1999). Overcoming expression and purification problems of RhoGDI using a family of "parallel" expression vectors. Protein Expr Purif *15*, 34-39.

Shen-Orr, S.S., Tibshirani, R., Khatri, P., Bodian, D.L., Staedtler, F., Perry, N.M., Hastie, T., Sarwal, M.M., Davis, M.M., and Butte, A.J. (2010). Cell type-specific gene expression differences in complex tissues. Nat Methods *7*, 287-289.

Swinney, D.C., and Anthony, J. (2011). How were new medicines discovered? Nat Rev Drug Discov *10*, 507-519.

Todi, S.V., Scaglione, K.M., Blount, J.R., Basrur, V., Conlon, K.P., Pastore, A., Elenitoba-Johnson, K., and Paulson, H.L. (2010). Activity and cellular functions of the deubiquitinating enzyme and polyglutamine disease protein ataxin-3 are regulated by ubiquitination at lysine 117. J Biol Chem *285*, 39303-39313.

Vanhaesebroeck, B., and Alessi, D.R. (2000). The PI3K-PDK1 connection: more than just a road to PKB. Biochem J *346 Pt 3*, 561-576.

Wang, J.C., Chu, P.Y., Chen, C.M., and Lin, J.H. (2012). idTarget: a web server for identifying protein targets of small chemical molecules with robust scoring functions and a divide-and-conquer docking approach. Nucleic Acids Res *40*, W393-399.

Wang, L., P. Wipf, and X.-Q. Xie (2012). HTDocking- identifying possible targets for small molecules by high throughput
docking algorithm.

Wishart, D.S., Knox, C., Guo, A.C., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z., and Woolsey, J. (2006). DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res *34*, D668-672.

Yamanishi, Y., Kotera, M., Kanehisa, M., and Goto, S. (2010). Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. Bioinformatics *26*, i246-254.

Ye, Z., Baumgartner, M.P., Wingert, B.M., and Camacho, C.J. (2016). Optimal strategies for virtual screening of induced-fit and flexible target in the 2015 D3R Grand Challenge. J Comput Aided Mol Des *30*, 695-706.

Zhang, H., Amick, J., Chakravarti, R., Santarriaga, S., Schlanger, S., McGlone, C., Dare, M., Nix, J.C., Scaglione, K.M., Stuehr, D.J.*, et al.* (2015). A bipartite interaction between Hsp70 and CHIP regulates ubiquitination of chaperoned client proteins. Structure *23*, 472-482.

Zhang, H., Neimanis, S., Lopez-Garcia, L.A., Arencibia, J.M., Amon, S., Stroba, A., Zeuzem, S., Proschak, E., Stark, H., Bauer, A.F.*, et al.* (2014). Molecular mechanism

bioRxiv preprint doi: https://doi.org/10.1101/254367; this version posted January 25, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

of regulation of the atypical protein kinase C by N-terminal domains and an allosteric small compound. Chem Biol *21*, 754-765.

Zhang, M., Windheim, M., Roe, S.M., Peggie, M., Cohen, P., Prodromou, C., and Pearl, L.H. (2005). Chaperoned ubiquitylation--crystal structures of the CHIP U box E3 ubiquitin ligase and a CHIP-Ubc13-Uev1a complex. Mol Cell *20*, 525-538.